FRACTIONAL PRINCIPAL COMPONENTS REGRESSION:

A GENERAL APPROACH TO BIASED ESTIMATORS

by

Wonwoo Lee

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

APPROVED:

Jeffrey B. Birch,
Chairman

Raymond H. Myers

Bradley O. Skarpness

Eric P. Smith

Klaus Hinkelmann

February, 1986

Blacksburg, Virginia

FRACTIONAL PRINCIPAL COMPONENTS REGRESSION:

A GENERAL APPROACH TO BIASED ESTIMATORS

by

Wonwoo Lee

Committee Chairman: Jeffrey B. Birch

Statistics

ABSTRACT

Several biased estimators have been proposed as alternatives to the least squares estimator when multicollinearity is present in the multiple linear regression model. Though the ridge estimator and the principal components estimator have been widely used for such problems, it should be noted that their performances in terms of mean square error are dependent upon the orientation of the unknown parameter vector and the magnitude of σ^2.

By defining the fractional principal components regression model as

$$\underline{Y} = Z\underline{\alpha} + \underline{\epsilon}$$

$$= ZF^{-}\underline{\alpha}_P + \underline{\epsilon},$$

where $\underline{\alpha}_P = F\underline{\alpha}$ and F^{-} is a generalized inverse of a diagonal matrix F, the resulting estimators of $\underline{\alpha}_P$, based on various forms of F, are shown to define the class of the fractional principal components estimators. In the fractional principal components framework, several new estimation techniques are developed. The performances of the new estimators are evaluated and compared with other commonly used biased estimators both theoretically and by simulation studies.

ACKNOWLEDGMENTS

There are numerous people I would like to acknowledge for their contributions to and participation in my graduate education and this dissertation.

First, I would like to thank the members of my committee. Special thanks must first go to Dr. Jeffrey B. Birch. As chairperson, Dr. Birch devoted much time and energy to me personally, as well as to this research. His critical guidance deserves special thanks and for all he has done for me, I extend my thanks.

Dr. Raymond H. Myers has allowed me to pursue my research interests by teaching several courses and providing invaluable comments. To Dr. Klaus Hinkelmann, thanks for his contribution to refinement of this dissertation. Two other members, Dr. Bradley O. Skarpness and Dr. Eric P. Smith, also deserve my special thanks for serving on my committee.

I would like to thank the faculty and graduate students in the Department of Statistics for their help before turning to other special acknowledgments.

There are two professors, Professor Ki-Joong Yoon and Dr. Suk-Bum Yoon at Yonsei University, Korea. They took me under their wings as an undergraduate and provided me with the basic Statistics. I would like to especially thank Professor Ki-Joong Yoon for his encouragement and guidance over the years. He also soothed me at critical periods in my career.

Last, but definitely not least, I would like to thank my family. To my mother, , I am appreciative of her patience, support, and sacrifices. To my wife, , my special thanks for her sacrifices and faith in me. It is to them that I dedicate this dissertation. For everything, I thank them with endless love.

Special thanks to Ginger Wenzlik for typing this dissertation with patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

The multiple linear regression (MLR) model, $\underline{y} = X\underline{\beta} + \underline{\epsilon}$ is one of the most popular methods for exploring linear relationships between a response variable and a set of regressor variables. Among some of the major goals of regression methodology, an important objective is to estimate the unknown parameters from the model. Under the assumptions of identically, independently distributed random errors, one of the parameter estimation techniques is the least squares (LS) method, which provides the well-known best linear unbiased estimator (BLUE). Unfortunately in many practical situations, the regressor variables are not orthogonal and hence, in such cases, the statistical properties of the LS estimator can be adversely affected. One of the harmful effects on the LS estimator is that its expected norm is too large. When there are near linear dependencies between the regressor variables, the problem of multicollinearity is said to exist.

Confronted with multicollinearity in the model, some alternative estimation techniques to the LS method have been developed to remedy the problems resulting from severe multicollinearity. Among the alternatives, the ridge and the principal components estimation methods are prevalent. Because the concepts of these methods are based on shrinking the norm of the LS estimator, they become biased estimators. Therefore, the ridge and principal components estimators provide improvement over the LS estimator in terms of mean square error (MSE). This dissertation will be mainly concerned with the principal components

estimation technique, which utilizes a reparameterized model based on the singular value decomposition or the eigenvalue decomposition.

The developments in principal components regression methodology have taken place for about 20 years. Massy (1965) suggested two deletion criteria when the problem of multicollinearity is involved in the MLR model. One of the criteria is the eigenvalue criterion which involves deleting the principal components whose associated eigenvalues are small. The other criterion is the t-value criterion which involves deleting the principal components whose squared correlation coefficients with the response variable are negligible. However, these traditional deletion criteria may result in the principal components estimator actually having larger MSE than the LS estimator in some situations. Indeed, it can be observed that the condition for which the MSE of the PC estimator is smaller than that of the LS estimator depends upon the orientation of the unknown parameter and the value of the error variance σ^2. Thus a new deletion criterion will be presented in this study.

Furthermore, since the two criteria are too strict in determining which principal components are deleted from the model, Marquardt (1970) proposed the generalized inverse estimation method. If there is a question about retaining a particular principal component, say, the (r+1)st component, Marquardt suggests that a fraction of the (r+1)st principal component should be considered while some principal components are retained and the others are deleted. Marquardt's method, however, is not adequate when more than one component is in question.

Therefore, principal components regression analysis can be gener-

alized by considering the importance of each principal component in reducing the MSE. That is, according to the contribution of the individual principal component to the reduction of MSE, different weights or fractions can be assigned to the principal components. Thus, the class of the fractional principal components (FPC) estimators can be obtained. This class includes the ridge, the generalized ridge, the PC, and some Stein-type estimators, and so on. In other words, most biased estimators can be expressed in the form of the FPC estimator.

In the fractional principal components framework, the ridge estimator will be studied and, interestingly, the situations where the ridge estimator does not provide an improvement over the LS estimator in MSE, will be found and demonstrated. In addition, utilizing the optimal values of the fractions, several iterative estimators will be considered.

In Chapter 2 the ridge and the Stein-type shrinkage estimators are explained, following by a review of the effects of multicollinearity on the LS estimator. Chapter 3 deals with extensive principal components analysis along with introducing a new deletion criterion. The new criterion, termed the unbiased optimal criterion, is based on the orientation of the unknown parameter vector and the value of σ^2. The fractional principal components estimator is defined and discussed in Chapter 4 and, furthermore, the problem related to the ridge estimator will be investigated. Chapter 5 is devoted to evaluating the biased estimators, discussed in this dissertation, by simulation studies. The conclusions and some suggestions for further study are given in Chapter 6.

CHAPTER TWO

MULTICOLLINEARITY AND ALTERNATIVES TO LEAST SQUARES

Linear regression analysis has frequently been used in statistical practice. One objective of the analysis is the estimation of the unknown parameters, typically by using the least squares method. However, due to the adverse effects that multicollinearity can have on least squares estimation, a number of alternative estimation procedures have been recommended. This chapter will present some methods for detection of multicollinearity and measurement of the performance of a biased estimator, introducing several biased estimation techniques designed to combat multicollinearity.

2.1 Linear Regression and Multicollinearity

2.1.1 Linear Regression Model

The general multiple linear regression (MLR) model is a statistical model which attempts to explain the response variable (Y) by a linear combination of p explanatory variables (regressors), $X_1, \ldots, X_p$, that is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i,$$

where β_j, $j=0, 1, \ldots, p$ are unknown parameters and ϵ_i, $i=1, \ldots, n$ are random errors. In the matrix notation, the MLR model is equivalent to

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}, \qquad (2.1)$$

where $\underline{y}$ is an $n\times 1$ vector of observed responses, X is an $n\times(p+1)$ full

(column) rank matrix of nonstochastic regressor variables, $\underline{\beta}$ is a $(p+1)\times 1$ vector of parameters, and $\underline{\epsilon}$ is an $n\times 1$ vector of ϵ_i's.

It is usual to assume that

$$\text{(i)} \quad E(\underline{\epsilon}) = \underline{0}$$

$$\text{(ii)} \quad Var(\underline{\epsilon}) = \sigma^2 I.$$

In other words, the random errors are independent and identically distributed (i.i.d.) with zero means.

The least squares (LS) method for estimating $\underline{\beta}$ is obtained by minimizing the error sum of squares, i.e.,

$$\underset{\underline{\beta}}{\text{Min}} \ (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta}). \tag{2.2}$$

Thus, the resulting normal equations

$$X'X\underline{b}_{LS} = X'\underline{y}$$

can be solved to obtain the LS estimator of $\underline{\beta}$ as

$$\underline{b}_{LS} = (X'X)^{-1} X'\underline{y}. \tag{2.3}$$

It is well known that the LS estimator of $\underline{\beta}$ has many optimal properties one of which is that it has the smallest variance among all linear unbiased estimators. Despite these desirable properties, the LS estimator can be unacceptable if multicollinearity is severe.

2.1.2 Multicollinearity and Its Detection

Multicollinearity refers to the near linear dependencies that may exist among the regressors. Severe multicollinearity is said to exist if there are nonzero constant e_j's such that

$$\sum_{j=1}^{p} e_j \underline{x}_j \approx \underline{0}, \tag{2.4}$$

where $\underline{x}_j$ is the j^{th} column vector of X. It is important to emphasize that multicollinearity is usually present in the regression model to a certain extent unless the columns of X are chosen to be orthogonal. Therefore, multicollinearity describes a condition in the regressor variables.

Before proceeding to the detection of multicollinearity, it will be convenient to have the regressor variables centered and scaled. Thus, the reparametrized linear regression model is

$$Y_i = \beta_o^* + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \ldots + \beta_p^* X_{ip}^* + \epsilon_i, \qquad i=1, \ldots, n, \tag{2.5}$$

where $X_{ij}^* = (X_{ij} - \bar{X}_j)/S_{Xj}$, $S_{Xj} = \left[\Sigma(X_{ij} - \bar{X}_j)^2\right]^{1/2}$, for $j=1, \ldots, p$.

The reparameterized model (2.5) is, of course, equivalent to the original-unit model (2.1). Since the centered and scaled model (2.5) can be rewritten as

$$Y_i = \left[\beta_o^* - \frac{\bar{X}_1}{S_{X1}} \beta_1^* - \ldots - \frac{\bar{X}_p}{S_{Xp}} \beta_p^*\right] + \frac{1}{S_{X1}} \beta_1^* X_{i1} + \ldots + \frac{1}{S_{Xp}} \beta_p^* X_{ip} + \epsilon_i,$$

the parameters have the following relationships:

$$\beta_o = \beta_o^* - \left[\frac{\bar{X}_1}{S_{X1}} \beta_1^* + \ldots + \frac{\bar{X}_p}{S_{Xp}} \beta_p^*\right]$$

$$\beta_j = \frac{1}{S_{Xj}} \beta_j^*, \qquad j=1, \ldots, p.$$

Note that the β_j^*'s are unit free while the β_j's are dependent upon the natural units of the regressors.

The centered and scaled model (2.5) without β_0^* will be used in the rest of this thesis since the estimator of β_0^* is $\Sigma Y_i/n$. Therefore, to simplify notation, the centered and scaled regressor variables will be denoted without the superscript.

In equation (2.4), the defining relationship ($\Sigma_j e_j \underline{x}_j \approx \underline{0}$) can be investigated by using the eigenvalue decomposition of the X matrix. That is,

$$V'X'X\,V = \Lambda,$$

where $V = [\underline{v}_1, \ldots, \underline{v}_p]$. The columns of V are normalized eigenvectors corresponding to the eigenvalues of $X'X$, which are the diagonal elements of the diagonal matrix Λ, $\lambda_1, \ldots, \lambda_p$. The problem of severe multicollinearity implies that at least one $\lambda_j \approx 0$. Thus, for at least one j,

$$\underline{v}_j'X'X\,\underline{v}_j \approx 0$$

which implies that

$$X\,\underline{v}_j = \sum_{k=1}^{p} v_{kj}\,\underline{x}_k \approx \underline{0}, \qquad (2.6)$$

where $\underline{x}_j$ is the j^{th} column of X and v_{kj} is the k^{th} element of $\underline{v}_j$. Therefore, the individual elements of $\underline{v}_j$ determine the near linear dependency among the regressors.

There are several diagnostic measurements used to detect and assess the degree of multicollinearity. The most popular and appealing measure is the condition number. The singular value decomposition (SVD) theorem

(Rao, 1973) allows X to be decomposed as

$$X = ULV', \tag{2.7}$$

where U and V are n×p and p×p matrices, respectively, such that $U'U = V'V = VV' = I$. The p×p matrix L is a diagonal matrix with non-negative diagonal elements, called the singular values of X. In fact, the columns of U are the eigenvectors associated with the positive eigenvalues of XX' and those of V are the eigenvectors of $X'X$. Furthermore, the singular values $\ell_1, \ldots, \ell_p$ are square roots of the eigenvalues of $X'X$ that is, $L^2 = \Lambda$, which are assumed, without loss of generality, to be ordered in magnitude, i.e.,

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p,$$

where λ_j is the j^{th} largest eigenvalue. Hence the condition number, κ, is defined as

$$\kappa = \frac{\ell_1}{\ell_p}, \tag{2.8}$$

which is equivalent to $\sqrt{\lambda_1} / \sqrt{\lambda_p}$. Then κ indicates the relative magnitude of the smallest singular value to the largest singular value and hence can be used as a yardstick to determine the degree of multicollinearity. Note that the ratio ℓ_1/ℓ_j, for j=1, ..., p is called the condition index. Belsley, Kuh, and Welsch (1980) found empirically that "weak dependencies are associated with condition indexes around 5 or 10, whereas moderate to strong relations are associated with condition indexes of 30 to 100."

2.1.3 Effects of Multicollinearity

Multicollinearity can have adverse effects on the LS solution and related procedures [Belsley, Kuh, and Welsch (1980); Tripp (1983); Myers (1985), etc.]. It may result in some LS coefficients having the wrong sign and being very large in magnitude, resulting in an inflated length of the estimated regression coefficient vector.

The expected squared norm of $\underline{b}_{LS}$ is

$$E(\underline{b}_{LS}'\underline{b}_{LS}) = E[\underline{y}'X(X'X)^{-1}(X'X)^{-1}X'\underline{y}]$$

$$= \underline{\beta}'\underline{\beta} + \sigma^2 tr(X'X)^{-1}$$

$$= \underline{\beta}'\underline{\beta} + \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}, \qquad (2.9)$$

which implies that the expected squared norm of $\underline{b}_{LS}$, under moderate to severe multicollinearity, is much larger than that of the true parameter vector $\underline{\beta}$, since there is at least one small eigenvalue λ_j. In other words, at least one of the LS coefficients $b_{j,LS}$ will be very large.

Furthermore, under moderate to severe multicollinearity, some of the LS estimated coefficients may have unreasonably large variances. Using the SVD, to see this, consider

$$Var(\underline{b}_{LS}) = \sigma^2(X'X)^{-1}$$

$$= \sigma^2 V\Lambda^{-1}V'.$$

Therefore, the variance of the j^{th} LS estimated coefficient is

$$Var(b_{j,LS}) = \sigma^2 \sum_{k=1}^{p} \frac{v_{jk}^2}{\lambda_k}, \qquad (2.10)$$

where v_{jk} is the j^{th} element of $\underline{v}_k$. Thus, at least one small λ_k may inflate the variance of $b_{j,LS}$ as long as v_{jk}, associated with λ_k, is not close to 0.

Inferential procedures, such as testing hypotheses, are also affected by the presence of multicollinearity. Under the assumption that the $\underline{\epsilon}$ are i.i.d. and follow normal ($\underline{0}$, $\sigma^2 I$) distributions, the test of the hypothesis H_0: $H\underline{\beta} = \underline{h}$ (vs. H_1: $H\underline{\beta} \neq \underline{h}$) is based on the statistic

$$F' = (H\underline{b}_{LS} - \underline{h})' \left[H (X'X)^{-1}H'\right]^{-1} (H\underline{b}_{LS} - \underline{h})/ms^2, \qquad (2.11)$$

where H is a m×p matrix with full (row) rank and s^2 is the estimator of σ^2, which is the mean squared error [Searle (1971)]. The test statistic F' has a non-central F distribution with m and (n-m) degrees of freedom and the non-centrality parameter

$$\eta = (H\underline{\beta} - \underline{h})' \left[H (X'X)^{-1}H'\right]^{-1} (H\underline{\beta} - \underline{h})/2\sigma^2.$$

In particular, setting $H = I$ and $\underline{h} = \underline{0}$, it is easily seen that the non-centrality parameter is a function of the eigenvalues, that is,

$$\begin{aligned} \eta &= \underline{\beta}'X'X\underline{\beta} \left[\frac{1}{2\sigma^2}\right] \\ &= \underline{\beta}'V \Lambda V'\underline{\beta} \left[\frac{1}{2\sigma^2}\right] \\ &= \sum_{j=1}^{p} \frac{\lambda_j(\underline{v}_j'\underline{\beta})^2}{2\sigma^2}. \end{aligned} \qquad (2.12)$$

Thus the non-centrality parameter can decrease as the degree of multicollinearity gets more severe, so that the power of the test may decrease.

It is worthwhile to consider testing hypotheses of individual elements of $\underline{\beta}$, i.e., H_{0j}: $\beta_j = 0$ (vs. H_{1j}: $\beta_j \neq 0$), for j=1, ..., p. Then the non-centrality parameters, η_j, are

$$\eta_j = \frac{\lambda_j(V_{jj}\beta_j)^2}{2\sigma^2}, \qquad j=1, \ldots, p, \tag{2.13}$$

and hence $\eta_j \to 0$ as $\lambda_j \to 0$. In this case, $H = H_j$, where H_j is p×p zero matrix but the j^{th} diagonal element is 1.

It is of interest to illustrate the effects of multicollinearity, discussed thus far, with the following canonical model:

$$\begin{aligned} \underline{y} &= X\underline{\beta} + \underline{\epsilon} \\ &= XVV'\underline{\beta} + \underline{\epsilon} \\ &= Z\underline{\alpha} + \underline{\epsilon}, \end{aligned} \tag{2.14}$$

where $Z = XV = [\underline{z}_1, \ldots, \underline{z}_p]$ and $\underline{\alpha} = V'\underline{\beta}$.

The LS estimator of $\underline{\alpha}$ is $\underline{a}_{LS} = (Z'Z)^{-1} Z'\underline{y}$ so that the expected squared norm of $\underline{a}_{LS}$ is

$$\begin{aligned} E(\underline{a}_{LS}'\underline{a}_{LS}) &= E\left[\underline{y}'Z(Z'Z)^{-1} (Z'Z)^{-1} Z'\underline{y}\right] \\ &= \underline{\alpha}'\underline{\alpha} + \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \end{aligned} \tag{2.15}$$

and the covariance matrix of $\underline{a}_{LS}$ is

$$\begin{aligned} \mathrm{Var}(\underline{a}_{LS}) &= \sigma^2(Z'Z)^{-1} \\ &= \sigma^2\Lambda^{-1}. \end{aligned} \tag{2.16}$$

Thus, the tranformation to the canonical model, consisting of orthogonal regressors $\underline{z}_1, \ldots, \underline{z}_p$, does not mean that the multicollinearity problem

is eliminated but only reexpressed in a more convenient form.

For the j^{th} LS estimator of $\underline{\alpha}$, it follows from (2.15) and (2.16) that the expectation of $a_{j,LS}^2$ and the variance of $a_{j,LS}$ are

$$E(a_{j,LS}^2) = \alpha_j^2 + \sigma^2 \frac{1}{\lambda_j} \tag{2.17}$$

and

$$Var(a_{j,LS}) = \sigma^2 \frac{1}{\lambda_j} \tag{2.18}$$

so that if λ_j is small, then $E(a_{j,LS}^2)$ makes a large contribution mainly to the size of $E(\underline{a}_{LS}'\underline{a}_{LS})$ and $Var(a_{j,LS})$ is seriously inflated.

Accordingly, for the test of hypothesis $H_0: \underline{\alpha} = \underline{0}$ (vs. $H_1: \underline{\alpha} \neq \underline{0}$), the test statistic

$$F' = \frac{\underline{a}_{LS}'\Lambda\,\underline{a}_{LS}}{ps^2} \tag{2.19}$$

has a non-central F distribution with the non-centrality parameter

$$\eta = \frac{\underline{\alpha}'\Lambda\,\underline{\alpha}}{2\sigma^2}$$

$$= \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^2}{2\sigma^2}. \tag{2.20}$$

Thus the power of the test will decrease, under moderate to severe multicollinearity, because there is at least one small λ_j. Also, for the tests of individual hypothesis $H_{0j}: \alpha_j = 0$ (vs. $H_{1j}: \alpha_j \neq 0$), for $j=1, \ldots, p$, the non-centrality parameters are

$$\eta_j = \frac{\lambda_j \alpha_j^2}{2\sigma^2}, \qquad j=1, \ldots, p. \tag{2.21}$$

In this particular case, only the j^{th} test of hypothesis will have reduced power if the j^{th} eigenvalue λ_j is small.

2.1.4 MSE: A Performance Measure for Biased Estimator

Due to the harmful effects of multicollinearity on the LS estimator, it is natural to consider some alternative estimation methods. Some of these alternatives lead to biased estimators whose performances can be measured by the mean square error (MSE).

Denoting the MSE of a biased estimator, $\tilde{\underline{b}}$, of $\underline{\beta}$ by $MSE(\tilde{\underline{b}})$, the expression of $MSE(\tilde{\underline{b}})$ can be decomposed into two components

$$\begin{aligned} MSE(\tilde{\underline{b}}) &= \sum_{j=1}^{p} MSE(\tilde{b}_j) \\ &= E(\tilde{\underline{b}} - \underline{\beta})'(\tilde{\underline{b}} - \underline{\beta}) \\ &= E(\tilde{\underline{b}} - E\tilde{\underline{b}})'(\tilde{\underline{b}} - E\tilde{\underline{b}}) + (E\tilde{\underline{b}} - \underline{\beta})'(E\tilde{\underline{b}} - \underline{\beta}). \end{aligned} \tag{2.22}$$

The first and second terms in (2.22) will be regarded as the "variance" and the "squared-bias" components, respectively. It is noted that by using a biased estimator, a large reduction in the "variance" component is expected while some increase in the "squared-bias" component is accepted. Hence, hopefully, the $MSE(\tilde{\underline{b}})$ will be smaller than $MSE(\underline{b}_{LS}) = \Sigma_j \, Var(b_{j,LS})$, resulting in an improvement over the LS estimator.

It is beneficial to use the standardized regressors since $MSE(\tilde{\underline{b}})$ can be interpreted as the average of the squared Euclidean distance between $\tilde{\underline{b}}$ and $\underline{\beta}$. Keeping this in mind, an estimator of $\underline{\beta}$ with smaller MSE is located closer on the average to the true parameter vector $\underline{\beta}$. This concept is illustrated in Figure 2.1 and 2.2 when $p = 2$. Figure 2.1 shows a biased estimator, $\tilde{\underline{b}}_A$, shortened from $\underline{b}_{LS}$; and Figure 2.2 presents a biased estimator, $\tilde{\underline{b}}_B$, designed to be located closer to $\underline{\beta}$ than $\underline{b}_{LS}$ to $\underline{\beta}$. In other words, $\tilde{\underline{b}}_A$ results from shortening $\underline{b}_{LS}$ uniformly in

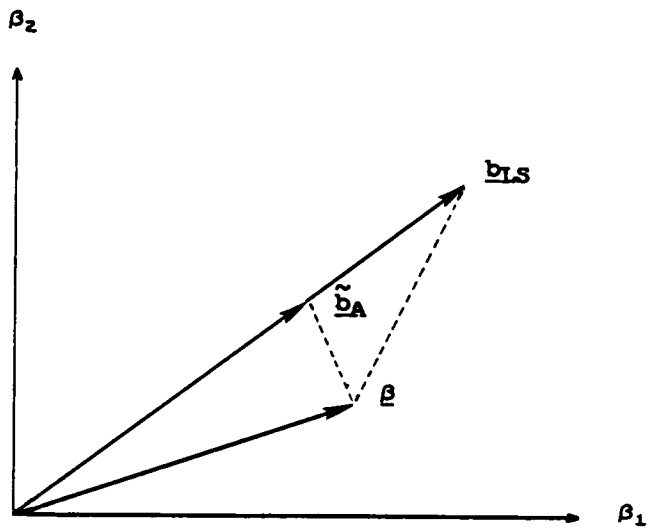

Figure 2.1 A Typical Biased Estimator with Smaller MSE than LS Estimator: Uniform Shortening Rate

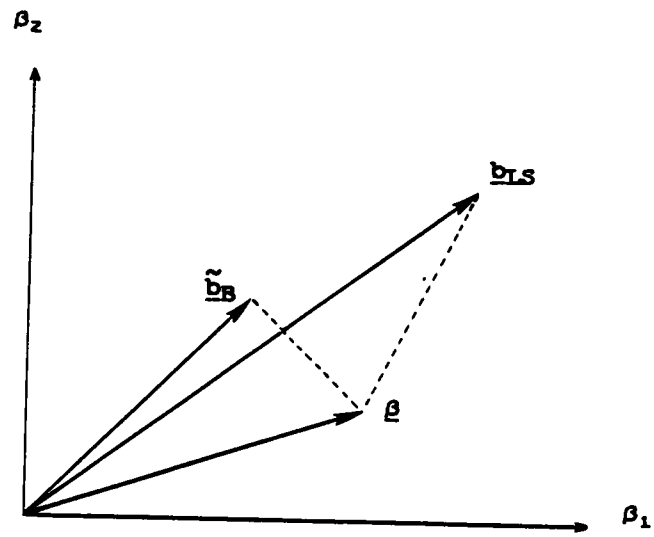

Figure 2.2 A Typical Biased Estimator with Smaller MSE than LS Estimator: Different Shortening Rate

$b_{j,LS}$ and $\tilde{b}_B$ uses different shortening rates on each $b_{j,LS}$, j=1, 2. Note that both $\tilde{b}_A$ and $\tilde{b}_B$ are typical biased estimates for which MSE($\tilde{b}_A$) and MSE($\tilde{b}_B$) are smaller than MSE(b_{LS}).

It is often of interest to consider the matrix form of MSE, denoted by MtxMSE($\tilde{b}$):

$$\text{MtxMSE}(\tilde{b}) = E(\tilde{b} - \beta)(\tilde{b} - \beta)'$$

$$= \text{Var}(\tilde{b}) + \text{Bias}(\tilde{b})\,\text{Bias}(\tilde{b})', \qquad (2.23)$$

where $\text{Bias}(\tilde{b}) = E(\tilde{b}) - \beta$.

In order to compare the performances of b_{LS} and $\tilde{b}$ in terms of MSE, either form, (2.22) or (2.23), can be used. That is, either MSE(b_{LS}) - MSE($\tilde{b}$) or MtxMSE(b_{LS}) - MtxMSE($\tilde{b}$) can be used since the former is non-negative if and only if the latter is positive semi-definite [see Theobald (1974)].

2.2 Ridge Estimation

Ridge regression has been examined extensively since its development by Hoerl and Kennard (1970a, 1970b). It is well known when the multicollinearity problem exists that the ridge estimator, which is a biased estimator of β, results from shortening or shrinking the length of the LS estimator, b_{LS}.

2.2.1 Ridge Estimator

Since the presence of multicollinearity can inflate the distance from b_{LS} to β, Hoerl and Kennard (1970a) developed the ridge estimator whose length is minimum for a fixed residual sum of squares.

They considered an estimate, $\underline{B}$, of $\underline{\beta}$ and the resulting residual sum of squares,

$$SS(\underline{B}) = (\underline{y} - X\underline{B})'(\underline{y} - X\underline{B})$$

$$= (\underline{y} - X\underline{b}_{LS})'(\underline{y} - X\underline{b}_{LS}) + (\underline{B} - \underline{b}_{LS})'(X'X)(\underline{B} - \underline{b}_{LS}). \qquad (2.24)$$

Hence, the ridge estimator is derived from the following formulation:

$$\underset{\underline{B}}{\text{Minimize}}\ \underline{B}'\underline{B}$$

$$\text{subject to } (\underline{B} - \underline{b}_{LS})'(X'X)(\underline{B} - \underline{b}_{LS}) = Q. \qquad (2.25)$$

The resulting ridge estimator can be written as

$$\underline{b}_R = (X'X + kI)^{-1} X'\underline{y}, \qquad (k > 0) \qquad (2.26)$$

where k is chosen to satisfy the constraint in (2.25). The positive constant k is commonly referred to as the biasing or shrinkage parameter. Using the canonical model (2.14), another expression of $\underline{b}_R$ is

$$\underline{b}_R = V(\Lambda + kI)^{-1} Z'\underline{y}.$$

Thus the ridge estimator is obtained by augmenting the eigenvalues of $X'X$ with k.

Some properties of the ridge estimator are as follows [Hoerl and Kennard (1970a) and Marquardt (1970)]. If k is treated as a constant, the bias of the ridge estimator is

$$-k(X'X + kI)^{-1} \underline{\beta}.$$

Since $\underline{b}_R$ is a biased estimator, the $MSE(\underline{b}_R)$, using (2.22), can be decom-

posed into the two components attributed to the "variance" and "squared-bias" parts as

$$MSE(\underline{b}_R) = \sum_{j=1}^{p} Var(b_{j,R}) + \sum_{j=1}^{p} (\text{bias in } b_{j,R})^2$$

$$= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \underline{\beta}'(X'X + kI)^{-2} \underline{\beta}$$

$$= \Psi_1(k) + \Psi_2(k), \qquad (2.27)$$

where $\Psi_1(k) = \sigma^2 \sum_{j=1}^{p} \lambda_j/(\lambda_j + k)^{-2}$ and $\Psi_2(k) = k^2 \underline{\beta}'(X'X + kI)^{-2} \underline{\beta}$. Note that $\Psi_1(k)$ is a continuous, monotonically decreasing function of k, while $\Psi_2(k)$ is continuous and monotonically increasing. Hoerl and Kennard (1970a) claimed, through investigation of the derivatives of $\Psi_1(k)$ and $\Psi_2(k)$, that $MSE(\underline{b}_R)$ is improved by a substantial reduction of $\Psi_1(k)$ at the expense of a slight increase in $\Psi_2(k)$. They also showed that $MSE(\underline{b}_R) < MSE(\underline{b}_{LS})$, whenever

$$0 < k < \frac{\sigma^2}{(\alpha_{max})^2}. \qquad (2.28)$$

Many authors have discussed a variety of methods for choosing the value of k. Among a number of ways for solving k, the original but subjective method suggested by Hoerl and Kennard (1970a, 1970b) is the ridge trace. Among the objective methods for choosing k, Hoerl, Kennard, and Baldwin (1975) suggested

$$k = \frac{ps^2}{\underline{b}_{LS}'\underline{b}_{LS}}. \qquad (2.29)$$

And, in addition, Hoerl and Kennard (1976) have proposed an iteration

method with

$$k^{(t+1)} = \frac{ps^2}{\underline{b}_R[k^{(t)}]'\underline{b}_R[k^{(t)}]}, \qquad t=0, 1, 2, \ldots \qquad (2.30)$$

where $k^{(t)}$ is the value of k after the t^{th} iteration, $\underline{b}_R[k^{(t)}]$ is the ridge estimate based on $k^{(t)}$ and the starting value, $k^{(0)} = ps^2/\underline{b}_{LS}'\underline{b}_{LS}$. They also suggested a stopping criterion as

$$\frac{k^{(t+1)} - k^{(t)}}{k^{(t)}} < 20\left[\frac{tr(X'X)^{-1}}{p}\right]^{-1.3}. \qquad (2.31)$$

However, many of the distributional properties of the ridge estimator assume that k is a constant. In practice, however, k is estimated from the data and, thus, is stochastic. Therefore, it is questionable if these properties still apply.

2.2.2 Generalized Ridge Estimator

While the ridge estimator results from augmenting the eigenvalues of X'X uniformly with k, the generalized ridge estimator is defined as:

$$\underline{b}_{GR} = (X'X + K)^{-1} X'\underline{y}, \qquad (2.32)$$

where $K = Diag(k_1, \ldots, k_p)$, $k_j \geq 0$ [Hoerl and Kennard (1970a)]. Since $\underline{b}_{GR}$ can be rewritten, by using (2.14),

$$\underline{b}_{GR} = V(\Lambda + K)^{-1} Z'\underline{y},$$

the different values of biasing parameters are added to the eigenvalues of X'X in the generalized ridge estimation. It can be easily shown that the optimal values for the k_j are

$$k_j = \frac{\sigma^2}{\alpha_j^2}, \qquad j=1, \ldots, p. \tag{2.33}$$

Hoerl and Kennard (1970a) also suggest an iterative procedure initiated at $k_j^{(0)} = s^2/a_{j,LS}^2$ with

$$k_j^{(t+1)} = \frac{s^2}{\left[a_{j,GR}^{(t)}\right]^2}, \qquad t=1, 2, \ldots \tag{2.34}$$

where $a_{j,GR}^{(t)}$ is the j^{th} generalized ridge estimate of α_j at the t^{th} iteration. Iteration is continued until there is stability achieved in $\underline{a}_{GR}^{(t+1)'}\underline{a}_{GR}^{(t+1)}$. An analytic solution to this iterative scheme is given by Hemmerle (1975).

It is difficult to choose either the ridge regression or the generalized ridge regression to combat the multicollinearity problem. Some authors [Hocking, Speed, and Lynn (1976)] prefer the latter to the former while Hoerl and Kennard (1970a, p.65) prefer the use of a common k by writing, "Based on experience, the best method for achieving a better estimator is to use $k_j = k$ for all j." However, no one has substantiated the preference.

The differences and the similarities of the two will be discussed although both of them are competitive. It will also be shown later in Section 4.1.3 that the ridge estimator is preferred in some situations and the generalized ridge estimator is preferred in others. The evaluation of the two based on a simulation study is in Section 5.2.2.

2.3 Shrinkage Estimation

The concept of shrinking the LS estimator, proposed by Stein (1960), has been justified as a broad class of alternative procedures to

LS estimation when multicollinearity is present in the MLR model.

2.3.1 Shrinkage Estimator (Stein)

The main objective for the shrinkage estimator, which is a biased estimator, is to have a MSE smaller than the variance of the LS estimator. Consider the class of estimators whose typical member is

$$\underline{b}_{SH} = \delta \underline{b}_{LS}, \qquad \delta \in [0,1] \tag{2.35}$$

where δ is called the shrinkage factor.

Investigating Figure 2.1, the range of $\underline{b}_{SH}$ in (2.35) for which $MSE(\underline{b}_{SH}) \leq MSE(\underline{b}_{LS})$ can be shown in Figure 2.3. The determination of the shrinkage factor is a major consideration. Perlman (1972) showed that $MSE(\underline{b}_{SH}) \leq MSE(\underline{b}_{LS})$ is satisfied for any $0 < \delta < 1$ such that

$$\frac{m_o - 1}{m_o + 1} \leq \delta, \tag{2.36}$$

where

$$\sigma^{-2}\underline{\beta}'(X'X)\underline{\beta} = m_o.$$

The Stein shrinkage estimator can be interpreted, in the regression context, by rewriting the model (2.14), based on SVD, as

$$\begin{aligned} \underline{y} &= X\underline{\beta} + \underline{\epsilon} \\ &= ULV'\underline{\beta} + \underline{\epsilon} \\ &= U\underline{\gamma} + \underline{\epsilon}, \end{aligned} \tag{2.37}$$

where $\underline{\gamma} = LV'\underline{\beta}$. Premultiplying this reparameterized model by U', a

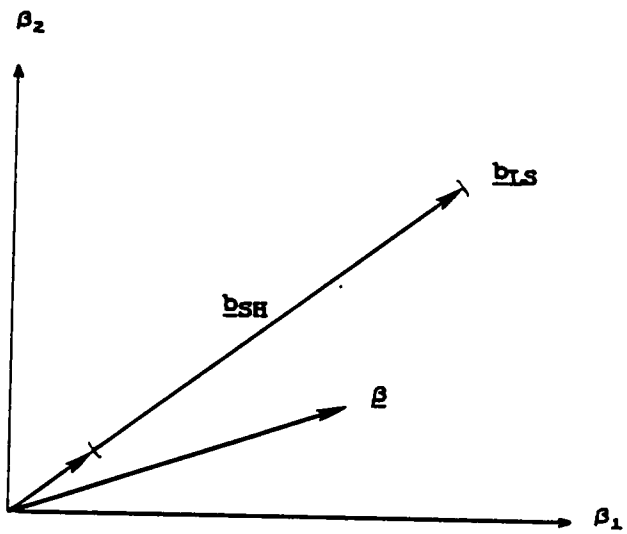

Figure 2.3 Range of $\underline{b}_{SH}$ for $MSE(\underline{b}_{SH}) \leq MSE(\underline{b}_{LS})$

simple mean-shift model can be obtained, that is,

$$U'\underline{y} = \underline{\gamma} + U'\underline{\epsilon}.$$

When σ^2 is known and $p > 2$, Stein (1955) proposed the solution for the estimator of $\underline{\gamma}$ as

$$\hat{\underline{\gamma}}_{ST} = \left[1 - \frac{c\,\sigma^2}{\underline{y}'U\,U'\underline{y}}\right] U'\underline{y}$$

$$= \left[1 - \frac{c\,\sigma^2}{\underline{b}_{LS}'X'X\,\underline{b}_{LS}}\right] \hat{\underline{\gamma}}_{LS}, \qquad (2.38)$$

where c is a positive constant and $\hat{\underline{\gamma}}_{LS} = U'\underline{y}$. Thus the Stein estimator of $\underline{\beta}$ [Stein (1960)] is

$$\underline{b}_{ST} = V\,L^{-1}\,\hat{\underline{\gamma}}_{ST}$$

$$= \left[1 - \frac{c\,\sigma^2}{\underline{b}_{LS}'X'X\,\underline{b}_{LS}}\right] V\,L^{-1}\,L\,V'\underline{b}_{LS}$$

$$= \left[1 - \frac{c\,\sigma^2}{\underline{b}_{LS}'X'X\,\underline{b}_{LS}}\right] \underline{b}_{LS}. \qquad (2.39)$$

One important property of the Stein estimator is that

$$MSE(\underline{b}_{ST}) < MSE(\underline{b}_{LS}),$$

for $0 < c < 2(p-2)$. In a subsequent paper, James and Stein (1961) proposed, for the case that σ^2 is unknown,

$$\underline{b}_{JS} = \left[1 - \frac{c\,SS}{\underline{b}_{LS}'X'X\,\underline{b}_{LS}}\right] \underline{b}_{LS}, \qquad (2.40)$$

for $0 < c < 2(p-2)/(n-p+2)$, where SS is the residual sum of squares. They also proved that $MSE(\underline{b}_{JS}) < MSE(\underline{b}_{LS})$ and showed that the $MSE(\underline{b}_{JS})$ is minimized at $c = (p-2)/(n-p+2)$. Note that the Stein estimator is unfortunately sensitive to a change of the origin of the response variable [see (2.38)]. Lindley (1962) suggested a modification to the Stein estimator, which is not sensitive with respect to a change of origin and scale. However, Lindley's modified Stein estimator is not always preferred over the Stein estimator so that attention may be paid to those situations where the origin change is meaningful.

The James-Stein estimator of $\underline{y}$ can be expressed in a manner equivalent to (2.38) as

$$\hat{\underline{y}}_{JS} = \left[1 - \frac{c\ SS}{\hat{\underline{y}}_{LS}'\hat{\underline{y}}_{LS}}\right] \hat{\underline{y}}_{LS},$$

or, equivalently, that of $\underline{\alpha}$ as

$$\underline{a}_{JS} = \left[1 - \frac{c\ SS}{\underline{a}_{LS}'A\ \underline{a}_{LS}}\right] \underline{a}_{LS}. \tag{2.41}$$

It will be beneficial to consider the Bayesian interpretation of the Stein estimator. Assume that the vector of the response variable, $\underline{y}$ is distributed as the multivariate normal $(Z\underline{\alpha}, \sigma^2 I)$, that is,

$$\underline{y} \sim N(Z\underline{\alpha}, \sigma^2 I) \tag{2.42}$$

and σ^2 is known. Taking a special prior distribution of $\underline{\alpha}$ as

$$\underline{\alpha} \sim N\left[\underline{\alpha}_0, \sigma_\alpha^2 (Z'Z)^{-1}\right], \tag{2.43}$$

the posterior density of $\underline{\alpha}$ is also the multivariate normal with the mean

vector,

$$\underline{a}_B = \left[\sigma^{-2} Z'Z + \sigma_\alpha^{-2} Z'Z\right]^{-1} \left[\sigma^{-2} Z'Z \underline{a}_{LS} + \sigma_\alpha^{-2} Z'Z \underline{\alpha}_o\right] \qquad (2.44)$$

and the covariance matrix

$$\Sigma_B = \left[\sigma^{-2} Z'Z + \sigma_\alpha^{-2} Z'Z\right]^{-1}.$$

Thus, the Bayes estimator of $\underline{\beta}$, which is optimal with respect to the expected quadratic loss function, is the posterior mean $\underline{a}_B$.

Assume, now, that $\underline{\alpha}_o = \underline{0}$. Then $\underline{a}_B$ in (2.44) becomes

$$\underline{a}_B = \frac{\sigma_\alpha^2}{\sigma^2 + \sigma_\alpha^2} \underline{a}_{LS}$$

$$= \left[1 - \frac{\sigma^2}{\sigma^2 + \sigma_\alpha^2}\right] \underline{a}_{LS}. \qquad (2.45)$$

Replacing the term $\sigma^2 + \sigma_\alpha^2$ by its unbiased estimator[1], $\underline{a}_{LS}'\Lambda\, \underline{a}_{LS}/p$ [see Vinod and Ullah (1981)], $\underline{a}_B$ can be rewritten as

$$\underline{a}_B = \left[1 - \frac{p\sigma^2}{\underline{a}_{LS}'\Lambda\, \underline{a}_{LS}}\right] \underline{a}_{LS}, \qquad (2.46)$$

which is the Stein estimator of $\underline{\alpha}$ when $c = p$. Furthermore, if the unbiased estimate of σ^2, s^2, is substituted for σ^2 then the Bayesian interpreted James-Stein estimator of $\underline{\alpha}$, denoted by $\underline{a}_{BJB}$, is

---

[1] $E\left[\underline{a}_{LS}'\Lambda\, \underline{a}_{LS}\right] = E\left[\left[\underline{\alpha} + (Z'Z)^{-1} Z'\underline{\epsilon}\right]' \Lambda \left[\underline{\alpha} + (Z'Z)^{-1} Z'\underline{\epsilon}\right]\right]$

$= E\left[\underline{\alpha}'\Lambda\, \underline{\alpha}\right] + E\left[\underline{\epsilon}'Z\, \Lambda^{-1} Z'\underline{\epsilon}\right] + 2E\left[\underline{\alpha}'Z'\underline{\epsilon}\right]$

$= p\, \sigma_\alpha^2 + p\, \sigma^2$

$$\underline{a}_{BJS} = \left[1 - \frac{ps^2}{\underline{a}_{LS}' \Lambda \underline{a}_{LS}}\right] \underline{a}_{LS}$$

$$= \left[1 - \left[\frac{p}{n-p}\right] \frac{SS}{\underline{a}_{LS}' \Lambda \underline{a}_{LS}}\right] \underline{a}_{LS}. \qquad (2.47)$$

If (2.47) is compared to (2.41), when $c = p/(n-p)$, then $\underline{a}_{BJS} = \underline{a}_{JS}$. Note that it is clear, through (2.47), that each element of the LS estimator $\underline{a}_{LS}$ is shrunk by the same proportion

$$\frac{ps^2}{\underline{a}_{LS}' \Lambda \underline{a}_{LS}}, \qquad (2.48)$$

where the shrinkage factor is just the reciprocal of the test statistic, F', for testing hypothesis H_0: $\underline{\alpha} = \underline{0}$ [see (2.19)]. Recalling that when the multicollinearity problem is present in the MLR model, the power of the test is so small that the F test may be unable to distinguish $\underline{\alpha}$ from $\underline{0}$, the value of F' can be expected to be small. In other words, the value of (2.48) is expected to be too large and hence the BJS estimator may result in being shrunk too much.

However, the Bayesian interpreted James-Stein estimator will be included later in the simulation studies. In Section 4.3.1, a related estimator, the generalized Bayesian interpreted James-Stein estimator, will be developed in a similar manner as has been done for $\underline{a}_{BJS}$.

2.3.2 Modified Stein-Shrinkage Estimators

A number of other types of shrinkage estimators have been studied by several authors including Baranchik (1964, 1970), Sclove (1968), Mayer and Willke (1973), Rolph (1976), and Vinod (1976).

Baranchik (1964) proved that the new estimator of $\underline{\alpha}$,

$$\underline{a}_{BA} = \left[1 - \frac{c\ SS}{\underline{a}_{LS}'\Lambda\ \underline{a}_{LS}}\right]^{+} \underline{a}_{LS}, \tag{2.49}$$

has smaller MSE than $\underline{a}_{JS}$, where

$$\left[1 - c\ SS/\underline{a}_{LS}'\Lambda\ \underline{a}_{LS}\right]^{+} = \max\left[0,\ 1 - c\ SS/\underline{a}_{LS}'\Lambda\ \underline{a}_{LS}\right].$$

In other words, the new estimator, $\underline{a}_{BA}$, is defined by taking the shrinkage factor as the positive part of that of $\underline{a}_{JS}$. Since $\underline{a}_{BA}$ can be rewritten as, denoting $F^{*} = \underline{a}_{LS}'\Lambda\ \underline{a}_{LS}/SS$,

$$\begin{aligned} \underline{a}_{BA} &= \underline{0}, && \text{if } F^{*} < c, \\ &= \left[1 - \frac{c\ SS}{\underline{a}_{LS}'\Lambda\ \underline{a}_{LS}}\right] \underline{a}_{LS}, && \text{if } F^{*} \geq c, \end{aligned} \tag{2.50}$$

the use of $\underline{a}_{BA}$ seems to correspond to a preliminary test of hypothesis H_0: $\underline{\alpha} = \underline{0}$ at a level of significance dictated by the value of c, where $0 < c < 2(p-2)/(n-p+2)$. Note that $\underline{a}_{LS}'\Lambda\ \underline{a}_{LS}/SS < c$ is identical to $F' < (n-p)c/p$. Thus, $\underline{\alpha}$ is estimated as $\underline{0}$ when the null hypothesis is not rejected and as $(1 - c\ SS/\underline{a}_{LS}'\Lambda\ \underline{a}_{LS})\underline{a}_{LS}$ when the hypothesis is rejected, in which case the shrinkage factor is non-negative.

Baranchik (1970) also generalized the James-Stein estimator by using a function $h(\cdot)$ in the place of c in (2.41) and defining

$$\underline{a}_{H} = \left[1 - \frac{h(F^{*})}{F^{*}}\right] \underline{a}_{LS}, \tag{2.51}$$

where $h(\cdot)$ is any monotone nondecreasing function such that $0 \leq h(\cdot) \leq 2(p-2)/(n-p+2)$. He proved that $MSE(\underline{a}_{H}) \leq MSE(\underline{a}_{LS})$. Note that the

James-Stein estimator is a special case in which $h(F^*) = c$. Furthermore, if the function $h(F^*)$ is taken to be [Figure 2.4]

$$h(F^*) = F^*, \qquad \text{if } F^* < c,$$

$$= c, \qquad \text{if } F^* \geq c, \tag{2.52}$$

then the resulting estimator, $\underline{a}_H$, is identical to $\underline{a}_{BA}$.

While Baranchik's (1964, 1970) modification still results in uniform shrinkage, that is, 0 for $F^* \leq c$ and $(1 - c/F^*)$ for $F^* > c$, Sclove (1968) proposed to apply a shrinkage factor to a subset of the elements of $\underline{a}_{LS}$. Suppose that the first r columns of the matrix Z are regarded, for some reason, as preferred variables. Then, partitioning $\underline{\alpha}$ as $\underline{\alpha} = (\underline{\alpha}_1', \underline{\alpha}_2')'$, where $\underline{\alpha}_1$ is a r×1 vector and $\underline{\alpha}_2$ is a (p-r)×1 parameter vector, the Sclove's shrinkage estimator is defined as

$$\underline{a}_{SC} = \begin{bmatrix} \underline{a}_{1,LS} \\ \hline \left[1 - \frac{c}{F_2^*}\right]^+ \underline{a}_{2,LS} \end{bmatrix}, \tag{2.53}$$

where $0 < c < (p-r-2)/(n-p+2)$, $\underline{a}_{i,LS}$ is the LS estimator of $\underline{\alpha}_i$, i=1, 2, and $F_2^* = \underline{a}_{2,LS}' \Lambda_2 \underline{a}_{2,LS} / SS$, $\Lambda_2 = \text{Diag}(\lambda_{r+1}, \ldots, \lambda_p)$. It can be easily shown that $MSE(\underline{a}_{SC}) < MSE(\underline{a}_{LS})$. Moreover, the test statistic under the null hypothesis, $H_{02}: \underline{\alpha}_2 = \underline{0}$

$$F_2' = \frac{\underline{a}_{2,LS}' \Lambda_2 \underline{a}_{2,LS}}{(p-r)\, s^2}$$

$$= \frac{n-p}{p-r} F_2^* \tag{2.54}$$

has the F distribution with (p-r), (n-p) degrees of freedom and hence

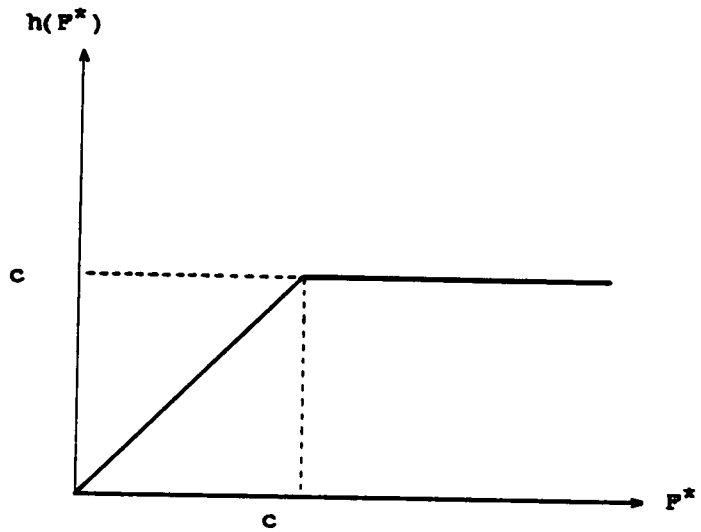

Figure 2.4 Special Form of $h(P^*)$

the Sclove estimator depends upon the test of hypothesis H_{02}: $\underline{\alpha}_2 = \underline{0}$ at a level of significance, ϕ_o, determined by c, that is,

$$\phi_o = \Pr\left[F_{(p-r),(n-p)} > \frac{n-p}{p-r}\, c\right]. \qquad (2.55)$$

Consequently, the Sclove estimator of $\underline{\alpha}_2$ is obtained as

$$\underline{a}_{2,SC} = \underline{0}, \qquad \text{if the hypothesis is not rejected,}$$

$$= \left[1 - \frac{c}{F_2^*}\right] \underline{a}_{2,LS}, \quad \text{if the hypothesis is rejected,}$$

in which case $(1 - c/F_2^*)$ is non-negative. It follows that (2.53) is

$$\underline{a}_{SC} = \begin{bmatrix} \underline{a}_{1,LS} \\ \hdashline \underline{0} \end{bmatrix}$$

when the hypothesis is not rejected and

$$\underline{a}_{SC} = \begin{bmatrix} \underline{a}_{1,LS} \\ \hdashline \left[1 - \frac{c}{F_2^*}\right] \underline{a}_{2,LS} \end{bmatrix}$$

when the hypothesis is rejected.

No one has found the optimal value of c. Therefore, in practice, there is difficulty in utilizing the Sclove estimator. Furthermore, the power of test is also damaged by the small eigenvalues of Λ_2 due to the multicollinearity problem. In other words, the F test may be unable to distinguish $\underline{\alpha}_2$ from $\underline{0}$.

However, the Sclove estimator is attractive since some shrinkage is applied only to the non-preferred variables. This concept, combined with the Bayesian interpreted James-Stein estimator and the principal components estimator, will be generalized later in this

paper.

2.4 Linear Transforms of the LS Estimator

Both the ridge estimator and the shrinkage estimators can be characterized by two different norms, the Euclidean norm and a design dependent norm, and hence are expressed in terms of a linear transform of the LS estimator.

2.4.1 Ridge and Shrinkage Estimators via Linear Transforms of $\underline{b}_{LS}$

Mayer and Willke (1973) define the class, T, of linear transforms of $\underline{b}_{LS}$ as follows: For an estimator of β, $\underline{b}$, to be contained in T, then $\underline{b} = A\underline{b}_{LS}$, for some p×p matrix A. Note that T contains, of course, the LS estimator and all the biased estimators discussed so far.

Since the residual sum of squares associated with $\underline{b}$ for fixed A is

$$SS(\underline{b}) = (\underline{y} - XA\underline{b}_{LS})'(\underline{y} - XA\underline{b}_{LS})$$

$$= (\underline{y} - X\underline{b}_{LS})'(\underline{y} - X\underline{b}_{LS}) + \underline{b}_{LS}'(A-I)'(X'X)(A-I)\underline{b}_{LS}, \qquad (2.56)$$

Mayer and Willke consider a subclass of estimators, $\underline{b}$, subject to

$$\underline{b}_{LS}'(A-I)'(X'X)(A-I)\underline{b}_{LS} = q$$

or, equivalently,

$$(A\underline{b}_{LS} - \underline{b}_{LS})'(X'X)(A\underline{b}_{LS} - \underline{b}_{LS}) = q, \qquad (2.57)$$

for any fixed constant q. The matrix A can be chosen to minimize a norm of $A\underline{b}_{LS}$. It was shown that if the Euclidean norm of $A\underline{b}_{LS}$ is used, then

the minimum Euclidean norm estimator, denoted by $\underline{b}_E$, in the equivalence class which satisfies (2.57) is just the ridge estimator. That is,

$$\text{Min } \underline{b}_{LS}'A'A\,\underline{b}_{LS}, \text{ subject to (2.57)},$$

yields $\underline{b}_E = A_E\underline{b}_{LS}$, where $A_E = \left[k(X'X)^{-1} + I\right]^{-1}$, $k > 0$. Note that

$$\begin{aligned}\underline{b}_E &= A_E\underline{b}_{LS}\\ &= \left[\left[X'X\right]^{-1}\left[kI + (X'X)\right]\right]^{-1}(X'X)^{-1}X'\underline{y}\\ &= (X'X + kI)^{-1}X'\underline{y}.\end{aligned}$$

In contrast, the design dependent norm of $\underline{b}$

$$\underline{b}_{LS}'A'(X'X)\,A\,\underline{b}_{LS}$$

gives a shrinkage estimator which is the minimum design norm estimator, denoted by $\underline{b}_D$, as

$$\underline{b}_D = A_D\underline{b}_{LS},$$

where $A_D = \delta I$, $0 \leq \delta \leq 1$. Thus, both the ridge and the shrinkage estimators are minimum norm estimators with respect to appropriate norms in the equivalence class which satisfies (2.57).

However, the class of linear transforms of the LS estimator, defined by Mayer and Willke, is unnecessarily too broad so that it may be reasonable to reduce the class by using the canonical model. Thus, in the following section, a class of linear transforms of $\underline{a}_{LS}$, which provides the restricted class of linear transforms of $\underline{b}_{LS}$, will be proposed.

2.4.2 Confined Linear Transforms of $\underline{b}_{LS}$

The class, T, defined in the previous section may be unnecessarily large since the defining matrix A is arbitrary. In order to reduce this class to a subclass which contains all the plausible biased estimators designed to remedy the multicollinearity problem, consider the canonical form of the model (2.14). Since the LS estimators of the α_j are uncorrelated, define a class of linear transforms of $\underline{a}_{LS}$ whose typical member is of the form,

$$\underline{a} = D\underline{a}_{LS}, \tag{2.58}$$

where D is a diagonal matrix and, for some $\underline{b} \in T$,

$$\underline{a} = V'\underline{b}, \tag{2.59}$$

where V is the matrix of the eigenvectors of $X'X$. Then, from (2.58) and (2.59), the subclass of T, T_R, consists of

$$\begin{aligned} \underline{b} &= VD\underline{a}_{LS} \\ &= VDV'\underline{b}_{LS}. \end{aligned} \tag{2.60}$$

Therefore, an estimator of $\underline{\beta}$ in the subclass is

$$\underline{b} = A\underline{b}_{LS},$$

where $A = VDV'$. In other words, for the restricted class T_R, a $p \times p$ matrix A is required to be such that

$$V'A\,V = D. \tag{2.61}$$

The condition (2.61) can be checked for A_E and A_D in the previous

section. That is, for example, for A_E

$$V'A_E V = V'\left[k(X'X)^{-1} + I\right]^{-1} V$$

$$= \left[kV'(X'X)^{-1} V + I\right]^{-1}$$

$$= \left[k\Lambda^{-1} + I\right]^{-1},$$

which is a diagonal matrix whose j^{th} element is $\lambda_j/(\lambda_j + k)$.

When this confined class is considered with respect to $\underline{a}_{LS}$, it is equivalent to the class of linear transforms of $\underline{a}_{LS}$, (2.58), which is large enough to deal with the multicollinearity problem. Therefore, with some constraints on the diagonal matrix D, fractional principal components analysis which provides a general approach to the biased estimators will be extensively considered in Chapter 4.

CHAPTER THREE

PRINCIPAL COMPONENTS ANALYSIS

Principal components regression has been used as an alternative estimation technique since Kendall (1957) recognized the potential usefulness of principal components to regression analysis. Despite some controversial deletion criteria it has been used frequently, especially in econometric models, because of its simplicity. The performance of the principal components (PC) estimator is discussed and, in addition, a new deletion criterion will be suggested in this chapter.

3.1 Principal Components Estimation

After Kendall's demonstration of the use of PC, Massy (1965) made a major contribution to principal components analysis, pointing out the necessity for this methodology in dealing with models with highly multicollinear regressors or with a large number of regressors.

3.1.1 Principal Components Estimator

The principal components regression model is the reparametrization of the standard MLR model as seen in (2.1), namely,

$$\begin{aligned} \underline{y} &= X\underline{\beta} + \underline{\epsilon} \\ &= XVV'\underline{\beta} + \underline{\epsilon} \\ &= Z\underline{\alpha} + \underline{\epsilon}, \end{aligned} \tag{3.1}$$

where $V = [\underline{v}_1, \ldots, \underline{v}_p]$, $\underline{v}_j$ is the j^{th} eigenvector of $X'X$, $Z = XV$, and

$\underline{\alpha} = V'\underline{\beta}$. Note again that all the explanatory variables are centered and scaled. The columns of Z, $\underline{z}_1, \ldots, \underline{z}_p$, are called the principal components. Without loss of generality, the principal components are ordered so that $\underline{z}_j$ is the j^{th} PC associated with the j^{th} largest eigenvalue of $X'X$. Thus, $V'V = VV' = I$ and $Z'Z = \Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_p)$, where the eigenvalues are ordered, i.e., $\lambda_1 \geq \ldots \geq \lambda_p$.

The PC estimation procedure simply amounts to deleting some principal components from the model (3.1) and applying LS method to the remaining components. That is, the model (3.1) is rewritten in the partitioned form as:

$$\underline{y} = \left[Z_1 \vdots Z_2\right] \begin{bmatrix} \underline{\alpha}_1 \\ \cdots \\ \underline{\alpha}_2 \end{bmatrix} + \underline{\epsilon}$$

$$= Z_1\underline{\alpha}_1 + Z_2\underline{\alpha}_2 + \underline{\epsilon}$$

where $Z_1 = [\underline{z}_1, \ldots, \underline{z}_r]$, $Z_2 = [\underline{z}_{r+1}, \ldots, \underline{z}_p]$, $\underline{\alpha}_1 = (\alpha_1, \ldots, \alpha_r)'$ and $\underline{\alpha}_2 = (\alpha_{r+1}, \ldots, \alpha_p)'$. Note that the columns of Z_2 represent the deleted principal components. Then, in the restricted model,

$$\underline{y} = Z_1\underline{\alpha}_1 + \underline{\epsilon}, \tag{3.2}$$

$\underline{\alpha}_1$ is estimated by LS method as $\underline{a}_{1,LS} = (Z_1'Z_1)^{-1} Z_1'\underline{y}$ and hence the PC estimator of $\underline{\beta}$ is

$$\underline{b}_{PC} = V_1\underline{a}_{1,LS}$$

$$= V_1(Z_1'Z_1)^{-1} Z_1'\underline{y}, \tag{3.3}$$

where $V_1 = [\underline{v}_1, \ldots, \underline{v}_r]$.

The PC estimator $\underline{b}_{PC}$ can be described, equivalently, in the form

$$\underline{b}_{PC} = V\underline{a}_{PC}$$

$$= VF\underline{a}_{LS}, \tag{3.4}$$

where $\underline{a}_{PC} = F\underline{a}_{LS}$ and $F = \text{Diag}(f_1, \ldots, f_p)$, $f_1 = \ldots = f_r = 1$, $f_{r+1} = \ldots = f_p = 0$. This is an attractive form that will be useful for further developments.

The mean vector of $\underline{b}_{PC}$ is

$$E(\underline{b}_{PC}) = E\left[V_1(Z_1'Z_1)^{-1}Z_1'\underline{y}\right]$$

$$= V_1\underline{\alpha}_1$$

$$= V_1V_1'\underline{\beta}$$

$$= \underline{\beta} - V_2V_2'\underline{\beta}, \tag{3.5}$$

where $V_2 = [\underline{v}_{r+1}, \ldots, \underline{v}_p]$, and the covariance matrix of $\underline{b}_{PC}$ is

$$\text{Var}(\underline{b}_{PC}) = \sigma^2 V_1\Lambda_1^{-1}V_1', \tag{3.6}$$

where $\Lambda_1 = \text{Diag}(\lambda_1, \ldots, \lambda_r)$. Thus, the PC estimator is a biased estimator, while the variance of the j^{th} PC estimator of $\underline{\beta}$,

$$\text{Var}(b_{j,PC}) = \sigma^2 \sum_{k=1}^{r} \frac{v_{jk}^2}{\lambda_k}$$

becomes much smaller than $\text{Var}(b_{j,LS})$ [see (2.10)] when the eigenvalues $\lambda_{r+1}, \ldots, \lambda_p$ are small.

It is also worth noting that the PC estimator is equivalent to the restricted LS estimator obtained by imposing the restrictions

$$V_2'\underline{\beta} = \underline{0} \text{ (equivalently, } \underline{\alpha}_2 = \underline{0}) \qquad (3.7)$$

[see, for example, Johnson, Reimer, and Rothrock (1973)]. These restrictions can be justified by testing hypothesis H_{02}: $\underline{\alpha}_2 = 0$. The non-centrality parameter is, from (2.20),

$$\eta_2 = \sum_{j=r+1}^{p} \frac{\lambda_j \alpha_j^2}{2\sigma^2}$$

so that η_2 may be damaged when the eigenvalues $\lambda_{r+1}, \ldots, \lambda_p$ are small. Therefore, in such a case, the F test may be unable to distinguish $\underline{\alpha}_2$ from $\underline{0}$ because of the low power of the F test. In other words, the model (3.2) is likely to be used by unduly deleting the principal components associated with $\underline{\alpha}_2$. Therefore, the criteria used to determine which components are deleted are important issues and will be discussed in more detail.

3.1.2 Deletion Criteria

Massy (1965) suggested two criteria for deleting components in principal components analysis:

(a) delete the components that are relatively unimportant as predictors of the original regressors (X), that is, the components having small eigenvalues should be dropped. This method will be referred to as the <u>eigenvalue criterion</u>.

(b) delete the components that are relatively unimportant as predictors of the response variable (Y), that is, the components having small values of the correlation

between Y and the components should be dropped. This method will be referred to as the t-value criterion.

Massy, furthermore, expected that the two criteria are likely to lead to different results since the response variable Y does not have to be highly correlated with principal components having large eigenvalues. Many other authors including Jeffers (1967), McCallum (1970), Marquardt (1970), Mittelhammer and Baritelle (1977), Pomby and Hill (1978), among others have discussed the above criteria.

The eigenvalue criterion, which is usually used in PC analysis, requires that the principal components to be deleted in principal components analysis be those corresponding to small eigenvalues. The major issue is to determine which eigenvalues are small. Pidot (1969) suggested that components be deleted when their associated eigenvalues are less than one. That is, he regarded 1 as an average value of eigenvalues ($\Sigma\lambda_j/p = 1$) and thought that a component should be dropped if its associated eigenvalue is less than the average in size. Similarly, Marquardt (1970) recommended the criterion of choosing the minimum value r for which

$$\sum_{j=p}^{p-r} \frac{\lambda_j}{p} < \omega,$$

where ω might be selected in the range 10^{-1} to 10^{-7}. Furthermore, Pomby and Hill (1978) proposed the criterion based on the percentage of variance reduction from deleting the (p-r) principal components corresponding to the (p-r) smallest eigenvalues. The percentage of variance reduction (PVR) is

$$PVR = \frac{1}{\sum_{j=1}^{p} \lambda_j^{-1}} \; tr\left[\sum_{j=r+1}^{p} \frac{\underline{v}_j \underline{v}_j'}{\lambda_j} \right] \times 100$$

so that the number of principal components retained in the analysis, r, can be chosen where PVR drops drastically. However, the PVR criterion is also dependent upon the subjective decision based on the percentage change. In general, it is difficult to choose an appropriate number of components to be retained in the analysis unless $\lambda_r \gg \lambda_{r+1}$ and $\lambda_{r+1} \approx 0$. In other words, when the last (p-r) condition indexes [see page 8] are large (say, > 30), the last (p-r) principal components may be deleted. Even though the eigenvalue criterion seems to be simple to use, Mittelhammer and Baritelle (1977) indicated that there is no control on the bias introduced to the resulting estimator when this criterion is used to delete components. The PC estimator of $\underline{\alpha}$, resulting from deleting the single component, $\underline{z}_j$, will be denoted by $\underline{a}_{PC(jth)}$ in this thesis.

A second criterion, called the t-value criterion, is based on testing the statistical significance of the hypotheses, H_{0j}: $\alpha_j = 0$, j=1, ..., p. The test statistics, from (2.19), are

$$t_j' = \frac{a_{j,LS}}{s/\sqrt{\lambda_j}}, \qquad j=1, \ldots, p. \tag{3.8}$$

When t_j' is small enough to fail to reject H_{0j} its associated principal component will be deleted. Thus, the PC estimator of $\underline{\alpha}$, resulting from this deletion criterion, denoted by $\underline{a}_{PCT}$, is of the form,

$$\underline{a}_{PCT} = \left[I - I_{(-d_j, d_j)}(a_{j,LS}) \right] \underline{a}_{LS}, \tag{3.9}$$

where $I_{(-d_j,d_j)}(a_{j,LS})$ is a pxp diagonal matrix of indicator functions,

$d_j = t^* s/\sqrt{\lambda_j}$, and t^* is the critical value for each test. When the common significance level ϕ_o is given, the critical value t^* is determined from the t-table with the (n-p) degrees of freedom.

However, since the power of test η_j [see (2.21)] is low for a small λ_j, the t-value criterion may delete too many principal components and give the misleading impression that the PC_t estimator is an improvement. Moreover, Mittelhammer and Baritelle, investigating the bias and the covariance matrix of $\underline{b}_{PCT}$, pointed out that it is not necessarily true that $Var(b_{j,PCT}) \leq Var(b_{j,LS})$ when the t-value criterion is used with a fixed significance level.

In addition to the above two criteria, McCallum (1970) proposed another method which deletes a principal component, $\underline{z}_j$, if deleting $\underline{z}_j$ reduces the MSE of $b_{k,LS}$, k=1, ..., p. In other words, the j^{th} component $\underline{z}_j$ is deleted if

$$MSE[b_{k,PC(jth)}] < MSE[b_{k,LS}], \qquad k=1, \ldots, p, \tag{3.10}$$

where $b_{k,PC(jth)}$ is the PC estimator of β_k when $\underline{z}_j$ is deleted. The MSE of $b_{k,PC(jth)}$ is

$$MSE[b_{k,PC(jth)}] = \sigma^2 \sum_{\ell \neq j} \frac{v_{k\ell}^2}{\lambda_\ell} + \left[E[b_{k,PC(jth)}] - \beta_k\right]^2. \tag{3.11}$$

From the equation (3.5),

$$E[\underline{b}_{PC(jth)}] = \underline{\beta} - \underline{v}_j \underline{v}_j' \underline{\beta},$$

so that (3.11) can be rewritten as:

$$MSE\left[b_{k,PC(jth)}\right] = \sigma^2 \sum_{\ell \neq j} \frac{v_{k\ell}^2}{\lambda_\ell} + \left[v_{kj} \sum_{\ell=1}^{p} v_{\ell j}\, \beta_\ell\right]^2 \qquad (3.12)$$

Therefore, comparing with $MSE(b_{k,LS})$, if

$$MSE(b_{k,LS}) - MSE\left[b_{k,PC(jth)}\right] = \sigma^2 \frac{v_{kj}^2}{\lambda_j} - \left[v_{kj} \sum_{\ell=1}^{p} v_{\ell j}\, \beta_\ell\right]^2$$

$$= v_{kj}^2\left[\frac{\sigma^2}{\lambda_j} - \alpha_j^2\right] \qquad (3.13)$$

is nonnegative, then the deletion of $\underline{z}_j$ improves the resulting MSE. McCallum also shows that, for $p = 2$, $MSE\left[b_{1,PC(2nd)}\right] \leq MSE(b_{1,LS})$ if and only if $(n/2\sigma^2)(\beta_2 - \beta_1)^2 < (1 - r_{12})^{-1}$, where r_{12} is the correlation coefficient between the two regressors and uses $(b_{2,LS} - b_{1,LS})$ and s^2 to assess the usefulness of the PC estimator. However, this is not recommended because $(b_{2,LS} - b_{1,LS})$ may be a highly unreliable estimator of $(\beta_2 - \beta_1)$ in the presence of multicollinearity [see Vinod (1976)].

The first two criteria, the eigenvalue criterion and the t-value criterion, would yield nearly identical results if the following condition is met:

$$R_{yz_1}^2 \geq R_{yz_2}^2 \geq \ldots \geq R_{yz_p}^2 , \qquad (3.14)$$

where $R_{yz_j}^2$ is the squared sample correlation coefficient between $\underline{y}$ and $\underline{z}_j$. Note that $R_{yz_j}^2 = \lambda_j a_{j,LS}^2$, $j=1, \ldots, p$ and $\lambda_1 \geq \ldots \geq \lambda_p$. In such a case, the t test statistics (3.8) have the same order as the eigenvalues so that the two criteria coincide when an appropriate significance level is chosen. However, the condition (3.14) may not, in general, hold for the canonical model (3.1), since the order of

$R_{yz_j}^2$ depends mainly upon the order of the unknown parameters, γ_j, j=1, ..., p. Note that, from (2.17),

$$E(\lambda_j a_{j,LS}^2) = \gamma_j^2 + \sigma^2, \qquad j=1, \ldots, p. \tag{3.15}$$

When the condition (3.14) is violated, neither of the two criteria should be used as a deletion criterion. Indeed, a criterion that also considers the orientation of $\underline{\gamma}$ will be followed. Furthermore, after extending the MSE properties of the PC estimators, the various orientations of the parameter vector will be investigated in the following sections.

3.1.3 New Criterion: Unbiased Optimal Deletion Criterion

The PC estimator results from applying the LS method to the reduced set of orthogonally transformed variables, that is, the principal components. Since it is a biased estimator, it is necessary to consider the MSE of $\underline{b}_{PC}$. The MSE matrix of $\underline{b}_{PC}$ is, from (2.23)

$$\begin{aligned} \text{MtxMSE}(\underline{b}_{PC}) &= E(\underline{b}_{PC} - \underline{\beta})(\underline{b}_{PC} - \underline{\beta})' \\ &= \sigma^2 V_1 \Lambda_1^{-1} V_1' + V_2 \underline{\alpha}_2 \underline{\alpha}_2' V_2'. \end{aligned} \tag{3.16}$$

In order for the MSE of $\underline{b}_{PC}$ to be smaller than that of $\underline{b}_{LS}$, the difference, $\text{MtxMSE}(\underline{b}_{LS}) - \text{MtxMSE}(\underline{b}_{PC})$ must be positive semi-definite [Toro-Vizcarrondo and Wallace (1968)]. Since the difference of the two is

$$\begin{aligned} \text{MtxMSE}(\underline{b}_{LS}) - \text{MtxMSE}(\underline{b}_{PC}) &= \sigma^2 V_2 \Lambda_2^{-1} V_2' - V_2 \underline{\alpha}_2 \underline{\alpha}_2' V_2' \\ &= \sigma^2 V_2 \left[\Lambda_2^{-1} - \sigma^{-2} \underline{\alpha}_2 \underline{\alpha}_2'\right] V_2', \end{aligned}$$

it suffices to show that $[\Lambda_2^{-1} - \sigma^{-2}\ \underline{\alpha}_2\underline{\alpha}_2']$ is positive semi-definite, that is,

$$\underline{\ell}'[\Lambda_2^{-1} - \sigma^{-2}\ \underline{\alpha}_2\underline{\alpha}_2']\ \underline{\ell}' \geq 0, \tag{3.17}$$

for any nonzero px1 vector $\underline{\ell}$. Since the relationship (3.17) can be rewritten as

$$\frac{\underline{\ell}'\underline{\alpha}_2\underline{\alpha}_2'\underline{\ell}}{\underline{\ell}'\Lambda_2^{-1}\underline{\ell}} \leq \sigma^2,$$

by the Cauchy-Schwarz inequality, the necessary and sufficient condition is given by

$$\underline{\alpha}_2'\Lambda_2\ \underline{\alpha}_2 \leq \sigma^2, \tag{3.18}$$

or, equivalently,

$$\sum_{j=r+1}^{p} \lambda_j\alpha_j^2 \leq \sigma^2. \tag{3.19}$$

The condition (3.18) or (3.19) depends upon the number of deleted components, the length of $\underline{\alpha}_2$, and the magnitude of σ^2. In other words, if the deleted components have the squared length of the associated parameters in γ_j's exceed σ^2, the MSE of the PC estimator will get larger than that of the LS estimator. In such cases, the primary contribution to the MSE for the PC estimator is due to the "squared-bias" term since bias($\underline{b}_{PC}$) = $V_2V_2'\underline{\beta} = V_2\underline{\alpha}_2$ and, thus, the "squared bias" = $\underline{\alpha}_2'\Lambda_2\ \underline{\alpha}_2 = \underline{\gamma}_2'\underline{\gamma}_2$. Similarly, if σ^2 is decreased then the MSE of the LS estimator tends to be reduced so that the existence of bias will be a burden on the MSE of the PC estimator. Note that this criterion does not have to delete the component associated with the smallest eigenvalue. There-

fore, the condition (3.19) can be rewritten as

$$\sum_{j \in D} \gamma_j^2 \leq \sigma^2, \tag{3.20}$$

where D is the set of the indexes whose principal components are supposed to be deleted. Since the unbiased estimators for the unknown parameters, γ_j^2, are, from (3.15),

$$\hat{\gamma_j^2} = \lambda_j a_{j,LS}^2 - s^2, \tag{3.21}$$

the substitution of (3.21) into (3.20) results in

$$\sum_{j \in D} (\lambda_j a_{j,LS}^2 - s^2) \leq s^2. \tag{3.22}$$

Thus, the inequality (3.22) can be used as a deletion criterion, named the unbiased optimal deletion criterion. That is, based on this criterion, delete those principal components, $\underline{z}_j$, $j \in D$ for which (3.22) is satisfied. The usefulness of the unbiased optimal deletion criterion will be seen in Chapter 5.

As mentioned in Section 2.1.4, if the condition (3.20) holds, then $MSE(\underline{b}_{PC}) \leq MSE(\underline{b}_{LS})$. It can be easily shown by the graphic method for $p = 2$. The PC estimator with $\underline{z}_2$ being deleted is compared with the LS estimator in Figure 3.1. Assuming $\lambda_1 >> \lambda_2$, the PC(2nd) estimator is closer to $\underline{\alpha}$ than the LS estimator, that is, $MSE(\underline{a}_{PC}) \leq MSE(\underline{a}_{LS})$ [equivalently, $MSE(\underline{b}_{PC}) \leq MSE(\underline{b}_{LS})$].

Since the necessary and sufficient condition (3.20) is dependent upon the magnitudes of the γ_j^2 for $j \in D$ and σ^2, the orientation of the unknown parameter vector and the size of σ are closely related to the evaluation of principal components analysis. The situations for which

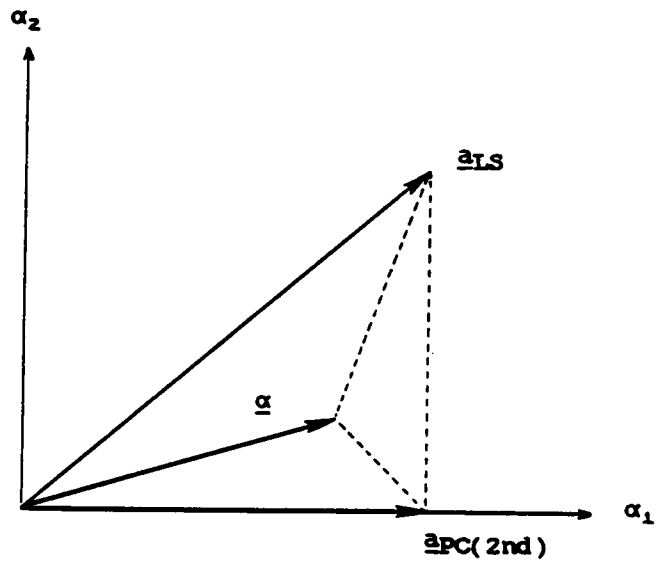

Figure 3.1 PC Estimator with Smaller MSE than LS Estimator

the usual PC estimator resulting from using either the eigenvalue criterion or the t-value criterion does not improve over the LS estimator will be introduced in the following section.

3.1.4 Problems of Principal Components Estimators

It has been mentioned in Section 3.1.2 that the order of the unknown parameters γ_j is also important in determining which component should be deleted in principal components regression. Since the magnitude of γ_j is determined by $\sqrt{\lambda_j}$ and α_j, the magnitudes of the α_j are sometimes crucial to achieve a proper PC estimator. For the simple case, p = 2, the problem related to the PC estimator is delineated in Figure 3.2. Note that the PC(1st) estimator has the smallest distance from $\underline{\alpha}$ among the three estimators: the LS, PC(2nd), and PC(1st) estimators.

Consider, now, the orientation of the unknown parameter vector in terms of $\underline{\alpha}$ or $\underline{\beta}$. Since $\underline{\alpha} = V'\underline{\beta}$, the squared length of $\underline{\alpha}$, denoted by a fixed constant C, is

$$\underline{\alpha}'\underline{\alpha} = \underline{\beta}'\underline{\beta} = C. \tag{3.23}$$

If the values of $\alpha_{r+1}, \ldots, \alpha_p$ which are associated with small eigenvalues $\lambda_{r+1}, \ldots, \lambda_p$ are large in magnitude relative to the others, then the PC estimator, through deletion of $\underline{z}_{r+1}, \ldots, \underline{z}_p$, will incur a large MSE due to a large contribution to the "squared-bias" component.

For illustrative purposes, consider the special case when p = 2. The matrix of the correlation form of the two regressors is

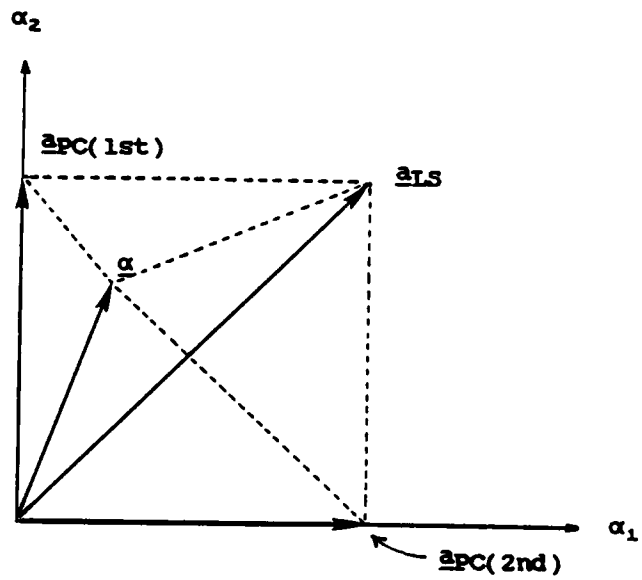

Figure 3.2 Smallest MSE$\left[\underline{a}_{PC(1st)}\right]$ Case

$$X'X = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix},$$

where r_{12} is the inner product of $\underline{x}_1$ and $\underline{x}_2$. Without loss of generality, r_{12} is assumed to be positive. The eigenvalues of $X'X$ are $1 + r_{12}$ and $1 - r_{12}$ with the matrix of the eigenvectors V, where

$$V = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}.$$

Assume also that r_{12} is close to 1, that is, there exists severe multicollinearity in this simple MLR model. Then, the usual PC estimator, based on the eigenvalue criterion, may be obtained by deleting the second principal component. Furthermore, the orientations of the unknown parameter vectors in terms of $\underline{\alpha}$ and $\underline{\beta}$, which satisfy (3.23), are described as the circle in Figure 3.3. Then the specific $\underline{\alpha}$ depicted in Figure 3.3 may cause the regular PC estimator, PC(2nd), to be inappropriate in that $MSE[PC(2nd)] > MSE[PC(1st)]$ since the absolute value of α_2 is relatively large compared to that of α_1. However, in Figure 3.3 the exact range of the orientations for which the regular PC estimator is improper, cannot be found.

In order to investigate the performances of the PC estimator, the corresponding orientation in terms of $\underline{\gamma}$ can be obtained by applying the SVD. Since $\underline{\gamma} = \Lambda^{1/2}\underline{\alpha}$, the orientation of $\underline{\gamma}$ corresponding to $\underline{\alpha}$ lies on the ellipse $\underline{\gamma}'\Lambda^{-1}\underline{\gamma} = C$ in Figure 3.4. Note that the ellipse is associated with the circle in Figure 3.3, and the dotted lines represent $\gamma_j = \pm \sigma$, $j=1, 2$.

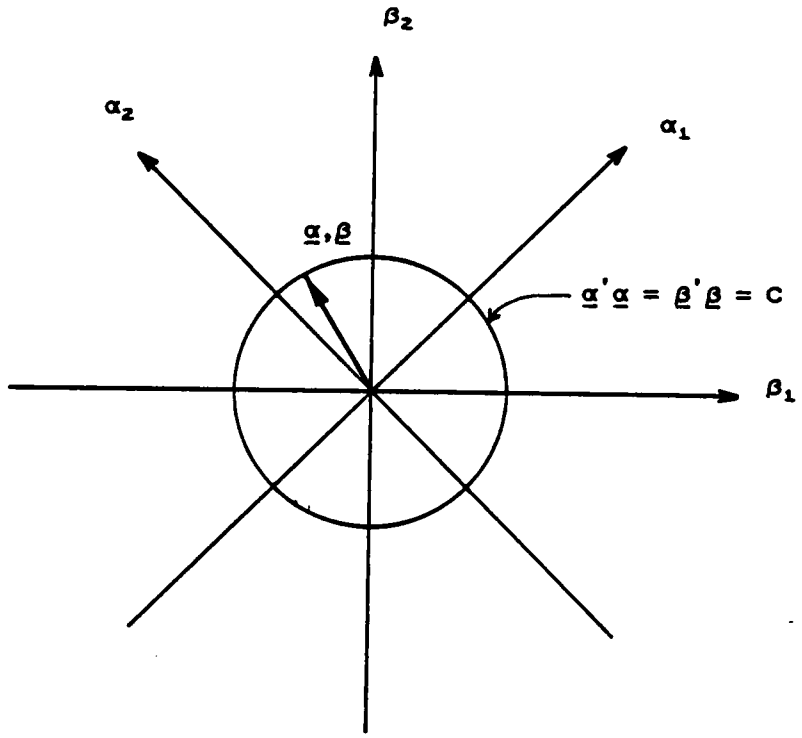

Figure 3.3 Specific Orientations, $\underline{\alpha}$ and $\underline{\beta}$

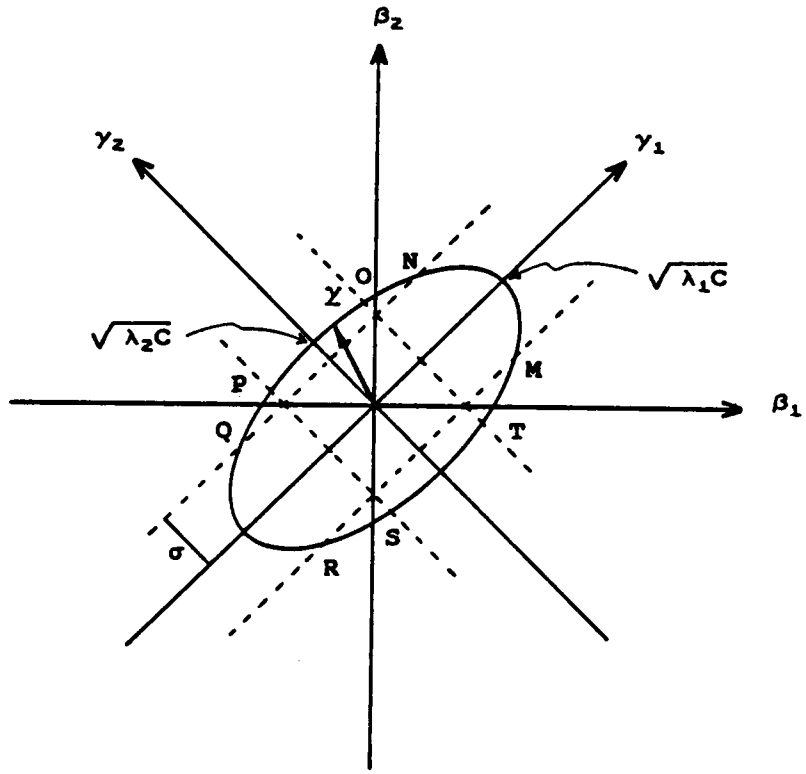

Figure 3.4 Ellipse with respect to γ: $\sigma < \sqrt{\lambda_2 C}$

Assume that σ is fixed such that

$$\sigma < \sqrt{\lambda_2 C}, \tag{3.24}$$

where $\pm\sqrt{\lambda_2 C}$ are the points of the ellipse at $\gamma_1 = 0$. Then, three cases of the condition (3.20) can be considered (see Table 3.1). First, when $\lambda_2\alpha_2^2 = \gamma_2^2 < \sigma^2 < \gamma_1^2 = \lambda_1\alpha_1^2$ (a), which can be described by the chord segments MN and QR in Figure 3.3, the deletion of $\underline{z}_2$ improves estimation of $\underline{\alpha}$ in that $MSE[\underline{a}_{PC(2nd)}] < MSE(\underline{a}_{LS})$. In other words, from (3.20), the usual PC estimator, obtained through the deletion of the component associated with the smallest eigenvalue, denoted by PC(2nd), will outperform the LS estimator in terms of MSE.

Secondly, if $\lambda_1\alpha_1^2 < \sigma^2 < \lambda_2\alpha_2^2$ (b), then the PC(2nd) estimator cannot have MSE as small as the LS estimator. The condition (3.20) requires the deletion of the first principal component even though $\lambda_1 > \lambda_2$. Since $\gamma_1^2 < \sigma^2 < \gamma_2^2$, the order of $R_{yz_j}^2$ will, on the average, be as $R_{yz_1}^2 < R_{yz_2}^2$ in this case. Thus any orientation of $\underline{\gamma}$ along the chord segments OP and ST leads to the improvement of the PC(1st) estimator over the LS estimator in terms of MSE.

Finally, at the other chord segments (NO, PQ, RS, and TM), the LS estimator has smaller MSE than any PC estimator since $\lambda_1\alpha_1^2 > \sigma^2$ and $\lambda_2\alpha_2^2 > \sigma^2$ so that the deletion of any principal component cannot guarantee smaller MSE. This result can be expected by the fact that the effect of multicollinearity may be eliminated by a relatively small σ^2. Note that as σ is decreased, the range of the orientations where PC estimation is preferred over LS estimation is also decreased. In other words, in Figure 3.4, the lengths of segments, NO, PQ, RS, and

Table 3.1 Condition (3.20) with respect to $\underline{\gamma}$

Condition	Region	Proper Decision	Result
(a) $\gamma_2 < \sigma < \gamma_1$	MN and QR	delete $\underline{z}_2$	MSE[PC(2nd)] < MSE(LS)
(b) $\gamma_1 < \sigma < \gamma_2$	OP and ST	delete $\underline{z}_1$	MSE[PC(1st)] < MSE(LS)
(c) $\gamma_1 > \sigma$ and $\gamma_2 > \sigma$	NO,PQ,RS, and TM	none	smaller MSE(LS)

TM, for which the LS estimator outperforms the PC estimator will be increased as σ is decreased.

In fact, only when the true parameter vector, in terms of $\underline{\gamma}$, lies on the limited chord segments, MN and QR, can the regular PC estimator, that is, PC(2nd) be used to combat the multicollinearity problem. Therefore, it is essential to examine the orientation of the parameter vector along the ellipse.

Furthermore, the magnitude of σ also affects the performances of the PC estimator. As long as σ is in the interval,

$$\sqrt{\lambda_2 C} < \sigma < \sqrt{\lambda_1 C} \quad \text{[see Figure 3.5]}, \tag{3.25}$$

there is no need to suspect the capabiity of the PC(2nd) estimator since any orientation $\underline{\gamma}$ satisfies the condition $\gamma_2^2 < \sigma^2$. That is, first of all, the second principal component should always be deleted under (3.25). In addition, if $\gamma_1^2 + \gamma_2^2 < \sigma^2$ then the trivial result that the PC estimator is $\underline{0}$ from (3.20) can be obtained by deleting both principal components. This case is depicted by the chords, KL and MN, in Figure 3.5. Therefore, for only the orientations along LM and NM, the PC(2nd) is recommended, and for the other orientations along the ellipse the PC(2) should be used, where the PC(2) refers to the PC estimator resulted from deleting the last two principal components. On the other hand, it is easily seen in Figure 3.6 that the trivial result is achieved when the length of the true parameter vector, which is $\sqrt{\underline{\alpha}'\underline{\alpha}}$, is shorter than that of the difference between the true parameter vector and the LS estimated parameter vector. In fact, in Figure 3.6 the length of $\underline{\alpha}$ is shorter than any length of $\underline{a}_{LS} - \underline{\alpha}$, $\underline{a}_{PC(2nd)} - \underline{\alpha}$, and

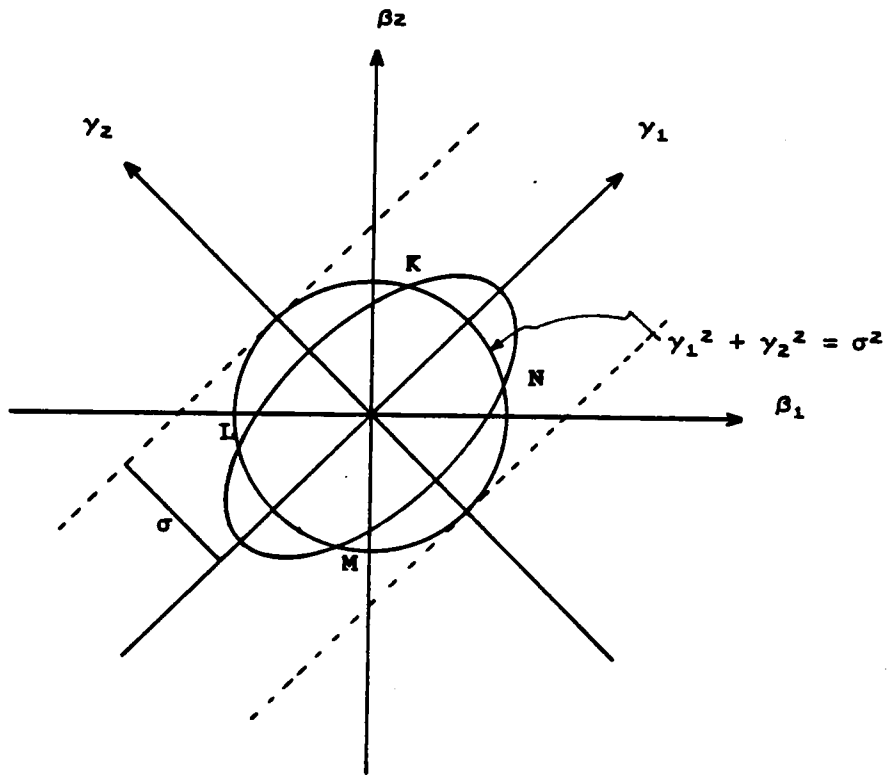

Figure 3.5 Ellipse with respect to $\underline{\gamma}$: $\sqrt{\lambda_2 C} < \sigma < \sqrt{\lambda_1 C}$

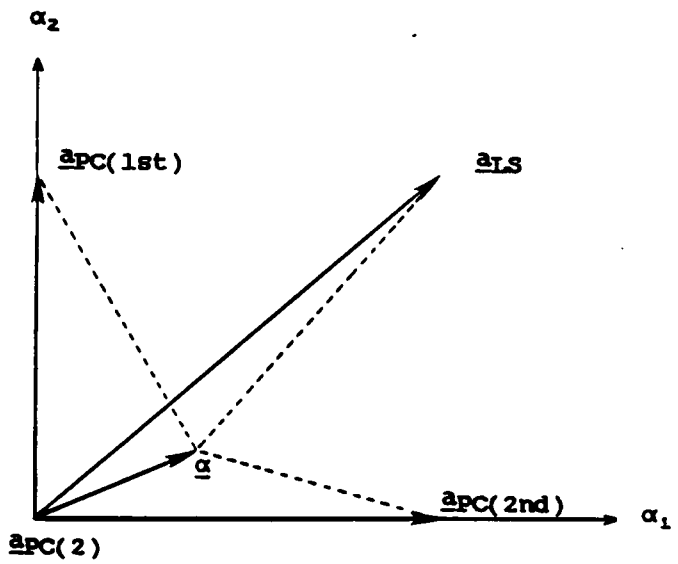

Figure 3.6 Trivial PC Estimator $\underline{0}$

$\underline{a}_{PC(1st)} - \underline{\alpha}$. It goes without saying that if

$$\sqrt{\lambda_1 C} < \sigma \tag{3.26}$$

then the PC estimator should always be $\underline{0}$.

Thus, the performances of the PC estimators depend upon not only the orientation of the unknown parameter vector but also the magnitude of σ. However, in practice, the true parameters are unknown and unbiased estimators would have to be substituted for them. The unbiased estimates of γ_j^2, $\hat{\gamma_j^2}$, j=1, 2 can be utilized and the unbiased estimate of $\underline{\alpha}'\underline{\alpha}$ (= C), denoted by C_U, can be used to compare the magnitudes of s and $\sqrt{\lambda_j C_U}$. Note that, from (2.15), the unbiased estimate of C is

$$C_U = \underline{a}_{LS}'\underline{a}_{LS} - s^2 \sum_{j=1}^{2} \frac{1}{\lambda_j}. \tag{3.27}$$

Therefore, for example, if the estimated inequality of (3.24),

$$s < \sqrt{\lambda_2 C_U} \tag{3.28}$$

holds, the orders of $\hat{\gamma_1^2}$, $\hat{\gamma_2^2}$, and s^2 should be investigated to determine an appropriate principal component to be deleted as in (3.22).

It is important to compare the magnitudes of the γ_j^2 and σ to achieve a proper PC estimator which indeed improves the LS estimator in MSE. Thus, principal components regression should be carefully used when the multicollinearity problem exists in the MLR model. Furthermore, when appropriate information is available, the PC_t estimator with a common significance level, $\underline{a}_{PCT}$ in (3.9), can be adjusted and hence improved to be as good a PC estimator as the one suggested in this section.

3.1.5 The Adjusted PC_t Estimator

For some special situations, as discussed in the previous section with p = 2, where the regular PC estimator would not apparently perform well, the t-value criterion often fails to delete the appropriate principal component because of the deterioration of the F tests (or t tests). Therefore, for example, for the situations where the first principal component should be deleted and the second one should not, the PC_t estimator, based on a common significance level, can be adjusted by using different significance levels. The resulting estimator will coincide with the PC(1st) estimator.

In general, if some principal components are determined to be deleted by the unbiased optimal deletion criterion, that is, the set D is fixed, then the significance levels for the hypotheses, H_{0j}: $\alpha_j = 0$, for $j \in D$ can be chosen to be close to 0. On the other hand, the significance levels for $j \notin D$ are enhanced in order to confirm the retention of the principal components $\underline{z}_j$, $j \notin D$. For example, for the indexes $j \in D$, take ϕ_j from .0001 to .0005 and for $j \notin D$, choose ϕ_j from .4 to .6. Then, the t-value criterion adjusted by the different significance levels will delete the same principal components as the unbiased optimal criterion. The resulting PC_t estimator, termed the <u>adjusted PC_t estimator</u>, can be defined as:

$$\underline{a}_{APCT} = \left[I - I_{(-d_j, d_j)}(a_{j,LS}) \right] \underline{a}_{LS}, \tag{3.29}$$

where $d_j = t_j^* \, s/\sqrt{\lambda_j}$, $j=1, \ldots, p$, and the t_j^* are the critical values corresponding to ϕ_j.

It can be shown that the different significance levels, ϕ_j, control

the power of the test for each hypothesis. That is, as the significance level is decreased, the power of test is also decreased. Since the power of test for the test statistic $F_j' = (t_j')^2$ [see (3.8)] with the noncentrality parameter, $\eta_j = \lambda_j \alpha_j^2/2\sigma^2$ is damaged due to either small λ_j or small α_j, using low significance level forces the F test to be less able to distinguish α_j from 0 and hence, H_{0j} is more likely to be accepted. The use of high significance level, on the other hand, enhances the power of test and hence, the null hypothesis is more likely to be rejected.

The appropriate significance levels play the key role of adjusting the PC_t estimator. They can be determined subjectively, in practice, via the discussions in the previous section. The different significance levels will be used directly to define the so-called generalized PC_t estimator in Section 4.3.2.

While the problems discovered in the previous two sections will be investigated again in Chapter 5 with simulated data sets, modified PC estimators will be introduced which assign "weights" to the principal components in subsequent sections of this dissertation. One such method is the fractional rank estimator which is based on Marquardt's generalized inverse estimation method. This method is discussed in the next section.

3.2 Fractional Rank Estimation

In some cases it is questionable to decide whether a particular principal component should be used or not used in principal component regression. A compromise solution is the use of the modified method of

generalized inverse proposed by Marquardt (1970), which is known as the fractional rank (FR) estimator. The practical use of the FR estimator has been examined by Hocking, Speed, and Lynn (1976).

3.2.1 Fractional Rank Estimator

Marquardt (1970) suggested that the assumption of an integral number of deleted components may be too restrictive. He wrote that "inspection of the eigenvalue spectrum usually suggests that there is no one rank clearly assignable to the matrix X. Rather, there is a range of ranks that may be reasonable choices." Thus, if the assigned rank of X is assumed to be r^*, where $r \leq r^* \leq r+1$, then the generalized inverse of $X'X$, denoted by $(X'X)^+$, is

$$(X'X)^+ = V_1\Lambda_1^{-1}V_1' + \frac{r^* - r}{\lambda_{r+1}} \underline{v}_{r+1}\underline{v}_{r+1}', \tag{3.30}$$

where $V_1 = [\underline{v}_1, \ldots, \underline{v}_r]$. Hence the fractional rank estimator of $\underline{\beta}$ is defined as:

$$\underline{b}_{FR} = (X'X)^+ X'\underline{y}. \tag{3.31}$$

Since, in this situation, the matrix of the eigenvectors of $X'X$ is $V^* = [V_1 | \underline{v}_{r+1} | 0]$, where 0 is the $p \times (p-r-1)$ zero matrix, the FR estimator can also be written as:

$$\underline{b}_{FR} = V_1\underline{a}_{1,LS} + \underline{v}_{r+1}(r^* - r)a_{r+1,LS},$$

where $\underline{a}_{1,LS}' = (a_{1,LS}, \ldots, a_{r,LS})$. Furthermore, a useful and succinct form of the FR estimator of $\underline{\alpha}$ [Montgomery and Peck (1982) and Hocking (1976)] is given by

$$\underline{a}_{FR} = (1 - c)\, \underline{a}_{r,LS} + c\, \underline{a}_{r+1,LS}, \tag{3.32}$$

where $c = r^* - r$ and $\underline{a}_{j,LS}' = (a_{1,LS}, \ldots, a_{j,LS}, 0, \ldots, 0)$, for $j = r$ and $r+1$.

Thus, the FR estimator of $\underline{\alpha}$, $\underline{a}_{FR}$ is obtained by assigning the positive decimal value of the rank of X to the $(r+1)$st element of $\underline{a}_{LS}$, $a_{r+1,LS}$, while retaining the first r principal components in the analysis. Therefore, it seems to be reasonable because only the specific component in question, $\underline{z}_{r+1}$, receives a fraction rather than zero (delete the component) or one (retain the component).

However, in practice, use of $\underline{b}_{FR}$ requires that the values r and r^* be determined. Some suggestions are given in the next section.

3.2.2 Choices of r and c

By considering a general class of biased estimators, Hocking, Speed, and Lynn (1976) set up a framework for determining the values of r and c in computing the FR estimator. A class of biased estimators of $\underline{\alpha}$ can be defined as:

$$\tilde{\underline{a}} = \sum_{j=1}^{p} d_j\, \underline{a}_{j,LS}, \tag{3.33}$$

where d_j are constants to be determined and $\underline{a}_{j,LS}' = (a_{1,LS}, \ldots, a_{j,LS}, 0, \ldots, 0)$, $j=1, \ldots, p$. According to the definition of $\underline{a}_{j,LS}$, (3.33) is equivalent to

$$\tilde{\underline{a}} = E\underline{a}_{LS}, \tag{3.34}$$

where $E = \mathrm{Diag}(e_1, \ldots, e_p)$ and $e_j = \sum_{i=j}^{p} d_i$. For example, $\tilde{\underline{a}} = \underline{a}_{LS}$ if d_i,

..., d_p are chosen as 0, ..., 0, 1.

It is natural to consider the MSE of $\tilde{\underline{\alpha}}$,

$$MSE(\tilde{\underline{\alpha}}) = \sigma^2 \sum_{j=1}^{p} \frac{e_j}{\lambda_j} + \sum_{j=1}^{p} (1 - e_j)^2 \alpha_j^2. \tag{3.35}$$

Then different sets of the constraints on d_j's (or e_j's) yield various types of estimators. The FR estimator (3.32) requires the following set of the constraints:

$$d_r + d_{r+1} = 1 \quad \text{and} \quad d_j = 0, \text{ for } j \neq r, r+1 \tag{3.36}$$

Therefore, substituting (3.36) into (3.35) and minimizing it with respect to d_{r+1} results in

$$d_{r+1} = \frac{\lambda_{r+1}\alpha_{r+1}^2/\sigma^2}{\lambda_{r+1}\alpha_{r+1}^2/\sigma^2 + 1}, \tag{3.37}$$

which is equal to c in (3.32) since $e_{r+1} = d_{r+1}$.

Hocking, Speed, and Lynn also suggested a method for the choice of r. They recommended that r could be chosen as the value for which $MSE(\underline{a}_{FR})$ is minimized, subject to (3.36) and (3.37), with respect to r. The $MSE(\underline{a}_{FR})$ subject to (3.36) and (3.37) is

$$\begin{aligned} MSE(\underline{a}_{FR}) &= \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} + \sigma^2 d_{r+1}^2 \lambda_{r+1}^{-1} + \sum_{j=r+2}^{p} \alpha_j^2 + \alpha_{r+1}^2 (1 - d_{r+1})^2 \\ &= \sigma^2 \left[\sum_{j=1}^{p} \lambda_j^{-1} - \sum_{j=r+2}^{p} (1 - \tau_j^2)\lambda_j^{-1} - (1 + \tau_{r+1}^2)^{-1}\lambda_{r+1}^{-1} \right] \\ &= MSE(\underline{a}_{LS}) - \sigma^2 \left[\sum_{j=r+2}^{p} (1 - \tau_j^2)\lambda_j^{-1} + (1 + \tau_{r+1}^2)^{-1}\lambda_{r+1}^{-1} \right], \end{aligned} \tag{3.38}$$

where $\tau_j^2 = \lambda_j \alpha_j^2/\sigma^2$. Thus, from (3.38), it is learned that the improvement of $\underline{a}_{FR}$ over $\underline{a}_{LS}$ in MSE relies on the values of τ_j^2, $j = r+1, \ldots, p$. In particular, if the assigned rank is actually in the interval $[p-1,p]$, then the FR estimator has always smaller MSE than the LS estimator since the second term in the block of (3.38) is always positive. Furthermore, if

$$\tau_j^2 < 1, \text{ for some } j \geq r+2 \quad \text{and} \quad r' = \max_r \left[\sum_{j=r+2}^{p} (1 - \tau_j^2)\lambda_j^{-1} \right] \tag{3.39}$$

then the FR estimator with $r = r'$ and c [see (3.37)] has minimum MSE among all of the FR estimators. Note that, in a special case where the τ_j^2's (or, equivalently $\lambda_j \alpha_j^2$'s) are ordered as

$$\tau_1^2 \geq \ldots \geq \tau_p^2,$$

the optimal value of r can be easily determined at $r = r'$ for which $\tau_{r'+2}^2 < 1$, where $\tau_{r'+2}^2$ is the largest τ_j^2 such that $\tau_j^2 < 1$.

Since the values of τ_j^2 are unknown, the LS estimates of α_j and σ^2 can be substituted in τ_j^2. However, the fact that $\underline{a}_{LS}$ may not be reliable suggests re-estimating c from the new estimate of $\underline{\alpha}$ obtained from (3.32) and hence iterating this procedure until convergence is achieved. Note that the initial value of r, r_0 is chosen by inspection of (3.38), that is,

$$r_0 = \max_r \left[\sum_{j=r+2}^{p} (1 - \tau_{j,LS}^2)\lambda_j^{-1} \right],$$

where $\tau_{j,LS}^2 = \lambda_j \, a_{j,LS}^2/s^2$.

The FR estimator assigns the fraction (3.37) to a specific component ($\underline{z}_{r+1}$) through minimization of the MSE($\underline{a}_{FR}$). The fraction is nothing but the (r+1)st optimal fraction in the following chapter [see (4.11)]. Furthermore, the FR estimator may be sensitive to the starting values and, therefore, can be improved by iteration.

In particular, since Marquardt's generalized inverse is mainly determined by the magnitudes of the small eigenvalues, this estimation method does not seem to be appropriate unless the order of the γ_j corresponds to the order of the λ_j and only one principal component is in question. However, the concept of the fraction to the (r+1)st component is valuable for extending further modifications of principal components analysis.

CHAPTER FOUR

FRACTIONAL PRINCIPAL COMPONENTS ANALYSIS

The alternative estimation techniques to the LS method that are presented in this dissertation result in biased estimators. Most of these estimators have utilized the eigenvalue decomposition (or the singular value decomposition) in order to combat the problem of multicollinearity. A general form that encompasses most of these biased estimators will be considered in this chapter. In particular, principal components analysis can be generalized by applying different weights to the different principal components. The weights are determined by the importance of the individual principal component in improving the MSE of the LS estimator. Thus, a new class of alternatives to the LS estimator will be introduced, the class of fractional principal components estimators.

4.1 Class of Fractional Principal Components Estimators

The concept of the fractional principal components (FPC) estimator is closely related to that of the shrinkage estimator in the canonical form of the MLR model. While both estimators shrink the length of the LS estimated vector of the parameters toward the origin, the FPC estimator accomplishes this by taking different portions of the individual principal components. In addition, the effects of the orientation of the parameter vector and the magnitude of σ on the ridge estimator, one of the FPC estimators, will be demonstrated in Section 4.1.3, as it has been done for the PC estimator in Section 3.1.4.

4.1.1 Definition of the Fractional Principal Components Estimator

In the canonical form, the MLR model (3.1) can be written as:

$$\begin{aligned} \underline{y} &= Z\underline{\alpha} + \underline{\epsilon} \\ &= ZF^{-}F\underline{\alpha} + \underline{\epsilon} \\ &= Z_F\underline{\alpha}_F + \underline{\epsilon}, \end{aligned} \tag{4.1}$$

where $Z_F = ZF^{-}$, $\underline{\alpha}_F = F\underline{\alpha}$, and F is a diagonal matrix whose diagonals are in [0,1]. Note that F is a fixed diagonal matrix and, in fact, F^{-} is a generalized inverse of F. Then, each principal component, $\underline{z}_j$, is weighted by the j^{th} diagonal element of F^{-}. The LS estimator of $\underline{\alpha}_F$ is

$$\begin{aligned} \underline{a}_{F,LS} &= \left[Z_F'Z_F\right]^{-}Z_F'\underline{y} \\ &= F(Z'Z)^{-1}Z'\underline{y} \\ &= F\underline{a}_{LS} \end{aligned}$$

Thus, the class of the FPC estimators is one whose typical member is of the form,

$$\underline{a}_{FPC} = F\underline{a}_{LS}, \tag{4.2}$$

where $F = \mathrm{Diag}(f_1, \ldots, f_p)$ and $0 \leq f_j \leq 1$, $j=1, \ldots, p$. The diagonal matrix F is termed the <u>fraction matrix</u> and the diagonals f_j are called the <u>fractions</u>. Note that the class of FPC estimators is a specific class of linear transforms of $\underline{a}_{LS}$ introduced in Section 2.4.2 because of the restrictions ($0 \leq f_j \leq 1$).

Table 4.1 Various Biased Estimators Expressed As Fractional Principal Components Estimators

	Fractions
LS	$f_1 = \ldots = f_p = 1$
Ridge	$f_j = \lambda_j/(\lambda_j + k)$, for all j
G. Ridge	$f_j = \lambda_j/(\lambda_j + k_j)$, for all j
Baranchik	$f_j = (1 - c\ SS/\underline{a}_{LS}'\Lambda\ \underline{a}_{LS})^+$, for all j
Sclove	$f_1 = \ldots = f_r = 1,\ f_{r+1} = \ldots = f_p = (1 - c/F_2')^+$
PC	$f_1 = \ldots = f_r = 1,\ f_{r+1} = \ldots = f_p = 0$
FR	$f_1 = \ldots = f_r = 1,\ f_{r+1} = \lambda_{r+1}\alpha_{r+1}^2/(\lambda_{r+1}\alpha_{r+1}^2 + \sigma^2)$, $f_{r+2} = \ldots = f_p = 0$

The class of FPC estimators is a rich class of estimators that includes the LS, ridge, generalized ridge, Baranchik, Sclove, principal components, fractional rank estimators and others. Table 4.1 displays the fractions associated with the estimators discussed so far. The fraction matrix of the LS estimator is, of course, the identity matrix. The fractions of the ridge estimator can be easily obtained by considering the ridge estimator of $\underline{\alpha}$:

$$\underline{a}_R = (\Lambda + kI)^{-1} Z'\underline{y}$$

$$= (\Lambda + kI)^{-1}\Lambda\, \underline{a}_{LS}.$$

Thus take $F_R = (\Lambda + kI)^{-1}\Lambda$. Similarly, the j^{th} diagonal element of the generalized ridge estimator is $f_j = \lambda_j/(\lambda_j + k_j)$ where k_j is the j^{th} diagonal element of K. The Baranchik estimator has simple uniform fractions for each principal component. The Sclove estimator assigns two different values of fractions to the two subsets of the components, respectively. That is, one of them is 1 to the first r components and the other is the same shrinkage factor, $(1 - c/F_z')^+$, to the remaining (p-r) components. The PC estimator applies 1 to the first r components and 0 to the rest. Finally, the fractional rank estimator attaches a fraction between 0 and 1 only to the (r+1)st component which might be on the border for deletion among the principal components.

As mentioned earlier, the uniform fractions such as in $\underline{a}_{LS}$, $\underline{a}_{BA}$ do not seem to be reasonable since they imply that the principal components are equally weighted in estimating the parameters. Thus, when the multicollinearity problem exists in the model (4.1), the utiliza-

tion rates of the principal components, or, the fractions should be different. The ridge and generalized ridge estimators have different fractions for the components to combat multicollinearity.

Furthermore, it is also interesting to consider the order of the f_j. Since the optimal value of f_j, which will be derived in the next section, is $f_j^o = \gamma_j^2/(\gamma_j^2 + \sigma^2)$, the order of the fractions may depend upon the order of the γ_j. Thus, the order of the fractions for the ridge estimator, $f_1 \geq \ldots \geq f_p$, is not appropriate to some situations since the orientation of the parameter vector is unknown. It should be noted, therefore, that the order property of the fractions is not a requirement for the FPC estimators. Note that the fractions of the generalized ridge estimator do not have to be ordered. In fact, under certain orientations of $\underline{\beta}$ (or $\underline{\alpha}$) it is desirable to have fractions which are not ordered as for the ridge estimator. This important concept will be seen in more detail in Chapter 5.

4.1.2 Properties and Extensions of FPC Estimators

It has been shown that various types of biased estimators can be characterized by different assignments of the fractions of the fraction matrix.

Several properties of the FPC estimator will now be presented. First of all, the FPC estimators are biased except when all the fractions are 1. Since

$$E(\underline{a}_{FPC}) = F\underline{\alpha} = \underline{\alpha} - (I - F)\,\underline{\alpha},$$

it follows that the bias is

$$- (I - F)\,\underline{\alpha}. \tag{4.3}$$

The covariance matrix of $\underline{a}_{FPC}$ is

$$\mathrm{Var}(\underline{a}_{FPC}) = \sigma^2 F \Lambda^{-1} F. \tag{4.4}$$

Moreover, the matrix form of $\mathrm{MSE}(\underline{a}_{FPC})$ is

$$\mathrm{MtxMSE}(\underline{a}_{FPC}) = \sigma^2 F \Lambda^{-1} F + (I - F)\,\underline{\alpha}\,\underline{\alpha}'(I - F). \tag{4.5}$$

Parallel to the condition (3.18), therefore, the condition for which

$$\mathrm{MtxMSE}(\underline{a}_{LS}) - \mathrm{MtxMSE}(\underline{a}_{FPC}) \tag{4.6}$$

is positive semi-definite can be obtained by showing that

$$\Lambda^{-1}(I - F^2) - \sigma^{-2}(I - F)\,\underline{\alpha}\,\underline{\alpha}'(I - F)$$

is positive semi-definite. Consequently, the necessary and sufficient condition is

$$\underline{\alpha}'\Lambda(I - F)^2(I - F^2)^{-}\underline{\alpha} \leq \sigma^2. \tag{4.7}$$

Note that $(I - F^2)^{-}$ is a generalized inverse. Note also that when $f_1 = \ldots = f_r = 1$ and $f_{r+1} = \ldots = f_p = 0$, as in the case for the PC estimator, the condition (4.7) is equivalent to (3.18).

According to the fraction matrices in Table 4.1, the conditions for the biased estimators are summarized in Table 4.2. Since the fraction matrix of the ridge estimator is $F_R = (I + k\Lambda^{-1})^{-1}$,

$$(I - F_R^2) = I - (I + k\Lambda^{-1})^{-2},$$

and

Table 4.2 Various Necessary and Sufficient Conditions (4.7)

	Conditions (4.7)
Ridge	$k\underline{\alpha}'\Lambda(2\Lambda + kI)^{-1}\ \underline{\alpha} \leq \sigma^2$
G. Ridge	$\underline{\alpha}'K\Lambda(2\Lambda + K)^{-1}\ \underline{\alpha} \leq \sigma^2$
Baranchik	$(1 - c\ SS/\underline{a}_{LS}'\Lambda\ \underline{a}_{LS})^{+}\ \underline{\alpha}'\Lambda\ \underline{\alpha} \leq \sigma^2$
Sclove	$(1 - c/F_z')^{+}\ \underline{\alpha}_z'\Lambda_z\ \underline{\alpha}_z \leq \sigma^2$
PC	$\underline{\alpha}_z'\Lambda_z\ \underline{\alpha}_z \leq \sigma^2$
FR	$\lambda_{r+1}\alpha_{r+1}^2(1 - f_{r+1})^2/(1 - f_{r+1}^2) + \underline{\alpha}_{z(r+1)}'\Lambda_{z(r+1)}\ \underline{\alpha}_{z(r+1)} \leq \sigma^2$

$$(I - F_R)^{-1} = \left[I - (I + k\Lambda^{-1})^{-1}\right]^{-1},$$

the inequality of (4.7) can be simplified, by considering the individual diagonals, as

$$k\underline{\alpha}'\Lambda\,(2\Lambda + kI)^{-1}\,\underline{\alpha} \leq \sigma^2. \tag{4.8}$$

The condition for the generalized ridge estimator is similarly obtained with the diagonal matrix K. For the Sclove estimator, since the shrinkage factor $(1 - c/F_2')^+$ is 0 when $1 \leq (c/F_2')$ so that the condition is identical to that for the PC estimator. Thus, when $1 > (c/F_2')$, the necessary and sufficient condition is

$$\underline{\alpha}'\Lambda\,(I - F_{SC})^2(I - F_{SC}^2)^{-1}\underline{\alpha} = \underline{\alpha}_2'\Lambda_2\left[I - (1 - g)I\right]^2\left[I - (1 - g)^2I\right]^{-1}\underline{\alpha}_2$$

$$= \frac{g}{2-g}\,\underline{\alpha}_2'\Lambda_2\,\underline{\alpha}_2 \leq \sigma^2,$$

where $g = (c/F_2') < 1$. Finally, the condition for the fractional rank estimator is easily found as

$$\frac{(1 - f_{r+1})^2}{(1 - f_{r+1}^2)}\;\lambda_{r+1}\alpha_{r+1}^2 + \underline{\alpha}_{2(r+1)}'\Lambda_{2(r+1)}\,\underline{\alpha}_{2(r+1)} \leq \sigma^2,$$

where $\underline{\alpha}_{2(r+1)}' = (\alpha_{r+2}, \ldots, \alpha_p)$ and $\Lambda_{2(r+1)} = \mathrm{Diag}(\lambda_{r+2}, \ldots, \lambda_p)$.

In particular, the condition (4.7) can be demonstrated graphically for the ridge estimator in a manner similar to that for the PC estimator in Section 3.1.4 when p = 2. This will be discussed in the following section.

From another viewpoint, the $\mathrm{MSE}(\underline{a}_{FPC})$ is

$$\sum_{j=1}^{p} MSE(a_{j,FPC}) = \sigma^2 \sum_{j=1}^{p} \frac{f_j^2}{\lambda_j} + \sum_{j=1}^{p} (1 - f_j)^2 \alpha_j^2. \qquad (4.9)$$

The set of the fractions, $f_1, \ldots, f_p$, implies that there is, indeed, a componentwise trade-off between the "variance" and "squared-bias" parts in (4.9). That is, as $f_j \to 0$, $Var(a_{i,FPC}) \to 0$ and $Bias^2(a_{i,FPC}) \to \alpha_j^2$. If $f_j \to 1$, then $Var(a_{j,FPC}) \to Var(a_{j,LS})$ and $Bias^2(a_{j,FPC}) \to 0$.

In conjunction with Table 4.1, comparisons of the biased estimation methods in terms of their "variance" and "squared-bias" components are made in Table 4.3. Note that the "squared-bias" components depend upon the unknown regression coefficient parameters and not at all upon the population variance σ^2.

Finally, the optimal values of the fractions are obtained by minimizing $MSE(\underline{a}_{FPC})$ with respect to the f_j's. Thus, evaluating $\partial MSE(\underline{a}_{FPC})/\partial f_j = 0$ simultaneously for j=1, ..., p, provides

$$f_j^{\circ} = \frac{\alpha_j^2}{\alpha_j^2 + \sigma^2/\lambda_j}$$

$$= \frac{\lambda_j \alpha_j^2}{\lambda_j \alpha_j^2 + \sigma^2}, \qquad j=1, \ldots, p, \qquad (4.10)$$

where f_j° denotes the optimal j^{th} fraction. Note that the f_j° are such that $0 \leq f_j^{\circ} < 1$. With the f_j°, the minimum of the $MSE(\underline{a}_{FPC})$ can be decomposed as:

$$MINMSE(\underline{a}_{FPC}) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^4}{(\lambda_j \alpha_j^2 + \sigma^2)^2} + \sum_{j=1}^{p} \frac{\sigma^4 \alpha_j^2}{(\lambda_j \alpha_j^2 + \sigma^2)^2}. \qquad (4.11)$$

Table 4.3 Decomposition of Various Forms of MSE($\underline{a}_{FPC}$)

	"Variance" Component	"Squared-Bias" Component
LS	$\sigma^2 \Sigma \lambda_j^{-1}$	0
Ridge	$\sigma^2 \Sigma \lambda_j/(\lambda_j + k)^2$	$k^2 \Sigma \alpha_j^2/(\lambda_j + k)^2$
G. Ridge	$\sigma^2 \Sigma \lambda_j/(\lambda_j + k_j)^2$	$\Sigma k_j^2 \alpha_j^2/(\lambda_j + k_j)^2$
Baranchik	$\sigma^2(1 - c\ SS/\underline{a}_{LS}'\Lambda\underline{a}_{LS})^{+2}\Sigma\lambda_j^{-1}$	$(1 - c\ SS/\underline{a}_{LS}'\Lambda\underline{a}_{LS})^{+2}\Sigma\alpha_j^2$
Sclove	$\sigma^2 \sum_{1}^{r} \lambda_j^{-1} + \sigma^2(1 - g)^2 \sum_{r+1}^{p} \lambda_j^{-1}$	$g^2 \sum_{r+1}^{p} \alpha_j^2$
PC	$\sigma^2 \sum_{1}^{r} \lambda_j^{-1}$	$\sum_{r+1}^{p} \alpha_j^2$
FR	$\sigma^2 \sum_{1}^{r} \lambda_j^{-1} + \sigma^2 f_{r+1}^2/\lambda_{r+1}$	$(1 - f_{r+1})^2 \alpha_{r+1}^2 + \sum_{r+2}^{p} \alpha_j^2$

Note that the MINMSE($\underline{a}_{FPC}$) is the weighted sum of the Var($a_{j,LS}$) with the weights f_j^o, that is, from (4.11),

$$\text{MINMSE}(\underline{a}_{FPC}) = \sum_{j=1}^{p} \frac{\sigma^2}{\lambda_j} \left[\frac{\lambda_j \alpha_j^2}{(\lambda_j \alpha_j^2 + \sigma^2)}\right] = \sum_{j=1}^{p} f_j^o \left[\frac{\sigma^2}{\lambda_j}\right]. \tag{4.12}$$

Furthermore, by using $f_j^o = 1 - (1 - f_j^o)$, (4.12) can be rewritten as

$$\begin{aligned} \text{MINMSE}(\underline{a}_{FPC}) &= \sum_{j=1}^{p} \frac{\sigma^2}{\lambda_j} - \sigma^2 \sum_{j=1}^{p} \frac{1 - f_j^o}{\lambda_j} \\ &= \text{MSE}(\underline{a}_{LS}) - \sigma^2 \sum_{j=1}^{p} \frac{1 - f_j^o}{\lambda_j}. \end{aligned} \tag{4.13}$$

Therefore, the quantity $\sigma^2 \sum_{j=1}^{p} (1 - f_j^o)/\lambda_j$ is the maximum amount of the reduction in MSE when the FPC estimator is used rather than the LS estimator, due to multicollinearity. Note that the quantity is always nonnegative for $0 \leq f_j \leq 1$, $j=1, \ldots, p$. Thus, the restrictions ($0 \leq f_j \leq 1$) are reasonable.

However, the theoretical, optimal fraction matrix cannot be used in practice since the f_j^o in (4.10) contain the unknown parameters α_j. If estimates for these unknowns, for example, $a_{j,LS}$, $j=1, \ldots, p$, and s^2, are used then the resulting estimates of the f_j^o depend upon $\underline{y}$ and hence, (4.9) is not valid (Obenchain, 1975). Further discussions related to the estimation of the f_j^o will be discussed later in this chapter.

4.1.3 Problems of Ridge Estimator

In the previous section, the necessary and sufficient condition

for the ridge estimator to have smaller MSE than the LS estimator is given in (4.8). Therefore, as has been done in Section 3.1.4, the effects of the orientation of the unknown parameter vector and the variance σ^2 to the ridge estimator can be illuminated when p = 2.

First of all, since $\gamma = \Lambda^{1/2}\,\underline{\alpha}$, the condition for the ridge estimator can be rewritten in terms of γ as:

$$\gamma'(2k^{-1}\Lambda + I)^{-1}\gamma \leq \sigma^2. \tag{4.14}$$

Note that k is an arbitrary positive fixed constant. Assuming the squared norm of γ is given as $\gamma'\gamma = C'$, where C' is fixed, if

$$\sqrt{C'} < \sigma\sqrt{1 + 2\lambda_2/k} \tag{4.15}$$

then the ellipse $\gamma'(2k^{-1}\Lambda + I)^{-1}\gamma = \sigma^2$ includes the circle $\gamma'\gamma = C'$ so that the ridge estimator improves over the LS estimator in MSE for any γ. This is shown in Figure 4.1, where the ellipse takes values on the γ_1 and γ_2 axes of $\pm\,\sigma\sqrt{1 + 2\lambda_1/k}$ and $\pm\,\sigma\sqrt{1 + 2\lambda_2/k}$, respectively.

If the norm of γ lies,

$$\sigma\sqrt{1 + 2\lambda_2/k} \leq \sqrt{C'} < \sigma\sqrt{1 + 2\lambda_1/k} \tag{4.16}$$

[see Figure 4.2], then for only the limited orientations, along the chords ST and UV which satisfy (4.16), the ridge estimator can be recommended instead of the LS estimator. It should be emphasized that when the inequality (4.16) holds, the LS estimator for any parameter vector from TU and VS of the circle outperforms the ridge estimator even though there exists the multicollinearity problem. Note that the points S, T, U, and V are found by solving the two simultaneous equations, $\gamma'(2k^{-1}\Lambda + I)^{-1}\,\gamma = \sigma^2$ and $\gamma'\gamma = C'$. The

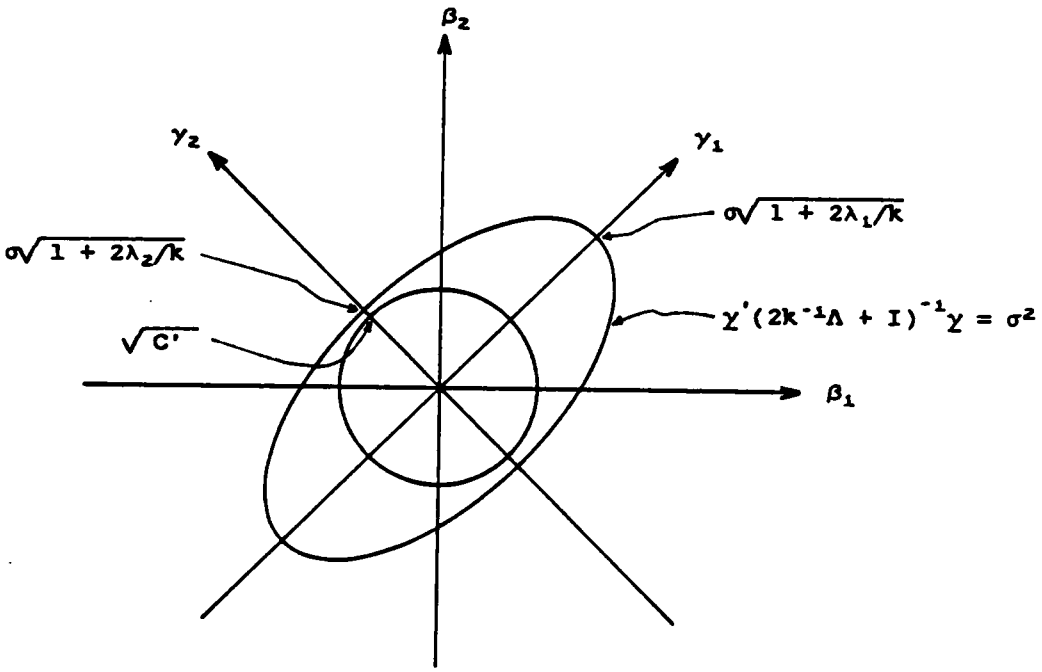

Figure 4.1 Ellipse of (4.14): $\sqrt{C'} < \sigma\sqrt{1 + 2\lambda_2/k}$

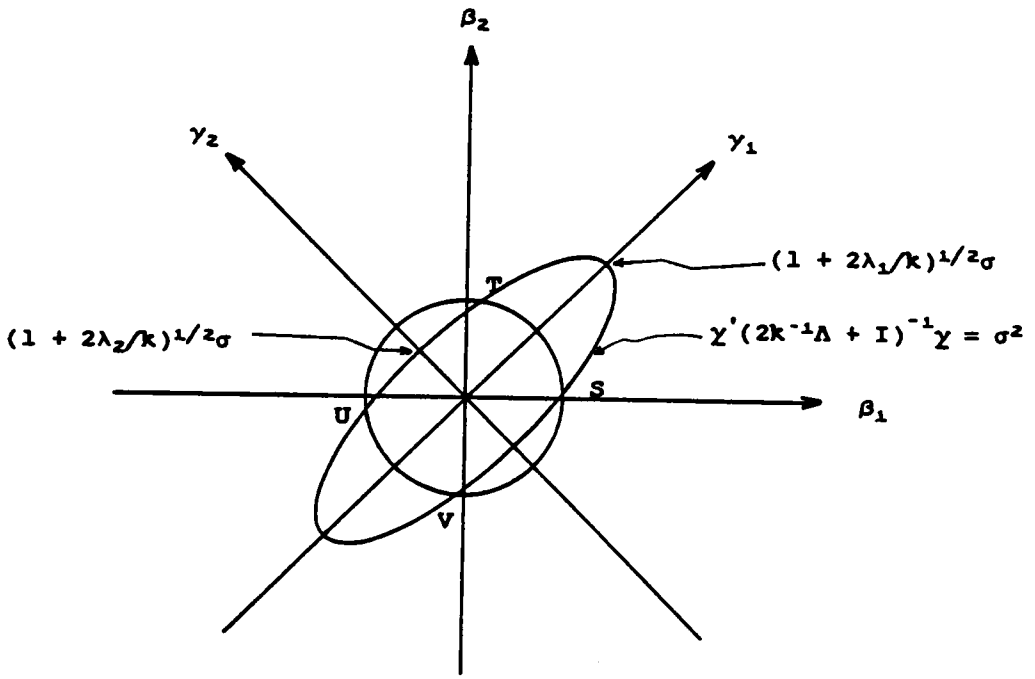

Figure 4.2 Ellipse of (4.14): $\sigma\sqrt{1 + 2\lambda_2/k} < \sqrt{c'} < \sigma\sqrt{1 + 2\lambda_1/k}$

coordinates of the four points are

$$\left[\pm\left[\sigma^2 - \frac{kC'}{s\lambda_2 + k}\right] \Big/ \left[\frac{k}{2\lambda_2 + k} - \frac{k}{2\lambda_1 + k}\right],\right.$$

$$\left.\pm\left[\sigma^2 - \frac{kC'}{2\lambda_1 + k}\right] \Big/ \left[\frac{k}{2\lambda_2 + k} - \frac{k}{2\lambda_1 + k}\right]\right], \qquad (4.17)$$

respectively.

Furthermore, if the ellipse is inside the fixed circle [see Figure 4.3], that is,

$$\sigma\sqrt{1 + 2\lambda_1/k} \leq \sqrt{C'}, \qquad (4.18)$$

then the ridge estimator should not be used. In other words, when the ratio (C'/σ^2), which is called the signal-to-noise ratio, is very large, the LS estimator does work well.

In addition, consider the generalized ridge estimator. Since the fractions for the generalized ridge estimator are, from Table 4.1,

$$f_{j,GR} = \frac{\lambda_j}{\lambda_j + k_j}, \qquad j=1, \ldots, p, \qquad (4.19)$$

and the optimal values for k_j [Hoerl and Kennard [1970(a)]] are

$$k_j = \frac{\sigma^2}{\alpha_j^2}, \qquad j=1, \ldots, p, \qquad (4.20)$$

the optimal fractions for the generalized ridge estimator, $f_{j,GR}^{o}$ are exactly the same as the optimal fractions in (4.10). Note that

$$f_{j,GR}^{o} = \frac{\lambda_j}{\lambda_j + \sigma^2/\alpha_j^2}, \qquad j=1, \ldots, p. \qquad (4.21)$$

Hence, the theoretical generalized ridge estimator is equivalent to the optimal FPC estimator.

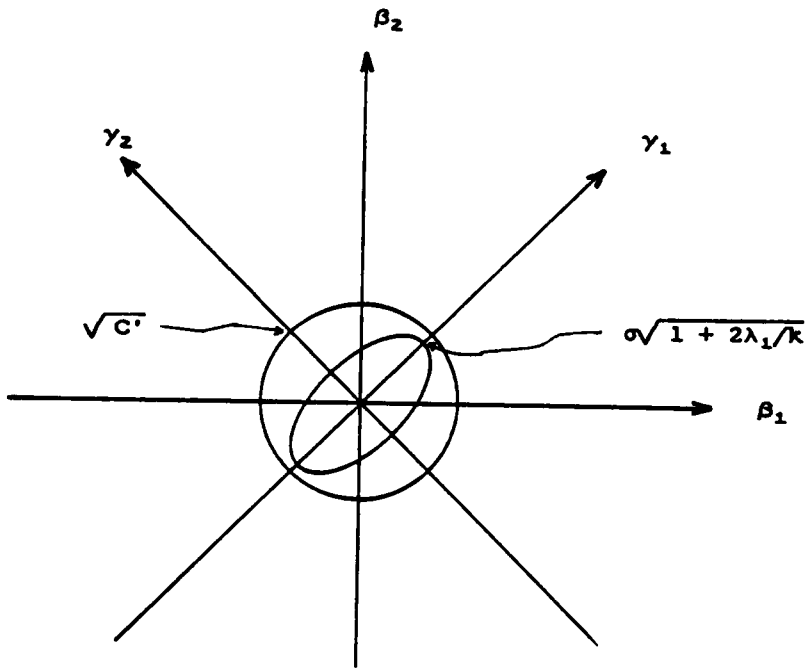

Figure 4.3 Ellipse of (4.14): $\sigma\sqrt{1 + 2\lambda_1/k} \leq \sqrt{c'}$

Therefore, from (4.13), it is concluded that the theoretically optimal generalized ridge estimator always outperforms the LS estimator for any orientation of the parameter vector and value of σ^2. This may be a great merit of the generalized ridge estimator even though the substitutions of estimates for the unknown parameters in (4.20) may not guarantee the minimum MSE. Thus, appropriate estimation methods for the f_j° will be discussed after the graphical demonstration of minimization of the $MSE(\underline{a}_{FPC})$.

4.2 Minimization of $MSE(\underline{a}_{FPC})$

The concept of the minimization of the $MSE(\underline{a}_{FPC})$ with respect to the fractions will be illustrated graphically. Since there are often some situations where some biased estimation techniques do not perform well, as discussed in Section 3.1.4 and the previous section, some new estimation techniques will be proposed in this section.

4.2.1 Review of Minimization of $MSE(\underline{a}_{FPC})$ via Graphs

The minimization of the $MSE(\underline{a}_{FPC})$ has provided the optimal fractions as in (4.10). The optimal fractions can also be obtained by minimizing the difference of the $MSE(\underline{a}_{LS})$ and the $MSE(\underline{a}_{FPC})$, denoted by Δ. Thus, it will be useful to express Δ as:

$$\Delta = MSE(\underline{a}_{LS}) - MSE(\underline{a}_{FPC})$$

$$= \sigma^2 \sum_{j=1}^{p} (1 - f_j^2)\, \lambda_j^{-1} - \sum_{j=1}^{p} (1 - f_j)^2\, \alpha_j^2$$

$$= \Delta_1 - \Delta_2,$$

where $\Delta_1 = \sigma^2 \sum_{j=1}^{p} (1 - f_j^2)\lambda_j^{-1}$ and $\Delta_2 = \sum_{j=1}^{p}(1 - f_j)^2 \alpha_j^2$. Note that Δ_1 is the reduction amount of the "variance" part.

It is of interest to investigate the contributions of the individual principal component to Δ_1 and Δ_2. Denote the "variance" reduction part and the "squared-bias" part of the j^{th} component by Δ_{1j} and Δ_{2j}, respectively. Then

$$\Delta_{1j} = \sigma^2(1 - f_j^2)\, \lambda_j^{-1}, \qquad j=1, \ldots, p, \tag{4.22}$$

and

$$\Delta_{2j} = (1 - f_j)^2\, \alpha_j^2, \qquad j=1, \ldots, p. \tag{4.23}$$

Since Δ_{1j} and Δ_{2j} are parabolas of f_j, it is convenient to graphically compare the individual contribution to Δ by graphs. Two representative cases determined by the magnitudes of σ^2/λ_j and α_j^2 are shown in Figure 4.4.

In the case that $\sigma^2/\lambda_j > \alpha_j^2$ [Figure 4.4(a)], deleting the j^{th} principal component, equivalently, $f_j = 0$ results in a reduction of Δ_j since $\Delta_j(= \Delta_{1j} - \Delta_{2j})$ is positive. In the case that $\sigma^2/\lambda_j < \alpha_j^2$ [Figure 4.4(b)], reversely, the deletion of the j^{th} principal component leads Δ_j to be negative so that the j^{th} component needs to be retained in the analysis. As long as $\Delta_{1j} > \Delta_{2j}$, from Figure 4.4, the use of the fraction will be beneficial to the reduction of the MSE($a_{j,LS}$). Note that the maximum of $(\Delta_{1j} - \Delta_{2j})$ occurs at $f_j^{o} = \lambda_j\alpha_j^2/(\lambda_j\alpha_j^2 + \sigma^2)$ as in (4.10). Furthermore, since the two parabolas meet at $f_j = (\lambda_j\alpha_j^2 - \sigma^2)/(\lambda_j\alpha_j^2 + \sigma^2)$ and 1, the range of f_j for $\Delta_{1j} > \Delta_{2j}$ is

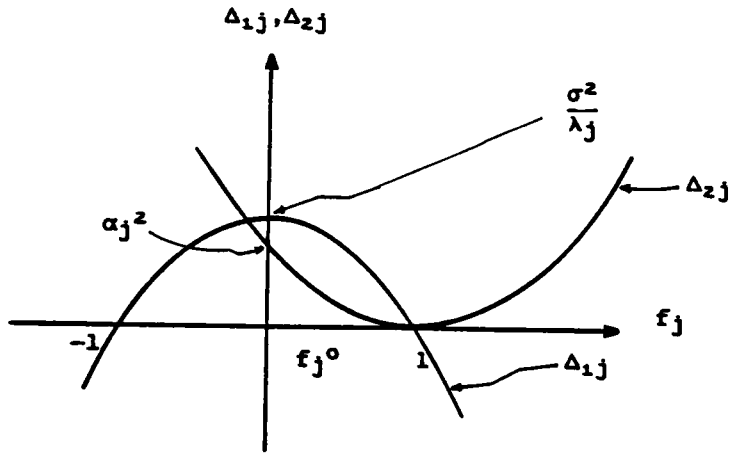

Figure 4.4(a) Parabolas Δ_{1j} and Δ_{2j}: $\sigma^2/\lambda_j > \alpha_j^2$

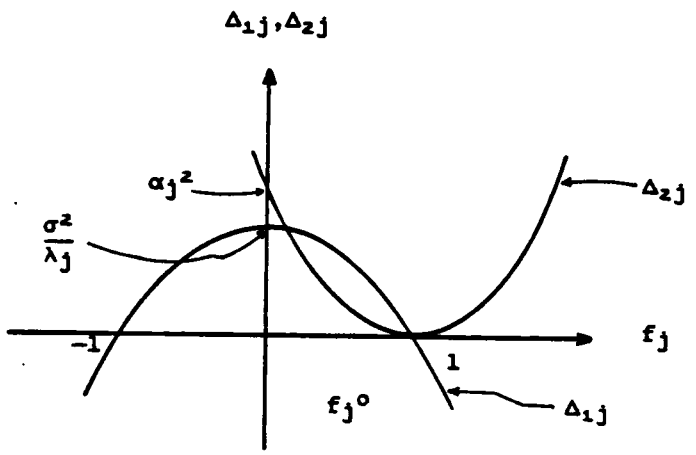

Figure 4.4(b) Parabolas Δ_{1j} and Δ_{2j}: $\sigma^2/\lambda_j < \alpha_j^2$

$$\frac{\lambda_j\alpha_j^2 - \sigma^2}{\lambda_j\alpha_j^2 + \sigma^2} < f_j < 1. \qquad (4.24)$$

The above condition is exactly the same as the condition, presented by Blight (1971), on the shrinkage factor for improvement over the unbiased estimator. However, the condition (4.24) should be changed to

$$\max\left[0, \frac{\lambda_j\alpha_j^2 - \sigma^2}{\lambda_j\alpha_j^2 + \sigma^2}\right] < f_j < 1 , \qquad (4.25)$$

since a negative fraction is not allowed in the fraction matrix. Apparently, the optimal fraction f_j^o lies in the interval of (4.25).

In practice, one needs to estimate the f_j^o. Since the fraction matrix of the generalized ridge estimator is just the optimal fraction matrix [see (4.21)], the f_j^o may be estimated iteratively just as the k_j are estimated by iteration [see Section 2.2.2].

4.2.2 Iterative FPC Estimator with PC Estimate

It is of interest to see that the iterative estimation of the optimal fractions can be described by the iteration of the k_j as in (2.34). By substituting the iterative device (2.34), initiated with the LS estimates, into the optimal fractions, the iterative estimates of the optimal fractions can be obtained. Consequently, the resulting estimator, that is, the iterative (optimal) FPC estimator with LS starting values, is just the iterative generalized ridge estimator. Note that the iterative scheme of the optimal fractions is

$$f_j^{o(t+1)} = \frac{\lambda_j}{\lambda_j + s^2/\left[a_{j,GR}^{(t)}\right]^2}, \qquad t=0, 1, 2, \ldots \qquad (4.26)$$

with the initial value $f_j^{o(i)} = \lambda_j/(\lambda_j + s^2/a_{j,LS}^2)$, where s^2 is the residual mean square error. The iteration continues until convergence is achieved.

However, though the MSE of the iterative FPC estimator with LS starting values is smaller than that of the LS estimator, the problem is how close is the resulting MSE($\underline{a}_{FPC}$) to the MINMSE($\underline{a}_{FPC}$) of (4.12). Therefore, because of the harmful effects of multicollinearity on $\underline{a}_{LS}$ and the crucial role of the starting values, it may be fruitful to use some appropriate biased estimates as initial values. For example, if the PC estimates are considered as the initial values, then the iterative scheme (4.26) becomes

$$f_{j,PCI}^{(t+1)} = \frac{\lambda_j}{\lambda_j + [s^2]^{(t)}/[a_j^{(t)}]^2},$$

$$= \frac{\lambda_j [a_j^{(t)}]^2}{\lambda_j [a_j^{(t)}]^2 + [s^2]^{(t)}}, \qquad t=0, 1, 2, \ldots, \qquad (4.27)$$

where $[s^2]^{(t)} = [\underline{y} - Z\underline{a}^{(t)}]' [\underline{y} - Z\underline{a}^{(t)}]/(n-p)$, $\underline{a}^{(t)}$ is the estimate of $\underline{\alpha}$ at the t^{th} iteration, and $\underline{a}^{(0)} = \underline{a}_{PC}$. Note that the subscript PCI refers to "use of the individual PC estimate as the starting value".

The final resulting estimator, denoted by $\underline{a}_{IFPCI}$, with the limiting fraction matrix, F_{PCI}^*, is

$$\underline{a}_{IFPCI} = F_{PCI}^* \underline{a}_{LS}, \qquad (4.28)$$

where $F_{PCI}^* = Diag(f_{1,PCI}^*, \ldots, f_{p,PCI}^*)$, $f_{j,PCI}^* = \lim_t [f_{j,PCI}^{(t)}]$, $j=1, \ldots, p$. The estimator $\underline{a}_{IFPCI}$, termed the iterative FPC estimator

with individual PC estimates as the starting values, has the characteristic that if the k^{th} principal component is determined to be deleted [see Chapter 3], then the limiting value of the k^{th} fraction, f_k^* is zero. Therefore, the IFPCI estimator is formed from the combined concept of the PC estimator and the iterative generalized ridge estimator.

Furthermore, the iterative ridge estimator in Section 2.2.1 can be equivalently described by using an iteration scheme in computing the fractions. Since the fractions for the ridge estimator are

$$f_{j,R} = \frac{\lambda_j}{\lambda_j + k}, \qquad j=1, \ldots, p, \tag{4.29}$$

and a reasonable choice for k, suggested by Hoerl, Kennard, and Baldwin (1975), is the harmonic mean of the k_j, that is,

$$k = \frac{p\,\sigma^2}{\underline{\alpha}'\underline{\alpha}}, \tag{4.30}$$

the iterative ridge estimator can be obtained by iterating the fractions with substituting (4.30) in (4.29). The iterative fractions of the ridge estimator are

$$f_{j,R}^{(t+1)} = \frac{\lambda_j}{\lambda_j + p\,s^2/\underline{a}_R^{(t)\prime}\underline{a}_R^{(t)}}, \qquad t=0, 1, 2, \ldots, \tag{4.31}$$

where $\underline{a}_R^{(t)}$ is the ridge estimate of $\underline{\alpha}$ at the t^{th} iteration with $\underline{a}_R^{(0)} = \underline{a}_{LS}$.

Comparing (4.31) with (4.26), the difference is that the $f_{j,GR}^{(t+1)}$ [equivalently, $f_j^{o(t+1)}$] uses the j^{th} estimated coefficient in the place of α_j^2 and, in contrast, the $f_{j,R}^{(t+1)}$ utilizes

$\underline{a}_R^{(t)'}\underline{a}_R^{(t)}/p$ for α_j^2, j=1, ..., p. Note that $\underline{a}_R^{(t)'}\underline{a}_R^{(t)}/p$ is the arithmetic average of the $[a_{j,R}^{(t)}]^2$. Thus, the iterative ridge estimator is not only based on the iteration of the forms of the optimal fractions but obtained by using a different starting value.

Similar to the definition of the (4.27), another iterative scheme related to (4.31) that can be developed by utilizing the appropriate PC estimate of $\underline{\alpha}$ as the starting value, is

$$f_{j,PCV}^{(t+1)} = \frac{\lambda_j}{\lambda_j + p\,[s^2]^{(t)}/\underline{a}^{(t)'}\underline{a}^{(t)}}, \qquad t=0, 1, 2, \ldots, \quad (4.32)$$

where the subscript PCV refers to "use of the PC estimated vector as the starting value". Thus, with the limiting values of (4.32), the estimator $\underline{a}_{IFPCV}$, the iterative FPC estimator with the vector of PC estimates as the starting value, can be defined as:

$$\underline{a}_{IFPCV} = F_{PCV}^{*}\,\underline{a}_{LS}, \qquad (4.33)$$

where $F_{PCV}^{*} = \mathrm{Diag}(f_{1,PCV}^{*}, \ldots, f_{p,PCV}^{*})$, $f_{j,PCV}^{*} = \lim_t [f_{j,PCV}^{(t)}]$, j=1, ..., p.

In summary, the four estimators in this section are based on the different estimates of the optimal fraction matrix. From heuristic results, the IFPCI and IFPCV estimators may be more reasonable than the iterative ridge and generalized ridge estimators, respectively. An empirical comparison of these estimators will be made in Section 5.2.2.

In the following section, several new members of the class of the FPC estimators will be elaborately defined by revisiting the Bayesian interpreted James-Stein estimator and the adjusted PC_t estimator and combining these two concepts.

4.3 New Members of the Class of FPC Estimators

The Bayesian interpretation, applied in Section 2.3.1, to the Stein estimator can also be used to define a generalized Bayesian interpreted James-Stein estimator that is a member of the class of FPC estimators. Also, further investigation of the adjusted PC_t estimator provides the, so-called, generalized PC_t estimator.

4.3.1 The Generalized Bayesian Interpreted James-Stein Estimator

If a prior distribution of $\underline{\alpha}$, of the form,

$$\underline{\alpha} \sim N\left[\underline{\alpha}_0,\ D_\alpha (Z'Z)^{-1}\right] \tag{4.34}$$

is assumed, where $D_\alpha = \mathrm{Diag}\left[\sigma_{\alpha_1}^2,\ \ldots,\ \sigma_{\alpha_p}^2\right]$, then the generalized Bayesian interpreted James-Stein (GBJS) estimator is defined as:

$$\underline{a}_{GBJS} = F_{GBJS}\ \underline{a}_{LS}, \tag{4.35}$$

where $F_{GBJS} = \mathrm{Diag}(f_1^+,\ \ldots,\ f_p^+)$, $f_j^+ = \max(0,\ 1 - s^2/\lambda_j\ a_{j,LS}^2)$, for all j.

This result can be proved by utilizing the steps similar to those employed in Section 2.3.1. Using the prior (4.34), the Bayes estimator of $\underline{\alpha}$, with the expected quadratic loss, is

$$\underline{a}_{GB} = (\sigma^{-2}\ Z'Z + D_\alpha^{-1}\ Z'Z)^{-1}\ (\sigma^{-2}\ Z'Z\ \underline{a}_{LS} + D_\alpha^{-1}\ Z'Z\ \underline{\alpha}_0)$$

$$= (\sigma^2 I + D_\alpha^{-1})^{-1}\ (\sigma^{-2}\ \underline{a}_{LS} + D_\alpha^{-1}\ \underline{\alpha}_0).$$

Then, the j^{th} element of $\underline{a}_{GB}$ is

$$a_{j,GB} = \frac{\sigma_{\alpha_j}^2}{\sigma^2 + \sigma_{\alpha_j}^2} a_{j,LS} + \frac{\sigma^2}{\sigma^2 + \sigma_{\alpha_j}^2} \alpha_{j,0}, \qquad j=1, \ldots, p.$$

By taking $\alpha_{j,0} = 0$ for all j, it follows that

$$a_{j,GB} = \frac{\sigma_{\alpha_j}^2}{\sigma^2 + \sigma_{\alpha_j}^2} a_{j,LS}$$

$$= \left[1 - \frac{\sigma^2}{\sigma^2 + \sigma_{\alpha_j}^2}\right] a_{j,LS}, \qquad j=1, \ldots, p. \qquad (4.36)$$

Now it can be shown that the unbiased estimator of $\left[\sigma^2 + \sigma_{\alpha_j}^2\right]$ is $\lambda_j \, a_{j,LS}^2$. The proof is as follows:

Define a matrix, $\Lambda_j(0)$, as

$$\Lambda_j(0) = \text{Diag}(0, \ldots, 0, \lambda_j, 0, \ldots, 0), \qquad j=1, \ldots, p.$$

Then the expectation of $\lambda_j \alpha_{j,LS}^2$ can be written as

$$E(\lambda_j \, a_{j,LS}^2) = E\left[\underline{a}_{LS}' \Lambda_j(0) \, \underline{a}_{LS}\right]$$

$$= E\left[\Lambda^{-1} Z' (Z' \underline{\alpha} + \underline{\epsilon})\right]' \Lambda_j(0) \left[\Lambda^{-1} Z' (Z' \underline{\alpha} + \underline{\epsilon})\right]$$

$$= E\left[\underline{\alpha}' \Lambda_j(0) \, \underline{\alpha} + \underline{\epsilon}' Z \, \Lambda^{-1} \Lambda_j(0) \, \Lambda^{-1} \, Z' \underline{\epsilon}\right].$$

Since $E\left[\underline{\alpha}' \Lambda_j(0) \underline{\alpha}\right] = \text{tr}\left[D_\alpha \, \Lambda^{-1} \Lambda_j(0)\right]$ and $E\left[\underline{\epsilon}' Z \, \Lambda^{-1} \Lambda_j(0) \Lambda^{-1} Z' \underline{\epsilon}\right] = \text{tr}\left[\sigma^2 \, \Lambda^{-1} \Lambda_j(0)\right]$, it follows that $E(\lambda_j \, a_{j,LS}^2)$ can be rewritten as

$$E(\lambda_j \, a_{j,LS}^2) = \sigma^2 + \sigma_{\alpha_j}^2, \qquad j=1, \ldots, p. \qquad (4.37)$$

Therefore, from (4.36), a generalized James-Stein estimator of α_j is

obtained by substituting $\lambda_j a_{j,LS}^2$ for $\left[\sigma^2 + \sigma_{\alpha_j}^2\right]$ and replacing σ^2 by s^2 gives

$$a_{j,GJB} = \left[1 - \frac{s^2}{\lambda_j a_{j,LS}^2}\right] a_{j,LS}, \qquad j=1, \ldots, p. \tag{4.38}$$

From (4.38), since $(1 - s^2/\lambda_j a_{j,LS}^2)$ could be negative, (4.35) is achieved by taking $\max[0, 1 - s^2/\lambda_j a_{j,LS}^2]$.

The GBJS estimator can also be obtained by means of minimizing the C_p-like statistic for fractional principal components regression (4.1). The C_p statistic, presented by Gorman and Toman (1966), is the estimate of the standardized total squared error, that is,

$$C_p = \frac{SS}{s^2} - (n-2p), \tag{4.39}$$

where SS is the residual sum of squares. Furthermore, Mallows (1973) developed a C_p-like statistic for ridge regression. It is possible to develop a C_p-like statistic for FPC regression by using the reparametrization model (4.1). The C_p-like statistic for FPC regression, denoted by C_F, turns out to be

$$C_F = 2\mathrm{tr}(H_F) - n + \frac{SS_F}{SS}(n-p), \tag{4.40}$$

where $H_F = ZF\Lambda^{-1}Z'$ and $SS_F = (\underline{Y} - ZF^{-1}\underline{a}_{F,LS})'(\underline{Y} - ZF^{-1}\underline{a}_{F,LS})$ [the proof of this result is deferred to Appendix]. Note that (4.40) becomes C_p in (4.39) when $F = I$. Then, by minimizing C_F with respect to F, the optimal values of the f_j are

$$f_j = 1 - \frac{s^2}{\lambda_j \alpha_{j,LS}^2}, \qquad j=1, \ldots, p, \tag{4.41}$$

which are identical to the shrinkage factors in (4.38). Therefore, the GBJS estimator is obtained by using the fractions (4.41).

This estimator seems to be appealing since the fractions in (4.35) may have different values, while the Bayesian interpreted James-Stein estimator (2.47) and the Baranchik estimator (2.49) have constant fractions or, in other words, a uniform shrinkage factor. The GBJS estimator is also an extended version of the Sclove estimator (2.53) which allows only two different shrinkage factors, 1 and $(1 - c/F_2^*)^+$.

Noticeably, the shrinkage portion, $s^2/\lambda_j\, a_{j,LS}^2$ is the reciprocal of the squared t-test statistic (or, equivalently $1/F_j'$) for the null hypothesis H_{0j}: $\alpha_j = 0$. Thus,

$$\begin{aligned} f_{j,GBJS} &= 0, && \text{if } F_j' \leq 1, \\ &= 1 - \frac{1}{F_j'}, && \text{if } F_j' > 1. \end{aligned} \tag{4.42}$$

In other words, (4.42) implies that the j^{th} principal component is deleted when the associated value of F_j' is less than 1, that is, delete the j^{th} principal component if

$$F_j' = \frac{\lambda_j\, a_{j,LS}^2}{s^2} \leq 1.$$

When the null hypothesis H_{0j} is rejected, the j^{th} fraction will be $(1 - 1/F_j')$.

The fraction $(1 - 1/F_j')$, which is shown in Figure 4.5, directly utilizes the value of the F test statistic which is sensitive to the problem of multicollinearity as discussed in Section 2.1.3. Thus, these fractions may not provide improved estimates over the LS estimates

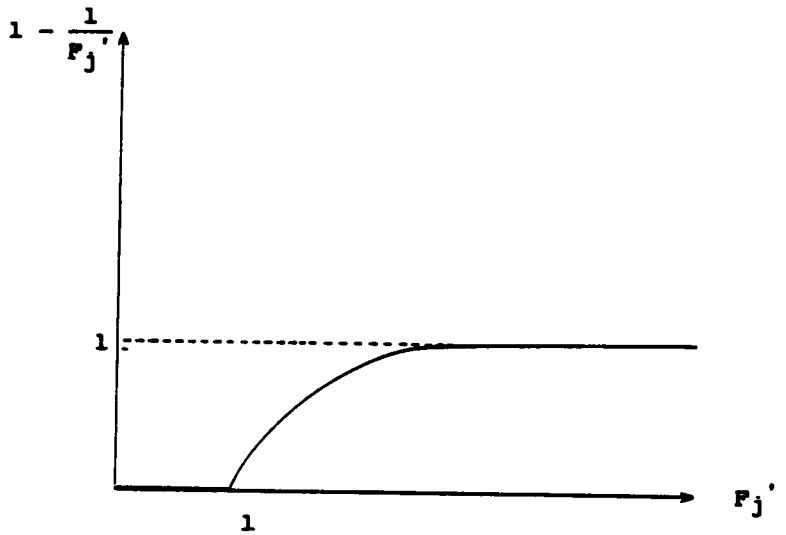

Figure 4.5 Fraction of GBJS Estimator: $(1 - 1/F_j')$

in terms of MSE. Furthermore, the common significance level, ϕ_o for all the hypotheses, H_{oj}: $\alpha_j = 0$, $j=1, \ldots, p$, is determined as

$$\phi_o = P\left[F_{1,n-p} \geq 1\right], \tag{4.43}$$

where $F_{1,n-p}$ is the F-distributed variable with the degrees of freedom, 1 and (n-p). Note that (4.43) is a special case of (2.55) when $\gamma = p-1$ and $c = 1/(n-p)$ and the value ϕ_o is no less than .25 for any value (n-p) from the F tables. The determination of the common ϕ_o is also inappropriate by the similar argument to the adjusted PC estimator [see Section 3.1.5].

In the following section, by investigating the flaws of the GBJS estimator and the adjusted PC_t estimator, another new member of the class of the FPC estimators will be proposed.

4.3.2 The Generalized PC_t Estimator

The fractions for the GBJS estimator are determined by the procedures of the testing hypotheses with the common significance level, ϕ_o. However, the common significance level in (4.43) is not appropriate at all since it is determined by only the value, (n-p). In addition, when $F_j' > 1$, the values of the fractions are functions of the F_j' whose values may be deteriorated due to multicollinearity. For the adjusted PC_t estimator, on the other hand, although the use of the different significance levels is reasonable to some situations as discussed in Section 3.1.5, its fraction is either 0 or 1. Since the significance levels are subjectively chosen, the decision for some principal components, either deletion or retention, may mislead the

performance of the adjusted PC_t estimator.

As discussed in this chapter so far, some fractions whose values are in the interval [0,1] can be utilized to obtain an improved estimator. Thus, the PC_t estimator can be generalized by choosing the different significance levels [the adjusted PC_t estimator] and defining the different fractions in the interval [0,1].

Consider, now, that the j^{th} fraction is a function of the F_j'. For example, the value f_j, can be determined by the linear function of the F_j' as in Figure 4.6. The broken straight line determines the fraction for the j^{th} principal component, denoted by $f_{j,GPCT}$, as:

$$\begin{aligned} f_{j,GPCT} &= 0, && \text{for } F_j' < F_{\phi_j;1,n-p}, \\ &= K_j + \frac{1-\phi_j}{F_{.000s;1,n-p} - F_{\phi_j;1,n-p}} F_j', \\ & && \text{for } F_{\phi_j;1,n-p} \leq F_j' < F_{.000s;1,n-p}, \\ &= 1, && \text{for } F_j' \geq F_{.000s;1,n-p}, \end{aligned} \tag{4.44}$$

where K_j is the intercept at $F_j' = 0$. The K_j is nothing but

$$K_j = 1 - \frac{1-\phi_j}{F_{.000s;1,n-p} - F_{\phi_j;1,n-p}} F_{.000s;1,n-p}.$$

Note that the subscript GPCT refers to "the generalized principal component based on the t-value criterion." The value $f_{j,GPCT}$ can be regarded as the j^{th} fraction after testing the j^{th} null hypothesis H_{0j}: $\alpha_j = 0$ with ϕ_j, where ϕ_j, $j=1, \ldots, p$ are the same significance levels as for the adjusted PC_t estimator. Note also that $F_{\phi_j;1,n-p}$ is the upper $100\ \phi_j$ percentile of the $F_{1,n-p}$ distribution and the value $F_{.0005;1,n-p}$

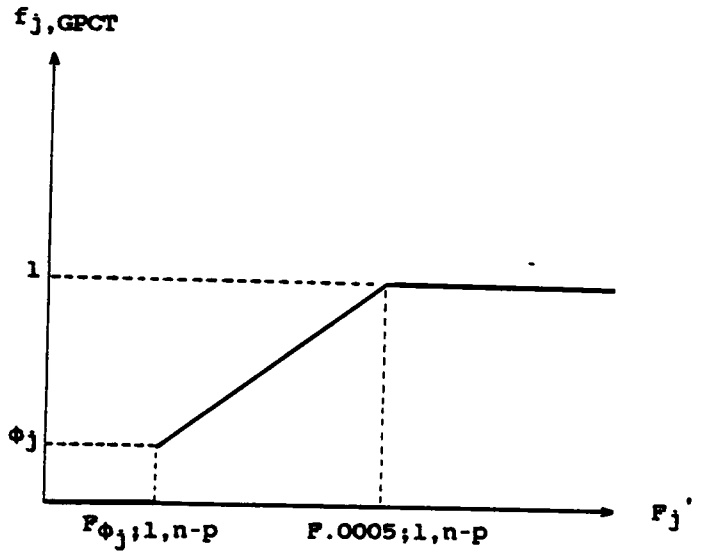

Figure 4.6 Fraction of the GPCT Estimator

is regarded as the upper 0 percentile point for convenience. Furthermore, at $F_j^* = F_{\phi_j;1,n-p}$, the fraction is chosen as ϕ_j via the straight line adjustment.

The resulting estimator with the fractions defined in (4.44) is termed the generalized PC_t (GPCT) estimator which is, in fact, a generalized version of the adjusted PC_t estimator in terms of the fractions in the interval [0,1]. Thus, the GPCT estimator of $\underline{\alpha}$ is defined as:

$$\underline{a}_{GPCT} = F_{GPCT}\, \underline{a}_{LS}, \tag{4.45}$$

where $F_{GPCT} = \text{Diag}(f_{1,GPCT}, \ldots, f_{p,GPCT})$.

It should be noted that the fractions in (4.44) may compromise the weakness of the adjusted PC_t estimator by applying values in [0,1] to some principal components whose optimal values are close to neitgher 0 or 1. Furthermore, the fractions are likely to be the linearized GBJS fractions when the different critical levels are regarded rather than $F_{\phi_0;1,n-p} = 1$. It should also be emphasized that the GPCT estimator, based on these simple linearized fractions, is an empirical result from a simulation study in the following chapter.

In the next chapter, the effects of the orientation of the parameter vector and the value of σ^2 on the performance of the PC estimators are examined and all the alternatives to the LS estimator including the commonly used biased estimators and the newly defined biased estimators in this research will be evaluated in terms of the empirical mean square error.

CHAPTER FIVE

SIMULATION STUDIES

The performance of many biased estimators has been studied in a number of Monte Carlo simulations. The common result of these studies is that the use of biased estimators leads to an improvement over the LS estimator, in terms of MSE, when multicollinearity is present. However, the advantages of some of these biased estimators, for example, the PC and the ridge estimators over the LS estimator, are functions of the orientation of the unknown parameter vector and the unknown σ^2.

5.1 Performance of PC Estimator

Different orientations of the parameter vector and different values of σ affect the performance of the PC estimator. This fact, demonstrated in Section 3.1.4 for p = 2, will be thoroughly illustrated in this section by using simulated data. In the simulation results that follow, two experimental factors will be controlled: (i) the orientation of $\underline{\beta}^*$, where $\underline{\beta}^*$ is the vector of the standardized coefficients, will change while $\underline{\beta}^{*'}\underline{\beta}^*$ will be held constant; and (ii) the values of σ^2 will vary.

5.1.1 Generated Data Sets: p = 2

Consider the two regressor variables situation where the 20 values of X_1 have been preselected, that is, n = 20, and X_2 is generated as a function of X_1 by

$$X_{i2} = 4.5 + 6.1\, X_{i1} + 1.1\, \epsilon_i, \qquad i=1, \ldots, 20,$$

where ϵ_i are independent standard normal random numbers. The variables X_1 and X_2 are then standardized so that $X^{*'}X^*$ is in correlation form. The generated values of X_1 and X_2 and the standardized data appear in Table 5.1. The two eigenvalues of $X^{*'}X^*$ are 1.99795 and .0020465, and hence the condition number is 31.2456.

Instead of choosing the orientations of the unknown parameters in terms of $\underline{\beta}^*$, it will be convenient to choose the orientations in terms of $\underline{\gamma}$, where $\underline{\gamma} = \Lambda^{1/2}\,\underline{\alpha} = \Lambda^{1/2}\,V'\underline{\beta}^*$. Thus two different orientations of $\underline{\gamma}$ were chosen as $(199.006,\ 4.5)'$ and $(4.5,\ 7.7971)'$. The orientations of the parameter vectors in terms of $\underline{\beta}^*$, $\underline{\alpha}$, $\underline{\gamma}$, and finally $\underline{\beta}$, are given in Table 5.2 with $\underline{\gamma}'\Lambda^{-1}\,\underline{\gamma} = \underline{\alpha}'\underline{\alpha} = \underline{\beta}^{*'}\underline{\beta}^* = 29717$. The first orientation of $\underline{\gamma}$, denoted by (A) has a relatively large value of γ_1 and a relatively small value of γ_2, while the values of γ_1 and γ_2 of the second orientation (B) are chosen as 4.5 and 7.7971, respectively. Note that the value of α_2 (172.357) of (B) is much larger than that of α_1 (3.1836).

The two sets of observations (i=1, ..., 20) are generated in terms of the original units by two different equations:

$$\text{(A)}\quad Y_i = 34.8 + 2.27165\,X_{i1} + 2.13966\,X_{i2} + \sigma\epsilon_i$$

$$\text{(B)}\quad Y_i = 34.8 - 9.30135\,X_{i1} + 1.56326\,X_{i2} + \sigma\epsilon_i,$$

where ϵ_i are independent standard normal random numbers and the value of σ is chosen as 5 or 10. Thus, the simulation will consider two parameter orientations, (A) and (B), and two values of σ, 5 and 10. Throughout this chapter, these four "treatment" combinations will be denoted by (A5), (A10), (B5), and (B10). Further, let γ_A and γ_B denote

Table 5.1 Generated X and X^* Matrices

X		X^*	
10.00000	64.46672	-0.29003	-0.30001
12.50000	79.51898	-0.09564	-0.11044
14.00000	89.58955	0.02099	0.01639
16.50000	106.61578	0.21538	0.23082
18.00000	113.92115	0.33201	0.32283
12.40000	79.57303	-0.10341	-0.10976
17.20000	108.57769	0.26981	0.25553
15.90000	102.24530	0.16873	0.17578
17.70000	113.28496	0.30869	0.31481
16.60000	104.83966	0.22316	0.20845
11.00000	70.87130	-0.21227	-0.21935
13.50000	84.66568	-0.01788	-0.04562
15.00000	97.16752	0.09875	0.11183
7.50000	50.48543	-0.48442	-0.47609
10.30000	69.33694	-0.26670	-0.23867
13.50000	86.42348	-0.01788	-0.02348
16.20000	105.97539	0.19206	0.22276
9.80000	64.20690	-0.30558	-0.30328
13.30000	86.17118	-0.03343	-0.02666
13.70000	87.82527	-0.00233	-0.00583

Table 5.2 Two Orientations of Parameter Vector

	(A)	(B)
$\underline{\gamma}$	$\begin{bmatrix} 199.006 \\ 4.5 \end{bmatrix}$	$\begin{bmatrix} 4.5 \\ 7.7971 \end{bmatrix}$
$\underline{\alpha}$	$\begin{bmatrix} 140.791 \\ 99.4738 \end{bmatrix}$	$\begin{bmatrix} 3.1836 \\ 172.357 \end{bmatrix}$
$\underline{\beta}^*$	$\begin{bmatrix} 29.2154 \\ 169.893 \end{bmatrix}$	$\begin{bmatrix} -119.624 \\ 124.126 \end{bmatrix}$
$\underline{\beta}$	$\begin{bmatrix} 34.8 \\ 2.27165 \\ 2.13966 \end{bmatrix}$	$\begin{bmatrix} 34.8 \\ -9.30135 \\ 1.56326 \end{bmatrix}$

the vectors of $\underline{\gamma}$ for each orientation A and B, respectively. Note that the values of γ_1 and γ_2 of (B) and the values of σ, 5 and 10 were chosen to reflect the various orientations of $\underline{\gamma}$ presented in Chapter 3. Since $\sqrt{\lambda_2 C} = 7.79845$ and $\sqrt{\lambda_1 C} = 243.7$, as having been discussed in Section 3.1.4, the conditions (3.24) and (3.25) can be considered with $\sigma = 5$ and $\sigma = 10$. With these given values, two figures are presented: Figure 5.1 for $\sigma = 5$, and Figure 5.2 for $\sigma = 10$.

In Figure 5.1, when $\sigma = 5$, the inequality (3.24) holds and the relationship $\gamma_2^2 < \sigma^2 < \gamma_1^2$ is satisfied with $\underline{\gamma}_A$ while the relationship $\gamma_1^2 < \sigma^2 < \gamma_2^2$ is satisfied with $\underline{\gamma}_B$. Thus, for the orientation $\underline{\gamma}_A$, the PC(2nd) estimator should be appropriate and for $\underline{\gamma}_B$, the PC(1st) estimator should be used to obtain smaller MSE than the LS estimator.

On the other hand, in Figure 5.2 with $\sigma = 10$, the inequality (3.25) and $\gamma_{2A}^2 < \sigma^2$ so that the PC(2nd) estimator is proper. For (B), since $\gamma_{1B}^2 + \gamma_{2B}^2 < 100$, both of the principal components should be deleted and hence the resulting PC estimator is $\underline{0}$.

These appropriate PC estimators will also be obtained via the simulated data in the following section.

5.1.2 Comparisons of PC Estimators with LS Estimator

As a measure of the performance of a biased estimator $\tilde{\underline{b}}$ of $\underline{\beta}$, in the simulation studies, the empirical mean square error (EMSE) can be used. The EMSE of $\tilde{\underline{b}}$ is defined as:

$$\text{EMSE}(\tilde{\underline{b}}) = \sum_{j=1}^{2} \sum_{\ell=1}^{50} \frac{(\tilde{b}_{j\ell} - \beta_j)^2}{50}, \qquad (5.1)$$

where $\tilde{b}_{j\ell}$ is the j^{th} estimated coefficient of $\tilde{\underline{b}}$ in the ℓ^{th} Monte Carlo

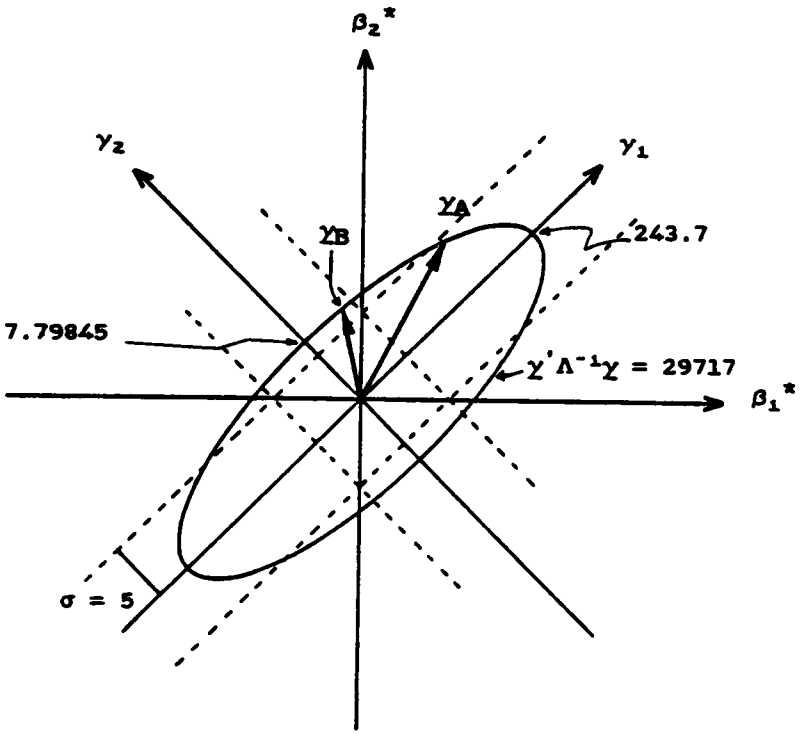

Figure 5.1 Orientations, γ_A and γ_B with $\sigma = 5$

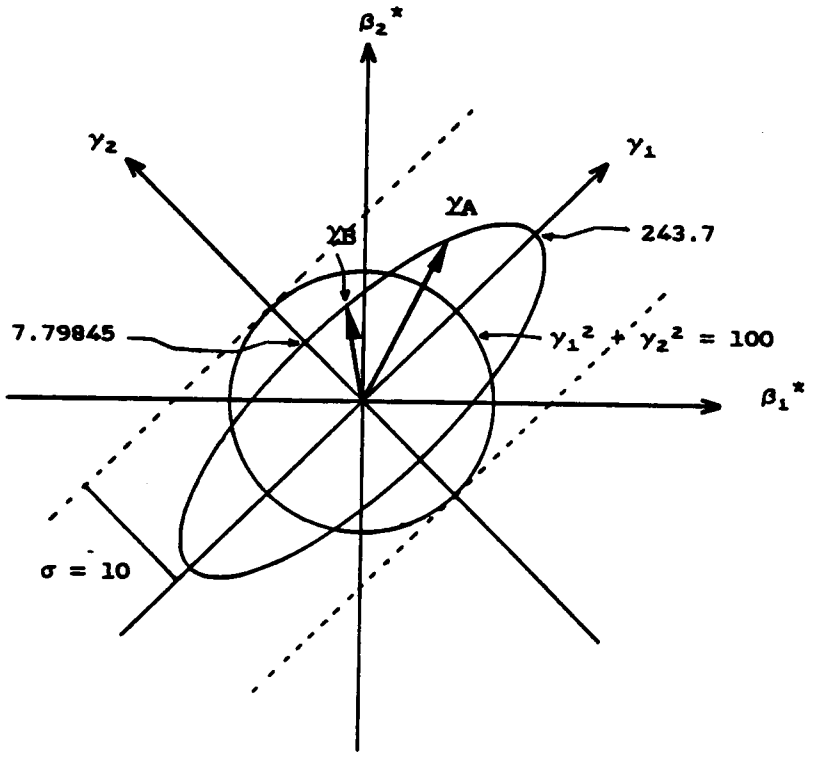

Figure 5.2 Orientations, γ_A and γ_B with $\sigma = 10$

repetition. Note that the number of Monte Carlo repetitions is 50 throughout this study.

In a manner similar to the MSE, the EMSE($\tilde{\underline{b}}$) can be decomposed into the empirical "variance" and "squared-bias" terms as, computationally,

$$\sum_{j=1}^{2} \sum_{\ell=1}^{50} \frac{(\tilde{b}_{j\ell} - \bar{\tilde{b}}_j)^2}{50}, \tag{5.2}$$

where $\bar{\tilde{b}}_j = \sum_{\ell=1}^{50} [\tilde{b}_{j\ell}/50]$ and

$$\sum_{j=1}^{2} [\bar{\tilde{b}}_j - \beta_j]^2, \tag{5.3}$$

respectively.

The comparisons of the possible PC estimators and the LS estimator for all the treatment combinations are displayed in Table 5.3(a) and (b). Note that the PC_t estimator uses the common significance level ϕ_o as .001. For the LS estimator, the EMSE($\underline{b}_{LS}$) can be described as:

$$\text{EMSE}(\underline{b}_{LS}) = \sum_{\ell=1}^{50} \frac{(\underline{b}_{\ell,LS} - \underline{\beta})'(\underline{b}_{\ell,LS} - \underline{\beta})}{50}$$

$$= \sum_{\ell=1}^{50} \frac{\underline{e}_\ell' X(X'X)^{-2} X'\underline{e}_\ell}{50}, \tag{5.4}$$

where $\underline{b}_{\ell,LS}$ is the LS estimate of $\underline{\beta}$ in the ℓ^{th} Monte Carlo repetition and $\underline{e}_\ell$ is the vector of the generated values of ϵ_i in the ℓ^{th} Monte Carlo repetition. Thus, the resulting values of the summary statistics for $\underline{b}_{LS}$ are equivalent for (A5) and (B5) in Table 5.3(a) and (b). The same result holds for the pair of (A10) and (B10). Furthermore, for the PC($\cdot$) estimators, the EMSE in terms of $\underline{b}_{PC}$ can be decomposed into

Table 5.3(a) EMSE of LS and PC Estimators of β

	LS	PC(2nd)	PC(1st)	PC(2)	PC_t
(A5)	48.97	29.85	87.14	-----	29.85
(A10)	195.86	29.11	211.70	-----	29.11
(B5)	48.97	90.66	48.94	-----	91.25
(B10)	195.86	89.27	196.75	88.96	88.96

Table 5.3(b) Decomposition of EMSEs in (a)

	LS		PC(2nd)		PC(1st)		PC(2nd)		PC_t	
	Var	$Bias^2$	Var	$Bias^2$	Var	$Bias^2$	Var	$Bias^2$	Var	$Bias^2$
(A5)	46.66	2.31	.04	29.81	46.89	40.25	-----	-----	.04	29.81
(A10)	186.63	9.23	.17	28.94	187.57	24.13	-----	-----	.17	28.94
(B5)	46.66	2.31	.04	90.62	46.89	2.05	-----	-----	11.01	80.24
(B10)	186.63	9.23	.17	89.10	187.57	9.18	.00	88.96	.00	88.96

$$\text{EMSE}(\underline{b}_{PC}) = \sum_{\ell=1}^{50} \frac{\left[\underline{b}_{\ell,PC} - \sum_{\ell=1}^{50} \underline{b}_{\ell,PC}/50\right]' \left[\underline{b}_{\ell,PC} - \sum_{\ell=1}^{50} \underline{b}_{\ell,PC}/50\right]}{50}$$

$$+ \left[\sum_{\ell=1}^{50} \underline{b}_{\ell,PC}/50 - \underline{\beta}\right]' \left[\sum_{\ell=1}^{50} \underline{b}_{\ell,PC}/50 - \underline{\beta}\right], \qquad (5.5)$$

where $\underline{b}_{\ell,PC}$ is the ℓ^{th} Monte Carlo repetition PC estimate of $\underline{\beta}$. Since $\left[\underline{b}_{\ell,PC} - \Sigma\underline{b}_{\ell,PC}/50\right] = VF\underline{a}_{\ell} - \Sigma VF\underline{a}_{\ell}/50 = (Z'Z)^{-1}Z'\underline{e}_{\ell} - \Sigma(Z'Z)^{-1}Z'\underline{e}_{\ell}/50$, the empirical "variance" part depends only on σ^2 while the empirical "squared-bias" part depends on σ^2 and the orientation of the parameter vector [see Table 5.3(b)].

From the results of the EMSEs in Table 5.3(a), it is of great interest to observe that the PC(1st) estimator should be used for (B5) though its contribution to the reduction of EMSE over that for the LS estimator is negligible. In fact, the appropriate PC estimator for each case coincides with the suggested PC estimators in the previous section.

Consider now the standardization of the EMSE($\tilde{\underline{b}}$) with respect to the true variances of the $b_{j,LS}$. The resulting standardized EMSEs of the LS estimator are all equal for all the cases in this present simulation study. The standardized empirical mean square errors (SEMSE) of $\tilde{\underline{b}}$ is carried out as

$$\text{SEMSE}(\tilde{\underline{b}}) = \sum_{j=1}^{2} \sum_{\ell=1}^{50} \frac{(\tilde{b}_{j\ell} - \beta_j)^2}{50\ \sigma_{j,LS}^2}, \qquad (5.6)$$

where $\sigma_{j,LS}^2$ is the variance of the j^{th} element of $\underline{b}_{LS}$. Then the SEMSE($\underline{b}_{LS}$) can be expressed as

$$\text{SEMSE}(\underline{b}_{LS}) = \sum_{\ell=1}^{50} \frac{(\underline{b}_{\ell,LS} - \underline{\beta})' \Sigma_D (\underline{b}_{\ell,LS} - \underline{\beta})}{50\,\sigma^2}$$

$$= \sum_{\ell=1}^{50} \frac{\underline{e}_\ell' X (X'X)^{-1} \Sigma_D (X'X)^{-1} X' \underline{e}_\ell}{50\,\sigma^2}, \qquad (5.7)$$

where Σ_D^{-1} is the diagonal matrix whose diagonal elements are the diagonal elements of $(X'X)^{-1}$. Since the numerator of (5.7) can be simplified as

$$\text{tr}\left[\underline{e}_\ell' X (X'X)^{-1} \Sigma_D (X'X)^{-1} X' \underline{e}_\ell\right] = \underline{e}_\ell' \Sigma_D (X'X)^{-1} \underline{e}_\ell,$$

the expression (5.7) is, consequently,

$$\text{SEMSE}(\underline{b}_{LS}) = \frac{1}{50} \sum_{\ell=1}^{50} \left[\frac{\underline{e}_\ell}{\sigma}\right]' \Sigma_D (X'X)^{-1} \left[\frac{\underline{e}_\ell}{\sigma}\right]. \qquad (5.8)$$

Therefore, the SEMSE($\underline{b}_{LS}$) is independent of both the orientation of the parameter vector and the value of σ. However, the SEMSE ($\underline{b}_{LS}$) will change with changes in the structure of the explanatory variables.

The SEMSEs for the possible PC estimators are in Table 5.4. From the table, it is easy to order the situations for which a specific PC estimator performs well. For example, it may be said that the PC(2nd) estimator performs three times better in situation (A10) than in situation (B10). Interestingly, the same can be said for the pair of (A5) and (B5). It is worth noting that even though the orientation of (B) is a <u>peculiar</u> one so that the usual PC estimator, PC(2nd), performs poorly for $\sigma = 5$, the PC estimation technique can be recommended for $\sigma = 10$.

Furthermore, by estimating the unbiased estimates of C, σ^2, and

Table 5.4 SEMSE of LS and PC Estimators

	LS	PC(2nd)	PC(1st)	PC(2)	PC_t
(A5)	3.6306	2.2569	8.2054	------	2.2569
(A10)	3.6306	.5677	4.7883	------	.5677
(B5)	3.6306	6.7648	3.6293	------	6.9243
(B10)	3.6306	1.6951	3.6267	1.6921	1.6921

the γ_j^2, the effects of the orientation and the error variance on the performance of the PC estimator can be examined. The values of the unbiased estimates are provided in Table 5.5. Note that they are the average values of the 50 Monte Carlo repetitions. Rewriting the unbiased estimate of C (= $\underline{\alpha}'\underline{\alpha}$), from (3.27), as

$$C_U = \underline{a}_{LS}'\underline{a}_{LS} - s^2 \sum_{j=1}^{2} \frac{1}{\lambda_j},$$

the inequality (3.28), which is an estimate of the inequality $\sigma < \sqrt{\lambda_2 C}$, holds. In particular, for (B5),

$$s = 4.9243 < \sqrt{\lambda_2 C_U} = 6.9944. \qquad (5.9)$$

Also, it follows by checking the order of s^2, $\hat{\gamma_1}^2$, and $\hat{\gamma_2}^2$ that

$$\hat{\gamma_1}^2 = 9.0644 < s^2 < \hat{\gamma_2}^2 = 48.9122 \qquad (5.10)$$

Thus, from (5.9) and (5.10), the PC(1st) estimator should be used for this special case, (B5). Note that $\hat{\gamma_j}^2$ is the unbiased estimator of γ_j^2 which is defined in (3.21).

For the case (B10), in addition, the unbiased estimate of σ, s lies in the interval $\left[\sqrt{\lambda_2 C_U}, \sqrt{\lambda_1 C_U}\right]$, that is,

$$\sqrt{\lambda_2 C_U} = 7.3341 < s = 9.8486 < \sqrt{\lambda_1 C_U} = 229.1569$$

and secondly,

$$\hat{\gamma_1}^2 + \hat{\gamma_2}^2 = 67.9763 < s^2.$$

Therefore, it would be expected that the PC(2) estimator improves over the LS estimator and, furthermore, has smaller SMSE than the PC(2nd)

Table 5.5 Unbiased Estimates of C, σ^2, γ_1^2, and γ_2^2 obtained through the simulations

	C_U	s^2	$\hat{\gamma_1^2}$	$\hat{\gamma_2^2}$
(A5)	27671.7	24.2486	38755.9	16.9324
(A10)	33816.5	96.9942	37924.8	30.3590
(B5)	23905.1	24.2486	9.06437	48.9122
(B10)	26283.4	96.9942	14.2024	53.7739

estimator. In other words, this is the case for which the trivial PC estimate, $\underline{a}_{PC} = \underline{0}$, is appropriate as in Figure 3.6.

Finally, in investigating the performance of the PC_t estimator from Table 5.4, the PC_t estimator with $\phi_0 = .001$ is not appropriate for (B5). For this specific case, the adjusted PC_t estimator would be recommended by applying different significance levels to the hypotheses, H_{0j}: $\alpha_j = 0$, j=1, 2 [see Section 3.1.5]. Since it is desirable to delete the first principal component and to keep the second PC in the PC analysis, two different significance levels ϕ_1 and ϕ_2 can be chosen as, for example, .0001 and .6, respectively. Thus, the power of test for H_{01} will be damaged, and that for H_{02} will be enhanced so that the first PC tends to be deleted and the second PC is retained. The resulting SEMSE of the adjusted PC_t estimator, $\underline{b}_{APCT}$ from the simulations is

$$\text{SEMSE}(\underline{b}_{APCT}) = 3.5399,$$

which is indeed close to the $\text{SEMSE}\left[\underline{b}_{PC(1st)}\right] = 3.6293$.

However, there are some difficulties, in practice, in choosing the values of ϕ_1 and ϕ_2. Though the choices of the different significance levels may be based upon experience, the comparisons of the magnitudes of s^2 and $\hat{\gamma_j}^2$, j=1, 2 may be helpful. For the interesting case (B5), from Table 5.5, since the unbiased optimal criterion ($\hat{\gamma_1}^2 < s^2$) suggests that the first PC should be deleted it is natural to take ϕ_1 close to 0. In addition, the unbiased optimal criterion suggests that the second PC should be retained so that the choice of ϕ_2, which is more subjective than that of ϕ_1, should be made large.

More precisely, since the PC estimation procedure is mainly based upon the decision of deletion or retention of each principal component, the adjusted PC_t estimator coincides with the suggested PC estimator resulting from applying unbiased optimal criterion. Thus, in particular, the value of ϕ_2 can be determined by applying various levels [for example, .3, .4, .5, etc.] in order to check whether the adjusted PC_t estimates are the same as the estimates of the suggested method in practice. Note that the adjusted PC_t are regarded only when the results of the regular PC_t estimator is far away from those of the suggested PC estimator.

The important problem related to principal components analysis which was introduced in Section 3.1.4 and 3.1.5 has been empirically viewed for p = 2. In the subsequent section, for p = 4, the comparisons of the various biased estimators will be made with some extended simulation data sets.

5.2 Comparisons of Biased Estimators

The performances of most biased estimators discussed in this paper will be compared, in terms of the standardized empirical MSE, by generating eight data sets for p = 4 in a similar manner as for p = 2. In addition to (i) and (ii) in the previous section, for more general comparisons, (iii) the change of the eigenvalue structures (equivalently, the degree of multicollinearity in the model) will be considered with the simulation procedures.

5.2.1 Generated Data Sets: p = 4

The explanatory variables X_1, X_2, X_3, and X_4 are established in the following way. Twenty values for X_1 and X_3 were independently preselected. And, by using these preselected values, X_2 and X_4 are generated as follows:

$$X_{i2} = 4.5 + 6.1\, X_{i1} + \epsilon_{i2}, \qquad i=1, \ldots, 20,$$

$$X_{i4} = .75 - 3.75\, X_{i2} + 5.25\, X_{i3} + \epsilon_{i4}, \qquad i=1, \ldots, 20,$$

where ϵ_{i2} and ϵ_{i4} are independent normal $(0, \sigma_2^2)$ and $(0, \sigma_4^2)$ random numbers, respectively.

Although the values of σ_2 and σ_4 determine the eigenvalue structure of $X'X$, the variation of σ_2 plays a key role. This is because the X_2 values are also used in generating X_{i4}. Two values of σ_2 are chosen as 2.1 and 4.1 and one value of σ_4 as 1.3. The resulting X matrix for $\sigma_2 = 2.1$ has one mild and one strong dependency ($\sigma_4 = 1.3$), while the X matrix for $\sigma_2 = 4.1$ has one weak and one strong dependency. The two sets of 20×4 X matrices along with the standardized X matrices, X^*, depending on the (σ_2, σ_4) combination, are given in Table 5.6(a) and (b). Also, the associated sets of the condition indexes of $X^{*\prime}X^*$ is supplied in Table 5.7 where the eigenvalues structure for ($\sigma_2 = 2.1$, $\sigma_4 = 1.3$) is denoted by (I) and the other by (II).

Since one major objective of this study is to investigate how different orientations of the parameter vector along the ellipsoid $\underline{\alpha}'\underline{\alpha} = C^*$, where C^* is fixed, affect the performances of biased estimators, four different choices of $\underline{\alpha}$ subject to $\underline{\alpha}'\underline{\alpha} = 304991$ were taken as in

Table 5.6(a) X and X^* Matrices for (σ_2 = 2.1, σ_4 = 1.3)

X				X^*			
10.00000	63.52738	25.70000	-104.00751	-0.29003	-0.29396	-0.09645	0.27104
12.50000	80.15732	34.10000	-119.08264	-0.09564	-0.09572	0.34031	0.22098
14.00000	89.17675	29.50000	-179.45786	0.02099	0.01179	0.10113	0.02050
16.50000	103.54194	33.20000	-212.33966	0.21538	0.18303	0.29351	-0.08869
18.00000	115.85583	31.10000	-271.52204	0.33201	0.32982	0.18432	-0.28522
12.40000	78.74884	20.50000	-189.51461	-0.10341	-0.11251	-0.36683	-0.01290
17.20000	111.64889	23.00000	-296.90512	0.26981	0.27967	-0.23684	-0.36951
15.90000	105.32143	25.50000	-260.83443	0.16873	0.20425	-0.10685	-0.24973
17.70000	117.53938	28.00000	-293.10904	0.30869	0.34989	0.02314	-0.35690
16.60000	106.79316	30.50000	-239.88858	0.22316	0.22179	0.15313	-0.18018
11.00000	72.85071	34.70000	-90.76078	-0.21227	-0.18282	0.37151	0.31503
13.50000	85.03747	28.80000	-165.89151	-0.01788	-0.03755	0.06473	0.06555
15.00000	101.04635	29.50000	-223.86227	0.09875	0.15328	0.10113	-0.12696
7.50000	49.31779	19.90000	-81.02018	-0.48442	-0.46334	-0.39802	0.34738
10.30000	64.45971	24.60000	-111.72797	-0.26670	-0.28284	-0.15365	0.24541
13.50000	83.79756	21.50000	-199.77653	-0.01788	-0.05233	-0.31483	-0.04698
16.20000	99.29425	24.00000	-245.03484	0.19206	0.13240	-0.18484	-0.19727
9.80000	64.08022	26.50000	-104.26253	-0.30558	-0.28737	-0.05486	0.27020
13.30000	86.78339	29.00000	-171.19733	-0.03343	-0.01674	0.07513	0.04793
13.70000	84.76920	31.50000	-152.40512	-0.00233	-0.04075	0.20512	0.11033

Table 5.6(b) X and X^* Matrices for (σ_2 = 4.1, σ_4 = 1.3)

X				X^*			
10.00000	61.64869	25.70000	-96.96243	-0.29003	-0.29364	-0.09645	0.27374
12.50000	79.59285	34.10000	-116.96591	-0.09564	-0.09463	0.34031	0.21182
14.00000	88.48794	29.50000	-176.87483	0.02099	0.00403	0.10113	0.02638
16.50000	102.01046	33.20000	-206.59661	0.21538	0.15400	0.29351	-0.06563
18.00000	117.33757	31.10000	-277.07858	0.33201	0.32399	0.18432	-0.28379
12.40000	77.42392	20.50000	-184.54617	-0.10341	-0.11868	-0.36683	0.00263
17.20000	113.77165	23.00000	-304.86545	0.26981	0.28444	-0.23684	-0.36981
15.90000	108.97041	25.50000	-274.51810	0.16873	0.23119	-0.10685	-0.27587
17.70000	122.36735	28.00000	-311.21395	0.30869	0.37977	0.02314	-0.38946
16.60000	107.77712	30.50000	-243.57844	0.22316	0.21796	0.15313	-0.18010
11.00000	74.04187	34.70000	-95.22761	-0.21227	-0.15619	0.37151	0.27911
13.50000	83.31124	28.80000	-159.41818	-0.01788	-0.05339	0.06473	0.08041
15.00000	105.85239	29.50000	-241.88494	0.09875	0.19661	0.10113	-0.17486
7.50000	48.42996	19.90000	-77.69084	-0.48442	-0.44025	-0.39802	0.33339
10.30000	61.72611	24.60000	-101.47695	-0.26670	-0.29278	-0.15365	0.25976
13.50000	80.89047	21.50000	-188.87495	-0.01788	-0.08023	-0.31483	-0.01077
16.20000	95.46020	24.00000	-230.65716	0.19206	0.08135	-0.18484	-0.14010
9.80000	63.88995	26.50000	-103.54902	-0.30558	-0.26878	-0.05486	0.25335
13.30000	87.88186	29.00000	-175.31658	-0.03343	-0.00270	0.07513	0.03120
13.70000	81.62558	31.50000	-140.61655	-0.00233	-0.07208	0.20512	0.13861

Table 5.7 Two Sets of Eigenvalues and Condition Indexes

	Eigenvalues	Condition Indexes
	2.96075	1.0000
(I)	1.02801	1.6971
(σ_2 = 2.1, σ_4 = 1.3)	.01112	16.3173
	.00012	157.0762
	2.95196	1.0000
(II)	1.01155	1.7083
(σ_2 = 4.1, σ_4 = 1.3)	.03638	9.0079
	.00011	163.8170

Table 5.8. The four orientations in terms of $\underline{\alpha}$ are, in fact, made from the orientations of $\underline{\gamma}$ by using the relationship $\underline{\gamma} = \Lambda^{1/2}\,\underline{\alpha}$. Thus, two sets of the first three elements of $\underline{\gamma}$ are fixed as:

$$(C)\quad \gamma_1 = 662.656, \quad \gamma_2 = 93.2298, \quad \gamma_3 = 35.1299$$

$$(D)\quad \gamma_1 = -88.5118, \quad \gamma_2 = -88.0519, \quad \gamma_3 = -2.97339.$$

Note that the magnitudes of γ_1 and γ_2 for both cases are so large that the first two principal components are regarded as important regressors. When the ellipsoid is fixed by $C^* = 304991$, the fourth element γ_4 is dependent upon the eigenvalue structure since $\gamma_4 = \sqrt{\lambda_4}\,\alpha_4$ and thus two different values of γ_4 with (C) [or (D)] are determined for (I) and (II). Therefore, denoting the orientation of the parameter vector for (I) with (C) by (IC), etc., all the orientations for (IC), (ID), (IIC), and (IID) are given in Table 5.8. Note that the absolute values of γ_4 for the (C)-oriented and (D)-oriented pairs are close to each other so that, for example, the γ_{IC} and γ_{IIC} are close together along the ellipsoid with $C^* = 304991$. Furthermore, the elements of $\underline{\gamma}$ related to (D) are ordered as $|\gamma_1| > |\gamma_2| > |\gamma_4| > |\gamma_3|$. Thus, the order of $|\gamma_3|$ and $|\gamma_4|$ may cause problems for some biased estimators. Table 5.8 also contains the true parameter vectors of the original-unit explanatory variables which are

$$\underline{\beta} = (\beta_0,\ X_{SD}^{-1}\underline{\beta}^*),$$

where $\beta_0 = 34.8$ and $X_{SD} = \mathrm{Diag}(S_{X1}, \ldots, S_{Xp})$, $S_{Xj} = \left[\Sigma_i (X_{ij} - \bar{X}_j)^2\right]^{1/2}$.

Finally, the observations of the response variable, for each case, are determined by

Table 5.8 Orientations of Parameter Vectors in Terms of χ, $\underline{\alpha}$, $\underline{\beta}^*$, and $\underline{\beta}$

	(IC)	(ID)	(IIC)	(IID)
χ	662.656	-88.5118	662.656	-88.5118
	93.2298	-88.0519	93.2298	-88.0519
	35.1299	-2.97339	35.1299	-2.97339
	-2.11488	5.94551	-3.49415	5.62222
$\underline{\alpha}$	385.112	-51.4399	385.685	-51.5164
	91.9509	-86.844	92.6961	-87.5478
	333.203	-28.2022	184.181	-15.589
	-192.875	542.225	-337.229	542.614
$\underline{\beta}^*$	-43.0837	-24.6925	76.8859	-26.4893
	209.724	351.367	54.0366	352.485
	56.5434	-49.7242	65.2993	-40.996
	-505.918	422.433	-540.274	422.33
$\underline{\beta}$	34.8	34.8	34.8	34.8
	-3.34998	-1.91997	5.97828	-2.05968
	2.5	4.18844	0.599306	3.90932
	2.93999	-2.58542	3.39525	-2.1316
	-1.68	1.40277	-1.67236	1.30728

$$Y_i = 34.8 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \quad i=1, \ldots, 20, \quad (5.11)$$

where ϵ_i are independent normal $(0, \sigma^2)$ random numbers. The two values of σ are chosen as 5 and 10 in order to vary the signal-to-noise ratio. Note that the generating equation of y (5.11) is equivalent to

$$Y_i = (34.8 + \bar{X}_1\beta_1 + \ldots + \bar{X}_4\beta_4) + \beta_1^* X_{i1}^* + \ldots + \beta_4^* X_{i4}^*.$$

For convenience, the eight simulated data sets are labeled as (IC5), (IC10), ..., (IID5), and (IID10), respectively. For example, (IIC10) indicates the set of data whose eigenvalue structure, the orientation of $\underline{\gamma}$, and the value of σ are based on (II), (D), and 10, respectively.

In the next section those biased estimators that have been presented in earlier chapters will be compared for eight simulated data sets [two exceptions are the Baranchik and the Sclove estimators, which are, in practice, difficult to compute]. At the end of the section, all estimators will be compared by using the standardized empirical MSE efficiencies relative to the LS estimator.

5.2.2 Evaluations of Biased Estimators

The performances of the biased estimators discussed in this study are evaluated in terms of the standardized empirical MSE(SEMSE) for each of the eight cases, (IC5) - (IID10). Recall from Section 5.1.2 that the standardization has been computed with respect to the $\mathrm{Var}(b_{j,LS})$ and hence the SEMSE for a biased estimator, $\tilde{\underline{b}}$, can be written as:

$$\mathrm{SEMSE}(\tilde{\underline{b}}) = \sum_{j=1}^{4} \sum_{\ell=1}^{50} \frac{(\tilde{b}_{j\ell} - \beta_j)^2}{50\ \sigma_{j,LS}^2}. \quad (5.12)$$

Comparisons of LS, R, GR, BJS, GBJS, and FR

The results of the SEMSEs for LS, ridge (R), generalized ridge (GR), Bayesian interpreted James-Stein (BJS), generalized Bayesian interpreted James-Stein (GBJS), and fractional rank (FR) estimators are provided in Table 5.9. Note that these estimators were presented and discussed in Sections (2.1.1), (2.2.1), (2.2.2), (2.3.1), (4.3.2), and (3.2.2), respectively. In addition, the optimal SEMSE is also included in the table, denoted MINMSE, which can be obtained by using the theoretically optimal fractions [see (4.12)].

From Table 5.9, the SEMSEs of the LS estimators are equivalent for the same eigenvalue structure since SEMSE ($\underline{b}_{LS}$) is dependent only upon the X matrix. The GR estimator is superior to most of the other estimators especially for the case (ID10) among all the cases with the eigenvalue system, (I). With the other eigenvalue structure, (II), the GR estimator still has the smallest SEMSE among all the other estimators, achieving a minimum value at the case (IIC5).

Comparing the SEMSEs of the R and the GR estimator, the GR estimator is superior to the R estimator for all but (IC5) and (IC10) cases. In other words, among the (C)-oriented cases, the performances of the two estimators are evenly split, while for all the (D)-oriented cases the GR estimator has smaller SEMSEs than the R estimator. Since the order of γ_3 and γ_4 is $|\gamma_3| < |\gamma_4|$ for (D), the order property of the fractions of the R estimator is not appropriate [see Section (4.1.1)]. Note that when the γ_j are well ordered, that is, $|\gamma_j| > |\gamma_{j+1}|$, for j=1, ..., p-1, the superiority of the two estimators seems to depend on the eigenvalue structure.

Table 5.9 SEMSE for Some Biased Estimators

	MINMSE	LS	R	GR	BJS	GBJS	FR
(IC5)	.6835	4.7435	1.7823	1.9742	4.7414	2.8969	1.6549
(IC10)	.3691	4.7435	1.2636	1.3592	4.7353	2.8120	1.7416
(ID5)	1.6866	4.7435	4.7670	4.2988	4.6923	4.0476	5.6090
(ID10)	.7456	4.7435	1.7019	1.0989	4.5417	2.1910	1.5745
(IIC5)	.9983	4.9363	2.4925	1.1067	4.9341	3.7535	2.9496
(IIC10)	.3939	4.9363	1.5071	1.1800	4.9275	3.2043	2.1627
(IID5)	1.5683	4.9363	4.8236	3.5139	4.8826	4.0602	5.1954
(IID10)	.6829	4.9363	1.8475	1.2368	4.7259	2.2009	1.4840

Furthermore, it is also important to observe that by comparing the pairs with respect to the change of σ, the greater the value of σ, then, generally, the smaller the SEMSE for each biased estimator. There are two exceptions: the FR estimator of the pair (IC5) and (IC10) and the GR estimator of the pair (IIC5) and (IIC10).

For the particular (D)-oriented cases with $\sigma = 5$, (ID5) and (IID5), it should be noticed that none of the biased estimators seem to perform particularly well.

Finally, from Table 5.9, the BJS and the GBJS estimators do not perform well compared with others, with the exception of the GBJS estimator for (ID5). Compared with others, the FR estimator performs extremely well for only the case (IC5) but never works for the particular cases, (ID5) and (IID5).

Comparisons of Possible PC Estimators

Consider, now, the performances of the various PC estimators through the simulations. For the usual PC estimators, that is, the PC(4th), PC(2), and PC_t with $\phi_o = .001$, the results in terms of SEMSE are shown in Table 5.10(a). In addition, Table 5.10(b) only contains the SEMSEs of the PC(3rd) and the adjusted PC_t estimator for the particular cases, (ID5) and (IID5), and the SEMSE of the adjusted PC_t estimator for (IC10) and (IIC10). Note that in Table 5.10(b) there are blank spaces for some cases where some usual PC estimators perform well.

In order to utilize the unbiased optimal deletion criterion, furthermore, the unbiased estimates of γ_3^2, γ_4^2, and σ^2 are obtained in Table 5.11. Note that the first two principal components are out of

Table 5.10(a) SEMSE of PC Estimators

	PC(4th)	PC(2)	PC_t
(IC5)	1.6108	44.8161	1.6108
(IC10)	1.1140	11.1510	7.9674
(ID5)	6.3654	5.2476	5.2476
(ID10)	2.3233	1.3120	1.3120
(IIC5)	2.9520	49.3543	2.9520
(IIC10)	1.5399	12.2279	8.4219
(IID5)	6.0476	4.7769	4.7769
(IID10)	2.3322	1.2009	1.2009

Table 5.10(b) SEMSE of PC(3rd) and Adjusted PC_t

	PC(3rd)	Adjusted PC_t
(IC5)	------	------
(IC10)	------	1.1140
(ID5)	4.1252	3.7155
(ID10)	------	------
(IIC5)	------	------
(IIC10)	------	1.5399
(IID5)	4.2259	3.8023
(IID10)	------	------

the question and hence the unbiased estimates of γ_1^2, γ_2^2 are not considered here. The true values of γ_3^2 and γ_4^2 from the orientations in Table 5.8 are also shown in the parentheses. Rewriting (3.22),

$$\sum_{j \in D} \hat{\gamma_j^2} \leq s^2,$$

the suggested PC estimators for each case are given in the last column of Table 5.11. The SEMSEs of the suggested PC estimators can be compared with others in Table 5.10(a) and (b). In particular, both (ID5) and (IID5) are interesting cases for which the usual PC estimators do not perform well.

From Table 5.10(a), the PC_t estimators with $\phi_0 = .001$ have large SEMSEs for the cases (IC10), (ID5), (IIC10), and (IID5). Keeping in mind the principal components suggested to be deleted from Table 5.11 for each of the four cases, the SEMSEs of the adjusted PC_t estimators can be obtained by applying different significance levels to the hypotheses H_{0j}: $\alpha_j = 0$, j=1, 2, 3, 4. Since the third PC is to be deleted and the fourth PC is not for the cases (ID5) and (IID5), take, for example, $\phi_1 = \phi_2 = .05$, $\phi_3 = .0005$, and $\phi_4 = .5$. Then, the resulting SEMSEs of the adjusted PC_t estimator are 3.7155 and 3.8023, respectively, [Table 5.10(b)], which are much less than those of the PC_t estimator. Similarly, since only the fourth PC should be deleted for (IC10) and (IIC10), take $\phi_1 = \phi_2 = .05$, $\phi_3 = .5$, and $\phi_4 = .0005$. Then the SEMSEs of the adjusted PC_t estimator are equivalent to those of the PC(4th) estimator. Note that the values of ϕ_1 and ϕ_2 are chosen as .05 since the first and second PCs will be almost sure to be included in the analysis. It should be emphasized that the adjusted

Table 5.11 Unbiased Estimates of γ_j^2, j=3, 4 and σ^2 and Suggested PC(·)

	$\hat{\gamma}_3^2$	$\hat{\gamma}_4^2$	$\hat{\gamma}_3^2 + \hat{\gamma}_4^2$	s^2	Suggested PC(·)
(IC5)	1189.27 (1234.11)	5.92 (4.47)	1195.19 (1238.58)	24.64	PC(4th)
(IC10)	1146.57	5.22	1151.79	98.55	PC(4th)
(ID5)	13.80 (8.84)	27.20 (35.35)	41.00 (44.19)	24.64	PC(3rd)
(ID10)	20.91	16.90	37.81	98.55	PC(2)
(IIC5)	1198.35	14.49	1212.84	24.64	PC(4th)
(IIC10)	1165.12	13.40	1178.52	98.55	PC(4th)
(IID5)	13.24 (8.84)	23.53 (31.61)	36.77 (40.45)	24.64	PC(3rd)
(IID10)	20.16	12.06	32.22	98.55	PC(2)

PC_t estimator can be used when the usual PC_t estimator does not provide the same estimates as the PC estimator suggested by the unbiased optimal criterion. Furthermore, the appropriate values of the ϕ_j can be determined when the two methods achieve the same estimates.

Comparisons of GPC_t, IFPCV, and IFPCI

The final partial table, Table 5.12 shows the simulation results of the SEMSEs for the new estimation techniques: the generalized PC_t (GPC_t), iterative fractional principal components estimators (IFPCV and IFPCI), which were introduced in Section (4.3.2) and Section (4.2.2), respectively. Note that the significance levels used in the GPC_t estimator are those in the adjusted PC_t estimator. Furthermore, the PC estimates that are used as the starting values are the suggested PC estimates for each case [see Table 5.11].

From Table 5.12, the GPC_t estimator has smaller SEMSE than either the IFPCV or the IFPCI estimator for the particular cases, (ID5) and (IID5). In comparisons between the two types of IFPC estimators, the IFPCV estimators have advantages in terms of SEMSE over the other for the (C)-oriented cases and the IFPCI estimators are preferred in the (D)-oriented cases. In other words, if the order of the γ_j is $|\gamma_j| > |\gamma_{j+1}|$, that is, the (C)-orientation case, then the IFPCV estimator will be more appropriate because the order of its fractions is corresponding to that of the γ_j. On the other hand, for the (D)-oriented parameter vector, the IFPCI estimator performs better than the IFPCV estimator due to the order of $|\gamma_3|$ and $|\gamma_4|$, that is, $|\gamma_4| > |\gamma_3|$.

Table 5.12 SEMSEs of GPC_t, IFPCV, and IFPCI

	GPC_t	IFPCV	IFPCI
(IC5)	1.5800	1.5596	1.6657
(IC10)	1.9036	1.2088	1.3510
(ID5)	2.8142	3.8906	3.3665
(ID10)	1.3678	1.3466	1.3335
(IIC5)	2.7926	2.3628	2.9623
(IIC10)	2.3193	1.4271	1.7453
(IID5)	2.7332	4.0519	3.3424
(IID10)	1.2757	1.4908	1.2257

It is noticeable that the GPC_t estimator has the smallest SEMSE among the three estimators in Table 5.12 for the particular cases, (ID5) and (IID5) which have been interesting cases so far in this study. However, for the cases (IC10) and (IIC10), the GPC_t estimator does not seem to work as well as the IFPC-type estimators, while for the rest of the cases it is competitive with them in improving over the LS estimator.

Overall Comparisons in Terms of Relative Efficiencies

In concluding the evaluations of the biased estimators, the efficiency of each biased estimator relative to the LS estimator can be defined as:

$$\text{SEMSE}(\underline{b}_{LS})/\text{SEMSE}(\tilde{\underline{b}}).$$

The relative efficiencies of all the biased estimators are given in Table 5.13. Since neither the BJS nor the GBJS estimator is competitive, they are excluded in the table. Note that the PC(·) represents the suggested PC estimator from Table 5.11 which is based on the unbiased optimal deletion criterion. Note also that the SEMSEs of the adjusted PC_t estimators from Table 5.10(b) are used for the cases (IC10), (ID5), (IIC10), and (IID5) of the PC_t column.

First of all, the largest relative efficiency for each case is asterisked in Table 5.13. Note that for the particular cases, (ID5) and (IID5), the GPC_t estimator is the most efficient and for the cases (ID10), (IIC5), and (IIC10), the GR estimator is the most efficient. The PC(4th) and the adjusted PC_t for (IC10) and the PC(2) and the PC_t

Table 5.13 Relative Efficiences of Biased Estimators to LS Estimator

	R	GR	FR	PC(·)	PC_t	GPC_t	IFPCV	IFPCI
(IC5)	2.6614	2.4027	2.8663	2.9448	2.9448	3.0022	3.0415*	2.8478
(IC10)	3.7540	3.4899	2.7236	4.2581*	4.2581*	2.4919	3.9241	3.5111
(ID5)	.9951	1.1034	.8457	1.1499	1.2767	1.6856*	1.2192	1.4090
(ID10)	2.7872	4.3166*	3.0127	3.6155	3.6155	3.4680	3.5226	3.5572
(IIC5)	1.9805	4.4604*	1.6735	1.6722	1.6722	1.7676	2.0892	1.6664
(IIC10)	3.2754	4.1833*	2.2825	3.2056	3.2056	2.1284	3.4590	2.8283
(IID5)	1.0234	1.4048	.9501	1.0334	1.2982	1.8061*	1.2183	1.4769
(IID10)	2.6719	3.9912	3.3263	4.1105*	4.1105*	3.8695	3.3112	4.0273

Notes: * refers to the largest relative efficiency for each case.

Adjusted PC_t estimators are used for cases: (IC10), (ID5), (IIC10), and (IID5) of the PC_t column.

with ϕ_o = .001 for (IID10) have the greatest relative efficiencies. Finally, for the case (IC5), the efficiency of the IFPCV estimator is the highest.

In general, the IFPCV estimator is more efficient than the R estimator for all cases, since the starting value of the latter is more inflated than that of the former. And the IFPCI estimator performs more efficiently than the GR estimator for all but the cases for which the GR estimator is curiously overwhelming.

CHAPTER SIX

CONCLUSIONS

The goals of this dissertation have been to discover the effects of both the orientation of the regression coefficient vector and the variance of the response variable on some biased estimators, in particular, the principal components estimator and the ridge estimator. Several heuristic biased estimators, which are members of the class of the fractional principal components estimators, have also been proposed for those situations where the MLR model suffers from the problem of multicollinearity.

6.1 Concluding Remarks

The present study, devoted to the multicollinearity problem, provides several conclusions:

(i) The orientation of the parameter vector and the value of σ^2 affects the performance of the PC estimator. When σ^2 is in a certain range, there are limited orientations of the parameter vector for which the regular PC estimator, that is, the principal components are deleted on the basis of the size of the eigenvalue, would remedy multicollinearity. It should be noted that either the PC($\cdot$) estimator or the PC_t estimator would not resolve the multicollinearity problem in those situations where the order of the γ_j does not agree with the order of the λ_j.

(ii) Furthermore, a new deletion criterion has been developed, named the unbiased optimal deletion criterion, where the associated PC estimator can be used even for those situations for which the regular PC estimator does not improve over the LS estimator in terms of MSE. This criterion can also be utilized to determine the different significance levels used in computing the principal components estimator based on t-tests. This estimator is called the adjusted PC_t estimator.

(iii) It has been known and demonstrated by many authors through simulation studies, that when the signal-to-noise ratio is very high the ridge estimator may not perform well. In the simple case of $p = 2$, the condition (4.18), that is, $(1 + 2\lambda_1/k) < \gamma'\gamma/\sigma^2$, indicates those situations where the signal-to-noise ratio is too high to expect good performance from the ridge estimator. In addition, if, from (4.16), $(1 + 2\lambda_2/k) < \gamma'\gamma/\sigma^2 < (1 + 2\lambda_1/k)$, then for only limited orientations of the parameter vector can the ridge estimator outperform the LS estimator in MSE.

(iv) The MSE of the theoretical generalized ridge estimator is always smaller than the MSE for the LS estimator regardless of any orientation of the parameter vector and σ^2. In the simulation study, except for one situation, the iterative generalized ridge estimator greatly improves over the LS estimator. In that one specific situation, the generalized ridge estimator offers only slight improvement over the LS estimator. Also in

this case, the iterative ridge estimator performs poorly. However, when the iterative ridge estimator is one of proper biased estimators, there is little difference between the ridge and the generalized ridge estimators.

(v) The most important conclusion of this study deals with the class of the fractional principal components estimators. The class contains all the plausible biased estimators useful in combating multicollinearity. The iterative ridge and generalized ridge estimators can be described in the framework of the optimal fractions. Furthermore, several new iterative estimation techniques corresponding to each of the iterative ridge-type estimators are suggested by using the PC estimates obtained under the unbiased optimal criterion as the starting values.

(vi) In fractional principal components analysis, several new estimators can be developed. One, the generalized Bayesian interpreted James-Stein estimator (GBJS), has been defined in this paper. Second, based on the concepts of this GBJS estimator and the adjusted PC_t estimator, the generalized PC_t estimator is also proposed. Even though the GBJS estimator is shown, in Section 4.3.1, to minimize the C_p-like statistic, its performance in terms of MSE is not competitive compared to the other biased estimators in the simulation study.

(vii) In order to investigate the effects of the orientation of the parameter vector and σ^2 on the biased estimators, two different 4×1 parameter vectors and two values of σ^2 are chosen in the simulation procedures. When the order of the γ_j is not in agreement with that of the eigenvalues and the value of σ is not large, most biased estimators, for example the ridge, fractional rank and the regular PC estimators, do not perform well. However, noticeably, the GPC_t estimator works very well in these circumstances.

(viii) Finally, from the results of the simulation study, even though any one of the four newly defined estimators is not a panacea for all the situations, they seem to perform substantially well.

6.2 Suggestions for Further Research

The investigation of the joint effects of influential cases and multicollinearity on linear regression analysis has been a prevalent research topic. For this purpose, the fractional principal components regression model (4.1), that is,

$$\underline{Y} = ZF^{-} \underline{\alpha}_F + \underline{\epsilon}$$

can be extended by premultiplying a diagonal matrix, W constructed to downweight the influential cases.

Postmultiplying Z by F^{-} and premultiplying W to Z results in

$$W Z F^{-} = \left[w_i z_{ij} f_j^{*} \right],$$

where w_i is the i^{th} diagonal element of W, z_{ij} is the i^{th} element of $\underline{z}_j$, and f_j^* is the j^{th} diagonal element of F^-. Under this scheme, the Z matrix can be simultaneously treated for the problems of multicollinearity and the influence of "extreme" observations. As a summary measure of the influence of the i^{th} observation, the "difference in fit standardized," denoted by $DFFITS_i$, was defined by Belsley, Kuh, and Welsch (1980) as:

$$DFFITS_i = \frac{\underline{z}_i'[\underline{a}_{LS} - \underline{a}_{LS(i)}]}{[Var(\underline{z}_i'\underline{a}_{LS})]^{1/2}},$$

where $\underline{a}_{LS(i)} = [Z_{(i)}'Z_{(i)}]^{-1} Z_{(i)}'\underline{y}_{(i)}$ and $Z_{(i)}$ is the $(n-1)\times p$ matrix excluding the i^{th} row of Z. Thus, the premultiplied matrix W can be constructed by utilizing the $DFFITS_i$, i=1, ..., n.

However, since the influential cases can be categorized into two types: inducing multicollinearity cases and masking multicollinearity cases, there exist some difficulties in adjusting the "weights" of the fractions.

Among other problems which have not been considered in this dissertation, several important suggestions are as follows: the first problem concerns the distributional properties of the FPC estimators. The problem is to find the distribution of the product of two random variables, the LS estimator $a_{j,LS}$ and the corresponding estimated fraction obtained by using the $a_{j,LS(s)}$. If its first and second moments can be solved, then the complete comparisons between the biased estimators may be observed. Second, broader simulation studies will supply more precise evaluation of the biased estimators. In particu-

lar, the change of the performances of the biased estimators may be more precisely investigated when the degree of multicollinearity varies systematically.

BIBLIOGRAPHY

Baranchik, A.J. (1964). "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution." Technical Report No. 51. Stanford: Dept. of Statistics, Stanford University.

Baranchik, A.J. (1970). "A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution." AMS 41:642-645.

Barnard, G.A. (1963). "The Logic of Least Squares." JRSSB 25:124-127.

Belsley, D.A., E. Kuh, and R.E. Welsch (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons, Inc.

Bhattacharya, P.K. (1966). "Estimating the Mean of a Multivariate Normal Population with General Quadratic Loss Function." AMS 37:1819-1824.

Blight, B.J.N. (1971). "Some General Results on Reduced Mean Square Error Estimation." American Statistician 25:24-25.

Dempster, A.P., M. Schatzoff, and N. Wermuth (1977). "A Simulation Study of Alternatives to Ordinary Least Squares." JASA 72:77-95.

Draper, N.R. and R.C.V. Nostrand (1979). "Ridge Regression and James-Stein Estimation: Review and Comments." Technometrics 21:451-466.

Farrar, D.E. and R.R. Glauber (1967). "Multicollinearity in Regression Analysts: The Problem Revisited." Review of Economics and Statistics 49:92-107.

Fomby, T.B. and R.C. Hill (1978). "Deletion Criteria for Principal Components Regression Analysis." American Journal of Agricultural Economics 60:524-527.

Gorman, J.W. and R.J. Toman (1966). "Selection of Variables for Fitting Equations to Data." Technometrics 8:27-51.

Graybill, F.A. (1976). Theory and Application of the Linear Model. North Scituate: Duxbury Press.

Greenberg, E. (1975). "Minimum Variance Properties of Principal Component Regression." JASA 70:194-197.

Gunst, R.F. and R.L. Mason (1977). "Biased Estimation in Regression: An Evaluation Using Squared Error." JASA 72:616-628.

Gunst, R.F. and R.L. Mason (1980). Regression Analysis and Its Application: A Data-Oriented Approach. New York and Basel: Marcel Dekker, Inc.

Hemmerle, W.J. (1975). "An Explicit Solution for Generalized Ridge Regression." Technometrics 17:309-314.

Hocking, R.R. (1976). "The Analysis and Selection of Variables in Linear Regression." Biometrics 32:1-49.

Hocking, R.R., F.M. Speed, and M.J. Lynn (1976). "A Class of Biased Estimators in Linear Regression." Technometrics 18:425-437.

Hoerl, A.E. and R.W. Kennard (1970a). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics 12:55-67.

Hoerl, A.E. and R.W. Kennard (1970b). "Ridge Regression: Applications to Nonorthogonal Problems." Technometrics 12:69-82.

Hoerl, A.E. and R.W. Kennard (1976). "Ridge Regression: Iterative Estimation of the Biasing Parameter." Commun. Statis. - Thoer. Meth. A5:77-88.

Hoerl, A.E., R.W. Kennard, and K.F. Baldwin (1975). "Ridge Regression: Some Simulations." Communications in Statistics 4:105-123.

James, W. and C. Stein (1961). "Estimation with Quadratic Loss." Proceedings Fourth Berkeley Symposium in Mathematical Statistics and Probability 1:361-379.

Jeffers, J.N. (1967). "Two Case Studies in the Application of Principal Component Analysis." Applied Statistics 16:225-236.

Johnson, S.R., S.C. Reimer, and T.P. Rothrock (1973). "Principal Components and the Problem of Multicollinearity." Metroeconomica 25:306-317.

Kendall, M.G. (1957). A Course in Multivariate Analysis. London: Charles W. Griffin & Co.

Lawless, J.F. and P. Wang (1976). "A Simulation Study of Ridge and Other Regression Estimators." Commun. Statis. - Theor. Meth. A5:307-323.

Lindley, D.V. and F.M. Smith (1972). "Bayes Estimates for the Linear Model." JRSSB 34:1-18.

Mallows, C.L. (1973). "Some comments on C_p." Technometrics 15:661-675.

Mandel, J. (1982). "Use of the Singular Value Decomposition in Regression Analysis." American Statistician 36:15-24.

Mansfield, E.R., J.T. Webster and R.F. Gunst (1977). "An Analytic Variable Selection Technique for Principal Component Regression." Appl. Statist. 26:34-40.

Marquardt, D.W. (1970). "Generalized Inverse, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." Technometrics 12:591-612.

Massy, W.F. (1965). "Principal Components Regression in Exploratory Statistical Research." JASA 60:234-256.

Mayer, L.S. and T.A. Willke (1973). "On Biased Estimation in Linear Models." Technometrics 15:497-508.

McCallum, B.T. (1970). "Artificial Orthogonalization in Regression Analysis." Review of Economics and Statistics 52:110-113.

McDonald, G.C. and D.I. Galarneau (1975). "A Monte Carlo Evaluation of Some Ridge-Type Estimators." JASA 70:407-416.

Mittelhammer, R.C. and J.L. Baritelle (1977). "On Two Strategies for Choosing Principal Components in Regression Analysis." American Journal of Agricultural Economics 59:336-343.

Montgomery, D.C. and E.A. Peck (1982). Introduction to Linear Regression Analysis. New York: John Wiley & Sons, Inc.

Myers, R.H. (1986). Classical and Modern Regression with Applications. Boston: Duxbury Press.

Newhouse, J.P. and S.D. Oman (1971). "An Evaluation of Ridge Estimators." Rand R-716-PR:1-28.

Obenchain, R.L. (1975). "Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis." Technometrics 17:431-441.

Perlman, M.D. (1972). "Reduced Mean Square Error Estimation for Several Parameters." Sankhyā 34:89-92.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications, 2nd ed. New York: John Wiley & Sons, Inc.

Rolph, J.E. (1976). "Choosing Shrinkage Estimators for Regression Problems." Commun. Statist. - Theor. Meth. A5:789-802.

Sclove, S.L. (1968). "Improved Estimators for Coefficients in Linear Regression." JASA 63:596-606.

Searle, S.R. (1971). Linear Models. New York: John Wiley & Sons, Inc.

Silvey, S.D. (1969). "Multicollinearity and Imprecise Estimation." JRSSB 31:539-552.

Stein, C. (1960). "Multiple Regression." Ch. 37 in I. Olkin (ed.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford: Stanford University.

Stuart, M. (1982). "A Geometric Approach to Principal Components Analysis." American Statistician 36:365-367.

Sujan, I. (1980). "Methods for Solving the Problem of Multicollinearity in Econometric Models: A Survey." Multicollinearity and Biased Esimation. Göttingen: Vandenhoeck & Ruprecht.

Swindel, B. (1981). "Geometry of Ridge Regression Illustrated." American Statistician 35:12-15.

Theobald, C.M. (1974). "Generalizations of Mean Square Error Applied to Ridge Regression." JRSSB 36:103-106.

Toro-Vizcarrondo, C. and T.D. Wallace (1968). "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression." JASA 63:558-572.

Tripp, R.E. (1983). "Nonstochastic Ridge Regression and Effective Rank of the Regressors Matrix." Unpublished Ph.D. dissertation, Dept. of Statistics, Virginia Tech.

Vinod, H.D. (1976). "Simulation and Extension of a Minimum Mean Squared Error Estimator in Comparison with Stein's." Technometrics 18:491-496.

Vinod, H.D. (1978). "A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares." Review of Economics and Statistics 60:121-131.

Vinod, H.D. and A. Ullah (1981). Recent Advances in Regression Methods. New York and Basel: Marcel Dekker, Inc.

Walker, E. (1984). "Influence, Collinearity and Robust Estimation in Regression." Unpublished Ph.D. dissertation. Dept. of Statistics, Virginia Tech.

Welsch, R.E. (1980). "Regression Sensitivity Analysis and Bounded-Influence Estimation." Evaluation of Econometric Models, eds. J. Kmenta, and J.B. Ramsey. New York: Academic Press, 153-167.

APPENDIX

For the completion of the equations, (4.40) and (4.41), the model (4.1) is rewritten,

$$\begin{aligned} \underline{y} &= Z\underline{\alpha} + \underline{\epsilon} \\ &= ZP^{-}P\underline{\alpha} + \underline{\epsilon} \\ &= ZP^{-}\underline{\alpha}_P + \underline{\epsilon}, \end{aligned}$$

where $\underline{\alpha}_P = P\underline{\alpha}$. Since $\underline{b}_P = V\underline{a}_P = VP\Lambda^{-1}Z'\underline{y}$, the residual sum of squares is

$$\begin{aligned} SSE_P &= (\underline{y} - X\underline{b}_P)'(\underline{y} - X\underline{b}_P) \\ &= (\underline{y} - ZP\Lambda^{-1}Z'\underline{y})'(\underline{y} - ZP\Lambda^{-1}Z'\underline{y}) \\ &= \underline{y}'\left[I - ZP\Lambda^{-1}Z'\right]^2\underline{y} \\ &= \underline{y}'\left[I - H_P\right]^2\underline{y}, \end{aligned} \tag{A.1}$$

where $H_P = ZP\Lambda^{-1}Z'$. Note that H_P is symmetric, but not idempotent. Now, considering the variance of $\hat{y}_{i,P}$ as follows:

$$\begin{aligned} Var(\hat{y}_{i,P}) &= Var(\underline{x}_i'\underline{b}_P) \\ &= \sigma^2\underline{x}_i'VP\Lambda^{-1}PV'\underline{x}_i, \end{aligned}$$

the standardized sum of $Var(\hat{y}_{i,P})$ is

$$\begin{aligned} \sigma^{-2}\sum_{i=1}^{n} Var(\hat{y}_{i,P}) &= tr(XVP\Lambda^{-1}PV'X) \\ &= tr(ZP^2\Lambda^{-1}Z') \\ &= tr(H_P^2), \end{aligned} \tag{A.2}$$

which is the "variance" part in the C_p-like statistic. Furthermore, consider the standardized sum of squares of biases in $\hat{y}_{i,F}$. That is,

$$\sigma^{-2}\sum_{i=1}^{n}(\text{Bias in } \hat{y}_{i,F})^2 = [E(X\underline{\beta} - X\underline{b}_F)]'[E(X\underline{\beta} - X\underline{b}_F)]$$

$$= (X\underline{\beta})'[I - H_F]^2(X\underline{\beta}). \qquad (A.3)$$

Since $E(SSE_F) = (X\underline{\beta})'[I - H_F]^2(X\underline{\beta}) + tr[(I - H_F)^2\sigma^2]$, the estimate for (A.3) is $SSE_F - s^2tr[(I - H_F)^2]$. Thus, we have

$$C_F = tr(H_F^2) + \frac{SSE_F - s^2tr[(I - H_F)^2]}{s^2}$$

$$= 2tr(H_F) - n + (n-p)\frac{SSE_F}{SSE}. \qquad (A.4)$$

Next, to obtain the conditions for minimizing C_F, consider the derivative with respect to F:

$$\frac{\partial C_F}{\partial_F} = 2\frac{\partial tr(H_F)}{\partial_F} + s^{-2}\frac{\partial SSE_F}{\partial_F}$$

$$= 2I + s^{-2}\left[-2\frac{\partial \underline{y}'H_F\underline{y}}{\partial_F} + \frac{\partial \underline{y}'H_F^2\underline{y}}{\partial_F}\right]$$

$$= 2I + s^{-2}\left[-2\begin{pmatrix}\lambda_1 a_{1,LS}^2 & & \\ & \ddots & \\ & & \lambda_p a_{p,LS}^2\end{pmatrix} + 2\begin{pmatrix}f_1\lambda_1 a_{1,LS}^2 & & \\ & \ddots & \\ & & f_p\lambda_p a_{p,LS}^2\end{pmatrix}\right]. \qquad (A.5)$$

Thus, setting $\partial C_F/\partial_F = 0$, we obtain

$$f_j = 1 - \frac{s^2}{\lambda_j a_j^2}, \quad j=1, \ldots, p.$$

The vita has been removed from the scanned document