

# Psychology 312

Spring, 2009

Final Exercise

(20 points each)

## 1. Factor Analysis – An Unusual Application from Marketing Research

(Note, you may find it convenient to use SPSS for this problem, but it can be done easily with R as well. Note: The cereal codes have value labels attached to them that give the names of the cereals. These are important for interpretation purposes, so be sure to tell SPSS to display them using the View->Value Labels command if you use SPSS.) The data in the file *rte\_cereal.sav* are a subset of the data that were presented in an article on consumer evaluation of ready-to-eat cereals by Roberts and Lattin (1991, *Journal of Marketing Research*, available online). This article contains an unusual application of a factor analysis approach. The analysis below partly replicates (and also extends) an analysis that appeared in that article. However, since the data set is a reduced version, the results you obtain will not agree precisely with those in the original article.

In this data set, a group of 116 people evaluated 12 cereals on 25 characteristics. Some people evaluated more than one cereal, so there are 235 vectors of observations and not all are independent.

Analyze the data with principal components analysis. A key choice is the number of components to retain, and my advice is probably not to retain borderline components with eigenvalues barely above 1. Rotate to varimax simple structure, and name your components by observing the names of the rating items and deciding what they have in common.

Once you have your rotated solution, obtain rotated component (“factor”) scores on each of your components, and have them added to your file. Once you save them in your file, obtain mean component scores for each of the 12 cereals by utilizing the cereal

code in the data file. (Treat the cereal codes as if they were group codes, and analyze the data by group. You may need to sort the data.) Examine these mean component scores by cereal and describe what these scores tell you about the cereals. Do the results make sense? Discuss them.

## 2. Multivariate Analysis from Scratch—Redundancy Analysis

In this question I am going to ask you to do something unusual. Instead of using a canned routine, you are going to create your own.

Dunham(1977) studied the effect of several organizational and work-related characteristics on several different measures of satisfaction. Overall, he surveyed 784 respondents, whose data are in the file *work\_satisfaction\_data.csv* which is connected to the lecture notes on canonical correlation. The 12 variables are divided into 7  $X$  variables and 5  $Y$  variables whose names are given as labels in the file. In class, we conducted a canonical correlation analysis of Dunham’s data using the **cca** package and examined the canonical weights. We used these to interpret the canonical variates (much the same as we would if they were factors). We saw that there were certain characteristics of job that had high canonical correlations with certain aspects of job satisfaction.

In class, I mentioned a potentially important failing of canonical correlation analysis, i.e., that canonical variates with a high canonical correlation need not explain large amounts of variance within their respective variable groups, despite the fact that they correlate highly across their respective groups of variables.

Van den Wollenburg (1977) proposed “redundancy analysis” as an approach to solving this problem. A copy of his paper is available in the statistical handouts section at the course website. At least some of it may make sense to you!

He states on page 209,

Given two sets of variables  $X$  and  $Y$  standardized to zero mean and unit variance, we seek a variate  $\xi = Xw$  with unit variance such that the sum of squared correlations of the  $Y$  variables with that variate is maximal, and a variate  $\zeta = Yv$  for which the same holds in the reverse direction. The correla-

The linear weights  $w$  and  $v$  that accomplish this are described near the bottom of page 210 as

first variates. In other words, the vectors  $w_j$  and  $v_j$  satisfying the above restrictions are proportional to the  $j^{\text{th}}$  eigenvectors of the characteristic equations

$$(19) \quad \begin{aligned} (R_{xy}R_{yx} - \mu R_{xx})w &= 0, \\ (R_{yx}R_{xy} - \nu R_{yy})v &= 0. \end{aligned}$$

If we premultiply the first equation by  $R_{xx}^{-1}$ , we obtain (note that Van Den Wollenburg does not follow our convention of putting matrices and vectors in boldface).

$$\left( R_{xx}^{-1}R_{xy}R_{yx} - \mu I \right) w = 0, \text{ or, equivalently, } (R_{xx}^{-1}R_{xy}R_{yx})w = \mu w$$

This is an equation of the form  $Aw = \mu w$  where  $A = R_{xx}^{-1}R_{xy}R_{yx}$ . Hence, each of the  $w_j$  producing “redundancy variates” on the  $X$  side corresponds to an eigenvector of  $R_{xx}^{-1}R_{xy}R_{yx}$ . These weights will need to be rescaled to produce variables with unit variances. (See my comments below in the context of the canonical correlation weights.) In a similar way, we can see that each of the  $v_j$  producing redundancy variates on the  $Y$  side is proportional to an eigenvector of  $R_{yy}^{-1}R_{yx}R_{xy}$ . By examining the linear weights in the  $w_j$  and  $v_j$ , we can interpret the “redundancy variates” in much the same way as canonical variates.

Here is what I want you to do. Van Den Wollenburg gives a small numerical example. On page 214, he gives the correlation matrix for 8 variables (4  $x$  and 4  $y$ ) in an  $8 \times 8$  matrix. The  $x$  variable correlations are in the upper left corner, the  $y$  variable intercorrelations in the lower right. So  $R_{yx}$  is in the lower left corner. **Note:** reading the matrix is made slightly confusing by the fact that the 1’s were left out of the correlation matrix. Be careful!

- a. Enter  $R_{xx}$ ,  $R_{yy}$ , and  $R_{yx}$ , then generate  $R_{xy}$  and reproduce all the results on page 215, using R. If you forget how to get the canonical weights for canonical

variates, you can get vectors that are proportional to them from equations 2 and 3 in the article! Remember, however, that the norm is to rescale these vectors so that the resulting canonical variates have unit variance. Hence, for example, if  $W$  contains eigenvectors of  $R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}$ , then the canonical variate weights are  $W^* = W \text{Diag}^{-1/2}(W'R_{xx}W)$ .

- b. By the time you finish (a) above, you will have learned a lot. So write an R function that returns the canonical weights and canonical correlations for the  $X$  and  $Y$  sets if you input  $R_{xx}$ ,  $R_{yy}$ , and  $R_{yx}$ . You can use the **cca** function library to check out your routine.
- c. Then, write your own routine to return the redundancy weights for the same input as in part (b) above.
- d. Finally, compute the canonical correlation and redundancy weights for the work satisfaction data. Compare the first two redundancy variates with the first two canonical variates by examining the weights for each.

### 3. Discriminant Analysis

The data set *discrimdata.sav* is available on the website. This data set has 100 observations on 3 variables, and the individuals are divided into 2 groups coded 1 and 2 with the Group variable in the data file. Using discriminant analysis (assume prior probabilities of .50 for group membership), find a linear function for discriminating between the two groups. What I am looking for is a function (with an intercept) of the form

$$D = a_1V_1 + a_2V_2 + a_3V_3 + a_0 \quad (1)$$

that will assign a person to group 1 if the score is greater than zero and group 2 if the score is less than zero. Look at the various discriminant functions output by SPSS, and look at the way people are assigned to groups, and see if you can come up with a function of the type I am looking for, with a cutoff point of zero, that agrees with the classifications SPSS made. (Hint. Look at the functions evaluated at the group centroid.

Draw a line between the two group centroids in the direction of the function. The cutoff is halfway between the two centroids.)

Is this function unique? Justify your answer (briefly, as this is not very profound). (If the function is not unique, then various computer programs may output different versions of the function.)

Look at the classification matrix output by SPSS. What percentage of the time did SPSS misclassify the people who were really in group 1? What percentage of the time did SPSS misclassify the people who were really in group 2?

Now let me tell you how I generated the data in problem number 2. I used a random number generator to generate 100 observations from the trivariate normal distribution, with mean vector  $\mathbf{0}$  and covariance matrix

$$\begin{bmatrix} 1 & & \\ .4 & 1 & \\ .5 & .6 & 1 \end{bmatrix}$$

Then, after obtaining the 100 observations, I divided them into two groups with the following rule: A single dimension differentiates perfectly between the two groups. A person is in group 2 if and only if  $2V_3 - V_1 - V_2 - 1$  is greater than zero. Note that, of course, the same linear function perfectly discriminates between the two groups in this sample of size 100. Confirm this for yourself by computing the above function and examining its values. Include this in your data file.

So—why did linear discriminant analysis fail to find the simple linear function that allows perfect discrimination between groups? The function exists, and since it works perfectly in all samples and all populations, it will also cross-validate perfectly.

(Strong hint: Note that I derived my rule by simply writing it, without any reference to the population mean and covariance structure for individual groups. Classical discriminant analysis assumes something about each *separate* population of scores, i.e., the population of scores in Group 1 and those in Group 2. Is this assumption incompatible in some sense with my rule? Is this assumption realistic with real world data? Looking at the present example, could this be important?)

#### 4. GLM – ANOVA

a. Suppose you have a  $3 \times 3$  fixed effects ANOVA with 3 observations per cell. The cell means are

$$\mathbf{U} = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \mu_{1,3} \\ \mu_{2,1} & \mu_{2,2} & \mu_{2,3} \\ \mu_{3,1} & \mu_{3,2} & \mu_{3,3} \end{bmatrix}$$

Suppose you stack the means in three rows of  $\mathbf{U}$  into a single column, i.e.,  $\boldsymbol{\mu} = \text{Vec}_r(\mathbf{U})$ . Write the hypothesis matrix  $\mathbf{H}'$  for expressing

- the null hypothesis of no row effects and
- the null hypothesis of no interaction effects in the form

$$\mathbf{H}'\boldsymbol{\mu} = \mathbf{0}$$

Show both  $\mathbf{H}'$  and  $\boldsymbol{\mu}$  explicitly in your answers.

#### 5. GLM - MANOVA

You have data representing two variables on 3 groups in the file “*Manova2.sav*”.

a. Perform 1-way univariate ANOVAs on the variables  $X$  and  $Y$ . Is either one significant? Are the covariance matrices significantly different?

b. The Tukey test is more powerful than the ANOVA  $F$ -test for detecting pairwise mean differences. Try computing the Tukey tests first on  $X$ , then on  $Y$ . Any significant differences?

c. Perform a 1-Way MANOVA. Are the 3 groups significantly different?

d. Using a technique we discussed in the course, find a linear combination of  $X$  and  $Y$  that is maximally different (i.e., discriminates maximally) between the groups, and compute scores on this linear combination. Do a 1-Way ANOVA on these new scores. How does the  $p$ -value compare to the ones reported by MANOVA in its “multivariate tests” output? Any comments?