Multivariate Methods in Finance and Marketing

Prof. Dr. Joachim Grammig and M. Sc. Miriam Sperl
Winter Term 08/09

Case Study in Discriminant Analysis: Analysis of Butter Customers

The file butter.txt contains data on the importance of spreadability and durability of butter evaluated by 12 customers of butter brand A and 12 customers of butter brand B on a 7-ary rating scale. The following variables are available:

Variable	Explanation
BRAND	categorial variable with A (B) for customer of brand A (B)
SPREADABILITY	metric variable with scores 1 to 7 for importance of spreadability
DURABILITY	metric variable with scores 1 to 7 for importance of durability

Use the online documentation http://support.sas.com/onlinedoc/913/docMainpage.jsp as help for the development of your SAS program.

1. Theoretical Tasks in Discriminant Analysis

1.1 What are the primary goals of discriminant analysis in general? And for which purpose do we conduct such an analysis in the case study?

1.2 Write down the canonical discriminant function for the case study and briefly explain the components of the model.

1.3 Name two desirable discrimination properties and explain how they are taken into account in the estimation procedure of the canonical discriminant function.

1.4 To solve the discriminant analysis's underlying optimization problem, the discriminant criterion $\Gamma = SS_b/SS_w$ can be expressed in matrix notation. Given the non-normalized discriminant function

$$Y_{gi} = v_1 X_{1gi} + v_2 X_{2gi}$$

show that for $g = A, B$ the sum of squares within groups

$$SS_w = \sum_{g=1}^{G} \sum_{i=1}^{I_g} (Y_{gi} - \bar{Y}_g)^2$$

can be written in matrix notation as follows

$$SS_w = \mathbf{v}'\mathbf{W}\mathbf{v}$$

with a vector of non-normalized discriminant coefficients

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

and a matrix of sum of squares and cross products within groups of the independent variables

$$\mathbf{W} = \begin{bmatrix} \sum_{g=1}^{G}\sum_{i=1}^{I_g}(X_{1gi} - \bar{X}_{1g})^2 & \sum_{g=1}^{G}\sum_{i=1}^{I_g}(X_{1gi} - \bar{X}_{1g})(X_{2gi} - \bar{X}_{2g}) \\ \sum_{g=1}^{G}\sum_{i=1}^{I_g}(X_{2gi} - \bar{X}_{2g})(X_{1gi} - \bar{X}_{1g}) & \sum_{g=1}^{G}\sum_{i=1}^{I_g}(X_{2gi} - \bar{X}_{2g})^2 \end{bmatrix}.$$

1.5 Explain the meaning of eigenvalue(s) and eigenvector(s) for the estimation of the optimal discriminant function.

1.6 Discuss and interpret in detail the relevant SAS output of the case study which is provided in the file discrim_case_study_output.pdf. The following questions may guide your analysis:

- What about descriptive statistics? $\rightarrow (1) - (4)$
- What about estimation results? $\rightarrow (7)$
- What about the classification check? $\rightarrow (9)$
- What about the discriminant criterion check? $\rightarrow (6), (7)$
- What about the independent variables check? $\rightarrow (5)$
- What about the classification of new observations? $\rightarrow (8)$

1.7 A new customer evaluates the importance of spreadability with score 5 and that of durability with score 3. Use the SAS output to classify this customer into either brand A or brand B customer group and explain your procedure.

2. Practical Tasks in Discriminant Analysis

2.1 Download the dataset butter.txt from Ilias and save it to the hard disk.

2.2 Open SAS, assign a library and import the data to that library.

2.3 Compute scatterplots of the variables SPREADABILITY versus DURABILITY for each category of the variable BRAND by using a `where` statement within `proc gplot`. Control for the values of the axes with the help of the `haxis` and `vaxis` options within the `plot` statement to ensure the comparability of the graphs. Interpret the results!

2.4 Use `proc univariate` to compute histograms of the variables SPREADABILITY and DURABILITY for each category of the variable BRAND by using a `where` statement. Make sure that you choose meaningful midpoints for the bars of the histograms with the help of the `midpoints` option within the `histogram` statement. Interpret the results!

2.5 Conduct a canonical discriminant analysis of the data by using the `can` option within `proc discrim`. Note that a successful run of `proc discrim` requires the specification of a `class` and a `var` statement.

2.6 Further include the following options within `proc discrim` and figure out which output they produce: `simple`, `bsscp`, `wsscp`, `tsscp`, `anova` and `manova`.

2.7 Go through the SAS output produced and find out which parts are relevant for interpretation. A comparison with the output delivered in discrim_case_study_output.pdf may me helpful.