Data Analysis of Tables of Content

Han-Ming Wu

Contents

1. Introduction

1.1 Data Description

A objective analysis of the contents of the Multivariate Analysis (MVA) books are presented in this report. Table 1 lists the proportion of the number of pages in each book according to the following seven subjects. The source of the data set is from Albert Gifi (1990).

MATH: Mathematics other than statistics, i.e. linear algebra, matrices, transformation group, sets, relations

CORR: Correlation and regression, including path analysis, linear structural and functional equations

FACT: Factor analysis and principal components analysis

CANO: Canonical correlation analysis

DISC: Discriminant analysis, classification, cluster analysis

STAT: Statistics, including distributional theory, hypothesis testing, and estimation; also statistical analysis of categorical data

MANO: MANOVA, and the general multivariate linear model

1.2 Correspondence Analysis

Albert Gifi (1990) applied correspondence analysis to this data set. Correspondence Analysis is a form of multidimensional scaling, which represents row and column

points in a low-dimensional joint space. Also, it is a singular value vector technique, in which the important of a dimension is defined as the value of the corresponding singular value. Some information can be found from the correspondence projection graph of the data,. For example:

1. Two books are close together in the space if they have similar contents.
2. Two subjects or topics are close if they occur in the same books in the same degree.
3. A book is close to a subject if the book pays relatively much attention to the subject.

Some interesting findings by Albert Gifi are stated as below.

1. The clearest, most pure representatives of the subjects CORR, STAT, and MATH are, respectively, THOR, ROY, and GREC.
2. Most of the books are on the line between CORR and STAT; these are the most classical books which vary, mainly, in degree of difficulty.
3. Typical data analysis books are near the line between CORR and MATH, at least if we define data analysis in this content as applied linear algebra.
4. Graphical data analysis as in GNAN fits better into the classical scale CORR-STAT.
5. COL1 gives much more attention to MANO than others on basis of its 'technical level'.

Table 1 Proportion of number of pages of MVA books devoted to several subjects

	MATH	CORR	FACT	CANO	DISC	STAT	MANO	Tatol
ROY	0.15	0.00	0.00	0.00	0.00	0.80	0.05	206.00
KEN1	0.00	0.11	0.38	0.13	0.19	0.09	0.10	142.00
KEN2	0.00	0.22	0.17	0.05	0.23	0.33	0.00	184.00
ANDE	0.06	0.00	0.11	0.06	0.09	0.52	0.16	316.00
COL1	0.09	0.05	0.23	0.15	0.11	0.00	0.37	151.00
COL2	0.07	0.25	0.26	0.12	0.20	0.00	0.11	281.00
MOR1	0.24	0.00	0.28	0.05	0.00	0.27	0.16	306.00
MOR2	0.23	0.00	0.23	0.01	0.05	0.30	0.17	345.00
GEE1	0.45	0.12	0.20	0.07	0.16	0.00	0.00	164.00
GEE2	0.31	0.26	0.26	0.06	0.11	0.00	0.00	259.00
DEMP	0.33	0.15	0.01	0.03	0.14	0.33	0.00	324.00
TATS	0.42	0.05	0.02	0.07	0.15	0.12	0.18	261.00
HARR	0.08	0.17	0.33	0.11	0.00	0.12	0.19	211.00
DAGN	0.09	0.28	0.20	0.02	0.16	0.16	0.09	302.00
GREC	0.92	0.03	0.02	0.00	0.03	0.00	0.01	316.00
CAPA	0.38	0.10	0.17	0.09	0.27	0.00	0.00	490.00
GIRI	0.09	0.00	0.00	0.00	0.13	0.67	0.10	313.00
GNAN	0.00	0.10	0.30	0.00	0.21	0.40	0.00	189.00
KSHI	0.00	0.05	0.10	0.09	0.13	0.50	0.13	458.00
THOR	0.09	0.40	0.28	0.09	0.15	0.00	0.00	324.00

1.3 Working Procedures

We analyze this data by following steps:

1. Summarize the basis statistics of the data to overview the data profiles.
2. Applied the Principal Component Analysis (PCA) to find the groups information. In this step, we decide the number of the group according to seven subjects covariates.
3. Applied Classification and Regression Tree (CART) to learn a classifier. By running a test sample of MVA books down to this classifier, we can calculate the accurate rate.

2. Methods and Results

2.1 Graphical Summaries of the Data

Twenty MVA books are chosen to analyze, each with seven subjects or topics. Figure 1 to 5 show the proportion of subjects in each book. Some interesting findings are:

1. Books ROY and GIRI pay much more attention to STAT topic which beyond 60% of the pages of the book.
2. GREC is the most pure representatives of the MATH subject.
3. GEE1 and GEE2 contain the same subjects with different proportion. The similar result also applied to books COL1 and COL2.

Summary plots of the data by subjects are shown in Figure 6 to 9. Some interesting findings are:

1. Subject CANO occupies an less proportion (no more than 25%) in these books.
2. FAC is the most common topic in these books except ROY and GIRI.

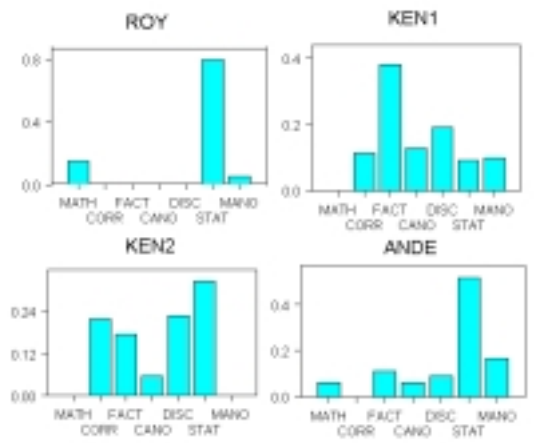

Figure 1

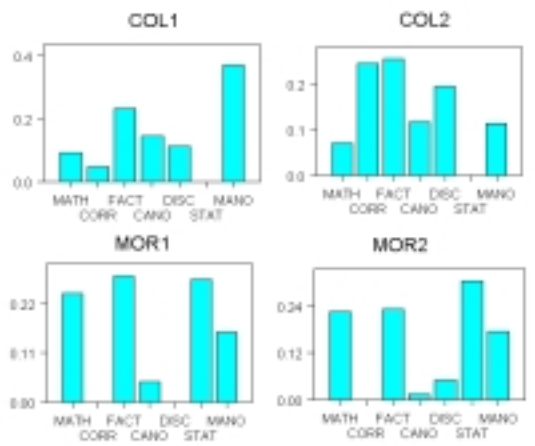

Figure 2

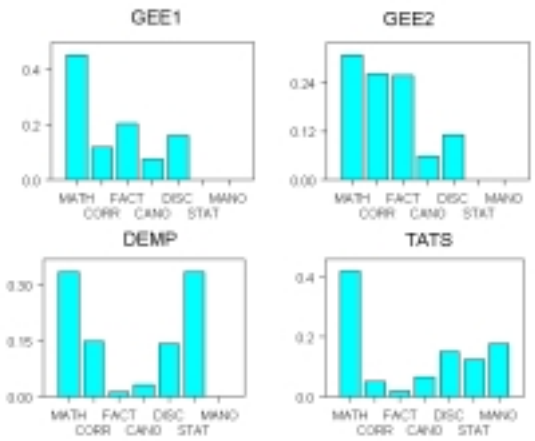

Figure 3

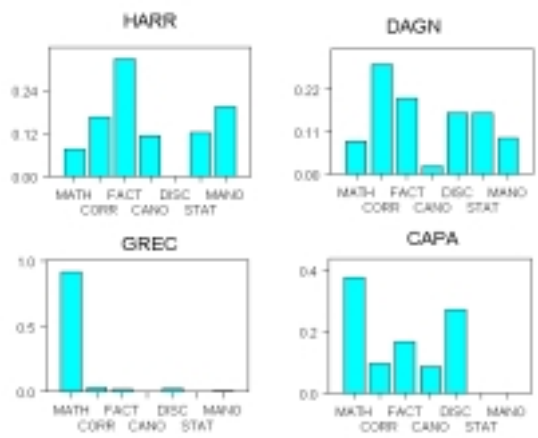

Figure 4

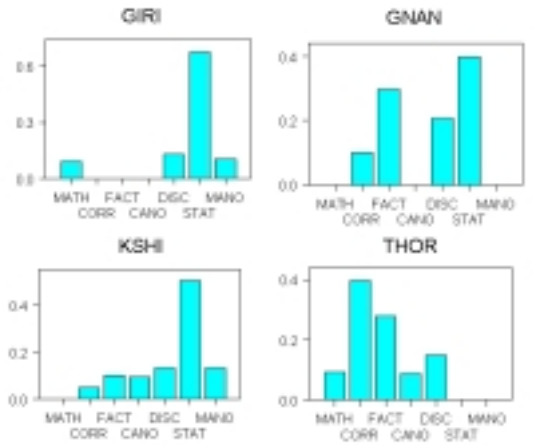

Figure 5

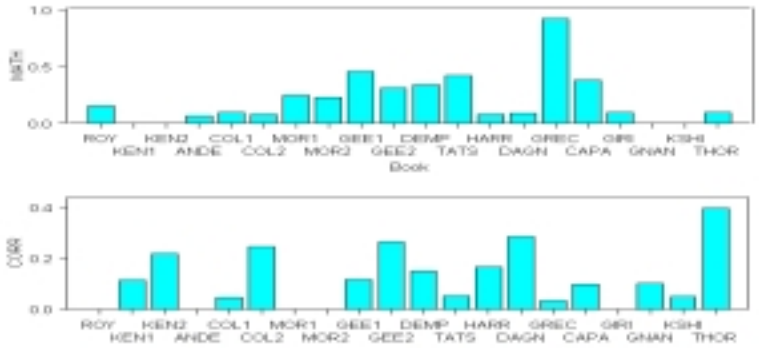

Figure 6

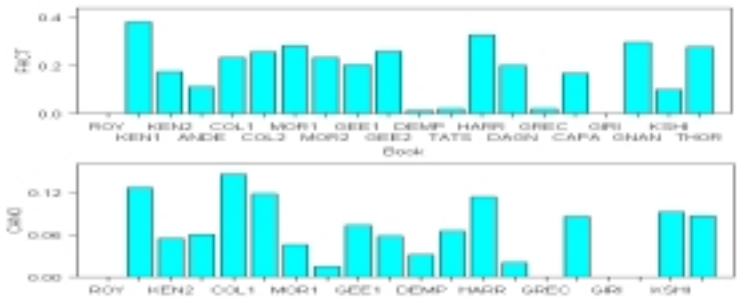

Figure 7

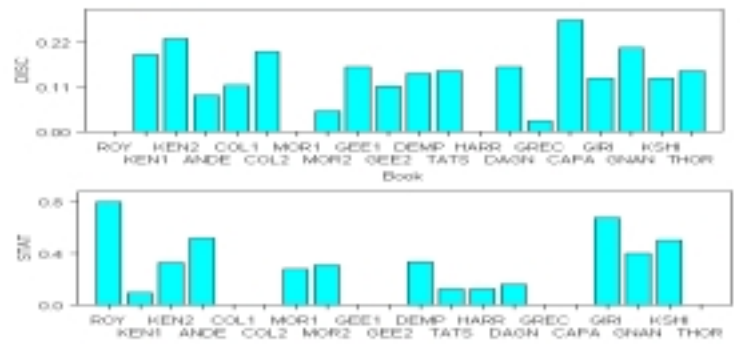

Figure 8

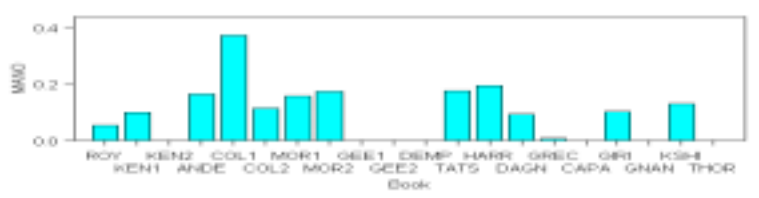

Figure 9

2.2 Principal Component Analysis

PCA finds a set of standardized linear combinations, called the principal components, which are orthogonal and taken together explain all the variance of the original data. We use books data with seven covariates as input data, to find its projection along first principal component. By looking its projection, we could roughly cluster the data by their projection distribution.

The results of the PCA are presented in Table 2. The first principal component is (-0.521, -0.155, -0.113, 0, 0, 0.827, 0). Figure 10 shows the relative important of principal components. The first two principals component have account for 80% importance.

Table 2 Results of Principal Component Analysis

```
    *** Principal Components Analysis ***
Standard deviations:
  Comp. 1   Comp. 2   Comp. 3    Comp. 4    Comp. 5    Comp. 6      Comp. 7
0.2731362 0.2201131 0.1189485 0.07424854 0.06450378 0.02567431 2.408837e-009

The number of variables is 7 and the number of observations is 20

Component names:

 "sdev" "loadings" "correlations" "scores" "center" "scale" "n.obs" "terms"

 "call" "factor.sdev" "coef"

Call:
princomp(x =  ~ MATH + CORR + FACT + CANO + DISC + STAT + MANO, data = book1,
    scores = T, cor = F, na.action = na.exclude)

Importance of components:
                      Comp. 1   Comp. 2    Comp. 3    Comp. 4    Comp. 5
   Standard deviation 0.2731362 0.2201131 0.11894854 0.07424854 0.06450378
Proportion of Variance 0.5056667 0.3283959 0.09590123 0.03736644 0.02820176
 Cumulative Proportion 0.5056667 0.8340627 0.92996390 0.96733034 0.99553210
                       Comp. 6      Comp. 7
   Standard deviation 0.025674307 2.408837e-009
Proportion of Variance 0.004467899 3.932971e-017
 Cumulative Proportion 1.000000000 1.000000e+000

Loadings:
     Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6 Comp. 7
MATH -0.521  -0.748           0.153                    0.378
CORR -0.155   0.305   0.552           -0.648   -0.123   0.378
FACT -0.113   0.428  -0.169   0.703    0.355    -0.112   0.378
CANO          0.130   -0.138  -0.182             0.886   0.378
DISC          0.142    0.322  -0.560    0.596   -0.250   0.378
STAT  0.827  -0.347    0.144   0.179                     0.378
MANO                  -0.723  -0.320   -0.313  -0.350    0.378
```

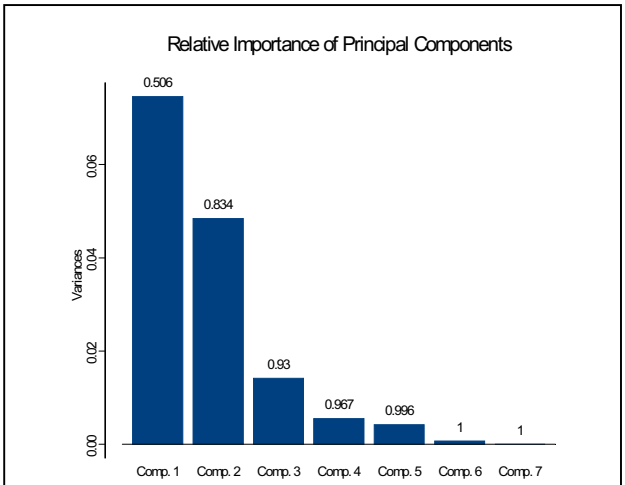

Figure 10

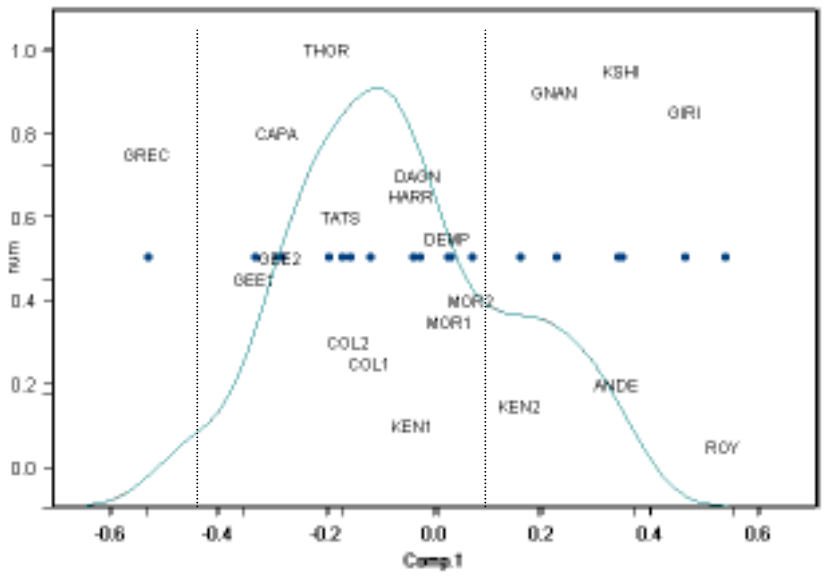

Figure 11

Figure 11 shows the projection of the twenty data points on the first component. The superimposed line is density estimation of the projection. Apparently there is only a single mode in the figure, we consider that this mixed distribution consists of three different distributions. Hence we cluster these data into three groups, which labeled by 1,2 and 3. Of course it is easy to criticize this clustering. We have chosen this clustering because we wanted to be able to apply CART technique.

Table 3 lists the clustering results. Thirteen books are clustered in one group. Six

books are in another group. Third group consists of only one book GREC. Use this grouping information, we could apply CART to the data in the next step.

Table 3 Cluster results for the books data

Group 1	GREC
Group 2	KEN1 COL1 COL2 MOR1 MOR2 GEE1 GEE2DEMP TATS HARR DAGN CAPA THOR
Group 3	ROY KEN2 ANDE GIRI GNAN KSHI

2.3 Classification and Regression Tree

Classification trees are constructed by dividing the sample space into descendant partitions recursively. The results is referred to as a decision tree. The best known methods for constructing classification tree are C4.5(Quinlan, 1993) and CART (Breiman, Friedman, Olshen and Stone, 1984). As in Figure 12, for a hypothetical three-class problem, the tree is grown by repeated splitting predictor space, starting with overall sample at root node. The subsets X1 and X3 are disjoint such that X=X2 ∪X3. Similaar outcome holds for X4 and X5, and so on. Internal and terminal nodes are represented by circles and square, respectively. Each terminal nodes is designed by a class label. A case x goes left if it satisfies the condition at the split, otherwise goes right. When x finally moves into a terminal nodes, its predicted class is given by the class label attached to that terminal node.

We use QUEST (Loh and SHIH 1997) software to perform CART-style exhaustive search. Three data points are random sampled as a test set, the remainder are as training set:

Training set : ROY, KEN1, KEN2, COL2, MOR1, GEE1, GEE2, DEMP, TATS, HARR, DAGN, CAPA, GIRI, GNAN, KSHI, GREC, MOR2

Test set: THOR, ANDE, COL1

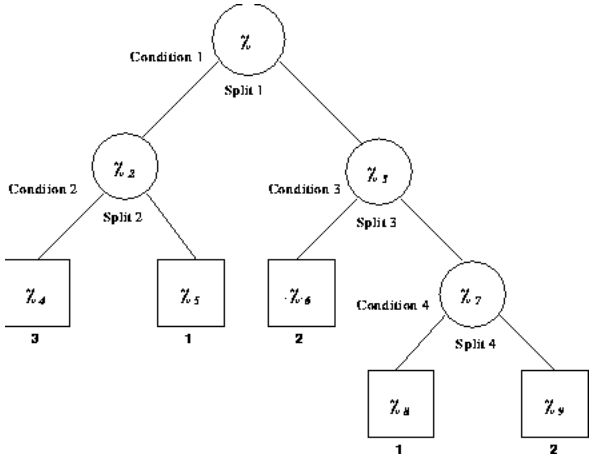

Figure 12

The results are summaries in Table 4. The decision tree is also presented in the table. CART first selected the MATH as splitting variable and 0.6580 as cut point. In node 2, the STAT is chosen as splitting point with cut point 0.3150. The terminal nodes are 3, 4, and 5 with class label 1,2, and 3, respectively.

Table 4 Selected result form QUEST output

```
 Classification tree program: QUEST version 1.8.4
 Copyright(c) 1997-9, by Shih, Yu-Shan
 This version was updated on: November 3, 1999
 Please send comments, questions, or bug reports to
 yshih@math.ccu.edu.tw

 Summary of response variable: BOOK
           class   frequency
               1       1
               2      11
               3       5
                  --------
                    17

Number of test samples: 3

 Size and CV misclassification cost and SE of subtrees:
 Tree   #Tnodes     Mean      SE(Mean)
   1       4      0.2353     0.1029
   2**     3      0.2353     0.1029
   3       1      0.3529     0.1159

 CART 0-SE tree is marked with *
 CART SE-rule using CART SE is marked with **


 Following tree is based on **
```

```
Structure of final tree
Node Left node Right node   Split variable
  1       2        3       MATH
  2       4        5       STAT
  4  * terminal node *
  5  * terminal node *
  3  * terminal node *

Number of terminal nodes of final tree
Total number of nodes of final tree = 5

Classification tree:

    Node 1: MATH <= 0.6850
      Node 2: STAT <= 0.3150
        Node 4: 2
      Node 2: STAT > 0.3150
        Node 5: 3
    Node 1: MATH > 0.6850
      Node 3: 1
```

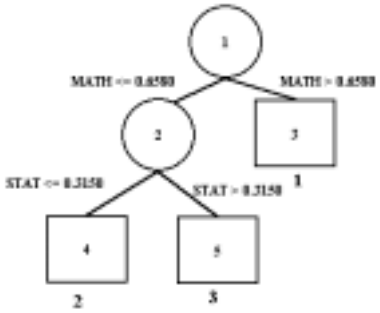

Table 5 and 6 present the predicted label for learning set and test set, respectively. In the learning sample, only case 9, DEMP, is misclassified. In test sample, three cases are all classified correctly. Because of small sample size in learning set and test set, the meaning of this report puts more emphasis on analytic procedures than the predicted results.

Table 5 Case ids, class label, terminal ids, and predicted label for the learning sample

case id	class label	terminal id	predicted label
1 ROY	3	5	3
2 KEN1	2	4	2
3 KEN2	3	5	3
4 COL2	2	4	2
5 MOR1	2	4	2
6 MOR2	2	4	2
7 GEE1	2	4	2
8 GEE2	2	4	2
9 DEMP	2	5	3*
10 TATS	2	4	2
11 HARR	2	4	2
12 DAGN	2	4	2
13 GREC	1	3	1
14 CAPA	2	4	2
15 GIRI	3	5	3
16 GNAN	3	5	3
17 KSHI	3	5	3

Table 6 Case ids, class label, terminal ids, and predicted label for the test sample

case id	class label	terminal id	predicted label
1 THOR	2	4	2
2 ANDE	3	5	3
3 COL1	2	4	2

3. Discussion

This report analyzes the MVA books data with seven subject covariates. The sample size consists of only 20 observations. We first look inside the data by subjects and then by books, and list some interesting findings. Next, we perform the PCA analysis to get the first principal component projections. According to this projection, we roughly cluster the data points into three groups. Finally, we use CART algorithm to grow a decision tree. Although we get 100% accurate rate about test sample, it is not our main point to emphasize. On the contrary, we provide another procedure to analyze this data besides correspondence analysis.

4. Reference

[1] Albert Gifi (1990). *Nonlinear Multivariate Analysis*, John Wiley & Sons.

[2] Breiman, L, et. al. (1984). *Classification and Regression Trees*, Wadsworth International Group.

[3] Yu-Shan Shih (1999). *QUEST User Manual*, Dept. of Mathematics, National Chung Cheng Uni, Taiwan.