1. Introduction

The major objective in many multivariate statistical analyses is to explain the complex relationship between groups of vector variables in as few terms as possible. Therefore, most techniques try to reduce the size of the set of variables concerned. Because of this, principal components, factor analysis, cluster analysis and similar ones are usually included in the analysis. In fact, the primary objective of the experiment may be to explain the relationship between groups of vector variables with only one linear combination of the vector variable. In this type of experiment, as many different variables may be included as can be feasibly measured, since any redundant variables will probably be eliminated early in the process. Many techniques have been developed to reduce the number of variables in an experiment using a variety of criteria. In some cases, the result desired is not necessarily that of reducing the dimensionality of the problem, but rather of detecting differences between groups or of discriminating between groups. In this type of analysis it is important that all of the variables be functionally independent as most test statistics are based on the estimate of the variance-covariance matrix. If any of the variables are redundant the resulting variance-covariance matrix will be singular, and any test statistic based on the inverse or determinant of this matrix will be meaningless.

In the experiment where the functional relationship is obvious, the problem can be handled simply by removing the variables which are redundant. If they cannot be readily identified, then the problem becomes more complex. The usual technique for handling this type of problem is to make a suitable transformation from the set of original variables to other variables, usually through principal-components analysis. Dempster (1) has such an example. The difficulty of this approach is that of interpretation of the results. This paper suggests an alternate approach to the problem which gives exactly the same results, but uses the original set of variables. This alternate approach uses generalized inverses and eigenvectors, all of which are easily obtained on SAS/76 through the FREQ MATRIX. The alternate approach will be demonstrated through an artificially generated data set which has a set of five variables, one of which is redundant.

2. Notation and Formulae

Suppose a multivariate analysis is being performed on a sample taken on a set of $p$ variables, $X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$, from $q+1$ groups. A sample of $n_1$ is taken from each group with a total sample of $n = n_1 - 1$ observations. Suppose that all $p$ of the variables are not functionally independent. That is, there exists a set of coefficients $a_i$ such that:

$$a_i^T x = f_i, \text{ for } i = p-k+1, \ldots, p$$

and $f_i$ is constant. This means that $k$ of the $p$ variables are combinations of some of the rest, or there are $k$ redundant variables. The usual approach to the problem would require a transformation from $x$ to some other set of variables. The matrix of transformation would be:

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_{p-k}^T \\ a_{p-k+1}^T \\ \vdots \\ a_p^T \end{bmatrix}$$

where $A'A = AA' = I$; and the transformed variables would be:

$$Z = AX = \begin{bmatrix} \frac{Z_1}{a_1^T x} \\ \vdots \\ \frac{Z_{p-k+1}}{a_{p-k+1}^T x} \\ \vdots \\ \frac{Z_p}{a_p^T x} \end{bmatrix}$$

The last $k$ values in $Z$ are constants and only the first $p-k$ variables in the set would be used. Thus the new set of variables is $Z$. The analysis of variance table for the new set of variables would be:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>$p-k$ by $p-k$ matrices of SS &amp; CP computed from standard formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>$q$</td>
<td>$H_x$</td>
</tr>
<tr>
<td>Within</td>
<td>$n-q$</td>
<td>$E_x$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
<td>$H + E_x$</td>
</tr>
</tbody>
</table>
Here $H$ is the hypothesis (between) matrix of sums of squares and cross products and $E_z$ is the error (within) matrix computed from $z$ only. The test statistics, canonical correlations and other values could be computed from these matrices as they would be of full rank. The computations for handling this type of problem could be easily accomplished via program statements and PROC GLM on SAS.76.

If the original variables could also be used in the standard formulae to calculate a between and within matrix. The result could be:

<table>
<thead>
<tr>
<th>MANOVA (Untransformed variable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Between</td>
</tr>
<tr>
<td>Within</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

If $z$ is a constant, then both $E_z$ and $H_x+E_z$ will be singular. Further, $E_z A_1 = 0$ and $(H_x+E_z) A_1 = 0$ for the $k$ coefficients $a_i$. Since these matrices are singular, test statistics using them will be meaningless. However, the standard test statistics can be modified as follows with the result being identically equal to those obtained using $H_x$ and $E_z$ (4):

(a) In formulae using natural inverses, replace with any form of a generalized inverse.

(b) In formulae using determinants, add

$$ F_{p-k+1}^{(p)} \text{diag}(A_1) $$

to the matrix before taking the determinant. The $A_1$ are the $k$ coefficient vectors which result in the redundancy.

Thus, the statistics for testing the hypothesis of equal means among the $q+1$ groups will be:

Wilks' lambda $= \frac{\det(H_x+E_z + A_1^t A_1)}{\det(H_x+E_z)}$.

Hotellings $T^2 = \text{Trace}(E_z H_x^{-1})$, and

Pillai's $V = \text{Trace}(H_x^{-1}(H_x+E_z)^{-1})$.

The canonical correlations will be solutions to the equation:

$$ E_z^{-1} H_x^{-1} \lambda_1 = 0, $$

and a test for the goodness of fit of a hypothetical discriminant function $g^t x$ will be:

$$ \text{Wilks' lambda} = \det(A_1 E_z^{-1} H_x^{-1} A_1^t) ^{1/p} $$

where $A_1$ is Wilks' lambda for the test of equal means given above.

Statistics for testing the direction and collinearity of the proposed discriminant function, and other test statistics can be modified in like manner and will give the same result as that obtained using transformed variables. Also, the choice of generalized inverse chosen is immaterial, so the inverse of $E_z^{-1} H_x^{-1} A_1$ will be a generalized inverse of $E_z$ and so on. This gives an easy method to compute the required generalized inverses using SAS.76. The value of the $A_1$'s can be obtained by using any of several orthogonalization techniques (such as the 'sweep-out' method in Rao (3)), or by noting that since $E_z A_1 = 0$, the $A_1$'s will be the eigenvectors corresponding to the eigenvalues of $E_z$ which are zero. Thus, the values needed can be obtained using the EIGEN option of PROC MATRIX.

3. An Example

The following artificial example is presented to demonstrate the technique and the ease of application using SAS.76. A sample of 100 observations is generated from each of three groups on five variables, $x_1$ through $x_5$. The variables were generated as follows:

**Group 1**

- $x_1 = 100 + .2*\text{Normal}(0)$
- $x_2 = 50 + .2*\text{Normal}(0)$
- $x_3 = 75 + .2*\text{Normal}(0)$
- $x_4 = x_1 + .1*\text{Normal}(0)$
- $x_5 = .2*\text{Normal}(0)$

**Group 2**

- $x_1 = 102 + .2*\text{Normal}(0)$
- $x_2 = 51 + .2*\text{Normal}(0)$
- $x_3 = 74 + .2*\text{Normal}(0)$
- $x_4 = x_1 + .1*\text{Normal}(0)$
- $x_5 = 1 + .2*\text{Normal}(0)$

**Group 3**

- Same as Group 1

Thus, $n+1 = 300$

Thus, $p+1 = 3$

PROC MATRIX was then used to perform the analysis using the proposed procedure. First, the hypothesis, error, and total matrices were constructed.
The EIGEN option was used to obtain the eigenvalues of \( E \times \) :

\[
\begin{pmatrix}
22.62 & 13.21 \\
11.73 & 10.99 \\
1.36 &
\end{pmatrix}
\]

Note that the smallest eigenvalue is only .02 of the total of the eigenvalues. Thus it could be considered effectively zero. A test for eigenvalues based on the normal distribution is available, and if applied here, the smallest eigenvalue would not be different from zero but the rest would (c.f. Kshirsagar (2)).

The eigenvector corresponding to the zero eigenvalue is:

\[
\begin{pmatrix}
-.75 \\
-.005 \\
.04 \\
.67 \\
.03
\end{pmatrix}
\]

The matrix \( E^{-1}H \) is computed, again using the options in PROC MATRIX. This matrix is the above eigenvector multiplied by its own transpose. This matrix is added to \( E \times \) and \( E^{-1}E \), then the determinant and inverse obtained. The resulting test statistics are then computed. The resulting values are:

Wilks' lambda = 0.0204

Hotelling's \( \hat{T}^2 \) = 46.910

Pillai's \( V \) = 1.003

Using the options in PROC MATRIX, the canonical correlations could be obtained, a test for one or more hypothetical discriminant functions could be computed, and other multivariate analysis techniques applied.

Since the data for this example was artificially generated, the cause of the singularity can be removed, and the results compared with the ones given here. The procedure was to drop \( x_4 \) from the analysis, and perform multivariate analysis of variance on the remaining variables. The statement

MODEL \( x_1 x_2 x_3 x_4 = \text{GRP} \)

was used in PROC GLM with the option MANOVA. The resulting hypothesis and error matrices were identical to those given previously less the fourth row and column (as they should be). The test statistics for the hypothesis of equal groups had the following values:

\[
\begin{align*}
\text{Wilks' lambda} & = 0.0202 \\
\text{Hotelling's } \hat{T}^2 & = 47.59 \\
\text{Pillai's } V & = .997
\end{align*}
\]

None of the test statistics so obtained differ from those obtained using PROC MATRIX and generalized inverses by more than 1%. Thus the results appear to be quite consistent.

The proposed procedure seems to give good results without the use of any transformations, and possibly will make interpretation easier, especially in testing the goodness of fit of a hypothetical discriminant function as the function in question does not have to be transformed prior to being tested.

4. References


