THE DISTRIBUTION OF THE RATIOS OF CHARACTERISTIC ROOTS 
(CONDITION NUMBERS) AND THEIR APPLICATIONS IN 
PRINCIPAL COMPONENT OR RIDGE REGRESSION 

BY 

C. G. TROSKIE AND W. C. CONRADIE 

TECHNICAL REPORT NO. 212 
MAY 1985 

PREPARED UNDER THE AUSPICES 
OF 
NATIONAL SCIENCE FOUNDATION 
DMS 84-11411 
INGRAM OLKIN, PROJECT DIRECTOR 

DEPARTMENT OF STATISTICS 
STANFORD UNIVERSITY 
STANFORD, CALIFORNIA
THE DISTRIBUTION OF THE RATIOS OF
CHARACTERISTIC ROOTS (CONDITION NUMBERS)
AND THEIR APPLICATIONS IN PRINCIPAL
COMPONENT OR RIDGE REGRESSION

C.G. Troskie*
University of Cape Town

and

W.C. Conradie
University of Stellenbosch

Key words: Characteristic roots; condition numbers;
Principal component regression; Ridge regression;
Asymptotic results.

Summary: In regression analysis large condition numbers
indicate the presence of multicollinearity. Principal
component regression and Ridge regression are used to
correct for the ill effects of such collinearities. In
this paper some distributional properties of the condition
number are considered.

*Part of this research was done while the author was a
Visiting Professor at the University of Stanford. The
CSIR is thanked for financial assistance.

February 1985
1. INTRODUCTION

Let us consider \((p+1)\)-random variables \((y, x_1, \ldots, x_p)\) which jointly have a multivariate normal distribution. If \(Z' = (y, x_1, \ldots, x_p)'\) then let \(Z\) be distributed as a multivariate normal with mean \(\mu\) and covariance matrix \(\Sigma\), i.e. \(Z \sim N(\mu, \Sigma)\).

where \(\mu' = (\mu_y, \mu_x)' = (\mu_y, \mu_1, \ldots, \mu_p)\)

and \(\Sigma = \begin{pmatrix} \sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}\)

Now the regression function is the conditional expectation of \(y\) given fixed values of \((x_1, \ldots, x_p)\). The conditional mean and variance of \(y\) are given by

\[
E(y|x) = \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mu_x) = \beta_0 + \beta^{(2)}' x \quad \text{with} \quad \\
\beta^{(2)} = \Sigma_{xx}^{-1} \Sigma_{xy}
\]

\[
V(y|x) = \sigma^2 = \sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}.
\]

Suppose now a random sample of size \(n\), \((Z_1, \ldots, Z_n)\) is taken from \(Z\), i.e.

\[
(Z_1^t) = (Y_1, X_{11}, \ldots, X_{1p}) = (Y, X) .
\]

\[
\begin{pmatrix}
Z_1^t \\
\vdots \\
Z_n^t
\end{pmatrix} = 
\begin{pmatrix}
Y_1, X_{11}, \ldots, X_{1p} \\
\vdots \\
Y_n, X_{n1}, \ldots, X_{np}
\end{pmatrix} = (Y, X).
\]
Then if

\[(1.3) \quad A = \Sigma (Z_{\alpha} - \bar{Z})(Z_{\alpha} - \bar{Z})', \quad \bar{Z} = \frac{1}{n} \Sigma Z_{\alpha}\]

the maximum likelihood estimates of $\beta = \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$ and $\sigma^2$ are given by

\[(1.4) \quad \hat{\beta} = \left( \bar{Y} - \hat{\beta}(2)' \bar{X} \right) \left( \begin{array}{c} \beta_0 \\ \beta \end{array} \right), \quad \bar{Z} = \left( \begin{array}{c} \bar{Y} \\ \bar{X} \end{array} \right)\]

and

\[(1.5) \quad \hat{\sigma}^2 = \frac{1}{n}(a_{yy} - A_{yx}A_{xx}^{-1}A_{xy})\]

where

\[(1.6) \quad A = \begin{pmatrix} a_{yy} & A_{yx} \\ A_{xy} & A_{xx} \end{pmatrix}, \quad A_{xx} = \Sigma (X_{\alpha} - \bar{X})(X_{\alpha} - \bar{X})'\]

If the variables $(x_1, \ldots, x_p)$ are considered to be fixed it is often convenient to write the above model as

\[(1.7) \quad Y = X\beta + e\]

where $E(e) = 0$, $E(ee') = \sigma^2 I$ and $e$ is distributed $N(0, \sigma^2 I)$. Here $X$ is an $(n \times (p+1))$ matrix with the first column consisting of the unit vector $1' = (1, \ldots, 1)$.

The model (1.7) is the standard linear model when the variables $(x_1, \ldots, x_p)$ are either considered to be fixed or known independent variables. The estimates of $\beta$ and $\sigma^2$ are usually given by
(1.8) \( \hat{\beta} = (X'X)^{-1}X'Y, \quad s^2 = \frac{1}{(n-p-1)} (Y-X\hat{\beta})'(Y-X\hat{\beta}). \)

2. SOME DISTRIBUTIONAL RESULTS

In what follows a brief review of some of the distributional results will be given. It is mainly the purpose to highlight the differences between the fixed model given by (1.7) and the random model given by (1.1) where \( Z \sim N(\mu, \Sigma) \). For detailed discussions of the differences between these two models see Trostkie (1971) or Sampson (1974).

For the fixed case

(1.9) \( \hat{\beta} \sim N(\beta, \sigma^2(X'X^{-1}) \) and \( \frac{(n-p-1)s^2}{\sigma^2} \sim \chi^2_{n-p-1} \)

and independently.

For the random case, if the variables \( (x_1, \ldots, x_p) \) are considered fixed then the distribution of \( \hat{\beta} \) and \( \frac{(n-p-1)s^2}{\sigma^2} \) will, of course, again be given by (1.9).

Although the unconditional distribution of \( (n-p-1)s^2/\sigma^2 \) remains unchanged for the random case, the unconditional distribution of \( \hat{\beta} \) as given by (1.9) or (1.4) is entirely different and very complicated. The density is given by (Kabe (1968)).
(1.10) \[ g(\hat{\beta}_0, \hat{\beta}(2)) = K |\Sigma^{-1}_{xx} + (\hat{\beta}(2) - \hat{\beta}(2))_\sigma YY(\hat{\beta}(2) - \hat{\beta}(2))_\sigma|^{-\frac{n+1}{2}} \]
\[ \cdot \exp(-\frac{1}{2}n(\hat{\beta}_0 - \hat{\beta}_0) + (\hat{\beta}(2) - \hat{\beta}(2))_\mu X) \cdot \{(\hat{\beta}(2) - \hat{\beta}(2))_\Sigma XX(\hat{\beta}(2) - \hat{\beta}(2)) \}
\[ + (\sigma YY)^{-1} \cdot ((\hat{\beta}_0 - \hat{\beta}_0) + (\hat{\beta}(2) - \hat{\beta}(2))_\mu X) . \]

where

\[ K = (2\pi)^{-\frac{n}{2}} n^{\frac{3}{2}} (\sigma YY)^{\frac{1}{2}} |\Sigma^{-1}_{xx}|^{-\frac{1}{2}} (n-2)^{\frac{p}{2}} \cdot \Pi_{i=1}^{p} (C(n-p-2+i))^{-1} , C(n) = (2\pi)^{n/2} / \Gamma(\frac{1}{2}n) \]

The marginal density of \( \hat{\beta}(2) \) is given by

(1.11) \[ g(\hat{\beta}(2)) = k' |\Sigma^{-1}_{xx} + (\hat{\beta}(2) - \hat{\beta}(2))_\sigma YY(\hat{\beta}(2) - \hat{\beta}(2))|^{-\frac{1}{2}} \]

where

\[ k' = \Sigma^{-\frac{1}{2}}_{xx} |\sigma YY|^{-\frac{1}{2}} (n-1)^{\frac{p}{2}} \cdot \Pi_{i=1}^{p} \Gamma(\frac{1}{2}(n+1-i)) \cdot \Gamma(\frac{1}{2}(n-p-i)) \cdot \Pi_{i=1}^{p+1} \Gamma(\frac{1}{2}(n-i))^{-1} \]

and is a multivariate t-density.

Now for all practical purposes, i.e. tests of hypotheses, confidence statements etc the model (1.7) i.e. the fixed or conditional model is sufficient for both cases. The difference is only in the power functions. For example suppose we test at level \( \alpha \) the hypothesis \( H_0 : \beta(2) = 0 \) against \( H_1 : \beta(2) \neq 0 \). Then the F-statistic for both cases would be

(1.12) \[ F = \frac{\hat{\beta}(2)' C_{22}^{-1} \hat{\beta}(2)}{(Y'Y - \hat{\beta}'X'Y)} \cdot \frac{n-p-1}{p} \]
which is an $F$-distribution under $H_0$ with $p$ and $n-p-1$ degrees of freedom. Under $H_1$ for the fixed case $F$ has a noncentral $F$-distribution with noncentrality parameter.

$$\lambda = \frac{1}{2\sigma^2} \beta(2)' C_{22}^{-1} \beta(2)$$

where $(X'X)^{-1} = C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$.

The density of $u = \frac{p}{n-p-1} F$ is then given by

$$(1.13) \ c \cdot u^{\frac{1}{2}p-1}(1+u)^{-\frac{1}{2}(n-p-1+p)} e^{-\lambda \cdot 1 F_1(\frac{1}{2}n-p-1+p); \frac{1}{2}p \lambda u/(1+u))}.$$  

Under $H_1$ for the random case the density of $F$ given by (1.11) is no longer noncentral $F$ but the density of $u = \frac{p}{n-p-1} F$ is now given by

$$(1.14) \ c \cdot u^{\frac{1}{2}p-1}(1+u)^{-\frac{1}{2}(n-p-1+p)(1-\rho^2)}^{\frac{1}{2}n}$$

$$\cdot 2 F_1(\frac{1}{2}(n-1), \frac{1}{2}(n-1); \frac{1}{2}p; \rho^2 u/(1+u))$$

where $\rho^2 = \beta(2)' \Sigma_{xx} \beta(2)/\sigma_{yy}$ and is the population multiple correlation coefficient between $y$ and $(x_1, \ldots, x_p)$. For both cases

$$c = \Gamma(\frac{1}{2}(n-p-1+p))/\Gamma(\frac{1}{2}p)\Gamma(\frac{1}{2}(n-p-1)).$$

3. THE PROBLEM OF MULTICOLLINEARITY

Whether the $x$ variables are fixed or random it often happens that there are linear dependencies between the $x$
variables causing multicollinearities to exist in the X-matrix. In future we will make the assumption that, either the X-matrix is centered, or standardized. If X is centered then \( X'X = A \) (compare with (1.3)) that is \( X'X \) is a Wishart matrix, if \( (x_1, \ldots, x_p) \) are considered to be random variables. If \( X'X \) is standardized then \( X'X = R \), i.e. the correlation matrix between the x-variables while \( X'Y \) is the correlation matrix between \( Y \) and \( (x_1, \ldots, x_p) \) assuming that \( Y \) is also standardized.

The effects of multicollinearities have been discussed by several authors in the literature, notably Hoerl and Kennard (1970), Marquardt (1970), Webster, Gunst and Mason (1974, 1976).

The effects could best be seen by examining the latent roots and vectors of \( X'X \).

Let \( \lambda_1 \geq \ldots \geq \lambda_p > 0 \) be the latent roots of \( X'X \) (correlation matrix) and \( V_1, \ldots, V_p \) the corresponding vectors. If there are near multicollinearities some of the roots will be small. For convenience assume only one near multicollinearity, i.e. assume that \( \lambda_p \) is close to zero.

Several authors have suggested a correction to the least squares estimates if some or one of these roots are "too small". In the following we will give a brief outline of some of the more important procedures.
If
\[(3.1) \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p), \quad V = (V_1, \ldots, V_p) \]
then
\[(3.2) \quad V'X'XV = \Lambda \]
and \( \lambda_j = V_j'X'XV_j, \quad j = 1, \ldots, p. \)

Also
\[(3.3) \quad (X'X)^{-1} = V\Lambda^{-1}V' \]
\[= \sum_{j=1}^{p} V_j V_j' / \lambda_j. \]

Thus the least squares estimate \( \hat{\beta} \) is given by
\[(3.4) \quad \hat{\beta} = \sum_{j=1}^{p} \frac{1}{\lambda_j} V_j V_j' X' Y = \sum_{j=1}^{p} C_j V_j / \lambda_j \]
with \( C_j = V_j'X'Y. \)

The covariance matrix of \( \hat{\beta} \) is given by
\[(3.5) \quad \text{Cov}(\hat{\beta}) = \sigma^2 \sum_{j=1}^{p} V_j V_j' / \lambda_j, \quad \text{and the mean square error by} \]
\[(3.6) \quad \text{MSE} = \text{E}(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}. \]

Thus if \( \lambda_p \) is small the effect on the estimate \( \hat{\beta} \) as well as on the \( \text{Cov}(\hat{\beta}) \) and \( \text{MSE} \) could be quite serious. It is for this reason that Hoerl and Kennard (1970) suggested that a constant \( k \) should be added to each root if say \( \lambda_p \) is considered to be too small. The Ridge procedure proposed by them is then
(3.7) $\hat{\beta}_R = \sum_{j=1}^{p} C_j V_j / (\lambda_j + k)$ with

(3.8) $\text{Cov}(\hat{\beta}_R) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} V_j V_j'$

and

(3.9) $\text{MSE} = \sigma^2 \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \sum_{j=1}^{p} (\lambda_j + k)^{-2} (V_j' \beta)^2$.

The main theoretical advantage for Ridge by Hoerl and Kennard is their theorem that a strictly positive $k$ exists for which $\text{MSE} (\hat{\beta}_R) < \text{MSE} (\hat{\beta})$.

The method of Principal Components proposed by Marquardt (1970) eliminates the effect of the small root $\lambda_p$; that is

(3.10) $\hat{\beta}_{PC} = \sum_{j=1}^{p-1} C_j V_j / \lambda_j$ with

(3.11) $\text{Cov}(\hat{\beta}_{PC}) = \sigma^2 \sum_{j=1}^{p-1} \frac{1}{\lambda_j} V_j V_j'$

(3.12) $\text{MSE} (\hat{\beta}_{PC}) = \sigma^2 \sum_{j=1}^{p-1} \frac{1}{\lambda_j} + (V_p' \beta)^2$.

There is a considerable reduction in the $\text{MSE}(\hat{\beta}_{PC})$ over $\text{MSE}(\hat{\beta})$ if $(V_p' \beta)$ is not too large.

A method called Latent Root Regression Analysis (LRRA) was also proposed by Webster, Gunst and Mason (1974) and independently by Hawkins (1973).

They argue that the dependent variable $y$ may be involved in the multicollinearity. For example if constants
\[ a_1, \ldots, a_p, a_{p+1} \] exists such that \[ a_1 x_1 + \ldots + a_p x_p + a_{p+1} y \approx 0 \]
and if \( a_{p+1} \neq 0 \) then there is a perfect predictor of \( y \).
On the other hand if \( a_{p+1} = 0 \) then the multicollinearity exists only among the \( x \)'s and should be eliminated. The procedure is first to calculate

\[
(3.13) \quad S = \begin{pmatrix} X'X & X'Y \\ Y'X & Y'Y \end{pmatrix}
\]

the correlation matrix between \( y \) and \( (x_1, \ldots, x_p) \) and then the roots \( (\lambda_1, \ldots, \lambda_{p+1}) \) and vectors \( (V_1, \ldots, V_{p+1}) \) of \( S \). Then if \( \lambda_{p+1} \) is "too small" and \[ |V_{p+1, p+1}| \] is "too small" \( (V_{p+1}' = (V_{1p+1}, \ldots, V_{pp+1}, V_{p+1, p+1}) \) then the last vector and root are eliminated to obtain the LRRA estimate of \( \hat{\beta} \). Cutoff values for \( \lambda_{p+1} \) and \[ |V_{p+1, p+1}| \] depend on the number of variables \( (x_1, \ldots, x_p) \) and the sample size \( n \).

No precise rules have yet been formulated for determining the cutoff values as a function of \( p \) and \( n \). Gunst, Webster and Mason (1976) have given sufficient evidence, however, about the usefulness of the LRRA estimate of \( \beta \).

Several authors (e.g. Forsythe and Moler (1967), Marshall and Olkin (1971), Forsythe, Malcolm and Moler (1977), Vinod (1978)) proposed the computation of the "condition number" to measure the instability of a matrix when solving for a system of linear equations. The condition number is usually defined as

\[ C_Q = Q(A) \cdot Q(A^{-1}) \quad \text{where} \]
Q is usually taken as the norm. For matrix $A = X'X$ in the linear model (being symmetric) the condition number is $C(A) = \lambda_1/\lambda_p$ where $\lambda_1 > \ldots > \lambda_p$ are the characteristic roots of $A$. The condition number is a better measure of the nearness of singularity than the determinant of $A$. For example if $|A|$ is a $(100 \times 100)$ matrix with 0,1 on the diagonal then $A = 10^{-100}$ which is usually regarded as a small number. But the condition number of $A$ is $C(A) = \lambda_1/\lambda_p = (0,1)/(0,1) = 1$.

For linear systems of the type $Ax = b$ an $A$ as above behaves more like the identity matrix than like a singular matrix.

Recently Belsley, Kuh and Welsch (1980) stated "Most of the experimental evidence shows that weak dependencies begin to exhibit themselves with 'condition indices' around 10. A number in the neighbourhood of 15-30 tends to result from an underlying near dependency with an associated correlation of 0.9. Condition numbers of 100 or more appear to be large indeed causing substantial variance inflation and great potential harm to regression estimates." (Note that the condition index as defined by Belsley et al is $\sqrt{\lambda_1/\lambda_p}$.)

If the $x$-variables are random variables then the roots $(\lambda_1, \ldots, \lambda_p)$ are also random variables and then one would be interested in the distribution of the condition number $\lambda_1/\lambda_p$. Another measure which may also be of interest is
\( \frac{\Sigma \lambda_i}{\lambda_p} = \text{tr}(A)/\lambda_p \), which can also be considered as a measure of the condition of a matrix. In the next section we will focus attention on the underlying distributions that will affect the distributions of the condition numbers.

4. DISTRIBUTIONAL RESULTS

It is important at this moment to review the basic assumptions made in Section 1. First we assumed that \( Z = (y, x_1, \ldots, x_p) \) have a joint multivariate normal distribution and our sample \( (Z_1, \ldots, Z_n) \) is from this distribution. In Principal Components the matrix \( X'X = R \) is now the correlation matrix and if \( \lambda_1 > \cdots > \lambda_p > 0 \) are the roots of \( X'X \) then the distributional properties of these roots, and in particular that of \( \lambda_p, \lambda_1/\lambda_p \) and \( \Sigma \lambda_i/\lambda_p \) are of interest. With the LRRA estimates the roots of the matrix \( S = \begin{pmatrix} X'X & X'Y \\ Y'X & Y'Y \end{pmatrix} \), say \( \ell_1 > \cdots > \ell_{p+1} > 0 \) are of interest.

Since both matrices \( R \) and \( S \) are correlation matrices the distribution results are the same except for the dimensions, i.e. \( p \) for \( R \) and \( (p+1) \) for \( S \).

The second assumption that we made was that the variables \( (x_1, \ldots, x_p) \) can be considered to be fixed or known quantities. Thus \( X \) and \( X'X \) are known or fixed quantities and also the roots \( \lambda_1, \ldots, \lambda_p \) are known or fixed quantities and their distributional results are no longer of interest. For this case it is probably sufficient to
determine a cutoff value for say $\lambda_1/\lambda_p$ and to correct the
least squares estimates if $\lambda_1/\lambda_p$ is larger than this cut-
off value. Belsley et al have suggested a cutoff value for
$\sqrt{\lambda_1/\lambda_p}$ of 5 to 10 i.e. 25 to 100 for the condition number
$\lambda_1/\lambda_p$.

Although $R = X'X$ is fixed for this situation the
matrix $S$ is still random since $Y$ is a random variable and
the roots ($\ell_1, \ldots, \ell_{p+1}$) are therefor also random variables.
What is now of interest is the conditional distribution of
$\ell_1, \ldots, \ell_{p+1}$ given that $(x_1, \ldots, x_p)$ are known or fixed
quantities. We will now consider some of the distributional
properties. Let us first deal with the random case, i.e.
when $(y, x_1, \ldots, x_p)$ have a multivariate normal distribution.

If $Z' = (y, x_1, \ldots, x_p) \sim N(\mu, \Sigma)$ let $Z_1, \ldots, Z_N$ be a
random sample of $Z$ and let

$$ A = \sum_{\alpha=1}^{N} (Z_\alpha - \bar{Z})(Z_\alpha - \bar{Z})' $$

be the Wishart matrix with
distribution $W(\Sigma, n)$, $n = N-1$, with

$$ \Sigma = \begin{pmatrix} \sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} a_{yy} & A_{yx} \\ A_{xy} & A_{xx} \end{pmatrix} $$

If $D = \text{diag}(\sigma_{yy}, \sigma_{11}, \ldots, \sigma_{pp})$, $S = \text{diag}(a_{yy}, a_{11}, \ldots, a_{pp})$
then the population and sample correlation matrices are
given by

$$ P = D^{-\frac{1}{2}} \Sigma D^{-\frac{1}{2}} \quad \text{and} \quad R = S^{-\frac{1}{2}} AS^{-\frac{1}{2}}. $$
Now when $P = I$, i.e. all correlations $\rho_{ij} = 0$ and the variables $(y, x_1, \ldots, x_p)$ are independent then the density of $R$ is given by

\[ t(R) = K \cdot \left| R \right|^{\frac{1}{2}(n-(p+1)-1)} \]

where

\[ K = \left[ \Gamma \left( \frac{1}{2}(n-1) \right) \right]^{p+1} / \pi^{\frac{1}{2}p(p+1)} \prod_{i=1}^{p+1} \Gamma \left( \frac{1}{2}(n-i) \right). \]

Let $H$ be an orthogonal matrix such that

\[ H^T = \text{diag}(r_1, \ldots, r_{p+1}) \]

where $r_1 > \ldots > r_{p+1} > 0$ are the characteristic roots of $R$, the correlation matrix. Even under the assumption of $P = I$ or $\Sigma = \sigma^2 I$ the joint density of $r_1, \ldots, r_{p+1}$ are unknown in the literature and therefore it is not possible to find the densities of the condition numbers $r_1/r_{p+1}$, and $\Sigma r_1/r_{p+1}$, or that of the smallest root $r_{p+1}$. Unfortunately the case where $\Sigma = \sigma^2 I$ is not of practical interest, because this would imply that $\beta(2) = 0$ ($\beta(2) = \Sigma_{yx} \Sigma_{xx}^{-1}$ and $\Sigma_{yx} = 0$, see (1.1)) and therefore no linear relationship exists.

When $P \neq I$ the density of $R$ is even more complicated and is given in an integral form by Fisher (1962).

From the above it is obviously clear that much research is still needed before any results of practical value could be given.
The difficulty of finding the joint density of the characteristic roots \((r_1, \ldots, r_{p+1})\) of the correlation matrix \(R\) has led the authors to try and solve the problem in some other way.

If \(A\) and \(R\) are defined as in (4.1) and (4.3) and if \(E\) is an orthogonal matrix such that

\[
EAE' = \text{diag}(\lambda_1, \ldots, \lambda_{p+1}) = \Lambda
\]

then the relationship between the roots of \(A\) and \(R = H'\Omega H\), \(\Omega = \text{diag}(r_1, \ldots, r_{p+1})\) is as follows

\[
(4.7) \quad \Lambda = EAE' = ES^{\frac{1}{2}}RS^{\frac{1}{2}}E' = ES^{\frac{1}{2}}H'\Omega HS^{\frac{1}{2}}E'.
\]

It is quite clear from (4.7) that no easy relationship exists between the roots of \(A\) and \(R\). But since \(A = S^{\frac{1}{2}}RS^{\frac{1}{2}}\) and \(S\) is nonsingular, then if multicollinearities exist among the \(y\) and \(x\)-variables, i.e. some of the roots \((r_1, \ldots, r_{p+1})\) of \(R\), are small, then the same multicollinearities must exist in the \(A\)-matrix. But it is not easy to determine the magnitude of the roots of the \(A\)-matrix from the magnitude of the roots of the \(R\)-matrix. For example the small roots of \(A\) may be very much larger, or smaller, than the small roots of \(R\). The magnitude of the different roots all depend on the variation explained by the individual roots. If the dimensions of the variables are widely different, with some variables having large variances, then for
the matrix $A$ the first few of the large roots could explain virtually all the variation, and the remaining roots would be close to zero. But it still may not imply that $A$ is singular. When the variables are standardized, the variation explained by the different roots are more evenly spread with the result that the small roots of $R$ can explain in comparison more variation than the small roots of $A$, and are therefore larger than the small roots of $A$. An example of this situation will be given in Section 6.

For the LRRA estimates Gunst et al have suggested that an adjustment be made if $r_{p+1}$ the smallest root of $R$ is less than 0.001. Now if $\lambda_{p+1}$ is the smallest root of $A$ then it is not at all clear what the magnitude cutoff value for $\lambda_{p+1}$ should be. In Ridge regression a constant $k$ is added to the roots so that the "Ridge" correlation matrix if of the form

$$R^* = [R+kI]$$

and since $A = S^{1/2}R S^{1/2}$ we have

$$A^* = S^{1/2}R S^{1/2} = S^{1/2}RS^{1/2} + kS$$

$$= A + kS$$

$$= [a_{ij} + ka_{ij}]; \quad i,j = 1, \ldots, p$$

so that the adjustments to the $A$ matrix is in fact $a_{ij} + ka_{ij}, \quad i = 1, \ldots, p$. No easy relationship exists between the roots of $A^*$ and $A$ like the relationship between $R^*$ and $R$. 
The difficulty in determining what the magnitude of the small roots of $A$ should be, focused attention more on the condition numbers $\lambda_1/\lambda_{p+1}$ or $\Sigma \lambda_1/\lambda_{p+1}$. Now clearly if $r_1/r_{p+1}$, say measures the condition of $R$ and adjustments are made if $r_1/r_{p+1}$ is "too large" then clearly $\lambda_1/\lambda_{p+1}$ also measures the condition of $A$ and adjustments should be made if this number is too large. In other words if the smallest roots of $A$ are too small in relation with the others then adjustments should be made to the least squares estimates.

At this point it is again important to stress the underlying assumptions. If $(y, x_1, \ldots, x_p)$ are all random variables having a joint normal distribution and if

$$A = \begin{pmatrix} a_{yy} & A_{xy} \\ A_{yx} & A_{xx} \end{pmatrix}$$

as defined in (4.2) then for LRRA we would be interested in the roots and condition numbers of $A$ and for Ridge and Principal Component regression we would be interested in the roots and condition numbers of $A_{xx}$. But since $A_{xx}$ is a submatrix of $A$, it will also have a Wishart distribution and the difference will only be in the dimension (i.e. $p$, instead of $p+1$). For the underlying distribution theory the assumptions made on $\Sigma$ are however critical. Since

$$\Sigma = \begin{pmatrix} \sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}$$
We cannot assume $\Sigma = \sigma^2 I_{p+1}$ since in this case $\Sigma_{xy} = 0$ and $\beta^{(2)} = \Sigma_{yx} \Sigma_{xx}^{-1}$

$= 0$

and therefore no linear relationship exists between $y$ and $(x_1, \ldots, x_p)$. Thus for LRRA we need the distribution of the roots of $(\lambda_1, \ldots, \lambda_{p+1})$ of $A$ under general $\Sigma$. But for Ridge and Principal Component regression we need the distribution of the roots say $(\ell_1, \ldots, \ell_p)$ of $A_{xx}$, which will depend of $\Sigma_{xx}$ and a gross simplification in the distribution of the roots will be obtained if $\Sigma_{xx}$ is assumed to be of the form $\sigma^2 I_p$. For all practical purposes this assumption may be too restrictive. For LRRA we may of course also assume that $\Sigma_{xx} = \sigma^2 I_p$ while $\Sigma_{yx} \neq 0$. This assumption will also affect the distributional results of the roots $(\lambda_1, \ldots, \lambda_{p+1})$ of $A$ but at present it appears very difficult to determine this effect.

As mentioned before if the variables $(x_1, \ldots, x_p)$ are considered fixed or known quantities then the matrix $A_{xx}$ is considered known or fixed and the roots $(\ell_1, \ldots, \ell_p)$ of $A_{xx}$ are also known or fixed quantities and their joint distribution is therefore not of interest. In this case it is not clear what a "cut off" value for the condition numbers $\ell_1/\ell_p$ or $\Sigma \ell_i/\ell_p$ should be. For the correlation matrix a cutoff value of 25 to 100 was suggested (Belsley et al).
For the above fixed case the joint distribution of the roots \((\lambda_1, \ldots, \lambda_{p+1})\) of \(A\) are of course still of considerable interest and hence the assumptions made on \(\Sigma\) will be critical.

5. THE DISTRIBUTIONS OF THE CONDITION NUMBERS

In this section we will consider the distribution of the condition numbers. Many of the results will be very theoretical and extremely complicated. For the interested reader the applications of these results will be considered in a separate section.

We assume that \(A\) is distributed \(W(\Sigma, n)\) and that the characteristic roots of \(A\) are \(\lambda_1 > \ldots > \lambda_p > 0\). The joint distribution of the roots \((\lambda_1, \ldots, \lambda_p)\) has been derived by James (1960) and is given by

\[
K(p, n) |\Sigma|^{-\frac{1}{2}n} |\Lambda|^{\frac{1}{2}(n-p-1)} \text{etr}(-\frac{i}{2} \Lambda) \alpha_p(\Lambda) \Phi_0(i(I_p - \Sigma^{-1}), \Lambda)
\]

for \(0 < \lambda_p < \ldots < \lambda_1 < \infty\)

where

\[
K(p, n) = \prod_{i} p^2 / \{2^{\frac{1}{2}pn} \Gamma_p(\frac{1}{2}n) \Gamma_p(\frac{1}{2}p)\}
\]

\(\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p), \alpha_p(\Lambda) = \prod_{i < j} (\lambda_i - \lambda_j),\)

and
(5.3) \[ F_{q}(a_1,\ldots,a_p;\ b_1,\ldots,b_q;\ S,T) \]
\[ = \sum_{k=0}^{\infty} \prod_{\kappa} [(a_1)_{\kappa},\ldots,(a_p)_{\kappa}]/[(b_1)_{\kappa},\ldots,(b_q)_{\kappa}] \]
\[ \cdot C_{\kappa}(S)C_{\kappa}(T)/(C_{\kappa}(I)\ k!) \]

where \( a_1,\ldots,a_p,\ b_1,\ldots,b_q \) are real or complex constants and the multivariate coefficient \((a)_{\kappa}\) is given by

\[ (a)_{\kappa} = \prod_{i=1}^{p} (a - \frac{1}{2}(i-1))_{k_i} \]

where \((a)_{k} = a(a+1)\ldots(a+k-1)\).

The partition \( \kappa \) of \( k \) is such that

\[ \kappa = (k_1,k_2,\ldots,k_p),\ k_1 > k_2 > \ldots > k_p > 0, \]

\[ k_1 + k_2 + \ldots + k_p = k \]

and the zonal polynomials \( C_{\kappa}(S) \), are expressible in terms of elementary symmetric functions (esf) of the latent roots of \( S \), James (1960).

Now let

(5.4) \[ u_i = \lambda_i/\sum_{i=1}^{p} \lambda_i, \quad i = 2,\ldots,p \]

and let

(5.5) \[ S = \sum_{i=1}^{p} \lambda_i. \]

The Jacobian of the transformation is easily seen to be \( S^{p-1} \).

The joint density of \((u_2,\ldots,u_p)\) and \( S \) is seen to be
(5.6) \( f(u_2, \ldots, u_p, S) = K(p, n) \left| \Sigma \right|^{-\frac{1}{2} n} \left| U \right|^{\frac{1}{2} (n-p-1)} \text{etr}(-\frac{1}{2} S) \]
\[ \times \left( \frac{1}{2} p(n-p-1) + \frac{1}{2} p(p-1) \right)^{-p-1} \alpha_p(U) \sum_{k=0}^{\infty} \Sigma_k S^k \]
\[ \cdot C \left( \frac{1}{2} (I - \Sigma^{-1}) \right) C_k(U) / (C_k(I)k!) \]

where

(5.7) \( U = \text{diag}(1 - \sum_{i=2}^{p} u_i, u_2, \ldots, u_p) \)

Integration over \( S, \ 0 < S < \infty \) we have the joint density of \( (u_2, \ldots, u_p) \) as

(5.8) \( (\frac{1}{2} p^2 / \Gamma_p(\frac{1}{2} n) \Gamma_p(\frac{1}{2} p)) \left| \Sigma \right|^{-\frac{1}{2} n} \left| U \right|^{\frac{1}{2} (n-p-1)} \alpha_p(U) \]
\[ \sum_{k=0}^{\infty} \Sigma_k \{C_k(I - \Sigma^{-1}) C_k(U) \Gamma(\frac{1}{2} np + k) / (C_k(I)k!) \} \]

James (1966) has considered the joint distribution of \( (\lambda_i / \bar{\lambda}) \ i = 2, \ldots, p \) where \( \bar{\lambda} = \frac{1}{p} \sum \lambda_i \). Krishnaiah and Waikar (1972) have derived the joint density of \( (u_2, \ldots, u_p) \) in the above form.

If \( A \) has a noncentral Wishart distribution then the joint density of \( (u_2, \ldots, u_p) \) can again be derived easily from the joint density of \( (\lambda_1, \ldots, \lambda_p) \) which is given by

(5.9) \( h(\lambda_1, \ldots, \lambda_p) = K(p, n) \text{etr}(-\frac{1}{2} \Omega) \left( \frac{1}{2} n \right) \Gamma_1(\frac{1}{2} n; \frac{1}{2} \Omega, A) \]
\[ |A|^{\frac{1}{2} n-p-1} \alpha_p(A) \text{etr}(-\frac{1}{2} A) \]

where \( (\lambda_1, \ldots, \lambda_p) \) are the roots of \( |A - \lambda \Sigma| = 0 \) and \( \Sigma \) is assumed to be known.
The joint density of \((u_2,\ldots,u_p)\) is then

\[
(5.10) \quad h(u_2,\ldots,u_p) = (\pi^{p/2} / \Gamma_p)^{1/p} \Gamma_p(\tilde{z}n) \exp(-\tilde{z}\Omega) \\
|U|^{\frac{1}{2}(n-p-1)} \alpha_p(U) \sum_{k=0}^{\infty} \sum_{\kappa} \left\{ C_{\kappa}(U) C_{\kappa}(\Omega) \Gamma(\tilde{z}np+k) / (\tilde{z}n) \right\} \kappa(\Gamma(\tilde{z}np+k))
\]

Since \(u_p = \lambda_p / \Sigma \lambda_i\) the marginal density of \(u_p\) can be found by integrating out \(u_2,\ldots,u_{p-1}\) over the range

\(0 < u_p < u_{p-1} < \ldots < u_2 < u_1 = 1 - \sum_{i=2}^{p} u_i\). To find an explicit expression for \(u_p\) does not seem to be possible.

The condition number \(\Sigma \lambda_i / \lambda_p\) is of course given by

\[
(5.11) \quad C_1 = \frac{1}{u_p} = \Sigma \lambda_i / \lambda_p.
\]

We will now derive the density of the condition number \(C_2 = \lambda_1 / \lambda_p\).

If the joint density of \((\lambda_1,\ldots,\lambda_p)\) is given by (5.1) make the transformation

\[
(5.13) \quad \ell_i = \frac{\lambda_1 - \lambda_i}{\lambda_1}, \quad i = 2,\ldots,p.
\]

The condition number is then

\[
(5.14) \quad C_2 = \lambda_1 / \lambda_p = 1/(1-\ell_p)
\]

The Jacobian is \(J(\lambda_1,\lambda_2,\ldots,\lambda_p; \lambda_1,\ell_2,\ldots,\ell_p) = \lambda_1^{p-1}\).

If \(\Lambda_\ell = \text{diag}(\ell_2,\ldots,\ell_p)\) with \(1 > \ell_p > \ldots > \ell_2 > 0\), then
etr(-\frac{1}{2}\Lambda) = e^{-\frac{1}{2}\lambda_1 p} \text{etr}(\frac{1}{2}\lambda_1 \Lambda_\ell), \prod \lambda_i^{\frac{1}{2}(n-p-1)} = \lambda_1^{\frac{1}{2}(n-p-1)p} |I-\Lambda_\ell|^{\frac{1}{2}(n-p-1)}

\alpha_p(\Lambda) = \prod_{i<j}(\lambda_i - \lambda_j) = \lambda_1^{\frac{1}{2}p(p-1)} |\Lambda_\ell| \prod_{i<j}(\ell_j - \ell_i)

and

C_k(\Lambda) = C_k(\lambda_1^4 (I-\Lambda_\ell)) = \lambda_1^k C_k(I-L)

where

'(I-A_\ell) = \text{diag}(1,1-\ell_2,\ldots,1-\ell_p).

Now using the well-known expansion (Constantine (1963))

\[ C_k(\ell(I-A_\ell)) = \sum_{k=0}^{\infty} \sum_{\tau} b_{\kappa,\tau} C_{\tau}(I-A_\ell) \]

we get the joint density

\[ f(\lambda_1, \ell_2, \ldots, \ell_p) = k \cdot |\Sigma|^{-\frac{1}{2}n} \lambda_1^{\frac{1}{2}np-1} e^{-\frac{1}{2}\lambda_1 p} \]

\[ \sum_k \Sigma_k (C_k(\ell(I-S^{-1}))) \lambda_1^k / k! C_k(I) \]

\[ \sum_{k=0}^{\infty} \sum_{\tau} b_{\kappa,\tau} C_{\tau}(I-A_\ell) \text{etr}(\frac{1}{2} \lambda_1 \Lambda_\ell) \]

\[ |I-\Lambda_\ell|^{\frac{1}{2}(n-p-1)} |\Lambda_\ell| \prod_{i<j}(\ell_j - \ell_i) \]

But

\[ C_{\tau}(I-\Lambda_\ell) = \sum_{\tau} \alpha_{\tau,\tau} C_{\tau}(\Lambda_\ell) \]

\[ \text{etr}(\frac{1}{2}\lambda_1 \Lambda_\ell) = \sum_{m} \sum_{\mu} (\frac{1}{2}\lambda_1)^m C_{\mu}(\Lambda_\ell) / m! \]

\[ C_{\tau}(\Lambda_\ell) C_{\mu}(\Lambda_\ell) = \sum_{\delta} g_{\tau,\mu} C_{\delta}(\Lambda_\ell) \]

and

\[ |I-\Lambda_\ell|^{-\frac{1}{2}(p+1-n)} C_{\delta}(\Lambda_\ell) = \sum_{s} \sum_{\sigma} [\frac{1}{2}(p+1-n)]_{\sigma} C_{\sigma}(\Lambda_\ell) / s! \]
\[
= \Sigma_s \Sigma_\sigma \Sigma_\nu \left[ \frac{1}{2} (p+1-n) \right] \sigma g^\nu, \delta C_\nu(\Lambda_\xi)/s!
\]

Thus

\[
(5.16) \quad |I-\Lambda_\xi|^{-\frac{1}{2}} (p+1-n) \text{etr}(\frac{1}{2} \lambda_1 \Lambda_\xi) C_{\tau}(I-\Lambda_\xi)
\]

\[
= \Sigma_{\tau} \Sigma_{\sigma} \Sigma_{\nu} \Sigma_{\mu} \Sigma_{\delta} \Sigma_s \Sigma_\sigma \Sigma_\nu \left( \frac{1}{2} \lambda_1 \right)^m a_{\sigma}, \mu g^\nu \delta \mu g^\nu \delta
\]

\[
\cdot (p+1-n) \sigma C_\nu(\Lambda_\xi)/m! s!
\]

and

\[
(5.17) \quad f(\lambda_1, \ell_2, \ldots, \ell_p) = K|\Sigma|^{-\frac{1}{2} \lambda_1 \lambda_1 p + 1} e^{-\frac{1}{2} \lambda_1 p}
\]

\[
\Sigma \Sigma_{C_\kappa} \left( \frac{1}{2} (I-\Sigma^{-1}) \right) \lambda_1^k m b_{\kappa}, \alpha_{\tau}, \frac{1}{2} (p+1-n) \sigma
\]

\[
g^\delta, \mu g^\nu \delta C_\nu(\Lambda_\xi)|\Lambda_\xi| \Pi_{1<i<j}(\ell_j-\ell_i)
\]

where

\[
(5.18) \quad \Sigma^* = \Sigma_k \Sigma_\kappa \Sigma_t = \Sigma_\tau \Sigma_{\sigma} \Sigma_\nu \Sigma_\mu \Sigma_\delta \Sigma_s \Sigma_\sigma \Sigma_\nu
\]

Now, let \( r_i = \ell_i/\ell_p, \ i = 2, \ldots, p-1 \) in (5.17) with Jacobian equal to \( \ell_p^{p-2} \); then \( |\Lambda_\xi| = \ell_p^{p-1}| \Lambda_{\ell} | \)

and

\[
C_\nu(\Lambda_\xi) = \ell_p^{s+t+m} C_\nu(\Lambda_{\ell}), \quad \Pi_{1<i<j}(\ell_j-\ell_i)
\]

\[
= \ell_p^{\frac{1}{2} (p-1)(p-2)} |I_p-2-\Lambda_\xi| \Pi_{1<i<j}(r_j-r_i)
\]

The joint density of \( (\lambda_1, r_2, \ldots, r_{p-1}, \ell_p) \) is then

\[
(5.19) \quad f(\lambda_1, r_2, \ldots, r_{p-1}, \ell_p) = K|\Sigma|^{-\frac{1}{2} \lambda_1 \lambda_1 p - 1}
\]

\[
e^{-\frac{1}{2} \lambda_1 p} \Sigma \Sigma_{C_\kappa} \left( \frac{1}{2} (I-\Sigma^{-1}) \right) \lambda_1^k m (\frac{1}{2} (p+1-n)) \sigma
\]

\[
b_{\kappa}, \alpha_{\tau}, \frac{1}{2} \ell_p^{(p-1)(p+2)+s+t+m-1}
\]

\[
g^\delta, \mu g^\nu \delta C_\nu(\Lambda_\xi)|\Lambda_\xi| \Pi_{1<i<j}(\ell_j-\ell_i)
\]
\begin{align*}
&\Lambda_r |C_v(\Lambda_r) | I_{p-2-\Lambda_r} | \Pi_{1<i<j} r_j-r_i \\
&\text{To find the marginal density of } (\lambda_1 \text{ and } \ell_p) \text{ we integrate over } (r_2, \ldots, r_{p-1}), \text{ i.e. (see also Sugiyama (1961)),}
\
&(5.20) \int r_p > \ldots > r_2 > 0 \Lambda_r |C_v(\Lambda_r) | I_{p-2-\Lambda_r} | \Pi_{i=2}^{p-1} dr_i \\
&= [\frac{1}{2}(p-1)(p+2)+s+t+m] \Gamma_{p-1}(\frac{1}{2}(p-1)) \Gamma_{p-1}(\frac{1}{2}(p+2)) \Gamma_{p-1}(\frac{1}{2}p)C_v(I_{p-1}) \\
&\cdot \frac{1}{(\frac{1}{2}(p+2))\sqrt{(\Pi \frac{1}{2}(p-1)^{2}\Gamma_{p-1}(p+1)(p+1)v)}}
\\
&\text{The marginal density of } (\lambda_1, \ell_p) \text{ is then}
\
&(5.21) f(\lambda_1, \ell_p) = \Pi \frac{1}{2} \Gamma_{p-1}(\frac{1}{2}(p-1)) \Gamma_{p-1}(\frac{1}{2}(p+2)) \\
&\cdot |\Sigma|^{-\frac{1}{2}n} \Gamma_{p-1}(\frac{1}{2}p)/(\Gamma_{p}(\frac{1}{2}p)\Gamma_{p}(\frac{1}{2}n)2^{\frac{1}{2}np} \Gamma_{p-1}(p+1)) \\
&\cdot \lambda_1^{\frac{1}{2}np-1} e^{-\frac{1}{2}\lambda_1 p} \sum_{k=0} C_{\kappa}(\frac{1}{2} I - \Sigma^{-1}) \lambda_1^k \\
&\lambda_p^{\frac{1}{2}(p-1)(p+2)+s+t+m-1}(\frac{1}{2}(p+1-n))^{\sigma} \\
&b_{\kappa, \tau} \alpha_{\tau} \cdot g_{\ell, \mu} g_{\sigma, \delta} (\frac{1}{2}(p-1)(p+2)+s+t+m) \\
&(\frac{1}{2}(p+1))v C_v(I_{p-1})/(k! s! m! 2^{m} C_{\kappa}(I_{p})(p+1)v) .
\\
&\text{Integrating w.r.t. } \lambda_1 \text{ i.e. } \int_{0}^{\infty} \lambda_1^{\frac{1}{2}np+k+m-1} e^{-\frac{1}{2}\lambda_1 p} d\lambda_1 \\
&= \Gamma(\frac{1}{2}np+k+m)/(\frac{1}{2}p)^{\frac{1}{2}np+k+m}
\\
&\text{we get the marginal density of } \ell_p \text{ as}
\end{align*}
\begin{equation}
(5.22) \quad g(\ell_p) = \frac{\pi^{p-\frac{1}{2}} \Gamma_{p-1}(\frac{1}{2}(p-1)) \Gamma_{p-2}(\frac{1}{2}(p+2)) \Gamma_{p-1}(\frac{1}{2}p) |\Sigma|^{-\frac{1}{2}n}}{\Gamma_p(\frac{1}{2}p) \Gamma_p(\frac{1}{2}n) 2^{\frac{1}{2}np} \Gamma_{p-1}(p+1)}
\end{equation}

\[ \Sigma^* C_\kappa(\frac{1}{2}(I-\Sigma^{-1})) \ell_p^{\frac{1}{2}}(p-1)(p+2) + s + t + m - 1(\frac{1}{2}(p+1-n))_\sigma \]

\[ b_{\kappa,\tau,\alpha,\gamma} g_{\tau,\mu,\sigma,\delta}(\frac{1}{2}(p+2)) (\frac{1}{2}(p-1)(p+2) + s + t + m) \]

\[ \Gamma(\frac{1}{2}np + k + m) C_\nu(I_{p-1}) / (k! m! s! C_\kappa(I_p)(p+1)) \nu 2^m \]

\[ (\ell_p)^{\frac{1}{2}np + k + m} \]

and \( \Sigma^* \) is given by (5.18).

Krishnaiah and Waikar (1972) also reported that they have derived the density of \( \lambda_i / \lambda_p \) but considered the expression so complicated that it is of no practical value and hence the density was not published.

It is quite clear from the above expressions that the densities of the condition numbers \( C_1 = \Sigma \lambda_i / \lambda_p \) and \( C_2 = \lambda_i / \lambda_p \) are extremely complicated for general \( \Sigma \) and would be of no practical value. Two alternatives are open to us, i.e. relax the assumptions on \( \Sigma \) or consider some asymptotic results. These will be considered in the next two sections.
6. THE DISTRIBUTIONS OF THE CONDITION NUMBERS WITH RESTRICTIONS ON THE COVARIANCE MATRIX $\Sigma$.

In Section 4 it was indicated that for Principal Component or Ridge Regression the distribution of the roots $(\lambda_1, \ldots, \lambda_p)$ of $A$ is also of interest if $\Sigma$ is assumed to be of the form $\sigma^2 I_p$. This would imply that the variables $(x_1, \ldots, x_p)$ are independent. If $\frac{1}{N} A$ and $R = \Sigma^{-\frac{1}{2}} A \Sigma^{-\frac{1}{2}}$ are the maximum likelihood estimates of $\Sigma = \sigma^2 I$ and $P = I$ respectively, then it is unlikely that multicollinearities will be present in the $A$ or $R$ matrix. If such multicollinearities are however present then the condition numbers $C_1 = \Sigma \lambda_i / \lambda_p$ and $C_2 = \lambda_1 / \lambda_p$ will, of course, be much larger than what would have been expected for an orthogonal system.

The joint distribution of the roots $(\lambda_1, \ldots, \lambda_p)$ for the case $\Sigma = \sigma^2 I$ is now much easier to handle and percentage points of the condition number $C_1 = \Sigma \lambda_i / \lambda_p$ can be found from the tables by Schuurman, Krishnaiah and Chattopadhyya (1973). They tabulated the percentage points of $\lambda_p / \Sigma \lambda_i = \frac{1}{C_1}$ for significance level $\alpha = (0.05; 0.01)$. More extensive tables are available in a technical report (ARL-730010) by the authors. The tables are given for $p = 3(1)5$ and $r = \frac{1}{2}(n-p-1) = 0(1)25$ but restricted to $r = 0(1)16$ for $p = 6$.

Tables of the percentage points of the condition number $C_2 = \lambda_1 / \lambda_p$ are given by Krishnaiah and Schuurman (1974).
We now consider two examples to illustrate some of the uses of these tables. The first example was reported by Troskie (1977). To illustrate the effect of multicollinearity Troskie (1977) performed a regression on eight securities of the Johannesburg Stock Exchange with one variable being the dependent variable. The characteristic roots of the sample covariance matrix $S = \frac{1}{N-1}A$ of the 7 independent variables was given by

$\lambda_{1S} = 26585.002; \lambda_{2S} = 6107.498; \lambda_{3S} = 4340.910; \lambda_{4S} = 864.108;
\lambda_{5S} = 303.634; \lambda_{6S} = 63.763; \lambda_{7S} = 23.537$

The condition number $C_{2S} = \lambda_{1S}/\lambda_{7S} = 1129$, while $C_{1S} = \Sigma \lambda_i/\lambda_7 = 1626$.

For the correlation matrix $R$ the characteristic roots are as follows:

$\lambda_{1R} = 3.393; \lambda_{2R} = 1.224; \lambda_{3R} = 1.004; \lambda_{4R} = 0.645;
\lambda_{5R} = 0.363; \lambda_{6R} = 0.067; \lambda_{7R} = 0.004$

The condition numbers are $C_{2R} = \lambda_{1R}/\lambda_{7R} = 948$, $C_{1R} = \Sigma \lambda_i/\lambda_7 = 1750$.

Notice that by inspection one would not consider that the smallest root of $S$ i.e. $\lambda_{7S} = 23.537$ is small. On the other hand the smallest root of $R$ i.e. $\lambda_{7R} = 0.004$ can certainly be considered to be too small. It is very interesting, however, to note that the condition numbers for both matrices are hardly different in magnitude and, in fact extremely large, reflecting the ill-condition of these
matrices. Under the assumption that \( \Sigma = \sigma^2 I \) the approximate critical values for these condition numbers are (from the tables reported above) with \( n = 104 \), \( C_1 S(0.01) = 21 \) and \( C_2 S(0.01) = 15 \).

Now obviously the assumption that \( \Sigma = \sigma^2 I \) is not feasible (and of course will be rejected if tested based on the sample roots). Nevertheless the difference in magnitude of the observed condition numbers and the critical values under the assumption \( \Sigma = \sigma^2 I \), is so large that one would immediately expect that the least squares estimates will be seriously affected if the matrices \( S \) or \( R \) are used without adjustment.

The second example is taken from data supplied by Thompson (1978).

The characteristic roots of the covariance matrix \( S \) and correlation matrix \( R \) are given by \( (p = 9 \) independent variables).
\[
\begin{array}{cc}
\lambda_{iS} & \lambda_{iR} \\
722.272 & 3.296 \\
201.667 & 3.154 \\
22.338 & 1.021 \\
6.829 & .808 \\
.562 & .363 \\
.443 & .215 \\
.134 & .106 \\
.029 & .034 \\
.001 & .004
\end{array}
\]

\[
C_{2S} = \frac{\lambda_1}{\lambda_p} = 722272 \\
C_{1S} = \Sigma \lambda_i / \lambda_p = 953110
\]

\[
C_{2R} = 849 \\
C_{1R} = 2250
\]

All the condition numbers are extremely large indicating that the matrices S and R are probably ill-conditioned. It is remarkable how much larger the condition numbers for the matrix S are compared to that of the matrix R. There is a definite indication that a strong multicollinearity exists among the independent variables and that a correction procedure is necessary for the least squares estimates. Because of the large number of variables \((p = 9)\) critical values for the condition numbers are not available in the tables given by Krishnaiah et al. More extensive tables are however available in a technical report.

The following table illustrates the difference in magnitude of the least squares estimates and some correction procedures as suggested in Section 3.
### Troskie

<table>
<thead>
<tr>
<th>O.L.S.</th>
<th>Ridge Trace</th>
<th>Principal Comp.</th>
<th>LRRA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k = 0.2</td>
<td>rank(X'X) = 6</td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>4625.7</td>
<td>3605.9</td>
<td>1382.1</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>6.9807</td>
<td>4.372</td>
<td>6.3295</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-8.8436</td>
<td>-3.3826</td>
<td>-9.6453</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.4058</td>
<td>-1.3295</td>
<td>0.6796</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>7.8023</td>
<td>7.1952</td>
<td>11.0451</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-5.2807</td>
<td>-0.2072</td>
<td>0.4712</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.4217</td>
<td>0.8562</td>
<td>0.3622</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>5.7400</td>
<td>-0.5619</td>
<td>-0.1760</td>
</tr>
</tbody>
</table>

### Thompson

<table>
<thead>
<tr>
<th>O.L.S.</th>
<th>Ridge Trace</th>
<th>Principal Comp.</th>
<th>LRRA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k = 0.1</td>
<td>rank(X'X) = 8</td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>2.1016</td>
<td>0.0835</td>
<td>6.1883</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.7088</td>
<td>0.4017</td>
<td>0.9558</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-1.9220</td>
<td>-0.1071</td>
<td>-4.6870</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>1.1256</td>
<td>0.0944</td>
<td>0.0650</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.1130</td>
<td>0.2485</td>
<td>0.0922</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.1795</td>
<td>0.1369</td>
<td>0.2218</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.0183</td>
<td>0.0158</td>
<td>-0.0187</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-0.0776</td>
<td>-0.0610</td>
<td>-0.0818</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>-0.0852</td>
<td>-0.0059</td>
<td>-0.0116</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-0.3432</td>
<td>-0.0906</td>
<td>-0.2740</td>
</tr>
</tbody>
</table>
For both examples the effect of only one small root on the Principal Component and LRRA estimates have been eliminated. Investigating the effect of the second smallest root one finds for the first example:

$$\lambda_{1S} = 26585.002 \quad \lambda_{1R} = 3.393$$

$$\lambda_{6S} = 63.763 \quad \lambda_{6R} = 0.067$$

$$C_{2S} = \lambda_{1S}/\lambda_{6S} = 417 \quad C_{2R} = \lambda_{1R}/\lambda_{6R} = 51$$

Although $C_{2S}$ is large $C_{2R}$ is not. Perhaps a modified fractional rank estimate (between 6 and 7) as suggested by Marquardt (1970) would be better.

For example 2 we have

$$\lambda_{1S} = 722.272 \quad \lambda_{1R} = 3.396$$

$$\lambda_{8S} = 0.029 \quad \lambda_{8R} = 0.034$$

$$C_{2S} = 24905 \quad C_{2R} = 99$$

Here again $C_{2S}$ is very large while $C_{2R}$ is not that large.

For LRRA, Webster et al. suggested that vectors should be eliminated for which the roots $\lambda_i < 0.3$ and weights of the characteristic vectors with respect to the dependent variable, i.e. say $V_{yi} < 0.1$.

For the above two examples we have for the roots and vectors of the augmented matrix

\[
\begin{pmatrix}
Y'Y & X'X \\
X'Y & X'X
\end{pmatrix}
\]
Example 1

\[ \lambda_{1R} = 4.478 \quad V_{y1} = 0.431 \]
\[ \lambda_{7R} = 0.025 \quad V_{y7} = -0.661 \]
\[ \lambda_{8R} = 0.003 \quad V_{y8} = -0.077 \]

Example 2

\[ \lambda_{1R} = 3.893 \quad V_{y1} = 0.445 \]
\[ \lambda_{9R} = 0.024 \quad V_{y9} = 0.360 \]
\[ \lambda_{10R} = 0.004 \quad V_{y10} = 0.083 \]

Thus, as suggested by Webster et al., only the last root and characteristic vector was eliminated for both examples.

The differences in magnitude between the OLS-estimates and the other estimates are quite alarming. For both examples the Principal Component and LRRA estimates are very close to each other.

One important point which emerges from the example given by Thompson (1978) is the very large condition numbers for the matrix S compared to that of the matrix R. The matrix S has very small roots and is possibly very unstable. It is clear that the first four roots of S virtually explains all the variation in the matrix S with the result that the remaining roots are small.
A third example of the effect of multicollinearity and the application of ridge regression has been reported by Hadgu (1984). The model is the following

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e \]

where \( Y \) = reported annual rates of congenital syphilis in the United States

\( X_1 \) = reported annual rates of primary and secondary syphilis in the United States

\( X_2 \) = reported annual rates of early latent syphilis in the United States

\( X_3 \) = reported annual rates of late and late latent syphilis in the United States (all rates are calculated per 500,000).

The correlation matrix for the 13 observations reported by Hadgu (1984) (years 1957-1969) is

\[
\begin{array}{ccccc}
Y & X_1 & X_2 & X_3 \\
Y & 1 & -0.629 & 0.867 & 0.990 \\
X_1 & -0.629 & 1 & -0.259 & -0.625 \\
X_2 & 0.867 & -0.259 & 1 & 0.895 \\
X_3 & 0.990 & -0.625 & 0.895 & 1 \\
\end{array}
\]

The condition number for the Wishart matrix is 1184 while the condition number for the correlation matrix is 122. Both these numbers are very large indicating severe multicollinearity.
The estimated regression coefficients are

\[
\begin{align*}
  k = 0 & \quad k = 0.1 & \quad k = 0.2 & \quad k = 0.3 \\
  \text{Constant} & \quad -2.7753 & \quad -2.4091 & \quad -1.8204 & \quad -1.2784 \\
  \chi_1 & \quad 0.0334 & \quad -0.0465 & \quad -0.0516 & \quad -0.0523 \\
  \chi_2 & \quad -0.1710 & \quad 0.1623 & \quad 0.1808 & \quad 0.1821 \\
  \chi_3 & \quad 0.1047 & \quad 0.0438 & \quad 0.0374 & \quad 0.0343
\end{align*}
\]

It is clear that even small ridging has considerable effect on the regression coefficients.

7. SOME ASYMPTOTIC RESULTS

Anderson (1965) gives the following expansion for the joint distribution of \((\lambda_1, \ldots, \lambda_p)\) of the roots of \(\frac{1}{n}A = S\) when the roots of \(\alpha_1, \ldots, \alpha_p\) of \(\Sigma\), say are all distinct.

\[
(7.1) \quad f(\lambda_1, \ldots, \lambda_p) = M(A) \prod_{i=1}^{p} \lambda_i^{(n-p-1)} e^{-n/2(\lambda_i/\alpha_i)}
\]

\[
\prod_{i<j} (\lambda_i - \lambda_j)^{\frac{1}{2}} F
\]

where \(M(A) = \prod_{i=1}^{n} (\alpha_i)^{-\frac{1}{2}} n \prod_{i<j} (\frac{1}{\alpha_j^2} - \frac{1}{\alpha_i^2})^{-\frac{1}{2}}\)

\[
(7.2) \quad M(A) = \frac{(n/2)^{-n}}{(n/2)^{-np/2-p(p-1)/4}} \prod_{i=1}^{p} \Gamma(i(n-i+1))
\]

and

\[
(7.3) \quad F = 1 + \frac{1}{2n} \Sigma_{i<j} \frac{1}{c_{ij}} + \frac{9}{8n^2} \Sigma_{i<j} \frac{1}{c_{ij}^2}
\]

\[
+ \frac{1}{4n^3} \Sigma_i \frac{1}{c_{ij} c_{ik}^2} + \ldots
\]
and

\[(7.4) \quad C_{ij} = \left(\frac{1}{\alpha_i} - \frac{1}{\alpha_j}\right)(\lambda_i - \lambda_j).\]

Assume \(n\) large enough so that \(F = 1\).

Then by again making the transformation \(u_i = \lambda_i/\Sigma\lambda_i, i = 2, \ldots, p\) and \(\ell_i = \lambda_i/\lambda_p, \ i = 1, \ldots, (p-1)\) we get the marginal densities of \((u_2, \ldots, u_p)\) and \((\ell_1, \ldots, \ell_{p-1})\) as

\[(7.5) \quad f(u_2, \ldots, u_p) = M(A) \cdot U^{-\frac{1}{2}(n-p-1)} \left(\frac{u_1}{\alpha_1} + \ldots + \frac{u_p}{\alpha_p}\right)^{-\frac{1}{4}(2n-p+1)} \prod_{i<j}(u_i-u_j)^{\frac{1}{2}} \Gamma\left(\frac{1}{4}(2n-p+1)\right) \left(\frac{n}{2}\right)^{\frac{1}{2}(2n-p+1)}\]

where \(U = \text{diag}(u_1, \ldots, u_p)\) and \(u_1 = 1 - \sum_{i=2}^{p} u_i\)

and

\[(7.6) \quad f(\ell_1, \ldots, \ell_{p-1}) = M(A) \cdot L^{-\frac{1}{2}(n-p-1)} \left(\frac{\ell_1}{\alpha_1} + \ldots + \frac{\ell_{p-1}}{\alpha_{p-1}} + \frac{1}{\alpha_p}\right)^{-\frac{1}{4}(2n-p+1)} \prod_{i<j} \ell_i - \ell_j^{\frac{1}{2}} \Gamma\left(\frac{1}{4}(2n-p+1)\right)\]

where \(L = \text{diag}(\ell_1, \ldots, \ell_{p-1})\).

Again it appears difficult to find the marginal densities of the condition numbers \(C_1 = 1/u_p\) and \(C_2 = \ell_1\).

Anderson (1965) also showed that for large \(n\)

\[\prod_{i<j}(\frac{\lambda_i - \lambda_j}{\alpha_j - \alpha_i})^\frac{1}{2} \rightarrow 1\] with probability one and hence the joint density of \((\lambda_1, \ldots, \lambda_p)\) then becomes
\[ (7.7) \quad C \prod_{i=1}^{\lambda_i} \left( \frac{n}{2\alpha_i} \right)^{\frac{1}{2}(n-p-1)} e^{-\frac{3n}{\alpha_i}} \]

with

\[ (7.8) \quad C = \prod_{i=1}^{p} \left( \frac{n}{2\alpha_i} \right)^{\frac{1}{2}(n-p+1)} \Gamma(\frac{1}{2}(n-p+1)) \]

Thus \( \frac{n\lambda_i}{\alpha_i} \) are independently distributed as \( \chi^2 \)-variates with \( (n-p+1)/2 \) degrees of freedom. Thus the density of \( C_2 = \ell_1 = \lambda_1 / \lambda_p \) is that of \( \frac{\alpha_1}{\alpha_p} F(\frac{1}{2}(n-p+1), \frac{1}{2}(n-p+1)) \) and therefore knowledge of the ratio \( \alpha_1 / \alpha_p \) is needed. The joint density of \( u_2, \ldots, u_{p-1} \) with the condition number \( C_1 = 1 / u_p \) is more complicated and is given by

\[ (7.9) \quad f(u_2, \ldots, u_p) = C' \prod_{i=1}^{p} u_i^{\frac{1}{2}(n-p-1)} \left( \frac{u_1}{\alpha_1} + \ldots + \frac{u_p}{\alpha_p} \right)^{-\frac{1}{2}(n-p-1)p+p} \]

with \( C' = C \cdot (\pi n)^{\frac{1}{2}(n-p-1)p+p} \Gamma(\frac{1}{2}(n-p-1)p+p) \)

and \( u_1 = 1 - \sum_{i=2}^{p} u_i \).

It is again difficult to find the marginal density of \( u_p = 1 / C_2 \).

Girschick (1939) has given the following normal approximation for the joint distribution of the roots \( (\lambda_1, \ldots, \lambda_p) \) as \( n \) becomes large. \( \sqrt{n-1} (\lambda_1 - \alpha_1) \) is normally distributed with mean zero and variance \( 2\alpha_i^2 \), and independent of each other for \( i = 1, \ldots, p \). This implies that
\[
\lambda_i \sim N(\alpha_i, \frac{2\alpha_i^2}{n-1})
\]

or

\[
(7.10) \quad \frac{(n-1)\lambda_i^2}{2\alpha_i^2} \sim \chi_i^2 \left( \mu \right) \text{ with } \mu = \frac{\frac{1}{2} \alpha_i^2}{(2\alpha_i^2)/(n-1)} = \left( \frac{n-1}{4} \right).
\]

Thus

\[
(7.11) \quad \frac{\lambda_i^2/\alpha_i^2}{\lambda_p/\alpha_p} \sim F_{1,1}(\mu, \mu)
\]

Now the doubly non-central F can be approximated by central F as follows

\[
(7.12) \quad F_{1,1}(\mu, \mu) = F_{v,v', v'}
\]

where \( v = v' = \left( 1 + \frac{n-1}{4} \right)^2 \left( 1 + \frac{2(n-1)}{4} \right) = \left( \frac{4+n-1}{4} \right) \left( \frac{2+n-1}{2} \right)^{-1} \).

From the fact that \( \lambda_i \sim N(\alpha_i, \frac{2\alpha_i^2}{n-1}) \) confidence limits can be found for the smallest root \( \alpha_p \). These are

\[
(7.13) \quad a = \frac{\lambda_p}{1 + Z_{\alpha} \sqrt{2/n-1}} \leq \alpha_p \leq \frac{\lambda_p}{1 - Z_{\alpha} \sqrt{2/n-1}} = b
\]

Press (1972) states that for large \( n \) it is approximately true that

\[
(7.14) \quad \frac{a}{\Sigma \lambda_j} \leq \frac{\alpha_p}{\Sigma \alpha_i} \leq \frac{b}{\Sigma \lambda_j}
\]

Thus for large \( n \) an approximate 1-\( \alpha \) confidence interval for the population condition number \( \Sigma \alpha_i/\alpha_p \) is given by

\[
(7.15) \quad \frac{\Sigma \lambda_j}{b} \leq \frac{\Sigma \alpha_i}{\alpha} \leq \frac{\Sigma \lambda_j}{a}
\]
8. SOME EXACT RESULTS FOR A CIRCULAR COVARIANCE MATRIX

Let \((X_{1\alpha}, \ldots, X_{p\alpha})\), \(\alpha = 1, 2, \ldots, N\) be a sample from a \(p\)-variate \(N(\mu, \Sigma)\) distribution. \(\Sigma\) is uniform and is given by

\[
\Sigma = \sigma^2(1-\rho)1 + \sigma^2 \rho e'e'
\]

where \(e' = (1, 1, \ldots, 1)\). Olkin and Pratt (1958) have shown that the minimal sufficient statistic of \((\mu, \tau_1, \tau_2)\) is given by

\[
(N^T \mathbf{x}, v_1, v_2) = (N^T \mathbf{x}, v_{11}, \Sigma_{i=2}^p v_{ii})
\]

where

\[
\tau_1 = \sigma^2(1+(p-1)\rho), \quad \tau_2 = \sigma^2(1-\rho)
\]

are the characteristic roots of \(\Sigma\) (the last \((p-1)\) roots being equal to \(\tau_2\))

\[
\Gamma = [Y_{ij}] = [p^{-\frac{1}{2}}(\cos 2\pi p^{-1}(i-1)(j-1) + \sin 2\pi p^{-1}(i-1)(j-1))]
\]

and

\[
V = \Gamma^T S \Gamma = [v_{ij}]; \quad i, j = 1, \ldots, p.
\]

Also \(\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_p})'\) and \(S/n, n = N-1\) is the sample covariance matrix where \(S = [s_{ij}]; i, j = 1, \ldots, p\) and

\[
s_{ij} = \sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j).
\]

The distribution of \(v_1/\tau_1\) is \(\chi^2_n\) and of \(v_2/\tau_2\) is \(\chi^2_n(p-1)\) and independent.

Furthermore

\[
\lambda_1 = \tau_1 = v_1/n
\]
Now \( \frac{\tau_1}{\tau_2} \) is the population condition number and

\[
\frac{\tau_1}{\tau_2} = \frac{1+(p-1)\rho}{1-\rho}
\]

Thus \( \frac{\nu_1/\tau_1 n}{\nu_2/\tau_2 (n(p-1))} \) is distributed as \( F_{n,n(p-1)} \)
or the sample condition number \( C_2 = \frac{\lambda_1}{\lambda_2} \) is distributed as

\[
\frac{1+(p-1)\rho}{1-\rho} \cdot F_{n,n(p-1)}
\]

The maximum likelihood estimate of \( \rho \) is given by

\[
\hat{\rho} = \frac{p(\sum_{i=1}^{p} \sum_{j=1}^{p} S_{ij})/[(p-1)\sum_{i=1}^{p} S_{ii}]}{\sum_{j\neq i}^{p} S_{ij}}
\]
REFERENCES


