A COMPARISON OF MULTIVARIATE STATISTICAL PROGRAMS AVAILABLE AT KANSAS STATE UNIVERSITY

by

ROBERT L. UMHOLTZ

B.S., Kansas State University, 1974

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1976

Approved by:

K. E. Kemp
Major Professor

TABLE OF CONTENTS

I. INTRODUCTION

In this paper I shall compare several multivariate statistical programming packages which are currently available at Kansas State University for performing multivariate variance, covariance, and regression analyses, discriminant analysis, and principal components and factor analysis. Among the programs considered are Multivariate AARDVARK, BMD, MANOVA, OSIRIS, SAS, SPSS, and FORTRAN SSP.

Several data sets were analyzed by similar procedures in different programs. Numerical results will be compared, as well as time and storage requirements of these analyses. The options available with each program will also be compared.

Short general descriptions of the program packages follow. Descriptions of the statistical methods used by each program will be found in the respective sections where these methods apply.

II. CHART OF PROGRAM CAPABILITIES

	Analysis of variance and covariance	Regression analysis	Discriminant analysis	Stepwise discriminant analysis	Factor analysis
AARDVARK	YES	NO	NO	NO	NO
BMD	NO	NO	YES	YES	YES *
MANOVA	YES	YES	NO	NO	NO
OSIRIS	YES	NO	NO	NO	YES
SAS	YES	YES	YES	NO	YES
SPSS	NO	NO	YES	YES	YES
SSP	NO	NO	YES	NO	YES

* separate program for principal components analysis

III. DESCRIPTION OF PROGRAMS

A. AARDVARK

The Multivariate AARDVARK program was obtained in late 1975 from the University of Rhode Island at Kingston, Rhode Island. It was written by Dr. William Hemmerle. The program is stored on tape and requires a rather large amount of JCL, which is available at the KSU statistical laboratory.

AARDVARK can perform both univariate and multivariate analysis of variance and covariance (single or multiple covariates) on balanced, complete data. The univariate analysis routine has capabilities for random and mixed effects models and for separate covariate analysis when several covariates are to be analyzed, although these options are not available for multivariate analysis.

B. BMD

BMD (Biomedical Computer Programs) were written at the University of California for the primary purpose of analyzing medical data. The programs available with this package are arranged in six classes. The multivariate programs comprise one of these classes.

Principal components analysis, factor analysis, and discriminant analysis can be performed. There are separate programs for discriminant analysis for two groups, discriminant analysis for several groups, and stepwise discriminant analysis.

Some standard transformations have been implemented and may be used by supplying a TRANSGENERATION card. This feature is explained fully on pages 18-20 of the BMD user's manual.

The procedures for factor analysis, discriminant analysis for several groups, and stepwise discriminant analysis are stored on disk at KSU and can be executed with the following procedure:

```
//jobname JØB (standard job card information),name
// EXEC BMD,PRØGRAM=program
control cards and data
/*
```

```
program=BMD03M for factor analysis
        BMD05M for discriminant analysis for several groups
        BMD07M for stepwise discriminant analysis
```

The following procedure is required for the remaining programs:

```
//jobname JØB (standard job card information),name
/*TAPE9
// EXEC BMDLDG,LØAD=YES,PRØGRAM=program
//LØAD.SYSIN DD *
CØPY INDD=BMD,ØUTDD=LIB
SELECT MEMBER=program
//BMD.SYSIN DD *
control cards and data
/*
```

```
program=BMD01M for principal components analysis
        BMD04M for discriminant analysis for two groups
```

C. MANOVA

MANOVA was written at the University of North Carolina in 1972, and can be used for univariate and multivariate analysis of variance, covariance, and regression. Contrasts among levels may be used, and data may be reanalyzed using different dependent variables, covariates, contrasts, or models.

The following card sequence is required for using MANOVA:

```
//jobname JØB (standard job card information),name
// EXEC MANØVA
//MANOVA.SYSIN DD *
control cards and data
/*
```

D. OSIRIS

OSIRIS III is a collection of programs written primarily for the analysis of social science data, and was written in 1973 at the University of Michigan.

OSIRIS can perform univariate and multivariate analysis of variance and covariance, and factor analysis. There is one procedure for analysis of variance and covariance, and one for factor analysis.

The format of an OSIRIS data set is common to all OSIRIS programs. Cases are stored by rows and variables are stored by columns.

The procedures for analysis of variance and factor analysis, as well as all other OSIRIS procedures, are in the form of subroutines, and hence several processing steps among different procedures cna be shared. The card sequence used depends on the procedure, and will be explained in the respective sections.

E. SAS

SAS (Statistical Analysis System) was designed by Anthony Barr and James Goodnight at North Carolina State University at Raleigh, North Carolina. It can be used for multivariate analysis of variance, covariance, regression, discriminant analysis, and factor analysis.

SAS consists of many different procedures within the package. Variance, covariance, and regression analyses are done with a least squares procedure, while discriminant analysis and factor analysis each use separate procedures.

The following card sequence is used:

```
//jobname JØB (standard job card information),name
// EXEC SAS
//SAS.SYSIN DD *
control cards
data
procedure information statements
/*
```

F. SPSS

SPSS (Statistical Package for the Social Sciences) is a system of programs designed for the analysis of social science data. The package was designed at the Department of Political Science and the National Opinion Research Center at the University of Chicago. SPSS can be used to perform discriminant analysis and factor analysis.

Calculations are done with a sequence of control cards which precede the data. These control cards are described in the respective sections of the SPSS user's manual. Various options and statistics are available for each routine, which will be described later.

SPSS has certain capabilities for handling missing data. Each variable may have up to three designated missing values, which are user specified. The missing data values are defined by using a MISSING VALUES card. Missing data capabilities will be covered in the respective sections for discriminant analysis and factor analysis because each procedure has different options for handling missing observations.

G. SSP

SSP (Scientific Subroutine Package) is a collection of FORTRAN subroutines. It can be used for discriminant analysis and factor analysis.

Each procedure consists of a main program and several subroutines, each of which perform a given function. The coding for the main programs and subroutines is supplied in the SSP user's manual, although decks are available at the KSU statistical laboratory.

The procedures may be executed using the following card sequence:

```
//jobname JØB (standard job card information),name
// EXEC FØRTGCLG
//FØRT.SYSIN DD *
main program
subroutines
//GØ.SYSIN DD *
data
/*
```

Alternatively, SSP may be run in WATFIV, requiring the following card sequence:

```
//WATFIV JØB (standard job card information),name
$JØB
$PRINTØFF      used if a source listing is not desired
main program
subroutines
$ENTRY
data
/*
```

IV. MULTIVARIATE ANALYSIS OF VARIANCE, COVARIANCE, AND REGRESSION PROGRAMS

A. AARDVARK

The multivariate AARDVARK program can be used to perform univariate and multivariate analyses of variance and covariance, with single or multiple covariates. Multiple covariates can be analyzed either separately or together in the univariate case, although they cannot be analyzed separately in the multivariate case.

The required input for a multivariate analysis includes a list of variates and covariates and the order in which they appear, what factors are in the model, and how many levels of each factor there are. The MODEL and LIMITS statements are used for this purpose. An ORDER OF VARIABLES card is also used if the order of variables in the input format is different from that of the model statement. The exact form of these cards and all others applicable to a multivariate analysis is described on pages 5-8 of the Multivariate AARDVARK reference manual.

Univariate output can be obtained in addition to the multivariate anlaysis. This includes analysis of variance tables for each variable. Sources of variance may be pooled and effects may be random for a univariate analysis, however these features are not available for multivariate analysis.

The data is preceded by an input format. This is a standard FORTRAN format for reading in the data. Classification variables are not read in. The data then appears one observation vector per read, and must be sorted properly in standard sort sequence with the last subscript in the error term in the model statement moving most rapidly.

The user may have a maximum of 20 variates and covariates in total, eight factors in the model, and a total of 40 terms in the model.

Output includes, first of all, the means for each variable for whatever factors are specified on the MEANS card. Next, if univariate output is desired, the usual analysis of variance tables are printed out for each of the variables.

For the multivariate analysis of variance, a table is printed which shows the results of tests for each effect in the model for three different criteria.

Let g=the number of degrees of freedom associated with a given effect in a univariate analysis,

u=the number of variates,

N=the total number of observations,

r=the number of cells in a corresponding univariate analysis,

H=the hypothesis sums of squares matrix associated with the effect being tested,

and E=the error sums of squares matrix.

The first criterion is the Hotelling-Lawley trace. This is the trace of HE^{-1}, and has an approximate chi-square distribution with gp degrees of freedom. The next criterion is Wilks' criterion. The test statistic is det(E)/det(H+E) and has an approximate chi-square distribution with gp degrees of freedom. The third criterion is the Heck characteristic root, which is the largest root of HE^{-1}. Charts for this test statistic are available in Morrison's multivariate methods text, and they use the following three quantities:

$s=\min(g,u)$,

$m=\frac{1}{2}(\text{abs}(g-u)-1)$,

and $n=\frac{1}{2}(N-r-u-1)$.

When covariance analysis is specified, the following output is obtained in addition to that just mentioned. The means for the covariates may be requested, and if univariate output is requested, then an analysis of variance table for each covariate is printed. Univariate analysis of covariance tables are printed.

Next is the multivariate analysis of covariance. The same three criteria are presented for testing the effect of covariates. For the Lawley and Wilks criteria, the test statistics have approximate chi-square distributions with pc degrees of freedom, where c is the number of covariates.

B. SAS

SAS performs multivariate analysis of variance, covariance, and regression using the REGR (least squares regression) procedure. The output consists of several univariate statistics as well as the multivariate analysis.

First are the simple descriptive statistics, including sums, means, variances, and standard deviations of non-classification variables. Next are the sums of crossproducts, covariances, and simple correlation coefficients for each pair of variables along with the significance probability for the correlation coefficient. An analysis of variance table, including the sources of regression, error, and total is printed. Regression refers to variation due to all independent variables, both classification and non-classification variables. Then there is a table which contains statistics for each of the sources comprising the regression source. These statistics include degrees of freedom, sequential sums of squares and F values, and partial sums of squares and F values. Sequential sums of squares are sums of squares adjusted for effects in the model preceding the effect being tested, while partial sums of squares are adjusted for all other effects in the model. Sequential sums of squares are, therefore, affected by the order of terms in the model.

Estimates of the regression coefficients for non-classification variables are printed along with a t-statistic and significance probability of the t-statistic. Also printed are the standard error of the coefficient and the standardized regression coefficients. Predicted values and confidence limits on these values are available. Adjusted means may be printed, and can be adjusted for covariates or any other effects. The matrix obtained by adjusting the Y'Y matrix for all independent variables can be printed. This is the error sums of squares matrix.

The multivariate analysis of variance statistics include the characteristic roots and vectors of HE^{-1}, where H is the hypothesis sums of squares matrix and E is the error sums of squares matrix. The Hotelling-Lawley trace, Pillai's trace, Wilks' criterion, and Roy's maximum root are printed.

Let p=the number of dependent variables,
q=the degrees of freedom associated with the effect being tested,
e=error degrees of freedom,
$s=\min(p,q)$,
$m=\frac{1}{2}((\mathrm{abs}(p-q))-1)$,
and $n=\frac{1}{2}(e-p-1)$.

The Hotelling-Lawley trace is the trace of HE^{-1}, and the quantity $(2(sn+1)tr(HE^{-1}))/(s^2(2m+s+1))$ has an approximate F distribution with $s(2m+s+1)$ and $2(sn+1)$ degrees of freedom. Pillai's trace is the trace of $H(H+E)^{-1}$ and is denoted by v. The quantity $v(2n+s+1)/((2m+s+1)(s-v))$ has an approximate F distribution with $s(2m+s+1)$ and $s(2n+s+1)$ degrees of freedom. Wilks' criterion is $\det(E)/\det(H+E)$ and has an approximate chi-square distribution with pq degrees of freedom. Roy's maximum root is the largest characteristics root of HE^{-1}, which is the same as the Heck criterion. Charts for this test statistic are in the Morrison multivariate methods text (1967).

Means of canonical variables and correlation coefficients between the dependent variables and canonical variablee are available.

Multivariate regression analysis is done in the same way as a covariance analysis, except that there are no classification variables, and no CLASSES statement.

The form for all of the parameter cards appears on pages 99-105 of the SAS manual.

C. MANOVA

Multivariate variance, covariance, and regression analyses can be performed with the MANOVA program. The output consists of the following items. First, the cell frequencies, means, and standard deviations are printed. Next, the reduced model matrix corresponding to the effects specified on significance test cards is printed. Error correlations of the variables are printed, along with the estimates of effects which are tested against the error term. These estimates are adjusted for covariates.

A regression analysis is performed if there are any covariates. The Wilks' lambda criterion is used as a significance test of effects. Rao's approximation is used to yield an approximate F test for significance of effects. Error canonical correlations between the variates and the covariates and the results of univariate F tests for regression adjusted for covariates are printed. Standardized discriminant functions corresponding to canonical weights for the dependent variables in a canonical correlation problem are printed. These functions are standardized by adjusted error standard deviations.

The results of both univariate and multivariate tests are printed for each effect. The order of effects in the multivariate printout is opposite from the order of significance test cards. That is, the highest order interaction comes first and the main effects come last.

A multivariate regression analysis is treated as an analysis of covariance with no classification variables.

Data may be reanalyzed using different models in the same run.

The following items are input:

1. A title,
2. The total number of variables, factors, and covariates, and an input format if the data is not punched according to the standard format,
3. Variable names (optional),
4. Variable subsets - used if not all of the variables are dependent variables,
5. Contrasts - There are three major types of contrasts. First are the devia-

tions of means from the grand mean for the usual significance tests. Second are the single degree of freedom contrasts, which are usually orthogonal. Third are orthogonal polynomial contrasts, and these may be at either equally or unequally spaced points. The spacing is specified on a special contrast card.,

6. Significance tests - tests of any effects, including main effects, interactions, nested effects, or pooled effects, and
7. Transformations - square root, natural log, and arc-sine transformations.

The form for all of the input is shown on pages 3-13 of the MANOVA writeup.

D. OSIRIS

OSIRIS uses a subprogram called MANOVA to perform multivariate analysis of variance and of covariance. Both univariate and multivariate analyses can be performed with up to eight factors. Multiple analyses may be done, using different sets of variates or covariates.

MANOVA uses a hierarchical model for analysis of variance. If the analysis has more than one classification variable and if the cells have unequal sample sizes, then the order in which effects are specified will affect the analysis. The variables listed first are analyzed with the effects of the subsequent variables removed.

A design matrix for specifying treatment effects, interactions, and covariates may be either generated by the program or supplied by the user. Contrasts are specified for each factor in the model. The deviations of row and column means from the grand mean are the contrasts usually used, and are the only ones used when only overall tests of significance are required. Single degree of freedom contrasts may also be supplied.

The printed output includes the following items:

1. Cell means and sample sizes,
2. The design matrix,
3. Correlations among coefficients of the normal equations,
4. The error correlation matrix, and the principal components of this matrix,
5. The error dispersion matrix and standard errors of estimation, adjusted for covariates, if any,
6. Rao's approximate F test for significance of the overall effect for all dependent variables simultaneously,
7. Eigenvalues of the hypothesis matrix,
8. Correlations between the variables and the components of the hypothesis matrix,
9. Scores of the hypothesis for contrasts used in the design,
10. Cumulative Bartlett's test on the roots, which is an approximate test for

the remaining roots after eliminating previous roots, and

11. F ratios for univariate tests.

The program can handle up to eight factors, 10 levels per factor, and 19 dependent variables. The total number of factors, covariates, and dependent variables cannot exceed 20. Up to 80 cells are allowed.

The following card sequence is used with the MANOVA procedure:

```
//jobname JØB (standard job card information),name
// EXEC ØSIRIS
//SETUP DD *
control cards
data
/*
```

Pages 542-548 of the OSIRIS manual, volume 1, show the detailed structure of the control cards.

E. Sample programs

The first example is a data set from page 179 of the Morrison (1967) multivariate methods text. It is a two factor design, with two dependent variables. The important statistics (F ratios, etc.) agree to several decimal places. Most of the optional output was requested in SAS, so that output is considerably longer than the rest.

Time and space on these runs are as follows:

	time(seconds)	space(k-bytes)
AARDVARK	4.68	113.77
SAS	6.12	200.00
MANOVA	1.38	125.22
OSIRIS	6.66	98.14

The second data set is from page 199, Morrison (1967). It is a one way analysis of covariance with one covariate and three dependent variables. All statistics agreed closely. Time and space requirements are as follows:

	time(seconds)	space(k-bytes)
AARDVARK	6.18	144.22
SAS	5.10	200.00
MANOVA	1.44	127.13
OSIRIS	6.90	98.29

The next problem is the same as the last one, except that there are two covariates. The time and storage are:

	time(seconds)	space(k-bytes)
AARDVARK	6.36	143.93
SAS	5.04	200.80
MANOVA	1.50	128.84
OSIRIS	7.02	97.86

The next example is from Morrison (1967), page 191. This is a one way analysis of variance problem with three dependent variables. There is again good agreement with numerical results. Time and space break down as follows:

	time(seconds)	space(k-bytes)
AARDVARK	4.80	120.53
SAS	5.10	200.39
MANOVA	1.26	129.00
OSIRIS	6.12	98.32

A data set from Morrison (1967), page 202, was used for the next set of runs. Different contrasts were used in the OSIRIS run. These are Helmert contrasts, which are means of deviations of effects from the sum of means one through r, where there are r levels. The example is a two way model with three dependent variables and one covariate. The data set caused errors in AARDVARK and MANOVA, and the numerical results differed considerably among programs. Following are the time and space statistics:

	time(seconds)	space(k-bytes)
AARDVARK	6.06	138.92
SAS	4.92	200.82
MANOVA	1.86	127.32
OSIRIS	8.16	97.64

The next example is from Morrison (1967), page 203. This example has three dependent variables, two covariates, and two factors. The usual statistics break down as follows:

	time(seconds)	space(k-bytes)
AARDVARK	6.60	140.13
SAS	5.58	198.92
MANOVA	2.22	128.57
OSIRIS	8.16	97.64

The last analysis of variance example is a three way model with two dependent variables and unbalanced data. The example was not run in AARDVARK because AARDVARK can only analyze balanced data. Numerical results agreed closely, as usual. Time and space break down as follows:

	time(seconds)	space(k-bytes)
SAS	5.58	198.92
MANOVA	1.80	128.00
OSIRIS	7.68	97.94

The remaining examples are multivariate regression problems. The data consists of 10 variables and each of the six runs used different subsets of these 10 variables in the analysis. Each example was run in SAS and MANOVA, and the numerical results (analysis of variance tables, regression coefficients, etc.) were in good agreement. The time and space of each run are as follows:

Example #1 - 10 variables total, 3 dependent variables

	time(seconds)	space(k-bytes)
SAS	5.28	101.42
MANOVA	1.38	127.83

Example #2 - 10 variables total, 5 dependent variables

	time(seconds)	space(k-bytes)
SAS	4.86	101.37
MANOVA	1.50	128.00

Example #3 - 5 variables total, 4 dependent variables

	time(seconds)	space(k-bytes)
SAS	2.88	100.94
MANOVA	1.38	127.09

Example #4 - 4 variables total, 2 dependent variables

	time(seconds)	space(k-bytes)
SAS	2.88	101.65
MANOVA	1.26	124.95

Example #5 - 7 variables total, 4 dependent variables

	time(seconds)	space(k-bytes)
SAS	3.42	102.30
MANOVA	1.44	125.33

Example #6 - 10 variables total, 2 dependent variables

	time(seconds)	space(k-bytes)
SAS	3.96	101.23
MANOVA	1.26	129.00

F. Summary

AARDVARK, SAS, and OSIRIS generally matched time requirements closely, while MANOVA took much less time. AARDVARK and MANOVA usually required about the same amount of space, while SAS used much more and OSIRIS used less.

Both AARDVARK and SAS are relatively easy to learn to use, and are more suitable for users without extensive programming experience. MANOVA and OSIRIS are more difficult to learn how to use and their manuals are difficult to understand at first, but the programs are not too difficult to use after getting a few successful runs with them.

SAS and OSIRIS have more options available than either AARDVARK or MANOVA, and should be used if the user wants output in addition to the most basic output for variance, covariance, or regression analysis. Of course, if a multivariate regression analysis is desired, then only SAS and MANOVA are available for that purpose.

The costs matched closely for all programs except for MANOVA, whose costs were much lower. The costs at KSU are based mostly on time, with some consideration given to the amount of space used.

V. DISCRIMINANT ANALYSIS PROGRAMS

A. SSP

Discriminant analysis is performed with SSP by using a main program named MDISC and three subroutines, namely DMATX, DMINV, and DISCR.

Three control cards are required as part of the data. The first of these is a title card. My main program differs from the sample main program in that a full card is used for the title. The title card must be included, even if it is blank. The other two cards contain the number of groups, the number of variables, the sample sizes of each group, and the input format for the data. The form for the card containing the number of groups, etc. is shown on page 426 of the SSP manual, version III.

The data cards follow the format card. Each observation begins on a new card. All observations from the first group are read in first, then those from the second group, etc.

The sample main program can handle up to five groups, 10 variables, and 250 observations. These limitations can be changed according to the following rules:

1. The dimension of array CMEAN must be greater than or equal to m, the number of variables,
2. The dimension of array N must be greater than or equal to k, the number of groups,
3. The dimension of array XBAR must be greater than or equal to mk.
4. The dimension of array C must be greater than or equal to k(m+1),
5. The dimension of array D must be greater than or equal to m^2,
6. The dimension of arrays P and LG must be greater than or equal to t, the total number of observations, and
7. The dimension of array X must be greater than or equal to tm.

The output with program MDISC includes the means of each variable within groups, the pooled dispersion matrix, common means, the generalized Mahalonobis D-square, the constant and coefficients of each discriminant function, and the probability associated with the largest discriminant function evaluated for each observation. The means of the variables in each group are calculated in subroutine DMATX. The pooled dispersion matrix is the sum of squares and crossproducts of deviations from means, and is computed for all groups combined. This matrix is also computed in the DMATX subroutine. The inverse of the pooled dispersion matrix is computed in subroutine DMINV. All remaining computations are done in subroutine DISCR.

The common means are the means of the variables combined for all groups. The generalized Mahalonobis D^2 statistic is defined as
$\sum_{a=1}^{m} \sum_{b=1}^{m} D^{-1}_{ab} \sum_{c=1}^{k} n_c(x_{c.a}-x_{..a})(x_{c.b}-x_{..b})$ where
D is the pooled dispersion matrix,

n_c is the sample size of group c,

$x_{c.a}$ is the mean of variable a within group c,

$x_{..a}$ is the common mean of variable a,

$x_{c.b}$ is the mean of variable b within group c, and

$x_{..b}$ is the common mean of variable b.

The constant and coefficients for each of the k discriminant functions are then computed. These functions are used for classifying observations into the groups. The coefficients are calculated as $c_{ji}=\sum_{a=1}^{m} D^{-1}_{ja}x_{i.a}$ where the subscript j refers to variable j and the subscript i refers to discriminant function i. The constants are calculated as $c_{0i}=-\frac{1}{2}\sum_{a=1}^{m} \sum_{b=1}^{m} D^{-1}_{ab}x_{i.a}x_{i.b}$ ·. The (i)th discriminant function is given by $f_i(z_1,z_2,\ldots,z_m)=\sum_{a=1}^{m} z_a c_{ai}+c_{0i}$ where $z_1,z_2,\ldots,z_m$ is any observation from the m variables.

The discriminant functions are then calculated for each observation. The largest of these functions and the probability associated with the largest function are printed for each observation. The probability associated with the largest function is given as $1/\sum_{i=1}^{k} \exp(f_i-\max(f_i))$.

B. SAS

Discriminant analysis is performed in SAS by using the DISCRIM procedure. The form for all of the parameter cards is shown on pages 192-195 of the SAS manual. Several options are available with the program. Simple descriptive statistics, including sample sizes, sums, means, variances, and standard deviations may be printed.

The generalized squared distance can be based on either the pooled covariance matrix or on within-group covariance matrices as the user chooses. The user may specify a TEST option along with a significance level to test the equality of the within-group covariance matrices. The pooled covariance matrix is used unless the within-groups covariance matrices are significantly different. The significance level is assumed to be .10 unless specified otherwise.

The within-group covariance and correlation matrices, the pooled covariance matrix, and the partial correlation matrix based on the pooled covariance matrix may be printed. The classification results for each observation may be output. Prior probabilities may be set equal or proportional to group sample sizes, or they may be specified on parameter cards.

The group into which an observation is classified is based on generalized squared distances. The generalized squared distance for group t is given as

$D_t^2(x)=g_1(x,t)+g_2(x,t)+g_3(x,t)$, where

$g_1(x,t)=(x-\bar{x}_t)'S_t^{-1}\ (x-\bar{x}_t)$ if within-group covariance matrices are used,

$\quad =(x-\bar{x}_t)'S^{-1}\ (x-\bar{x}_t)$ otherwise;

$g_2(x,t)=\ln(\det(S_t))$ if within-group covariance matrices are used,

$\quad =0$ otherwise;

$g_3(x,t)=-2\ \ln($prior probability for group t$)$ if prior probabilities are not equal,

$\quad =0$ otherwise.

An observation is classified into the group which has the smallest generalized squared distance from all groups.

The test for equality of covariance matrices is a chi-square test. The quantity $-2r \ln(n^{pn/2}v/\prod_{i=1}^{k} n_i^{pn_i/2})$ has an approximate chi-square distribution with $(k-1)p(p+1)/2$ degrees of freedom, where

k=the number of groups,

p=the number of variables,

n=the total number of observations,

n_i=the number of observations from group i,

$v=(\sum_{i=1}^{k}$ (det(within-group covariance matrix))$^{n_i/2}$)/(det(pooled matrix))$^{n/2}$, and

$r=1-\sum_{i=1}^{k} 1/(n_i-1)-1)-1/(n-k))(2p^2+3k-1)/6(p+1)(k-10)$.

A table showing how many observations would be classified correctly or incorrectly according to the classification criteria is printed.

If an observation has a missing value for one or more variables, then that observation will be ignored. If all values are present except for the group number, then the observation will be classified according to the classification criteria developed by the program.

C. BMD

BMD has three separate programs for performing discriminant analysis. There is one program for discriminant anlaysis with two groups, one for discriminant analysis with several groups, and one for stepwise discriminant analysis, where the discriminating variables are selected one at a time in a forward solution.

1. Two groups program (BMD04M)

The two groups program computes a linear discriminant function of the p variables measured on each individual of the two groups, and computes the following statistics:

a. The group means,

b. The differences between group means,

c. Within-group covariance matrices,

d. The pooled covariance matrix,

e. The inverse of the pooled covariance matrix,

f. The coefficients $L_1,L_2,\ldots,L_m$ of the discriminant function $L=A^{-1}(x_{1.}-x_{2.})$, where $L'=(L_1\ L_2\ \ldots\ L_m)$, A is the pooled covariance matrix, $x_{1.}$ and $x_{2.}$ are the means for groups one and two, respectively, and m is the number of variables,

g. The Mahalonobis $D^2=(n_1+n_2-2)\sum_{i=1}^{m}\sum_{j=1}^{m}a_{ij}(x_{1.i}-x_{2.i})(x_{1.j}-x_{2.j})$ where n_1 and n_2 are the group sample sizes, $x_{1.i}$ is the mean of variable i within group one, and $x_{2.i}$ is the mean of variable i within group 2,

h. The F statistic is computed as $F(m,n_1+n_2-1-m)=(D^2n_1n_2(n_1+n_2-m-1))/(m(n_1+n_2)(n_1+n_2-2))$. This statistic is used to test the equality of the two group means.,

i. The mean, variance, and standard deviation of the quantities $Z_{ia}=\sum_{j=1}^{m}L_jx_{iaj}$ $a=1,2,\ldots,n_i$ $i=1,2$, and

j. The n_1+n_2 values of Z, arranged in descending order.

The program can handle up to 25 variables and up to 300 observations per group. Any subset of the variables can be selected for a subsequent analysis. The card input for the program is shown fully on pages 186-189 of the BMD user's manual.

2. Several groups program (BMD05M)

The program for several groups computes a set of linear functions in order to classify an observation into one of several groups. A multivariate normal distribution with equal covariance matrices is assumed. Therefore, the pooled covariance matrix is used in the analysis.

The following quantities are computed and printed out. Most of these statistics are computed in the SSP discriminant analysis procedure and have already been described. Therefore, only those which are not computed by SSP will be explained here.

a. Group means,

b. Within-group covariance matrices,

c. The pooled covariance matrix,

d. The inverse of the pooled covariance matrix,

e. The matrix product DD^{-1} as a check on the accuracy of D^{-1}, where D is the pooled covariance matrix,

f. Common means,

g. Generalized Mahalonobis D^2,

h. Coefficients and constants of the discriminant functions,

i. Largest probability and function number for the largest probability for group membership,

j. The classification matrix which shows how many observations would be classified correctly and incorrectly according to the classification criteria.

The data input is basically the same as for the two groups program, except there is no capability for variable subset selection. The input form is described fully on pages 197-200 of the BMD manual.

3. Stepwise discriminant analysis program (BMD07M)

The stepwise discriminant analysis program performs a multiple discriminant analysis by entering one variable at each step. The variable with the largest F to enter is selected. Variables are deleted if their F value becomes less than a specified value. Canonical correlation coefficients may be requested.

The program computes the following quantities:

a. Group means and standard deviations,

b. Within-group covariance and correlation matrices,

c. Discriminant functions and a classification matrix,

d. The posterior probability of coming from each group and the square of the Mahalonobis distance from each group for each case,

e. Eigenvalues, canonical variables, and coefficients of canonical variables,

f. Plot of the first canonical variable versus the second, and

g. Residuals and canonical coefficents.

The following information is printed at each step:

a. Variables already included and their F values,

b. Variables not included and their F values,

c. A U statistic and approximate F statistic for testing equality of group means,

d. A matrix of F statistics to test the equality of means between each pair of groups,

e. The variable which has just been entered or removed, and

f. The number of variables included.

Let p=the number of variables,

g=the number of groups used in the analysis,

t=the total number of groups,

n_m=the number of observations in group m,

n=the total number of observations, and

x_{mkj}=the value of variable i for observation k of group m.

First, the data are read and the overall group means and group standard deviations are computed. In addition, the within and total crossproduct matrices, within-groups covariance matrices, and within-groups correlation matrices are calculated. At each step the variables are divided into two groups - those that are included in the analysis and those that are not. Assume that the first r are included.

Let $W=\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$ and $T=\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$,

where W is the within-groups crossproducts matrix and T is the total crossproducts matrix, and W_{11} and T_{11} are r by r matrices. Let

$A=(a_{ij})=\begin{bmatrix} W_{11}^{-1} & W_{11}^{-1}W_{12} \\ W_{21}W_{11}^{-1} & W_{22}-W_{21}W_{11}^{-1}W_{12} \end{bmatrix}$ and $B=(b_{ij})=\begin{bmatrix} T_{11}^{-1} & T_{11}^{-1}T_{12} \\ T_{21}T_{11}^{-1} & T_{22}-T_{21}T_{11}^{-1}T_{12} \end{bmatrix}$

Coefficients of the discriminant functions are computed as $C_{ki}=(n-g)\sum_{j=1}^{r}\bar{x}_{kj}a_{ij}$ i=1,2,...,r and k=1,2,...,g. Constants are computed as $C_{k0}=-\frac{1}{2}\sum_{i=1}^{r}C_{ki}\bar{x}_{ki}$ i=1,2,...,g. The square of the Mahalonobis distance between each pair of groups is given as $D_{mL}=\sum_{i=1}^{r}(C_{mi}-c_{Li})(\bar{x}_{mi}-\bar{x}_{Li})$ m=1,2,...,g, L=1,2,...,g, where $\bar{x}_{mi}$ and $\bar{x}_{Li}$ are the group means for group m and group L, respectively. F values for testing differences between each pair of groups are also calculated.

F statistics are calculated, and these are compared to the F to remove in the case of a variable which has already been included, or to the F to include in the case of a variable which is not included. If a variable has already been entered (variable j, say), then $F_j=(a_{jj}-b_{jj})/b_{jj})((n-r-g+1)/(g-1))$ with g-1 and n-r-g+1 degrees of freedom. If variable j has not been entered, then $F_j=((b_{jj}-a_{jj})/a_{jj})((n-r-g)/(g-1))$ with g-1 and n-g-r degrees of freedom.

A U statistic and approximate F statistic are calculated to test the equality of group means. $U=\det(W_{11})/\det(T_{11})$ and $F=((1-U^{1/s})/U^{1/s})((ms+1-rq/2)/rq)$, where $s=\sqrt{(r^2q^2-4)/(r^2+q^2-5)}$ if $r^2+q^2\neq 5$,

= 1 otherwise, m=n-(r+q+3)/2 and q=g-1. F has rq and ms+1-rq/2 degrees of freedom.

Tolerance values (W_i) are computed, where $W_i=a_{ij}/t_{ii}$, i=r+1,r+2,...,p, and t_{ii} is the (i)th diagonal element of T. A variable passes the tolerance test if W_i and t_{ii} are greater than or equal to the value specified on the subproblem card.

At each step one variable will be added or deleted according to the following rules:

a. Out of the variables which have been entered, if any have a control value of one and have a computed F value lower than the F to remove specified by the user, the one with the smallest F value will be deleted.

b. From those variables which have the greatest control value, one is selected (if no variable satisfies rule 1) according to the option stated on the problem card. The default option is to enter the variable with the greatest F to enter. If option 1 is stated, the entering variable is the one which minimizes $C_1=2/(g(g-1)) \sum_{L=1,L\neq m}^{g} 1/(1+D_{Lm}/4)$. If option 2 is specified, the variable entered is the one which minimizes $C_2=1/h \sum_{L=1,L\neq m}^{g} A_{Lm}/4)$, where $h=\sum_{L=1,L\neq m}^{g} A_{Lm}$. The A_{Lm} are specified on an ALPHA card. The formula for C_2 is a generalization of C_1. If option 3 is selected, the entering variable is the one which maximizes the smallest F between pairs of groups.

The following statistics are computed:

a. The value of the (m)th discriminant function evaluated at observation k of group L $(S_{Lmk}=c_{m0}+\sum_{j=1}^{r} C_{mj}x_{mkj})$,

b. The posterior probability of case k in group L having come from group m $(P_{Lmk}=\exp(S_{1mk})/\sum_{i=1}^{q} \exp(S_{1mk}))$,

c. The square of the Mahalonobis distance of case k in group m from group L $(D^2_{Lmk}=\sum_{i=1}^{r} \sum_{j=1}^{r} (x_{mki}-\bar{x}_{Li}a_{ij}(x_{mkj}-\bar{x}_{Lj}))$,

Coefficients, U_i, of canonical variables and the amount of dispersion, L_i, explained by each variable are computed. Let W and T be the within and total sum of crossproduct matrices of the p variables already entered, and let B=T-W. Then U_i and L_i are found from $BU_i=L_iWU_i$ for i=1,2,...,p. These vectors are normalized so that $U_i'WU_j=S_{ij}$. Canonical correlations $r_1,r_2,\ldots,r_p$ are computed, where $r_i=L_i/(1+L_i)$. The first three canonical variables are computed for each observation. $Z_{imk}=\sum_{j=1}^{p} U_{ji}(x_{mkj}-\bar{x}_j)$ i=1,2,3. j=1,2,...,g k=1,2,...,n_m. The first two of these are plotted.

The form for the input cards may be found on pages 214 b - 214 g of the BMD manual.

D. SPSS

Discriminant analysis is performed in SPSS by using a subprogram called DISCRIMINANT. Like BMD, all variables may be used, or they may be entered one at a time, although SPSS has five different stepwise selection procedures available.

The method by which all variables are entered concurrently is called the direct method (METHOD=DIRECT specification), and is the default method.

The first stepwise method is the Wilks procedure. This procedure produces a multivariate F ratio for testing for differences between group means. The variable which maximizes the F ratio and minimizes Wilks' lambda is the next variable to be entered.

The second method is the Mahalonobis procedure, which finds the maximum Mahalonobis distance between the two closest groups.

With METHOD=MAXMINF, the next variable selected is the one which maximizes the smallest F ratio between pairs of groups. This method should produce equivalent results to METHOD=MAHAL when the group sizes are equal.

The criterion $R=\sum_{i=1}^{g}\sum_{j=1}^{g}{}_{i\neq j}\ 1/(1+D_{ij}/4)$ is used with METHOD=MINRESID, where g is the number of groups included and D_{ij} is the Mahalonobis distance between groups i and j. The objective is to minimize R.

With METHOD=RAO, Rao's V, a generalized distance measure, is computed. The variable which causes the largest increase in V is selected because the objective is to separate the groups as much as possible. The change in V has a chi-square distribution with one degree of freedom and can be tested when there is a large number of observations.

Variables are selected in any of the stepwise procedures only if their multivariate F ratio is larger than a specified value. Also, a variable will be removed only if its F ratio is less than a specified value.

The keywords and parameters are summarized on page 457 of the SPSS user's manual, and their use is illustrated on pages 448-456 of that manual.

The order of entry of variables with the stepwise methods may be controlled by specifying inclusion levels. An inclusion level must be an integer number from zero to 99 with default value of one. If an inclusion level is even, all variables with that level are entered concurrently. If the level is odd, variables are entered according to the criteria specified by the selection method. Variables with higher inclusion levels are entered before those with lower inclusion levels.

There are 19 options available using the program, which are summarized on pages 456-459 of the user's manual. Options 13-19, however, are not currently operational but will be in the near future.

The following statistics may be calculated and output:

1. Group means and common means,
2. Standard deviations,
3. Pooled covariance and correlation matrices,
4. Matrix of pairwise F ratios, or significance tests for the Mahalonobis distance between groups,
5. Univariate F ratios, which are tests for the equality of group means on one discriminating variable,
6. A test for equality of covariance matrices, and
7. Within-group total covariance matrices.

Also, covariance matrices may be input, rather than raw data, to save excessive input for large data sets that are to be rerun several times.

E. Sample programs

The first example is a set of data from page 154 of the Morrison (1967) multivariate methods text, and was run on all programs. There are two groups with sample sizes 10 and 12, respectively, and three variables. Most numerical results (discriminant function coefficients, Mahalonobis D^2, etc.) agree to at least three decimal places. For those programs which show classification results (SSP, SAS, BMD05M, and BMD07M) the classification matrix is the same. There were two observations from group one incorrectly classified into group two. All others were correctly classified.

Both the BMD and SPSS stepwise procedures were used. The MAXMINF method was used in SPSS. BMD included all three variables, while SPSS entered only the first two variables. The time and storage requirements break down as follows:

	time(seconds)	space(k-bytes)
SSP	1.56	27.26
SAS	2.34	197.44
BMD04M	2.64	208.82
BMD05M	.60	136.80
SPSS	1.38	203.00
BMD07M	1.14	160.00
SPSS(stepwise)	1.20	209.45

The next example comes from page 155 of Morrison's text, and contains two groups with 10 and six members, respectively. There are four variables. The within-group covariance matrices were used in SAS, but the classification results came out the same in all programs. All observations were correctly classified. BMD07M was the only stepwise procedure used. It included only the first variable. Time and storage are as follows:

	time(seconds)	space(k-bytes)
SSP	1.26	27.26
SAS	2.46	200.00
BMD04M	1.86	183.58
BMD05M	.54	149.33
SPSS	1.20	200.85
BMD07M	1.26	157.48

The next example is a modification of the previous one, and has four groups with two variables. The example could not, of course, be run on the BMD two groups program. It also could not be run on BMDO5M because that program requires the number of variables to be greater than or equal to the number of groups.

The POOL=TEST option was specified in SAS, and the within-group covariance matrices were found to be significantly different, so they were used in the analysis rather than the pooled covariance matrix. All the programs except SAS showed the same classification matrix. The SAS matrix differed considerably, mostly because of the use of within-group covariance matrices. The SAS results were worse than others for some groups and better than other results for other groups.

The BMD and SPSS (MINRESID) stepwise procedures were used. BMD included both variables, while SPSS included only the second. Time and storage are:

	time(seconds)	space(k-bytes)
SSP	1.32	27.26
SAS	2.94	198.63
SPSS	1.08	209.83
BMDO7M	1.38	160.00
SPSS(stepwise)	1.20	207.70

The fourth example is from page 156, Morrison (1967). There are two groups each with 10 members, and all of the programs were used. Numerical results were in closer agreement than usual, and no observations were misclassified.

The BMDO7M stepwise procedure was used, as well as the WILKS and RAO methods in SPSS. BMD and the RAO method resulted in variables two through five being included in the final analysis, while the WILKS method selected variables two, four, and five. Time and storage are as follows:

	time(seconds)	space(k-bytes)
SSP	1.38	27.26
SAS	2.79	197.98
BMDO4M	2.10	187.69
BMDO5M	.66	137.45
BMDO7M	4.02	159.21
SPSS(WILKS)	1.44	204.58
SPSS(RAO)	1.38	203.00

This data set was revised to contain five groups and two variables. Each group has five observations.

The SAS program used the POOL=TEST option. The covariance matrices did not differ significantly, so the pooled covariance matrix was used in the analysis. Equal classification matrices were produced for all programs except BMD, which differed slightly for groups two through five. The BMD stepwise analysis included both variables, as did the SPSS analysis, which used the Mahalonobis method. The usual statistics on the jobs break down as follows:

	time(seconds)	space(k-bytes)
SSP	1.50	27.26
SAS	3.00	198.66
BMD07M	1.44	162.21
SPSS	1.38	203.00

The sixth and final example is a data set with four groups and six variables from page 426 of the SSP user's guide, volume III. It was run on all programs except BMD04M. All classification results agreed except for SAS, whose results differed slightly in the first three groups. The only stepwise analysis was in BMD07M, which resulted in all six variables being entered. Time and space are as follows:

	time(seconds)	space(k-bytes)
SSP	1.98	27.26
SAS	3.06	200.65
BMD05M	.84	144.00
BMD07M	4.14	159.61
SPSS	1.20	206.00

The BMD07M run took considerably longer than usual because plots of canonical variables were requested. The SSP program was run in WATFIV because this produced substantial savings in both time and space over FORTRAN G.

F. Summary

The SSP program used far less space than any of the others. SAS generally needed the largest amount of time. SSP, SAS, and BMD generally cost the most.

SSP, SAS, and BMD are easy to learn to use, while SPSS is a bit more difficult. SPSS has many more options available than any of the other programs and is especially useful for stepwise discriminant analysis because there are five different stepwise selection procedures.

VI. PRINCIPAL COMPONENTS AND FACTOR ANALYSIS

A. SSP

SSP performs factor analysis using a main program called FACTO along with six subroutines: (1) CORRE - to compute means, standard deviations, and correlations, (2) EIGEN - to compute eigenvalues and eigenvectors of the correlation matrix, (3) TRACE - to compute the cumulative percentage of the eigenvalues, (4) LOAD - to compute factor loadings, (5) VARMX - to perform varimax rotation of the factor matrix, and (6) DATA - to read in the sample data.

SSP uses principal component analysis to determine the number of variables needed to account for most of the variance among the total set of variables. A varimax rotation simplifies columns (factors) of the factor matrix.

The example main program can handle up to 35 variables and 99,999 observations, but these restrictions can be changed according to the following rules:

1. The dimension of arrays B, D, S, T, and XBAR must be greater than or equal to p, the number of variables.
2. The dimension of array V must be greater than or equal to p^2.
3. The dimension of array R must be greater than or equal to $p(p+1)/2$.

The form for the required input is shown on pages 429-430 of the SSP manual. There are two major differences from what is shown in the manual. First, an entire card is used for problem identification. This card must be included, even if it is blank. Second, a variable format card follows the card with the number of observations, etc.

The following quantities are computed and printed out:

1. Means and standard deviations of each variable,
2. A matrix of correlation coefficients between pairs of variables,
3. Eigenvalues and eigenvectors associated with the eigenvalues of the correlation matrix,
4. The cumulative percentages of the eigenvalues,
5. The factor matrix, which is a p by q matrix, where p is the number of variables and q is the number of factors. Denote the factor matrix by A. Then

$A_{ij}=\sqrt{L_j}V_{ij}$ $i=1,2,\ldots,p$ $j=1,2,\ldots,q$ where L_j is the (j)th eigenvalue, V_{ij} is its associated eigenvector, and q is determined by the number of eigenvalues ues greater than or equal to the minimum value specified,

6. Communalities: $h_i^2=\sum_{j=1}^{q} a_{ij}^2$ $i=1,2,\ldots,p$,
7. A normalized factor matrix B, where $b_{ij}=a_{ij}/\sqrt{h_{ij}^2}$,
8. The variance of the factor matrix at each iteration. The variances for the factors are computed as $S_j=(p\sum_{i=1}^{p} b_{ij}^4-(\sum_{i=1}^{p} b_{ij}^2)^2)/p^2$ and variances for the matrix are computed as $V_c=\sum_{j=1}^{q} S_j$.
9. The rotated factor matrix - The factor matrix is rotated as a function of $\emptyset$, the angle of rotation. First, a different angle, $\emptyset'$, is computed as a function of the following quantities:

x_i and y_i are factor loadings $i=1,2,\ldots,p$,

$A=\sum_{i=1}^{p}(x_i+y_i)(x_i-y_i)$,

$B=2\sum_{i=1}^{p}x_iy_i$,

$C=\sum_{i=1}^{p}((x_i+y_i)(x_i-y_i)+2x_iy_i)((x_i+y_i)(x_i-y_i)-2x_iy_i)$,

$D=4\sum_{i=1}^{p}(x_i+y_i)(x_i-y_i)x_iy_i$,

$NUM=D-2AB/p$, $DEN=C-(A+B)(A-B)/p$, $\emptyset'=\frac{1}{4}\tan^{-1}(NUM/DEN)$.

If DEN is positive, then $|\cos(\emptyset)|=\cos(\emptyset')$ and $|\sin(\emptyset)|=\sin(\emptyset')$.

Otherwise, $|\cos(\emptyset)|=\frac{1}{2}\sqrt{2}(\cos(\emptyset')+\sin(\emptyset'))$ and $|\sin(\emptyset)|=\frac{1}{2}\sqrt{2}(\cos(\emptyset')-\sin(\emptyset'))$.

If NUM is positive, then $\cos(\emptyset)=|\cos(\emptyset)|$ and $\sin(\emptyset)=|\sin(\emptyset)|$.

Otherwise, $\cos(\emptyset)=|\cos(\emptyset)|$ and $\sin(\emptyset)=-|\sin(\emptyset)|$.

Single plane rotations are made on each pair of normalized factors. This completes one iteration, and the program returns to the step for calculating V_c. If $|V_c-V_{c-1}|$ is less than 10^{-7} four successive times, then convergence has been reached. The final rotated factor matrix A is computed as $A_{ij}=B_{ij}h_i$ and then printed.

10. Original communalities, final communalites, and the difference between the two. The difference should theoretically be zero.

We have the relationship
$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_p & y_p \end{bmatrix} \begin{bmatrix} \cos(\emptyset) & -\sin(\emptyset) \\ \sin(\emptyset) & \cos(\emptyset) \end{bmatrix} = \begin{bmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ X_p & Y_p \end{bmatrix}$$

B. SAS

The procedure named FACTOR is used to perform factor analysis. Every numeric variable read in is analyzed unless the user instructs the program to do otherwise. A matrix of factor loadings is computed and the technique of varimax rotation is used on this matrix.

The following quantities are always printed out, except as noted:

1. Means and standard deviations of each variable,
2. The correlation matrix,
3. The eigenvalues of the correlation matrix,
4. The cumulative proportions of eigenvalues,
5. The eigenvectors associated with each eigenvalue,
6. Transformed responses (optional) - The transformed response is the product of a matrix of eigenvectors and a vector consisting of the values of the variables.,
7. The factor matrix (factor loadings),
8. The results of the varimax rotation for each iteration, and
9. A communality check, or differences between communalites from the rotated factor matrix and communalities from the original factor matrix.

The number of factors to be used in the analysis can be controlled by two factors, namely N and V. N is the maximum number of factors that may be used. The number of factors actually used may be less than N if the number is further restricted by V. V is the smallest value that an eigenvalue of the correlation matrix may have and still be used in the analysis. N and V are specified on the PROCEDURE FACTOR statement, whose complete form is shown on page 202 of the SAS user's manual.

C. BMD

BMD has two programs for factor analysis. One of these is a principal component analysis routine and the other is a general factor analysis routine.

1. Principal components program (BMD01M)

The principal components program computes the principal components of standardized data (data with a mean of zero and a variance of one) and ranks the standardized data by the size of each principal component separately.

There may be from two to 25 variables used, and from three to 400 observations. Instructions for input of these quantities appear on page 152 of the BMD manual.

Computations are done in the following order:

First, the matrix W is computed as $W_{ij}=(x_{ij}-x_{.j})/\sum_{i=1}^{n}(x_{ij}-x_{.j})^2$, where $x_{.j}=(1/n)\sum_{i=1}^{n}x_{ij}$, n=the number of observations, $i=1,2,\ldots,n$ $j=1,2,\ldots,p$ and p=the number of variables. The data is standardized in this step.

Next, the correlation matrix of the W_{ij} values is calculated. This is the correlation matrix among the p variables. Also, the eigenvalues and corresponding eigenvectors of this correlation matrix are computed.

The matrix Z is computed. $Z_{ij}=\sqrt{n-1}\ WB$ $i=1,2,\ldots,n$ $j=1,2,\ldots,p$ where B is the matrix whose columns are the previously computed eigenvectors, and the W's are transformed into their orthogonal components. A covariance matrix between the columns of Z is computed as an accuracy check. This matrix should be a diagonal matrix with eigenvalues on the diagonal. Each standardized observation is ranked in descending order for the p eigenvectors such that

$R_{mc_1}=\max_i\left(\sum_{k=1}^{p}(W_{ik})(B_{km})\right)$, $i=1,2,\cdots,n$ $m=1,2,\ldots,p$,

$R_{mc_2}=\max_j\left(\sum_{k=1}^{p}(W_{ik})(B_{km})\right)$ over the remaining n-1 observations,

$\vdots$

$R_{mc_n}=\min_i\left(\sum_{k=1}^{p}(W_{ik})(B_{km})\right)$, $i=1,2,\ldots,n$. The c_i's are the observation numbers having the (i)th ranks. The n components for each variable along with their ranks are printed.

2. General factor analysis program (BMD03M)

The general factor analysis program is used in much the same way as the SSP factor analysis program. The major difference is that in BMD a correlation matrix or a factor matrix may be input instead of raw data.

The program can handle from two to 80 variables and from two to 9999 observations. The maximum number of factors to be rotated will be at least two and at most the number of variables. The number of factors to be rotated is determined by the smaller of these two numbers: (1) the number of eigenvalues greater than or equal to a control value specified by the user, and (2) the number of factors specified by the user.

Communality estimates may be supplied by the user and in this case they will replace the ones on the diagonal of the correlation matrix. A special COMMUN card is required in this case. Alternatively, multiple correlation coefficients or maximum absolute row values may be used as initial communality estimates.

The details for parameter and data input appear on pages 170-174 of the BMD manual. The output from this program includes all of the items described in the SSP factor analysis routine, and additionally the factor scores for each observation are computed and can either be printed out or written on tape. These are standardized scores and are computed as $Z_{ij}=(X_{ij}-\bar{X}_i)/A_j$, $i=1,2,\ldots,n$, $j=1,2,\ldots,q$ where q is the number of factors retained.

D. SPSS

SPSS performs factor analysis using a subprogram named FACTOR. There are five different methods of factoring available. These are: (1) principal factoring without iteration, (2) principal factoring with iterations, (3) Rao's canonical factoring, (4) alpha factoring, and (5) image factoring.

Principal factoring without iteration consists of extraction of principal components of the correlation matrix among variables. The principal components solution is used to determine the number of factors to be rotated. Alternatively, the principal components solution can be modified by using communality estimates on the diagonal of the correlation matrix instead of ones. Usually in this case either the squared multiple correlation coefficient of one variable with the others is used or the absolute value of the highest element in each column of the correlation matrix is used.

With the method of principal factoring with iterations, the diagonal elements of the correlation matrix are automatically replaced by communality estimates, these being squared multiple correlation coefficients. Also, this method uses an iterative process for improving communality estimates and continues until two successive estimates are equal. The iterative principal factoring method is automatically used by the program unless specific instructions are given otherwise.

Rao's canonical factoring method attempts to find a factor solution where the correlation between a set of hypothesized factors and the data variables is maximized. It assumes that the hypothesized factors can be determined by a linear combination of the common variance of the observed variables.

With alpha factoring, factors are defined which have maximum generalizability, measured by Cronbach's alpha. The process starts with the squared multiple correlation communality estimates. The correlation matrix is readjusted assuming that the variables used are a sample from a larger population of variables. The number of factors used is determined by the number of factors with positive alpha values.

The image factoring method seeks to find the portion of a variable associated with common factors and the unique part not associated with other variables. Factor extraction is performed on an image covariance matrix containing squares of images on the main diagonal and correlation coefficients elsewhere. The number of factors retained is determined by the eigenvalues of the image-covariance matrix. Factors with eigenvalues greater than one are retained. Four methods of factor matrix rotation are available, namely (1) quartimax, (2) varimax, (3) equimax, and (4) oblique rotation. The first three of these are orthogonal rotation methods.

The goal of quartimax rotation is to rotate initial factors so that a variable loads high on one factor but nearly zero on all others. The method simplifies rows of the factor matrix.

The varimax method maximizes the variance of factor loadings in each column of the factor matrix. Columns are simplified rather than rows. This method is automatically used unless another is specified.

The equimax method is a cross between the quartimax and varimax methods in the sense that it simplifies both the rows and columns of the factor matrix.

With oblique rotation, the factors are not orthogonal to each other. The method minimizes the crossproducts of factor loadings on reference axes to simplify primary factor loadings. Factors may be correlated with each other.

A summary of parameters for the FACTOR program appears on page 499 of the SPSS manual and the exact form of the parameter cards may be found on pages 490-498.

Correlation matrices or factor matrices may be entered instead of raw data. In this case a READ MATRIX card is used instead of a READ INPUT DATA card. Other program options include inclusion of missing data, pairwise deletion of missing data, input and output of the correlation matrix and factor matrix along with communalities, output of the factor score matrix, output of means and standard deviations, specifying the order of variables on input correlation matrices on the variable list card, weighted factor scores with missing data, and sequencing of factor scores. The rotated factor matrix, transformation matrix, and a plot of rotated factors may also be output. Factor scores may not be output if matrix input is used. The maximum number of variables for the program is 100.

E. OSIRIS

OSIRIS performs factor analysis by using a subprogram called FACTAN. The input can consist of either raw data or a correlation matrix.

If raw data is input, then the sum, standard deviation, mean, maximum and minimum for each variable are printed. Optionally, the sums of squares and crossproducts, both adjusted and unadjusted for means, and correlation matrices may be printed. Only the correlation matrix may be obtained if matrix input is used.

The Hotelling principal axes method is used to extract factors. The procedure solves for all eigenvalues and eigenvectors of the correlation matrix simultaneously with a given level of accuracy, and ranks the eigenvalues.

Principal components analvsis is used to determine the number of factors to be rotated and to estimate communalities. The number of factors may be specified in terms of either Kaiser's criterion, which is the number of eigenvalues greater than or equal to one, or the minimum percentage of variance which should be explained by the factors. Initial communality estimates may be either ones, squared multiple correlations, or estimates supplied by the user.

Varimax and oblimin rotations of the factor matrix may be performed. The varimax rotation is done first if both are specified. Different analyses on different subsets of variables may be performed in one run. The output includes the following statistics, in addition to those already mentioned:

1. Eigenvalues of the correlation matrix, and percentages and cumulative percentages of these eigenvalues,
2. Eigenvectors of the correlation matrix,
3. The inverse and determinant of the correlation matrix,
4. The correlation matrix with communality estimates on the diagonal, along with eigenvalues and eigenvectors of this matrix,
5. Final communality estimates,
6. Factor score coefficients both in raw score and standard score form,
7. Multiple correlations ,

8. Varimax rotated factor matrix (raw and normalized solutions),
9. Contributions of each factor to the total variance,
10. The transformation matrix,
11. A list of criterion values and differences at each cycle of oblimin rotation,
12. Reference structure (raw and normalized solutions),
13. Correlations between reference factors,
14. Correlations between primary factors and contributions of each one, and
15. Factor matrices.

If there is missing data, the user can have the program either stop, treat the missing data as valid data, or delete cases with missing data. Up to 60 variables may be input to FACTAN and up to 30 factors may be used if oblimin rotation is desired. Detailed instructions for the FACTAN procedure are on pages 606-613 of the OSIRIS user's manual, volume 1.

The following card sequence is used with FACTAN when raw data is read in:

```
//jobname JØB (standard job card information),name
// EXEC ØSIRIS
//SETUP DD *
control cards
$DICT
$PRINT
input dictionary - list of variables
$DATA
$PRINT
data
/*
```

The following card sequence is used for FACTAN when matrix input is being used:

```
//jobname JØB (standard job card information),name
// EXEC ØSIRIS
//FT02F001 DD SYSØUT=A
//SETUP DD *
control cards
$MATRIX
$PRINT
dictionary
correlations
means
standard deviations
/*
```

F. Sample programs

The first sample data set comes from page 154 of Morrison (1967). There are three variables and 23 observations. A minimum eigenvalue of 2.0 was specified on the SSP and SAS runs, and because all of the eigenvalues were less than that value, the factor matrix was not rotated. In the BMD run, diagonal elements of the correlation matrix were replaced by the maximum absolute value in each row. No rotation was done because the problem card specified rotation of three factors, and only two were retained. The NOROTATION option was chosen in SPSS. In the OSIRIS program, the number of factors was chosen so that at least 90% of the total variance was to be explained. Both varimax and oblique rotations were done in this problem. Rotation was done only in this run, but the other results (correlations, factor matrices, etc.) agreed closely, usually to four decimal places. The time and space of each run are as follows:

	time(seconds)	space(k-bytes)
SSP	1.20	35.90
SAS	1.80	90.97
BMD01M	1.80	201.00
BMD03M	.66	129.82
SPSS	1.68	208.46
OSIRIS	4.08	96.00

The next example is from Morrison (1967), page 155. The data set has four variables and 16 observations. A maximum of two factors and a minimum eigenvalue of 0.9 were specified in the SSP, SAS, and BMD programs. No rotation was done because there was only one eigenvalue high enough. Only means, standard deviations, and correlation coefficients were computed in SPSS because the BYPASS option was used for factoring. Oblique rotation was used in OSIRIS, with two factors and squared multiple correlations on the diagonal of the correlation matrix. This was the only run where rotation was done, but the other results agreed closely. Following are the time and space requirements of each run:

	time(seconds)	space(k-bytes)
SSP	1.26	35.90
SAS	1.62	92.00
BMD01M	1.98	195.88
BMD03M	.60	119.70
SPSS	.96	208.13
OSIRIS	3.42	96.56

Next is an example from Morrison (1967), page 156, with five variables and 20 observations. The SSP, SAS, and BMD runs used a maximum of three factors and a minimum eigenvalue of 1.0. This resulted in two factors being used for the rotated factor matrix. Also, the correlation matrix and original factor matrix were entered in BMD and SPSS, and the correlation matrix was input to OSIRIS. This was done in separate runs, and the results all agreed closely. The time and space requirements break down as follows:

	time(seconds)	space(k-bytes)
SSP	1.44	35.90
SAS	1.92	91.03
BMD01M	2.04	190.94
BMD03M(raw data)	.48	129.13
BMD03M(correlations)	.48	129.13
BMD03M(factor matrix)	.36	130.83
SPSS(raw data)	1.80	206.00
SPSS(correlations)	1.38	195.65
SPSS(factor matrix)	1.02	204.00
OSIRIS(raw data)	3.84	96.25
OSIRIS(correlations)	3.00	100.34

The fourth example is from page 14 of the Harman (1967) factor analysis textbook. There are five variables and 12 observations. SSP and BMD had a minimum eigenvalue of zero, so there were five factors used for rotation. SAS specified a minimum eigenvalue of 1.0, which resulted in two factors being retained. The SPSS run used quartimax rotation with three factors. The OSIRIS run used the assumption that the data made up the entire population instead of a sample from that population. Rotation was done with two factors. The number of factors rotated varied among programs, but the other results agreed well, as usual. The time and space requirements are as follows:

	time(seconds)	space(k-bytes)
SSP	1.80	35.90
SAS	1.50	92.60
BMD01M	1.92	200.53
BMD03M	.96	123.38
SPSS	2.70	203.71
OSIRIS	4.02	95.28

The next example is from Harman (1967), page 132. There are six variables and 24 observations. Two factors were rotated in all of the programs except SPSS, which rotated three. The results agreed closely. Rao factoring and equimax rotation were used by SPSS. The time and space statistics are as follows:

	time(seconds)	space(k-bytes)
SSP	1.62	35.90
SAS	1.56	93.19
BMD01M	2.22	193.16
BMD03M	.48	126.63
SPSS	2.94	206.69
OSIRIS	3.84	96.50

Next is an example from page 208 of Harman's (1967) text with five variables and 24 observations. Rotation was performed with four factors only in the OSIRIS run. The other runs specified a minimum eigenvalue of 1.5, and only one eigenvalue met that criterion. Time and space are as follows:

	time(seconds)	space(k-bytes)
SSP	1.44	35.90
SAS	1.74	92.52
BMD01M	2.16	190.22
BMD03M	.66	122.18
SPSS	1.74	203.62
OSIRIS	4.14	95.54

The final example is from page 429 of the SSP manual, volume III. A minimum eigenvalue of 1.0 was used in the SSP, SAS, and BMD runs and the first of the SPSS and OSIRIS runs. Four eigenvalues satisfied this requirement, so four factors were rotated. Varimax rotation was used in all of these cases, and the results agreed closely. The OSIRIS run included the output of factor scores for each observation on cards.

The next SPSS and OSIRIS runs used a correlation matrix with an altered diagonal. Initial communality estimates of .73, .74, .81, .80, .83, .76, .92, .86, and .76 were used. Three factors were rotated, and the factor matrices, correlations, etc. agreed very well between the two runs.

Oblimin (oblique) rotation was specified in two other SPSS and OSIRIS runs. Four factors were rotated in the SPSS program. OSIRIS attempted to perform a rotation with these four factors, but convergence was not reached. Time and storage requirements are as follows:

	time(seconds)	space(k-bytes)
SSP	2.88	35.98
SAS	2.16	90.72
BMD01M	2.46	195.46
BMD03M	.90	119.87
SPSS	4.02	205.49
OSIRIS	4.38	95.78
SPSS(altered diagonal)	2.70	205.24
OSIRIS(altered diagonal)	4.26	98.23
SPSS(oblimin rotation)	2.10	206.97
OSIRIS(oblimin rotation)	4.08	98.00

G. Summary

SSP used far less space than any of the other programs. BMD01M, SPSS, and OSIRIS generally agreed closely in time requirements, although OSIRIS used considerably less space. SAS and BMD03M were usually faster than the other programs. Time and storage did not seem to be affected much by the type of data input (raw data, correlation matrix, or factor matrix) because the example data sets are so small. Considerable savings in time can be achieved by using matrix input in the case of a large data set. SSP and OSIRIS generally cost the most.

SSP, SAS, and BMD are the easiest programs for an inexperienced programmer to use. I found OSIRIS to be very difficult to learn how to use at first.

SPSS and OSIRIS have more options available than the other programs. SPSS is particularly useful for experimenting with different factoring and rotation methods.

VII. REFERENCES

Anderson, T.W. An Introduction to Multivariate Statistical Analysis
John Wiley and Sons, Inc., New York, New York 1958

Center for Political Studies ISR, Inter-university Consortium for Political Research, and Survey Research Center ISR OSIRIS III User's Manual, vols. 1-6
University of Michigan, Ann Arbor, Michigan 1973

Dixon, W.J. BMD, Biomedical Computer Programs
University of California Press, Berkeley, California 1970

Harman, Harry H. Modern Factor Analysis
The University of Chicago Press, Chicago, Illinois 1967

Hemmerle, W.J., Carney, E.J., and D'Silva, A.B.
Multivariate AARDVARK Reference Manual
Computer Laboratory, University of Rhode Island, Kingston, Rhode Island 1969

International Business Machines, Inc.
System/360 Scientific Subroutine Package Version III Programmer's Manual
White Plains, New York 1968

Kansas State University Computing Center MANOVA
Manhattan, Kansas 1972

Kansas State University Computing Center Newsletter
Volume IV, Number 5, January, 1976

Morrison, Donald F. Multivariate Statistical Methods
McGraw Hill Book Company, New York, New York 1967

Nie, Norman H., Hull, C. Hadlai, Jenkins, Jean G., Steinbrenner, Karin, and Bent, Dale H. SPSS, Statistical Package for the Social Sciences 2nd edition
McGraw Hill Book Company, New York, New York 1975

Service, Jolayne A User's Guide to the Statistical Analysis System
Department of Statistics, North Carolina State University,
Raleigh, North Carolina 1972

VIII. ACKNOWLEDGEMENTS

I would like to thank Dr. Kenneth Kemp for serving as my major professor and directing this report. I would also like to thank Dr. Arthur Dayton and Dr. Dallas Johnson for serving on my graduate committee.

Kris Arheart, Kenneth Laws, and Charles Buckley of the KSU computing center gave me much helpful programming assistance.

I would like to thank Dr. William Hemmerle of the Department of Computer Science and Statistics, and Ronald Ferri of the computing center, both at the University of Rhode Island, Kingston, Rhode Island, for sending the Multivariate AARDVARK tape to us, and David Culp for assisting in copying this tape.

A COMPARISON OF MULTIVARIATE STATISTICAL PROGRAMS AVAILABLE AT KANSAS STATE UNIVERSITY

by

ROBERT L. UMHOLTZ

B.S., Kansas State University, 1974

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1976

Several programs are presently available at Kansas State University for performing multivariate statistical analyses, particularly analysis of variance and covariance, regression analysis, discriminant analysis, principal component, and factor analysis.

Among these programs are the following:

1. Multivariate AARDVARK - for analysis of variance and covariance; from the University of Rhode Island, Kingston, Rhode Island.
2. BMD (Biomedical Computer Programs) - for discriminant analysis and factor analysis; from the University of California, Los Angeles, California.
3. MANOVA - for analysis of variance, covariance, and regression; from the University of North Carolina, Chapel Hill, North Carolina.
4. OSIRIS - for analysis of variance and covariance and factor analysis; from the University of Michigan, Ann Arbor, Michigan.
5. SAS (Statistical Analysis System) - for analysis of variance, covariance, regression, discriminant analysis, and factor analysis; from North Carolina State University, Raleigh, North Carolina.
6. SPSS (Statistical Package for the Social Sciences) - for discriminant analysis and factor analysis; from the University of Chicago, Chicago, Illinois.
7. SSP (Scientific Subroutine Package) - for discriminant analysis and factor analysis; a collection of FORTRAN main programs and subroutines.

Different options are available with different programs. The options which a user needs and the comparative efficiency of programs should be considered in choosing a program for a particular computing task.