The Multivariate Dustbin

Phil Ender

UCLA Statistical Consulting Group (Ret.)

Stata Conference Baltimore - July 28, 2017

Back in graduate school...

My advisor told me that the future of data analysis was multivariate.

By multivariate he meant...

- MANOVA

By multivariate he meant...

- MANOVA
- Linear Discriminant Function analysis (LDA), and

By multivariate he meant...

- MANOVA
- Linear Discriminant Function analysis (LDA), and
- Canonical Correlation analysis (CCA)

Why didn't he mention factor analysis?

My advisor wasn't interested in factor analysis. He didn't use factor analysis.

So, I will not include factor analysis in this presentation.

Further...

At that time statistical training in psychology was very ANOVAcentric. MANOVA is very ANOVA like, so many psychologists liked it. Further, MANOVA provides some of the most powerful tests of group differences that are available.

For Software we ran NYBMUL by Jeremy Finn.

NYBMUL stands for New York university Buffalo Multivariate analysis.

And so it came to pass...

In spite of my advisor's ringing endorsement, newer fancier methods came along and MANOVA, discriminant function analysis (LDA) and canonical correlation (CCA) were put in the back of the closet and were somewhat forgotten.

In fact...

In the last fifteen plus years in UCLA's Stat Consulting there have been only a few questions concerning MANOVA. And, no questions about linear discriminant function analysis or canonical correlation analysis.

Let's look at each method beginning with MANOVA

MANOVA is either a multivariate generalization of univariate ANOVA,

or univariate ANOVA is a restricted form of MANOVA.

MANOVA uses information simultaneously from each of the response variables to examine differences in group centroids.

Example Data

Three response variables; four groups; N = 200

```
. tabstat read write math, by(program)

Summary statistics: mean
  by categories of: program

 program |      read     write      math
---------+------------------------------
       1 |  49.41026  50.97436  49.84615
       2 |  56.41975  56.30864  57.06173
       3 |  46.10417   46.4375   46.0625
       4 |     54.25  55.53125     54.75
---------+------------------------------
   Total |     52.23    52.775    52.645
----------------------------------------
```

Stata MANOVA Example

```
. manova read write math = program

Number of obs =          200
W = Wilk's lambda        L = Lawley-Hotelling trace
P = Pillai's trace       R = Roy's largest root

    Source | Statistic    df   F(df1,    df2) =  F   Prob>F
-----------+------------------------------------------------
   program |W   0.7267     3      9.0   472.3    7.36 0.0000 a
           |P   0.2752            9.0   588.0    6.60 0.0000 a
           |L   0.3735            9.0   578.0    8.00 0.0000 a
           |R   0.3665            3.0   196.0   23.94 0.0000 u
           |------------------------------------------------
  Residual |               196
-----------+------------------------------------------------
     Total |               199
------------------------------------------------------------
  e = exact, a = approximate, u = upper bound on F
```

Four multivariate criteria testing group differences

Wilks' Lambda: $\text{Det}(\mathbf{W})/\text{Det}(\mathbf{H}+\mathbf{E})$

Pillai's Trace: $\text{trace}\{\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\}$

Lawley-Hotelling Trace: $\text{trace}\{\mathbf{HE}^{-1}\}$

Roy's largest root: maximum eigenvalue of $\{\mathbf{HE}^{-1}\}$

Critical values of the multivariate criteria

Although tables of critical values have been derived for various multivariate criteria, they are extremely large and very cumbersome to use.

The common practice these days is to convert the multivariate criteria into F-ratios.

exact, approximate and upper bound for F-ratios

When converting the multivariate criteria to F-ratios the results may be exact, approximate or an upper bound depending on the number of response variables and number of groups.

For example, Rao's largest latent root reduces to an exact F-ratio when the number of response variables (p) equals 1 or 2, or when the number of levels (k) equals 2 or 3.

Which multivariate criteria is best?

Answer: It depends.

Schatzoff (1966): • Roy's largest-latent root was the most sensitive when population centroids differed along a single dimension, but was otherwise least sensitive.
• Under most conditions it was a toss-up between Wilks' and Hotelling's criteria.

Olson (1976): • Pillai's criteria was the most robust to violations of assumptions concerning homogeneity of the covariance matrix.
• Under diffuse noncentrality the ordering was Pillai, Wilks, Hotelling and Roy.
• Under concentrated noncentrality the ordering is Roy, Hotelling, Wilks and Pillai.

Final "Best": • When sample sizes are very large the Wilks, Hotelling and Pillai become asymptotically equivalent.

How does one interpret MANOVA results?

Many researchers fall back on separate univariate ANOVAs to interpret the results.

It would be better to be able to do multivariate post-hoc comparisons.

Multivariate post-hoc comparisons?

In general, there are no multivariate multiple group comparisons in the sense of **pwcompare** in the major stat packages. **pwcompare** itself does work in **manova** but only on one response variable at a time.

It is possible to do "true" MANOVA post-hoc pairwise comparisons using multivariate simultaneous confidence intervals but this requires custom programming.

I computed simultaneous confidence intervals and found, for example, that 2 vs 3 was significant while 2 vs 4 was not.

What about **manovatest**?

It is possible to manually compute pairwise and other contrasts using **manovatest**. However, **manovatest** does not compute adjustments for multiplicity.

Here is the test for 2 vs 3 and 2 vs 4 using **manovatest**:

```
. matrix c1 = (0,-1,1,0,0)
. matrix c2 = (0,-1,0,1,0)

. manovatest, test(c1)
. manovatest, test(c2)
```

manovatest partial output

```
(1)     - 2.program + 3.program = 0
              Statistic   df   F(df1,   df2)    F    Prob>F
manovatest |W 0.7542     1       3.0 194.0 21.08 0.0000 e
           |P 0.2458             3.0 194.0 21.08 0.0000 e
           |L 0.3260             3.0 194.0 21.08 0.0000 e
           |R 0.3260             3.0 194.0 21.08 0.0000 e
  Residual |           196

(1)     - 2.program + 4.program = 0

manovatest |W 0.9890     1       3.0 194.0  0.72 0.5432 e
           |P 0.0110             3.0 194.0  0.72 0.5432 e
           |L 0.0111             3.0 194.0  0.72 0.5432 e
           |R 0.0111             3.0 194.0  0.72 0.5432 e
  Residual |           196
```

Linear Discriminant Function Analysis (LDA)

LDA is really just a variation of MANOVA. It looks at different facets of the same multivariate associations that are analyzed by MANOVA. I often run LDA along with MANOVA as an aid in interpreting the results.

In addition to tests of group differences, LDA provides information on the dimensionality of the multivariate group differences along with the weights (coefficients) used to create the latent discriminant functions (variates).

An early form of discriminant analysis was developed by R.A. Fisher in the 1930's. He demonstrated it with his famous Iris example.

LDA Example

candisc is a convenience command that automatically includes many of the **discrim lda** post estimation results. By an amazing coincidence SAS also has a proc named candisc. The following two sets of commands perform the same analysis.

```
. candisc read write math, group(program)
```

```
. discrim lda read write math, group(program)
. estat canontest
. estat loadings
. estat structure
. estat grmeans, canonical
. estat classtable
```

```
Canonical linear discriminant analysis
      | Canon.   Eigen-     Variance
  Fcn | Corr.    value    Prop.   Cumul.
  ----+-----------------------------------
    1 | 0.5179  .366505  0.9812  0.9812
    2 | 0.0831  .006945  0.0186  0.9998
    3 | 0.0087  .000076  0.0002  1.0000
  ---------------------------------------
  Ho: this and smaller canon. corr. are zero;
      Likelihood
  Fcn | Ratio      F       df1     df2  Prob>F
  ----+----------------------------------------
    1 | 0.7267  7.3558      9   472.3  0.0000 a
    2 | 0.9930  .34172      4     390  0.8497 e
    3 | 0.9999   .0149      1     196  0.9030 e
  ---------------------------------------------
   e = exact F, a = approximate F
```

Concerning the previous slide

Although three dimensions are possible, only the first dimension is statistically significant. This is not a big surprise since the three predictor variables are standardized test scores administered in an academic setting.

Also note that the F-ratio for the first dimension is the same as the R-ratio for the Wilks' lambda in the earlier MANOVA example.

LDA Output 2

Coefficients (loadings, weights) used with standardized variables to create each of the discriminant functions.

```
Standardized canonical discriminant function coefficients

             | function1  function2  function3
-------------+---------------------------------
        read |  .2355628    .579575   1.123113
       write |  .3523274  -1.171814   .1070375
        math |  .5956301   .5208397  -1.024233
```

LDA Output 3

Correlations of variables with each of the discriminant functions.

```
Canonical structure

             | function1  function2  function3
-------------+---------------------------------
        read |  .7600931   .2820111   .5854299
       write |  .7827538  -.604529    .1477876
        math |   .915274   .2460601  -.3189482
```

LDA Output 4

Group means on canonical variables

program	function1	function2	function3
1	-.3463043	-.1060211	-.0126342
2	.568244	.0571809	-.0025857
3	-.8876842	.0676699	.0047275
4	.315217	-.1170306	.0148518

LDA Output 5

Resubstitution classification summary

True program	Classified 1	2	3	4	Total
1	7	6	18	8	39
	17.95	15.38	46.15	20.51	100.00
2	12	41	12	16	81
	14.81	50.62	14.81	19.75	100.00
3	11	5	31	1	48
	22.92	10.42	64.58	2.08	100.00
4	5	13	6	8	32
	15.62	40.62	18.75	25.00	100.00
Total	35	65	67	33	200
	17.50	32.50	33.50	16.50	100.00
Priors	0.2500	0.2500	0.2500	0.2500	

There was a time in history...

before the emergence of logistic regression that 2-group discriminant function analysis was used for analyses with binary response variables.

And now, on to Canonical Correlation Analysis.

Canonical Correlation Analysis (CCA)

CCA looks at the relations between two sets of variables, which Stata calls the u- and the v-variables. Like discriminant analysis CCA also provides information on the dimensionality of the multivariate associations.

CCA creates two canonical variates (latent variables) for each dimension. The correlation between these variates are the canonical correlations.

```
. canon (read write math)(science socst), test(1 2)
/* redacted output */
Canonical correlation analysis
Number of obs =         200

Canonical correlations:
  0.8123  0.1384
Test of significance of canonical correlations 1-2
                Statistic  df1   df2          F   Prob>F
Wilks' lambda     .333617    6   390    47.5353   0.0000 e
--------------------------------------------------------
Test of significance of canonical correlation 2
Wilks' lambda     .980832    2   196     1.9152   0.1501 e
--------------------------------------------------------
e = exact, a = approximate, u = upper bound on F
```

Typical Canonical Correlation Example Slide 2

```
. canon (read write math)(science socst), stderr first(1)

Linear combinations for canonical correlations  N = 200
-----------------------------------------------------------
        |    Coef. Std. Err.    t  P>|t| [95% Conf Int]
--------+--------------------------------------------------
u1      |
   read | .0467307   .007043  6.64 0.000 .0328422 .0606191
  write | .0394098 .0072566  5.43 0.000 .0251001 .0537194
   math |  .031928 .0078627  4.06 0.000 .0164231 .0474329
--------+--------------------------------------------------
v1      |
science | .0609812 .0058361 10.45 0.000 .0494726 .0724898
  socst |  .052568 .0053823  9.77 0.000 .0419544 .0631816

(Standard errors estimated conditionally)
```

```
Canonical correlations:
  0.8123  0.1384

Tests of significance of all canonical correlations

                  Statistic df1   df2    F      Prob>F
     Wilks' lambda   .333617   6   390  47.5353 0.0000 e
    Pillai's trace   .679031   6   392  33.5839 0.0000 a
Lawley-
   Hotelling trace   1.95953   6   388  63.3582 0.0000 a
Roy's largest root   1.93999   3   196 126.7460 0.0000 u

e = exact, a = approximate, u = upper bound on F
```

```
. tabulate program, generate(p)
. canon (read write math)(p2 p3 p4)
  /* canon does not allow the use of factor variables */

Canonical correlation analysis     N = 200

Raw coefficients for the first variable set
                |        1         2         3
    ------------+------------------------------
           read |  -0.0216    0.0620    0.1206
          write |  -0.0352   -0.1365    0.0125
           math |  -0.0622    0.0634   -0.1250
```

Remember the MANOVA example? Slide 2

```
Raw coefficients for the second variable set

             |        1         2         3
-------------+------------------------------
          p2 |  -1.5222    1.9732    1.1614
          p3 |   0.9011    2.1000    2.0066
          p4 |  -1.1010   -0.1331    3.1767
--------------------------------------------

Canonical correlations:
  0.5179  0.0831  0.0087
```

```
Tests of significance of all canonical correlations

                  Statistic df1      df2        F Prob>F
     Wilks' lambda   .726691   9  472.296   7.3558 0.0000 a
    Pillai's trace   .275179   9      588   6.5980 0.0000 a
Lawley-
   Hotelling trace   .373526   9      578   7.9962 0.0000 a
Roy's largest root   .366505   3      196  23.9450 0.0000 u
---------------------------------------------------------
 e = exact, a = approximate, u = upper bound on F
```

Same results as MANOVA

Conclusion

These three multivariate methods may not be used as much as my advisor expected. But, nonetheless, they remain an interesting phase in the development of data analysis.

This concludes my presentation.