

Discriminant Analysis via Statistical Packages

Carl J Huberty and Laureen L. Lowman, The University of Georgia, Athens GA 30602

Abstract

The two aspects of discriminant analysis are briefly reviewed: predictive discriminant analysis (PDA) and descriptive discriminant analysis (DDA). Use of three popular statistical packages (BMDP[®], SAS[®], and SPSS[®]) to obtain computational results for a PDA and for a DDA is generally reviewed. Results yielded by two BMDP procedures (7M and 5M) are discussed, as are four SAS procedures (DISCRIM, STEPDISC, CANDISC, and GLM), and two SPSS procedures (DISCRIMINANT and MANOVA). It is pointed out which procedures are used to obtain PDA results and which are used for DDA results. Examples of printout information is given. Some of these examples pertain to misleading information, one to incorrect information, and some, of course, to very useful information. The need for nonpackage analyses to generate specific PDA and DDA information is also mentioned.

Introduction

Suppose one has three groups of analysis units (e.g., students, fish, businesses, institutions) on each of which there are measures available on a collection of 13 unit characteristics/attributes. With this situation, we have $k = 3$ groups and $p = 13$ response variables. There are two different research questions that may be associated with such a context. One question pertains to prediction: How well can group membership of units be predicted (using the 13 predictor variables)? This question may apply to the units on hand in that interest may be on assessing prediction accuracy for the three groups and 13 predictors, or it may apply to new units whose group membership is not known. Whatever, the set of analysis techniques useful in addressing this research question is known as *predictive discriminant analysis* (PDA).

The second question pertains to description: Assuming the grouping variable is found to have an effect on the collection of 13 outcome variables (i.e., the three vectors of 13-element means are different), how can this effect be described in terms of outcome variable composites? The set of analysis techniques useful in addressing this research question is known as *descriptive discriminant analysis* (DDA). DDA techniques are also applicable to the research context in which there is more than one grouping variable -- that is, in a so-called multiple-factor design.

It should be noted that for a PDA situation, the grouping variable plays the role of a criterion variable, and the p response variables are *predictor* variables. For a DDA situation, on the other hand, the grouping variable plays the role of a predictor variable, and the p response variables are *outcome* variables. With a two-factor (say, A -by- B) design, for

example, there would be three sets of effects of potential interest -- those pertaining to the $A \times B$ interaction, to A , and to B . Simply put, we have different analyses (PDA and DDA) for different questions; one is for prediction of group membership and one is for description of grouping variable effects. For more on the distinction of PDA and DDA, see Huberty (1994, pp. 252-253).

Whether we are dealing with a PDA or a DDA, there are multiple response (predictor or outcome) variables involved; that is, we have a *multivariate* situation. As such, the associated analyses will necessarily have to be done via computer programs. The purpose of this paper is to review the computer output available from statistical packages that is useful in conducting a PDA and a DDA. It is assumed that the reader is familiar with at least one of the statistical packages, and has some background in discriminant analysis in general.

In the next section the data set to be used to illustrate the use of the package is briefly described. Then, an indication is made of the statistical procedures to be discussed in this paper. The following two sections are devoted to describing the basic information obtainable via the packages for a PDA and for a DDA. The final section pertains to some general comments.

Data Set

A data set used by Huberty (1994) will be used to illustrate the conduct of a PDA and a DDA. With this "Bisbey" data set, there are $k = 3$ groups of $N = 153$ college French students: beginning ($n_1 = 35$), intermediate ($n_2 = 81$), and advanced ($n_3 = 37$). And there are $p = 13$ response measures: five high school cumulative grade-point averages (EGPA, MGPA, SGPA, NGPA, FGPA), number of semesters of high school French (SHSF), four measures of academic aptitude (ACTE, ACTM, ACTS, ACTN), two scores on a French test (ETSA, ETSG), and number of semesters since last high school French course (SLHF). For the PDA results to be illustrated, the 13 response variables are considered predictor variables, and for the DDA results they are considered outcome variables.

Statistical Packages

Output obtainable from three popular statistical packages will be considered for this paper. The package acronyms and releases are:

	<u>Mainframe</u>	<u>Windows</u>
BMDP	88	-
SAS	6.08	6.11
SPSS	4.1.2	6.1.3

The SAS Windows™ package will not be explicitly discussed.

SPSS for Windows is a menu-driven program that allows users to request optional outputs through dialog boxes. However, it also includes a syntax window that can be used to run most of the commands that are available on the mainframe. All syntax commands given in this paper for use on the SPSS mainframe can be used in the syntax window. To access a syntax window, the "File" menu is selected followed by the "New" and "SPSS Syntax" submenus. Then, the command syntax is typed in the same manner as for the mainframe. Available commands/options can be found in Norusis (1993a, 1993b) or in the *SPSS Base System Syntax Reference Guide*. Certain commands/options only available through the syntax windows are mentioned later.

As a side-note, Norusis (1993a) describes two additional macros, one for canonical correlation and one for ridge regression. Also, SPSS makes additional macros available on the internet. Users can access these macros using the following uniform resource locator (URL): <http://www.spss.com/support/11page.html#macro>.

The procedures within the packages that pertain to discriminant analysis vary in terms of the appropriateness of the output for PDA or for DDA. A brief summary of the package-analysis fit is as follows:

	<u>BMDP</u>	<u>SAS</u>	<u>SPSS</u>
	<u>7M 5M</u>	<u>DISC. STEP. CAN. GLM</u>	<u>DISCRIM. MANOVA</u>
PDA	X	X	
DDA		X	X
Both	X		X

Three of these procedures carry out what is called a stepwise discriminant analysis: BMDP 7M, SAS STEPDISC, and SPSS DISCRIMINANT. When using any of these procedures, it is necessary that all variables are entered into the analysis and that none are removed. [It is *not* recommended that a stepwise analysis be used to delete variables or to rank-order variables (Huberty, 1994, pp. 227, 261-262).] This may be accomplished via some commands/options in the program used for each package (e.g., in the context of the Bisbey data set):

BMDP 7M	FORCE=13	(command in DISCRIMINANT paragraph)
SAS STEPDISC	INCLUDE=13	(Option in PROC STEPDISC statement)
SPSS DISCRIMINANT	ANALYSIS=EGPA TO SLHF	(subcommand in DISCRIMINANT command file)

Comments in the following sections pertain almost exclusively to output specific to a discriminant analysis (PDA or DDA). Although very relevant in some analysis situations, output for data transformations and data imputation is not discussed. See Huberty and Julian (1995) for a way of handling missing data in a PDA context. Also not discussed is the *initial* search for potential outlying units, and output for descriptive information.

The use of terms related to discriminant analysis is somewhat inconsistent across the many manuals for the statistical packages. There are three descriptors in particular that take on somewhat varied meanings in the manuals (and in other writings as well!). One is "discriminant analysis." In some manual writing with regard to some of the procedures mentioned earlier, what is being referred to is PDA, and yet some numerical results yielded by the very same procedures pertain exclusively to DDA. With reference to a specific research question, the term "discriminant analysis" has limited interpretive value and may, practically speaking, be noninformative -- the same may be stated regarding the (non)descriptor "discriminant function analysis." The third descriptor that is interpreted differently is "discriminant function." The word "function" in both aspects of discriminant analysis refers to some type of composite of the response variables, often linear. There is a linear composite of the predictor variables in a PDA, and there is a linear composite of the outcome variables in a DDA, but these are *different* composites. So, one must take caution when reading discussions pertaining to discriminant analysis in general in the manuals (and in some books and articles as well!).

Basic PDA Information

Preliminary Analyses

[It is assumed that missing data problems and outlier problems have been resolved.] The assessment of multivariate normality of the collection of predictor variable scores for each group can be accomplished via all three packages. Mahalanobis distances can be used to assess whether this condition has been met. For BMDP, Mahalanobis distances can be saved using the 7M program. A SAVE paragraph can be inserted after the GROUP paragraph (e.g., in the context of the Bisbey data):

```
/SAVE FILE = 'BISBEY2.DAT'.
FORMAT = '(I1, 2F4.1, I1, 3F4.1)'.
KEEP = GRP.
MAHAL.
NEW.
```

The FILE and NEW commands create a new data file to which seven variables are saved. The first variable saved is the original (not predicted) grouping variable and is requested by the KEEP = GRP command. Next, by default, BMDP saves the scores for each of the two linear discriminant functions (LDFs) followed by the predicted group membership. Finally, the command MAHAL requests that the Mahalanobis

distances be saved. These are saved as variables MAHAL1, MAHAL2, and MAHAL3. The FORMAT command describes the format in which the data is saved. Then, by using the 5D program, a normal plot of Mahalanobis distances for each group can be printed. Program paragraphs and commands (for group 1 distances) are as follows:

```
/INPUT FILE IS 'BISBEY2.DAT'.
VARIABLES = 2.
FORMAT IS '(I1, 9X, F4.1)'.
CASES = 35.

/VARIABLE NAMES = GRP, MAHAL1.
USE = MAHAL1.
MAXIMUM = (GRP)1.
MINIMUM = (GRP)1.

/PLOT TYPE = NORMAL.

/END
```

Hence, the variables read by the FORMAT command are GRP and MAHAL1.

A Q-Q Plot for assessing normality can be obtained using the PROC PRINCOMP statement and USS command in the SAS package. The following program can be downloaded from the World Wide Web at the following URL: <http://www.sas.com>. From the SAS menu, click-on "Customer Services," followed by "Documentation/Publications," "On-line Code Samples," and "Applied Multivariate Statistical and SAS software (Pubcode #55234)." This program is listed in the publication by Khattree and Nalk (1995). The example is for the Bisbey data, group 1.

```
TITLE1 J=L 'NORMAL P-PLOT';
OPTIONS LS=64 PS=45 NODATE
NONUMBER;

DATA A;
INFILE 'A:BISBEY.DAT';
INPUT Y1-Y13;
TOTN=35.0; /* TOTN IS THE NUMBER OF
OBSERVATIONS

/*
PROC PRINCOMP DATA=A COV STD OUT=B
NOPRINT;
VAR Y1-Y13;
DATA QQ;
SET B;
DSQ=USS(OF PRIN1-PRIN13);
/*

DATA QQ;
SET QQ;
NDSQ=DSQ/2;
PROC CAPABILITY NOPRINT;
QQPLOT NDSQ/GAMMA(ALPHA=2);
PROBPLOT NDSQ/GAMMA(ALPHA=2);
```

```
RUN;
*/

PROC SORT;
BY DSQ;
DATA QQ;
SET QQ;
CHISQ=CINV(((N-5)/TOTN),13);
PROC PLOT;
PLOT DSQ*CHISQ='*';
TITLE2 'Q-Q PLOT FOR ASSESSING
NORMALITY';
LABEL DSQ='MAHALANOBIS D SQUARE'
CHISQ='CHI-SQUARE QUANTILE';
RUN;
```

A normal probability plot assessing normality (of the Mahalanobis distances "d1") that can be obtained using the NPLOT command on SPSS is given below. Mahalanobis distances are obtained using the DISCLASS Macro that will be discussed later.

```
NPLOT
/VARIABLES=d1
/TYPE=P-P .
```

When using SPSS for Windows, after selecting data for the group being plotted (e.g., group 1), select "Graphs" followed by "Normal P-P..." on the (sub)menus. Then, moved the variable "d1" from the source variable list to the "Variables" dialog box and click "OK."

Another preliminary analysis involves the assessment of the equality of the k group covariance matrices. This assessment is typically accomplished via a statistical test using any one of the three packages. Using BMDP 5M, SAS DISCRIM, and SPSS MANOVA, results of a chi-squared test are reported, whereas F test results may be obtained via SPSS DISCRIMINANT and SPSS MANOVA. The test is obtained by default via BMDP 5M, SPSS DISCRIMINANT, and SPSS MANOVA, whereas the option POOL=TEST is required to get the test results via SAS DISCRIM. The printout headings for this test are explicit and make it easy to find results. The referent distribution -- chi-squared -- is, however, *not* specified by BMDP 5M; only the test statistic value itself is printed: $2 \times \text{Rho} \times \text{Log}(\text{Lambda}) = 211.54$ (this is the same value, within rounding, output by SAS DISCRIM and SPSS MANOVA).

Prior Probabilities

An initial step in conducting a PDA is to specify the probability that any analysis unit is a member of each group *without* consideration of the unit's predictor variable vector of scores, that is, *prior* to data collection for the study on hand. Such a probability is called a *prior probability*. There are three types of prior probabilities (or priors) that may be used with the packages: equal priors, proportional priors (n_i / N), and specified priors. It may be argued that the priors should reflect the relative sizes of the corresponding populations; this is the case regardless of the group sizes. Thus, it is expected

that for nearly all applications of PDA, priors should be specified by the researcher. How to specify priors for each package is indicated below.

BMDP 7M, 5M PRIOR=.25, .50, .25.
 (Command in GROUP paragraph)
 SAS DISCRIM PRIORS '1'=.25 '2'=.50 '3'=.25;
 (Option in PROC DISCRIM
 statement)
 SPSS DISCRIMINANT
 PRIORS=.25 .50 .25
 (Subcommand in DISCRIMINANT
 command)

When using the SPSS for Windows syntax, users can indicate priors according to only two options available through the dialog boxes, equal priors or priors based on sample size. However, the user specified option can be accessed two different ways. First, as discussed above, the same program subcommands used on the mainframe would be typed into the syntax file. Second, decide on the procedure and option desired through the dialog boxes and select "Paste." SPSS will automatically paste the correct program commands to the syntax window. Then, the PRIORS subcommand can be modified by removing the specification "Equal" and replacing it with specified priors. An example of specifying the priors .25, .50, and .25 for groups 1 to 3 respectively is as follows:

/PRIORS = .25, .50, .25

Classification Functions

If the researcher concludes that the k covariance matrices are "in the same ballpark," then *linear* classification functions (LCFs) are utilized, otherwise, *quadratic* classification functions (QCFs) may be used. [It is assumed that the researcher has specified the k prior probabilities of group membership at the outset -- these priors are automatically incorporated into the constants for the classification function composites.] Numerical values of the k classification functions (i.e., composites of the predictor variables) may be used to assign a unit (e.g., a student) to one of the k criterion groups -- a unit is assigned to that group with which is associated the largest function value.

The weights and constant for each of the k *linear* classification functions (LCFs) are printed by default with BMDP 7M, SAS DISCRIM (POOL=YES is used, or POOL=TEST is used and the P value for the covariance matrix homogeneity test is smaller than the default value -- .10 -- or smaller than a value specified by the researcher), and SPSS DISCRIMINANT. The printout heading used by BMDP 7M is CLASSIFICATION FUNCTIONS; it is assumed that the user recognizes that they are *linear*. The headings used by SAS DISCRIM and SPSS DISCRIMINANT are less clear and it may be argued that they are inappropriate. With SAS DISCRIM the heading is Linear Discriminant Function -- the current writers consider this expression is a DDA, not a PDA, context. The main heading used by SPSS DISCRIMINANT, CLASSIFICATION FUNCTION COEFFICIENTS, is

acceptable, but not the parenthetical expression, FISHER's LINEAR DISCRIMINANT FUNCTIONS -- R.A. Fisher's functions are the composites used in a DDA, *not* in a PDA.

The weights and constant for each of the k *quadratic* classification functions (QCFs) are obtainable via BMDP 5M and SAS PROC DISCRIM. The k sets of linear weights, sets of quadratic weights, and sets of cross-product weights may be obtained as part of the output of BMDP 5M (printed by default under heading QUADRATIC DISCRIMINANT ANALYSIS followed by the subheading COEFFICIENTS FOR GROUP ...) and SAS DISCRIM (with the option POOL=NO when OUTPUT = file name (no quotes) is added to the DISCRIM paragraph and a PROC PRINT data = file (no quotes) name paragraph is added.)

Suppose one has determined a classification rule in the form of a set of LCFs or QCFs using a given data set. Knowing the weights and constant for each function, then, enables a researcher to classify a set of new units -- more on classification of new units is provided later in this section.

Posterior Probabilities

Another way to determine group assignment for a unit is to use the probability of belonging to a group, given the p -element vector of predictor scores -- these are the (estimated) *posterior probabilities*. These probabilities may be estimated, in general, using either an internal analysis or an external analysis. An *internal* analysis is one in which a classification rule (i.e., a set of k classification functions) is determined using the data on hand, and then the functions are applied to the same data to make group assignments and to estimate hit rates. For an *external* analysis, a rule is built on one data set and then applied to a different data set. It is, perhaps, clear that an external analysis would provide less biased hit rate estimates. In short, the current writers suggest that an external analysis *always* be conducted. Two ways to do an external analysis are to use a holdout sample, and to use a leave-one-out (L-0-0) analysis. For the latter, BMDP uses the expression, "jackknife" and SAS uses "cross-validation." Suppose one has a total sample of $N = 153$. To accomplish a L-0-0 analysis, then, one unit is set aside and a classification rule (linear or quadratic) is built on the remaining 152 units; this rule is used to classify the left-out unit (and the result of the classification is tabulated). This process is carried out a total of 153 times, thus producing k L-0-0 posterior probabilities for each unit. The L-0-0 method is strongly preferred over the holdout-sample method by the current writers.

Values of (estimated) posterior probabilities are provided by all three PDA package procedures. BMDP also outputs a squared Mahalanobis distance from the unit's score vector to the respective group centroid. Posterior probabilities for four French students are given in Table 1 for each of four procedures. Some comments about these results follow.

- BMDP distances do not involve the priors.
- To get L-0-0 (or jackknife) results with BMDP 7M, use the JACK command in the DISCRIMINANT paragraph.
- L-0-0 results are different from internal results.
- No matter what the value of k is, SPSS DISCRIMINANT outputs only the two largest posterior probabilities -- $P(G/D)$ values.
- Group assignment is more decisive for some units than for others.
- Internal results are different from external (L-0-0) results.
- Results yielded by SPSS DISCRIMINANT are linear internal results; correct quadratic results are not available.

Leave-one-out results are available via the SPSS for Windows package using a matrix macro (DISCLASS). To run the macro, the following sequence of (sub)menus is used: "File," "New," "SPSS syntax." This will open a Syntax Window. Then, the following (sub)commands are typed:

```
INCLUDE 'DISCLASS.SPS'.
DISCLASS VARS = VARIABLE LIST
/GROUPS=GROUP
/HOLDOUT = YES
/CASE = N
/PRIORS (VALUE LIST)
```

The first line calls up the DISCLASS macro. The subcommand "holdout = yes" requests the L-0-0 analysis and "case = n" identifies the number of groups used in the analysis. The macro saves both posterior probabilities and Mahalanobis distances into a file name "dc__tmp2.sav." When using this macro on a computer that does not give write permission to the current directory containing both output files, the macro must be modified to save the output files on floppy disk. This can be accomplished by changing six lines from either "dc__tmp1.sav" or "dc__tmp2.sav" to "a:dc__tmp1.sav" or "a:dc__tmp2.sav." The quotation marks must be included. The use of this macro is not recommended with large data files without a math co-processor.

Outliers

Suppose an analysis unit is assigned to group 2. A natural question, then, is: To what extent does that unit "look" like the typical member of group 2? The "typical" group member would be one whose score vector is close to the group centroid. So, it would be informative to have an index that reflects that closeness of a unit's score vector to the centroid of the assigned group. Such an index is the *typicality probability*. This index may be useful in identifying units that are possible outliers. The SPSS DISCRIMINANT procedure is the only one that outputs the typicality probability for each unit -- it is denoted $P(D/G)$ on the printout. From Table 1 we see that

student 106 would be assigned to group 1, but the student's score vector is very distant from the centroid of group 1 -- $P(X_{106} | 1) = .0072$. Also, even though student 5 has $P(1 | X_5) = 1.0000$, $P(X_5 | 1) = .0384$. Thus, student 106 and student 5 may be considered as possible outliers.

Table 1

Some classification results for Bisbey data from each package

BMDP 7M (linear L-0-0)

INCORRECT CLASSIFICATION	CASE	MAHALANOBIS D-SQUARE FROM AND POSTERIOR PROBABILITY FOR GROUP -					
		GRP1		GRP2		GRP3	
	5	16.9	1.000	39.3	0.000	61.7	0.000
	106 GRP1	17.8	0.893	23.5	0.107	45.1	0.000
	109 GRP3	46.8	0.000	24.8	0.338	22.1	0.662
	123 GRP2	33.3	0.007	24.8	0.955	29.8	0.038

SAS DISCRIM (linear L-0-0)

Posterior Probability of Membership in GRP: From Classified

Obs	GRP	into GRP	1	2	3
5	1	1	1.0000	0.0000	0.0000
106	2	1	0.8935	0.1065	0.0000
109	2	3	0.0000	0.3375	0.6625
123	3	2	0.0067	0.9531	0.0381

SPSS DISCRIMINANT (linear internal)

CASE	ACTUAL	HIGHEST	PROBABILITY		2ND	HIGHEST
SEONUM	GROUP	GROUP	P(D/G)	P(G/D)	GROUP	P(G/D)
5	1	1	0.0384	1.0000	2	0.0000
106	2	1	0.0072	0.7459	2	0.2541
109	2	2	0.0748	0.5205	3	0.4795
123	3	2	0.7868	0.8560	3	0.1409

SAS DISCRIM (quadratic L-0-0)

From Classified Posterior Probability of Membership in GRP:

Obs	GRP	into GRP	1	2	3
5	1	1	0.9956	0.0044	0.0000
106	2	2	0.4374	0.5626	0.0000
109	2	2	0.0000	0.8847	0.1153
123	3	2	0.0001	0.9999	0.0000

BMDP 5M (quadratic internal)

INCORRECT CLASSIFICATIONS	CASE	POSTERIOR PROBABILITIES FOR GROUP		
		GRP1	GRP2	GRP3
	5	1.000	0.000	0.000
	106	0.017	0.983	0.000
	109	0.000	0.993	0.007
	123	0.000	0.041	0.959

Fence-riders

Suppose, for $k = 3$, a L-0-0 analysis yields three respective posterior probabilities of .44, .46, and .10 for a particular unit. According to these numbers, the unit would be assigned to group 2. But this unit is "close" to being assigned to group 1 as well; that is, the unit may be identified as a *fence-rider* or an in-doubt unit. A researcher could very well scan the list of k posterior probabilities to identify possible fence-riders. If it is judged that a *minimum* posterior probability is needed for group assignments to be made, then an option may be used in the SAS PROC DISCRIM statement, THRESHOLD. It turns out that with the Bisbey data, a fairly high THRESHOLD value (of about .55) would have to have been used to detect possible fence-riders (in that way). It should be noted that use of THRESHOLD will *not* detect all potential fence-riders. If THRESHOLD = .35 is used, one could still get two posterior probabilities of, say, .46 and .45. A visual inspection of the printed posterior probabilities will well serve the researcher in identifying some fence-riders. If the data set is extremely large, some data file manipulation may be called for.

Hit Rate Estimates

There are two common approaches to estimating hit rates -- separate-group or across-group. One is to use a *counting estimator*. With this approach, one simply counts the number of units from each group that are, based on classification function scores or posterior probabilities, assigned to each group. Such counting is typically summarized in the form of a $k \times k$ *classification table*, in which the main diagonal elements are the hits and the off-diagonal elements are the misses. [This is, without a doubt, the more popular approach to hit rate estimation.] Because some of the procedures output multiple classification tables, it is imperative that the researcher identify the table of interest. There are four different analyses that yield classification tables. Headings used by four procedures are given in Table 2.

Table 2
Classification table headings

	BMDP		SAS DISCRIM	SPSS DISCRIMINANT
	7 M	5 M		
Linear				
Internal	classification Matrix	linear classification matrix	resubstitution summary using linear discriminant Function	classification results
L-0-0	jackknifed classification	-	cross-validation summary using Linear Discriminant Function	-
Quadratic				
Internal	-	Quadratic classification Matrix	Resubstitution summary using Quadratic Discriminant Function	classification results
L-0-0	-	-	Cross-Validation Summary Using Quadratic Discriminant Function	-

It should be pointed out that classification results yielded by SPSS DISCRIMINANT when separate covariance matrices are used (as opposed to an error covariance matrix) that the so-called quadratic results are incorrect. This is recognized by Norusis (1993b): because "SPSS uses the discriminant function values to classify cases, not the original variables, it is not possible to obtain the optimum quadratic rule" (p. 37). The second approach to obtaining hit rate estimates is the maximum-posterior-probability (M-P-P) method, available only through SAS DISCRIM. With this approach, hit rate estimates are not based on counts but on the actual posterior probabilities of group membership. The M-P-P hit rate estimate for group 1 is determined by finding the mean of the posterior probabilities for all those units assigned to group 1 -- these would be the *largest* posterior probabilities for those units. Group 2 and group 3 (for $k = 3$) M-P-P hit rate estimates are found in a similar manner. This approach to hit rate estimation still has not "caught on" in applied research. See Huberty (1994, pp. 90-93) for a discussion of the M-P-P method (and M-P-P/L-0-0 too).

Predictor Selection

Not one of the three packages has any commands that would yield information to determine the best (in the sense of highest hit rate of interest) subset of a given size. With p predictors, one would need to conduct $2^p - 1$ PDAs. This number of analyses could be drastically reduced if the researcher "forces" in, say, q ($< p$) predictors -- the number of analyses reduces to $2^{p-q} - 1$. So, all subsets to be considered may be determined by "taking the bull by the horns" and conducting a "weekend analysis." A useful program has been written, however, to conduct an all-possible-subset analysis by J. D. Morris (Florida Atlantic University) -- this program does allow for "forcing in" some predictors. The classification accuracy criterion used in determining the best subsets (of size 15 or less) is a linear L-0-0 separate-group hit rates or across-group hit rate. Such an analysis is very relevant, because a subset of predictors will very often do a better job -- in terms of a higher hit rate of interest -- than the complete set of p predictors.

Predictor Ordering

Once the final set of predictors is determined, it may be of interest to assess the relative contribution of the predictors to the hit rate of interest -- a separate-group hit rate or the across-group hit rate. Here, too, none of the package procedures directly yield such information. The variable ordering approach suggested here is based on the question: How well can I do without a variable? The "well" is, of course, taken in a relative sense. The "do" refers to the hit rate of interest. An equivalent question is: How much (in a relative sense) does a variable contribute in addition to the remaining variables? A predictor ordering for this approach may be determined using the package of choice. What needs to be done to order the predictors is to simply conduct p analyses, each with $p-1$ predictors. Suppose the hit rate of interest is that for group 2. Then the most important variable is the one, when deleted, that is associated with the *lowest* group 2 hit rate based on the

remaining $p-1$ predictors. As might be expected, some judgment calls may need to be made because hit rates associated with two or more deleted predictors may be "close."

Classification of New Units

Suppose one has a data set with $N = 153$ for which predictive accuracy is acceptably high. Suppose, further, that a purpose of the study is that of implementing the derived rule in practice. By "practice" is meant the application of the rule with new analysis units. Consider having 10 new units with complete vectors of predictor scores, but group membership is unknown and is to be predicted. There are two methods that can be utilized to determine predicted group membership for the 10 new units. One method is to use the derived classification functions (linear or quadratic). Group assignment for a new unit is accomplished by finding the k function scores and assigning the unit to the group associated with largest function score. This method may be implemented by creating a file containing the function weights and constants, and then preparing a fairly simple computer program to generate the function scores (and maybe even to make the group assignments).

The second method to determine group membership of the 10 new units is to simply annex the 10 predictor score vectors to the original 153 vectors, but with no group membership indicated. This may be accomplished via BMDP 5M or 7M, SAS DISCRIM, or SPSS DISCRIMINANT.

Even though both methods will yield identical group assignments (assuming the same rules are used), the latter method may be preferred -- assuming one has access to the original data set. This is the preference because one gets a clearer idea of fence-riders and of group assignment confidence using posterior probabilities than using classification function scores, and, if SPSS DISCRIMINANT is used, one can determine if any of the new units are potential outliers.

Basic DDA Information

Preliminary Analyses

The assessment of multivariate normality of the collection of outcome variable scores for each group and of group covariance matrix equality for DDA would be done using the same methods as for PDA. There is another analysis that may be considered preliminary to DDA, or as part of DDA; via, multivariate analysis of variance (MANOVA). One of four test criteria may be considered for the MANOVA test of the equality of the k p -element mean vectors: Bartlett-Pillai, Hotelling-Lawley, Roy, and Wilks. The reporting of these criteria by the three packages is given in Table 3. In addition to the P value for the statistical test, it is advisable to report some type of effect size index value. At least five effect size indexes have been proposed (see Huberty, 1994, pp. 194-196). The SPSS MANOVA procedure is the only one considered herein that explicitly yields effect size values -- three of them are under the heading "Multivariate Effect Size."

Table 3

MANOVA criteria reported by the packages

	BMDP 7M 5M	SAS DISC. STEP. CAN. GLM				SPSS DISCRIM. MANOVA	
Bartlett-Pillai		X	X	X	X		X
Hotelling-Lawley	X ¹	X		X	X	X	X
Roy		X		X	X		X
Wilks	X X	X X	X X	X X	X X	X X	X X

¹Obtainable via BMDP 4V.

Linear Discriminant Functions

It is the variance shared between the grouping variable and the linear discriminant functions (LDFs) that is of primary interest in conducting a MANOVA. That is, the linear composites of the outcome variables -- the LDFs -- is the focus of a DDA. In a k -group design, the total number of LDFs is equal to $m = \min(p, k-1)$. Although LDF weights are judged to be of very limited interpretive value, they are given by all the packages. The labeling of the weights (and constants) varies somewhat across the packages:

BMDP 7M & 5M	COEFFICIENTS FOR CANONICAL VARIABLES
SAS CANDISC & GLM	Raw Canonical Coefficients
SPSS DISCRIMINANT	UNSTANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS
MANOVA	Raw discriminant function coefficients

LDF Dimensions

All m LDFs often do not contribute to the interpretation of the resultant group separation. To best describe the grouping variable effects, it is important to focus only on "relevant" LDFs. Relevance may be determined in one or more of three ways: statistical tests, cumulative proportion of variance in the system of variables that is explained by the grouping variable, and LDF plots. In this subsection of the paper we focus on the first two ways. Statistical tests are done by SAS CANDISC and SPSS DISCRIMINANT and MANOVA. Test results are given in Table 4. As with any statistical test, some judgment calls will need to be made. When making multiple tests, as one would be in this situation, each tail probability area should be multiplied by the number of tests conducted. Particular attention should be paid to some heading information provided by SAS CANDISC: "Test of HO: The canonical correlations in the current row and all that follow are zero." This heading explicitly tells the user that what is *not* being tested is the "significance" of *each* LDF. Results of the multiple statistical tests (F by SAS CANDISC and SPSS MANOVA, and chi-

squared by SPSS DISCRIMINANT) are clearly indicated. The cumulative variance proportions are labeled "Cumulative" by SAS CANDISC, "CUMULATIVE PERCENT" by SPSS DISCRIMINANT, and "Cum. Pct." by SPSS MANOVA.

Table 4

Test information for LDF dimensionality

SAS CANDISC

Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation	Eigenvalue	Difference	Eigenvalues of $INV(E)^*H = CanRsq/(1-CanRsq)$	
						Proportion	Cumulative
1	0.860026	0.847725	0.021118	0.739645	2.8409	2.7154	0.9577
2	0.333886	0.239379	0.072069	0.111480	0.1255		0.0423
							1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.23133105	11.4555	26	276	0.0001
2	0.88852018	1.4533	12	139	0.1492

SPSS DISCRIMINANT

CANONICAL DISCRIMINANT FUNCTIONS									
Function	Eigenvalue	Percent Variance	Cumulative Percent	Canonical Correlation	After Function	Wilks' Lambda	Chi-squared	D.F.	Significance
1*	2.84090	95.77	95.77	0.8600259	0	0.2313311	210.80	26	0.0000
2*	0.12547	4.23	100.00	0.3338859	1	0.8885202	17.021	12	0.1488

* Marks the 2 Canonical Discriminant Functions Remaining in the Analysis.

SPSS MANOVA

Eigenvalues and Canonical Correlations

Root No.	Eigenvalue	Pct.	Cum. Pct.	Canon Cor.
1	2.84090	95.77036	95.77036	.86003
2	.12547	4.22964	100.00000	.33389

Dimension Reduction Analysis

Roots	Wilks L	F	Hypoth. DF	Error DF	Sig. of F
1 TO 2	.23133	11.45545	26.00	276.00	.000
2 TO 2	.88852	1.45332	12.00	139.00	.149

LDF Plots

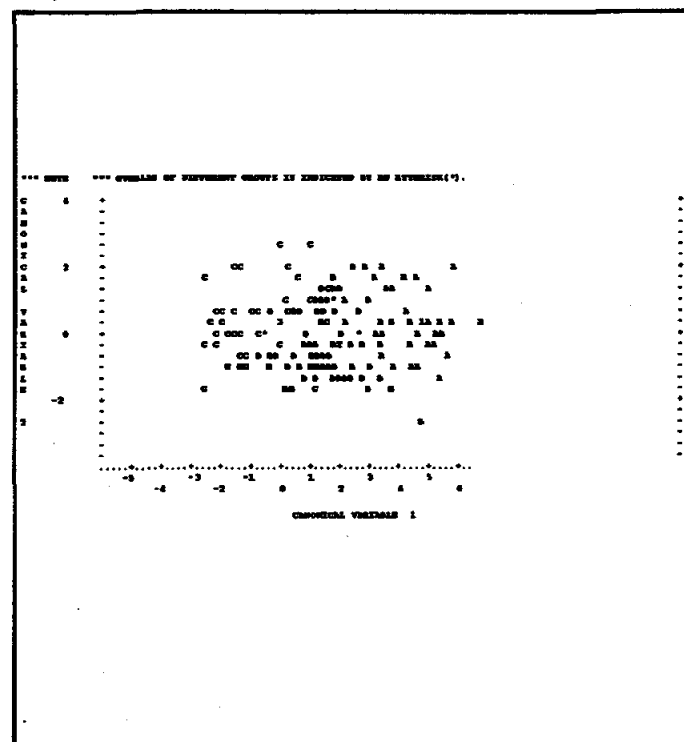
Another bit of information that may be useful in determining the number of relevant LDFs is obtainable by examining plots of the LDF centroids; that is, plots of the means of the LDFs. Two-dimensional plots are yielded by all three packages. Rectangular coordinates are used in all instances -- see Huberty (1994, p. 215f). Two plots of the LDF centroids for the Bisbey data are given in Figure 1. When comparing the two plots, a distortion is noted. The location of the group 2 centroid is "reversed" from one plot to the other. Although not reported in this paper, the signs of the LDF weights in the BMDP 7M output are opposite those in the

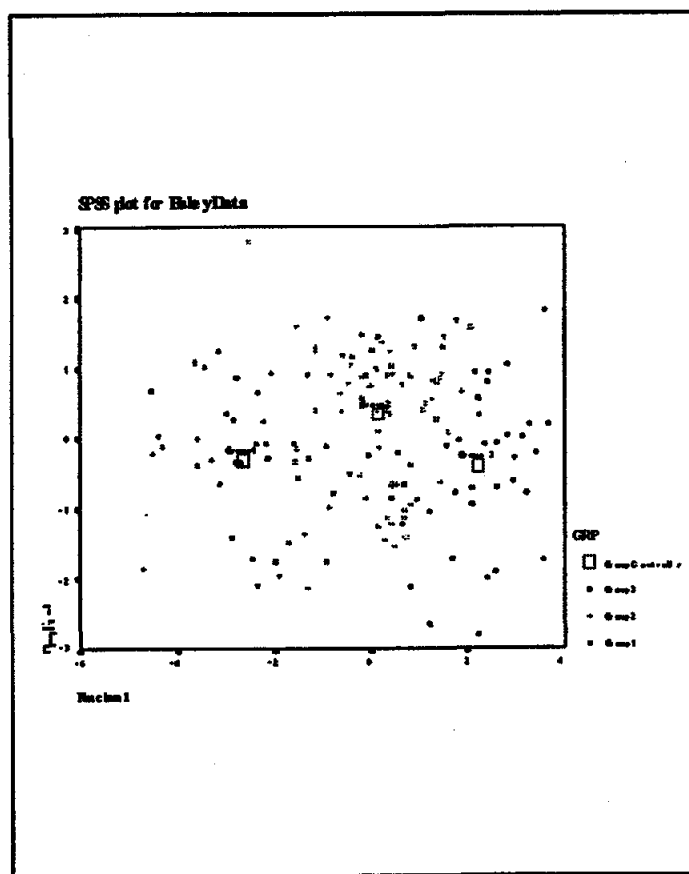
SPSS DISCRIMINANT output. Nothing is wrong (!), because the LDF weights are only unique up to a constant of proportionality. The relevance of the second LDF beyond that of the first one may be judged by projecting the group centroids onto the second LDF axis. [The relevance of the first LDF has already been assessed via the initial MANOVA.] If there is "substantial" group separation on the second LDF, then it may be retained for interpretation purposes -- another judgment call (!). Another purpose of examining LDF plots is that of learning something about the relative topological locations of the groups in the space of the LDFs. Once some interpretive "sense" is made of the respective LDFs, the group topology may be informative. Making interpretive sense of LDF is turned to next.

Figure 1

LDF plot for Bisbey Data

BMDP





Structure r 's

As mentioned earlier, the focus of a study of multivariate group separation is on the linear composites, that is, on the LDFs. It is with respect to these composites that the mean vector group differences are tested, and it is with respect to these composites that we focus our interpretation of the analysis results. The question then becomes, "What do these composites represent in a substantive sense?" If a score on an LDF is calculated, what does such a score represent? An LDF is, in fact, a type of principal component. As such, then, a reasonable approach to naming or labeling an LDF is to use the p variable-LDF correlations, or *structure r 's*. Because there are multiple groups of units in a DDA context, structure r 's may be based on total or error covariances. The current writers favor the *error* structure r 's. Error structure r 's are reported by SAS CANDISC, SAS GLM, SPSS DISCRIMINANT, and SPSS MANOVA. The (sub)commands or options needed for these four procedures to output the structure r 's are:

SAS CANDISC	BY DEFAULT (Except when using SHORT option)
SAS GLM	MANOVA H= GRP/CANONICAL
SPSS DISCRIMINANT	BY DEFAULT
SPSS MANOVA	DISCRIM CORR

The headings used are:

SAS CANDISC
SAS GLM
SPSS DISCRIMINANT

Pooled Within Canonical Structure
Within Canonical Structure
STRUCTURE MATRIX: POOLED
WITHIN-GROUPS
CORRELATIONS BETWEEN
DISCRIMINATING VARIABLES
AND CANONICAL
DISCRIMINANT FUNCTIONS

SPSS MANOVA

Correlations between DEPENDENT
and canonical variables

The structure r 's may be used, as in factor analysis, to arrive at LDF names or labels. An LDF label is determined primarily by the larger (in absolute value) structure r 's, and secondarily by moderate structure r 's -- of course, some judgment calls will have to be made.

Outcome Variable Ordering

The approach suggested for variable ordering in a DDA context is based on the same notion used for the PDA context: How well can I do without a variable? The "do" in a DDA context pertains to group separation. One way to address the DDA variable ordering problem is to carry out a set of all-but-one-variable analyses, the same as was mentioned for PDA variable ordering. Even though this could be accomplished enough by conducting the p analyses, each with $p-1$ outcome variables, results from package output may be utilized. This output is some last-step information in a stepwise analysis. [Otherwise, stepwise information is of very limited usefulness (Huberty, 1994, pp. 227, 261-262).] To use stepwise output for variable ordering purposes, it is imperative that all p outcome variables be entered into the analysis -- how this is accomplished was discussed earlier in this paper in the section, Statistical Packages. The information of interest is provided by the following procedures with headings as indicated (for the Bisbey data, $p=13$):

BMDP 7M	STEP NUMBER 13
SAS STEPDISC	Statistics for Removal
SPSS DISCRIMINANT	VARIABLES IN THE
	ANALYSIS AFTER STEP 13

The index of interest is labeled "F TO REMOVE" by BMDP 7M and SPSS DISCRIMINANT, and simply "F" by SAS STEPDISC. These F values are really partial F values in that they are transformations of a *partial lambda*, the lambda for a given variable when the remaining variables are partialled out. The rank ordering of these F values will be the same rank-ordering obtained by using an all-but-one-variable set of analyses. As so happens in many data analysis situations, some judgment calls will need to be made -- in this case because some F values may be "close." For the Bisbey data, the following F-to-Remove values resulted:

Variable	F	Possible Rank	
EGPA	0.17	12.5	9
MGPA	1.17	10.5	9
SGPA	1.73	7	9
NGPA	6.85	3	3.5
SHSF	5.07	4	3.5
FGPA	1.31	7	3.5
ACTE	1.36	7	9
ACTM	0.20	12.5	9
ACTS	1.84	7	9
ACTN	0.84	10.5	9
ETSA	16.48	2	2
ETSG	23.45	1	1
SLHF	1.53	7	9

It is clear from this listing that ETSG and ETSA are most important, followed by NGPA and SHSF, with the remaining variables playing a lesser role in contributing to overall group separation.

A final comment is in order. As mentioned above, the "do" in the question related to variable ordering, "How well can I do without a variable?," pertains to separation. The implication in this subsection is that "separation" refers to omnibus group separation. The "do," however, may very well pertain to separation reflected in a *contrast*.

Contrast Analyses

A multivariate group *contrast* may be viewed as a linear composite of the group mean vectors with the sum of the mean vector weights being zero. For example, with $k = 3$, two sample contrasts are $\bar{Y}_1 - \bar{Y}_3$ and $\bar{Y}_1 - \bar{Y}_2 - 2\bar{Y}_3$. The null hypothesis for a contrast would be that the population contrast is equal to zero. For pairwise contrasts (e.g., $\bar{Y}_1 - \bar{Y}_3$), some statistical test information is reported by default for BMDP 7M, BMDP 5M, SAS CANDISC, and SPSS DISCRIMINANT. The test information is based on the squared Mahalanobis distances (D^2) among the groups, and an F transformation of D^2 . A summary of the obtainable pairwise contrast information is given below:

	D^2	F	P
BMDP 7M		X	
BMDP 5M	X	X	X
SAS CANDISC	X	X	X
SPSS DISCRIMINANT		X	X

Because $\eta^2 = F / (F + v)$ with $v = N - p - 2$, one can use an adjusted effect size index, η^2_{adj} , in conjunction with the P value in deciding upon pairwise group differences --, η^2_{adj} is given in Huberty (1994, p. 197).

Any form of a contrast -- pairwise or complex -- may be tested via any of the three packages: BMDP 7M, SAS GLM, and SPSS MANOVA. The commands/options used for two contrasts are as follows:

```

BMDP 7M      CONTRAST = 0, 1, -1
              (in DISCRIMINANT paragraph)
              CONTRAST = 1, 1, -2.
SAS GLM      CONTRAST '2 V5 3' GRP 0 1 -1;
              CONTRAST '1 AND 2 V5 3' GRP 1 1 -2
SPSS MANOVA  CONTRAST (GRP) =
              SPECIAL (1 1 1, 0 1 -1, 1 1 -2)

```

When more than one contrast is specified in the BMDP 7M DISCRIMINANT command, the specified contrasts are used simultaneously. Note that the "SPECIAL" keyword in the SPSS MANOVA command is only one of several keywords that may be used. In addition, the first set of weights indicates how to calculate the mean or constant effect -- this is usually a vector of one. Also, the number of degrees of freedom for "GRP" ($k - 1 = 2$ for the Bisbey data) indicates the number of sets of contrast weights that need to be specified -- if only one contrast is of interest, then a second one must be specified as a "filler."

If a contrast P value and effect size value jointly suggest a "real" contrast effect, then the (error) structure r 's for the lone LDF may be examined to label the construct that underlies the contrast effect. Also, the p outcome variables may be ordered with respect to (relative) contribution to specific contrast effects. Again, F -to-Remove values from the last step of a BMDP 7M stepwise analysis may be used to rank-order the outcome variables in terms of their relative contribution to the contrast effect. The other two procedures, SAS GLM and SPSS MANOVA, do not yield variable ordering information relative to contrast effects.

When doing a contrast analysis via SPSS for Windows, contrast weights must be specified through the syntax window using a MANOVA command. For example:

```

MANOVA EGAP TO SLHF BY GRP(1 3)
      /CONTRAST(GRP) = SPECIAL (1 1 1, 0 1 -1, 1 1 -2)
      /DESIGN = GRP(1) GRP(2).

```

Comments

An attempt was made in this paper to provide some details regarding the completion of analyses to obtain information for the conduct of a predictive discriminant analysis (PDA) and a descriptive discriminant analysis (DDA). Some details pertaining to commands/subcommands/options/statements for procedures in the three statistical packages were not included in this paper. Such details may be found in Huberty (1994) -- in an appendix that has some annotated printouts as well as in the text proper. Although one procedure in each package is broadly viewed as stepwise discriminant analysis procedures, stepwise analyses, per se, are viewed as yielding very limited useful information (Huberty, 1989).

Some pluses and minuses of the package discriminant analysis procedures are briefly indicated below:

	Plus	Minus
BMDP	Linear L-0-0	No quadratic L-0-0 Only one MANOVA criterion No typicality probability No structure r 's
SAS	Linear and quadratic L-0-0 M-P-P hit rates	No typicality probability No F transform of Box M
SPSS	Typicality probability F transform of Box M	No L-0-0 No quadratic PDA

In doing either type of discriminant analysis, PDA or DDA, it is suggested that multiple analyses be strongly considered. Multiple analyses may be considered within a given package -- use of multiple procedures and multiple commands/options -- as well as across packages. Across-package analyses might be done to obtain additional information (e.g., typicality probabilities) or to compare results (e.g., linear versus quadratic classifications). Given a research situation that calls for a discriminant analysis (PDA or DDA), expecting to "get by" with a single analysis would be analogous to buying a car from the first dealership you visit without checking on the list price -- you may not be getting the best deal for your money.

There is an analysis related to a PDA situation that is highly desirable to do but has not been discussed to this point. This simple analysis involves a statistical test to determine if an obtained hit rate is better than one that could be obtained by "chance." If it is judged that an obtained hit rate is better than chance via this test, then it is suggested that the value of an index be reported that indicates how much better than chance. See Huberty (1994, pp. 103-108) for a discussion of the test and the better-than-chance index.

There is an analysis that some readers associate with a DDA that was not discussed in this paper, but is often used by applied researchers. This analysis pertains to outcome variable selection/deletion. What is often used by some researchers (and generally condemned by some methodologists) is a stepwise discriminant analysis. Setting aside a stepwise analysis, some researchers might use the more appropriate all-possible-subset analysis. But, why is it desirable to discard some outcome variables in a DDA context? Some would "default" to the parsimony argument. In designing a study whose purpose suggests a DDA, it is assumed that considerable care and thought were used in choosing the initial set of outcome variables. After some preliminary variable screening that may be deemed desirable, there appears no real substantive reason for deleting variables. If some variables are found not to contribute substantially to identifying constructs that underlie grouping variable effects, such a finding may be worth noting. But there is no apparent need to delete some outcome variables to redefine the system of variables to be studied.

References

- Dixon, W. J. (Ed.) (1992). BMDP statistical software manual. Berkeley, CA: University of California Press.
- Huberty, C. J. (1989). Problems with stepwise methods: Better alternatives. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1) (pp. 43-70). Greenwich, CT: JAI Press.
- Huberty, C. J. (1994). Applied discriminant analysis. New York: Wiley.
- Huberty, C. J., & Julian, M. W. (1995). An ad hoc analysis strategy with missing data. Journal of Experimental Education, 63, 333-342.
- Khattree, Ravindra & Naik, Dayanand N (1995). Applied Multivariate Statistics with SAS software. Cary, NC: SAS Institute Inc.
- Norusis, M. J. (1993a). SPSS for windows: Advanced statistics, release 6.0. Chicago: SPSS Inc.
- Norusis, M. J. (1993b). SPSS for windows: Professional statistics, release 6.0. Chicago: SPSS Inc.
- SAS Institute Inc. (1989). SAS/STAT user's guide. Cary, NC: SAS Institute Inc.

SAS is a trademark or registered trademark of SAS Institute Inc. in the USA and other countries.

SPSS and BMDP are trademarks of SPSS Inc. in the USA and other countries.

Windows is a trademark of Microsoft Corp. in the USA and other countries.