2003 BELTWIDE COTTON CONFERENCES, NASHVILLE, TN – JANUARY 6-10

COMPUTATIONAL EXPLORATORY DATA ANALYSIS IN COTTON SPINNING

Maria Elisabete Cabeço Silva and Antonio Alberto Cabeço Silva
School of Engineering
University of Minho
Guimarães, Portugal

Abstract

Discriminant analysis is the statistical technique that is most commonly used to solve complex problems. Its use is appropriate when you can classify data into two or more groups, and when you want to find one or more functions of quantitative measurements that can help you discriminate among the known groups. The objective of the analysis is to provide a method for predicting which group a new case is most likely to fall into, or to obtain a small number of useful predictor variables. In this work, the purpose is to classify cotton bales into well define groups or categories based on a training set of similar samples for grading cotton blends. A new algorithm has been implemented using discriminant analysis and its advantages in quality design of cotton blends are demonstrated.

Introduction

Discriminant analysis is used to determine which variables discriminate between two or more naturally occurring groups. Discriminant analysis is capable to handling either two groups or multiple groups (three or more). When three or more classifications are identified, it is known as multiple discriminant analysis. The concept of discriminant analysis involves forming linear combinations of independent (predictor) variables, which become the basis for group classifications.

Discriminant analysis is appropriate for testing the hypothesis that the group means for two or more groups are equal. Each independent variable is multiplied by its corresponding weight, then the products are added together, which results in a single composite discriminant score for each individual in the analysis.

Averaging the scores derives a group centroid. If the analysis involves two groups there are two centroids; in three groups there are three centroids; and so on. Comparing the centroids shows the distance of the groups along the dimension you are testing.

Applying and interpreting discriminant analysis is similar to regression analysis, where a linear combination of measurements for two or more independent variables describes or predicts the behavior of a single dependent variable. The most significant difference is that you use discriminant analysis for problems where the dependent variable is categorical versus regression where the dependent variable is metric.

The objectives for applying discriminant analysis include:

- determining if there are statistically significant differences among two or more groups,
- establishing procedures for classifying units into groups,
- determining which independent variables account for most of the difference in two or more groups.

Discriminant analysis involves three steps:

- derivation,
- validation, and
- interpretation.

In derivation, we must first choose the variables, test the validity of the discriminant function, determine a computational method, and assess the level of significance.

Validation involves determining the reason for developing classification matrices, deciding how well the groups are classified into statistical groups, determining the criterion against which each individual score is judged, constructing the classification matrices, and interpreting the discriminant functions to determine the accuracy of their classification.

Interpretation involves examining the discriminant functions to determine the importance of each independent variable in discriminating between the groups, then examining the group means for each important variable to outline the differences between the groups.

The analysis assumes that the variables are drawn from populations that have multivariate normal distributions and that the variables have equal variances.

Computationally, discriminant function analysis is very similar to analysis of variance.

To summarize the discussion so far, the basic idea underlying discriminant function analysis is to determine whether groups differ with regard to the mean of a variable, and then to use that variable to predict group membership.

The discriminant function problem can be rephrased as a one-way analysis of variance (ANOVA) problem. Specifically, one can ask whether or not two or more groups are significantly different from each other with respect to the mean of a particular variable.

In the case of a single variable, the final significance test of whether or not a variable discriminates between groups is the *F* test. *F* is essentially computed as the ratio of the between-groups variance in the data over the pooled (average) within-group variance. If the between-group variance is significantly larger then there must be significant differences between means.

Usually, one includes several variables in a study in order to see which one(s) contribute to the discrimination between groups. In that case, we have a matrix of total variances and covariances; likewise, we have a matrix of pooled within-group variances and covariances. We can compare those two matrices via multivariate *F* tests in order to determine whether or not there are any significant differences (with regard to all variables) between groups. This procedure is identical to multivariate analysis of variance.

When interpreting multiple discriminant functions, which arise from analyses with more than two groups and more than one variable, one would first test the different functions for statistical significance, and only consider the significant functions for further examination.

Next, we would look at the standardized *b* coefficients for each variable for each significant function. The larger the standardized *b* coefficient, the larger is the respective variable's unique contribution to the discrimination specified by the respective discriminant function. In order to derive substantive "meaningful" labels for the discriminant functions, one can also examine the factor structure matrix with the correlations between the variables and the discriminant functions.

Finally, we would look at the means for the significant discriminant functions in order to determine between which groups the respective functions seem to discriminate.

Linear Discriminant Model

For a set of p variables X_1, X_2, …, X_p, the general model is:

$$Z_i = \sum_{j=1}^{p} a_{ij} X_j$$

Where, the X_{js} are the original variables and the a_{ijs} are the discriminant function coefficients.

The principles to find the discriminant functions are:

The first discriminant function is that which maximizes the differences between groups compared to the differences within group which is equivalent to maximizing *F* in a one-way ANOVA.

$$F(Z) = \frac{MS_B(Z)}{MS_W(Z)},$$
$$Z_1 = \max\{F(Z)\}$$

The second discriminant function is that which maximizes the differences between groups compared to the differences within groups unaccounted for by Z_1 which is equivalent to maximizing F in a one-way ANOVA given the constraint that Z_1, Z_2 are uncorrelated.

$$F(Z) = \frac{MS_B(Z)}{MS_W(Z)},$$

$$Z_2 = \max\{F(Z) \,|\, r_{Z_1, Z_2} = 0\}$$

The total (T) SSCP matrix (based on p variables X_1, X_2,…, X_p) in a sample of objects belonging to m groups G_1, G_2,…, G_m with sizes n_1, n_2, …, n_m can be partitioned into within-groups (W) and between-groups (B) SSCP matrices:

$$T = B + W$$

where:

$$t_{rc} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ijr} - \overline{x}_r)(x_{ijc} - \overline{x}_c)$$

$$w_{rc} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ijr} - \overline{x}_{jr})(x_{ijc} - \overline{x}_{jc})$$

and:

x_{ijk} Value of variable X_k for ith observation in group j,

$\overline{x}_{jk}$ Mean of variable X_k for group j

$\overline{x}_k$ Overall mean of variable X_k

t_{rc}, w_{rc} Element in row r and column c of total (T, t) and within (W, w) SSCP

Analytic procedures to find discriminant functions:

- Calculate total (T), within (W) and between (W) SSCPs

$$T = B + W$$

- Determine eigenvalues and eigenvectors of the product W^{-1} B.

$$\lambda(\mathbf{B}^{-1}\mathbf{W}) = (\lambda_1, \lambda_2, \mathrm{K}, \lambda_p)$$

- λ_i is ratio of between to within SSs for the ith discriminant function Z,

$$\lambda_i = \frac{SS_B(Z_i)}{SS_W(Z_i)}$$

and the elements of the corresponding eigenvectors are the discriminant function coefficients.

$$\xi_i(\mathbf{B}^{-1}\mathbf{W}) = (a_{i1}, a_{i2}, \mathrm{K}, a_{ip})$$

Textile Approach

Traditionally the blends were defined by the grouping of different cottons, leading in account the colour and length parameters, looking for always homogeneity of micronaire. The classification of these cottons was initially done using an algorithm developed in the Excel.

It was verified, however that these blends were not homogeneous of a point of view of regularity and constancy of properties related with its length and its uniformity, as well as of the resistance of the fibres.

In this study we present the information's of the raw materials that are organized in knowledge bases, distributed by databases.

One became, therefore, important to look to an accurate method or algorithm that allowed according to make the grouping of the different bales cottons following priority: micronaire, span length 2.5, strength, uniformity ratio, yellow degree, and reflectance.

The database is composed by the parameters showed in the Table 1 that has been evaluated by the HVI systems and representing 29 bales of different african cottons.

Using the conventional approach we defined 4 different blends, M1, M2, M3, M4 (the groups).

Results and Discussion

The most common application of discriminant function analysis is to include many measures in order to determine the ones that discriminate between groups.

A common result that one looks at in order to determine how well the current classification functions predict group membership of cases is the classification matrix.

A common misinterpretation of the results of the discriminant analysis is to take statistical significance levels at face value. By nature, the procedures will capitalize on chance because they "pick and choose" the variables to be included in the model so as to yield maximum discrimination.

The classification matrix shows the number of cases that were correctly and and those that were misclassified (Table2), the linear discriminant function for groups (Table 3), and the misclassified observations (Table 4) in the first phase.

In this phase we can see that only 79.3 % of the blends are correctly classified. So, we correct this classification by introduction the new data information (predicted groups).

The Tables 5, 6, and 7 present the results of the second phase of the discriminant analysis. We can verify now, that 93.1 % of the observations (bales) are correctly classified; only the observations 11 and 27 are misclassified.

Finally, in the third step, we meet the goal of the correct classification (100 %) (Table 8) The Table 9 shows the coefficients of the linear discriminant function for groups.

Conclusions

We build a "model" of how we can best predict to which group a case belongs. The term "in the model" in order to refer to variables that are included in the prediction of group membership, and we will refer to variables as being "not in the model" if they are not included.

Those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

Discriminant Analysis is a very useful tool for detecting the variables that allow the researcher to discriminate between different (naturally occurring) groups, and for classifying cases into different groups with a better than chance accuracy, as we demonstrated in data analysis of the raw cotton blends.

References

Baxter, M., *Exploratory Multivariate Analysis in Archaeology*, pp. 167-170, Edinburgh University Press, Edinburgh, 1994.

Brandt S., *Statistical and Computation Methods in Data Analysis*', Northholland, Publishing Company, 5th edition, 1989.

Cabeço Silva, M. E., Cabeço Silva, A. A., *Multivariate Analysis in Quality Design of Cotton Blends*, 1996 Beltwide Cotton Conferences, Vol. 2, p. 1467 - 1471, Nashville, USA, 1996, January.

Cabeço Silva, M. E., Marques, M. J. A., Cabeço Silva, A. A., *Involving Statistical Factor Analysis for Disposable Surgical Clothing Properties Prediction*, EPMESC-Computational Methods in Engineering and Science, ELSEVIER SCIENCE Ltd, Edited by J. Bento, E. Arantes e Oliveira, E. Pereira, Vol. 2, p. 1085-1093, ISBN 008043570 X HC, Macau, 1999.

Cabeço Silva, M. E., *Análise Exploratória de Dados em Sistemas HVI de Metrologia Têxtil*, X Congreso Latino-Iberoamericano de Investigación de Operaciones y Sistemas - X CLAIO, México, D. F., Edição em CD, México, 2000.

Cooley, W. W., & Lohnes, P. R., *Multivariate Data Analysis*, New York: John Wiley & Sons, 1971.

Darlington, R. B., *Regression and Linear Models*, New York: McGraw-Hill, 1990.

Draper, N. R., & Smith, H., *Applied Regression Analysis*, New York: Wiley, 1981.

Hoaglin, D. C., Mosteller, F. Tuckey, J. W., *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York, 1983.

Huberty, C. J., *Applied Discriminant Analysis*, New York: Wiley and Sons, 1994.

Huberty, C. J., & Wisenbaker, J. M., *Discriminant Analysis: Potential Improvements in Typical Practice*, In B. Thompson (Ed.), Advances in social science methodology, (Vol. 2, pp. 43-70), Greenwich, CT: JAI Press, 1992.

Jennrich, R. I., *Stepwise Discriminant Analysis*, In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), Statistical methods for digital computers, Vol. 3, pp. 76-96, New York, Wiley, 1977.

Press, S. J., & Wilson, S., *Choosing between Logistic Regression and Discriminant Analysis*, Journal of the American Statistical Association, *73*, 699-705, 1978.

Thompson, B., *Why Multivariate Methods are Usually Vital in Research: Some Basic Concepts*, Southwestern Society for Research in Human Development, Austin, TX. (ERIC Document Reproduction Service No. ED 367 687), 1994.

Tuckey, J. W., *Exploratory Data Analysis*, Addison Wesley, 1977.

Table 1. Fibre Properties.

PROPERTY/ VARIABLE	**UNITY**
Micronaire Index (Mic)	•g / "
Upper Half Mean Length (UHML)	mm
Span Length 50 (SL 50)	mm
Span Length 2.5 (SL 2.5)	mm
Uniformity Index (UI)	-
Uniformity Ratio (UR)	-
Tenacity (ST)	cN/tex
Elongation (EL)	%
Reflectance (Rd)	-
Yellow Degree (+b)	-
Colour Grade (CGrd)	-
Leaf (LF)	-

Table 2. Classification (Phase 1).

GROUP	**M1**	**M2**	**M3**	**M4**
M1	3	0	0	0
M2	0	4	0	4
M3	0	0	12	1
M4	0	0	1	4
TOTAL	3	4	13	9
N Correct	3	4	12	4
Proportion	1.000	1.000	0.923	0.444

N = 29, N Correct = 23, Prop. Correct = 0.793

Table 3. Linear Discriminant Function for Group (Phase 1).

	M1	M2	M3	M4
Constant	- 6135.0	- 6278.7	- 6463.5	- 6357.0
Mic	- 73.9	- 76.8	- 80.9	- 78.9
SL 2.5	275.5	278.4	280.0	278.4
ST	0.4	1.6	1.7	1.8
UR	37.6	39.1	41.1	40.7
+b	133.2	137.5	136.5	138.6
Rd	21.4	20.4	21.3	20.3

Table 4. Misclassified Observations (Phase 1).

Observation	True Group	Pred. Group	Group	Square Distance	Probability
1	M4	M2	M1	21.094	0.000
			M2	3.680	0.787
			M3	18.487	0.000
			M4	6.300	0.212
3	M4	M2	M1	12.409	0.003
			M2	1.861	0.565
			M3	10.122	0.009
			M4	2.443	0.423
5	M4	M2	M1	10.170	0.010
			M2	2.320	0.518
			M3	3.762	0.252
			M4	4.031	0.220
8	M4	M2	M1	21.094	0.000
			M2	3.680	0.787
			M3	18.487	0.000
			M4	6.300	0.212
12	M4	M3	M1	15.088	0.008
			M2	10.466	0.079
			M3	6.348	0.618
			M4	7.824	0.295
15	M3	M4	M1	20.371	0.000
			M2	8.952	0.028
			M3	4.154	0.308
			M4	2.614	0.664

Table 5. Classification (Phase 2).

GROUP	M1	M2	M3	M4
M1	3	0	0	0
M2	0	8	0	0
M3	0	0	11	0
M4	0	0	2	5
TOTAL	3	8	13	5
N Correct	3	8	11	5
Proportion	1.000	1.000	0.846	1.000

N = 29, N Correct = 27, Prop. Correct = 0.931

Table 6. Linear Discriminant Function for Group (Phase 2).

	M1	M2	M3	M4
Constant	- 8756.2	- 8948.5	- 9549.2	- 9654.5
Mic	- 521.9	- 525.1	- 563.3	- 570.8
SL 2.5	158.2	160.8	154.1	150.0
ST	43.5	44.9	48.1	49.1
UR	225.7	228.4	244.0	249.0
+b	73.5	79.3	72.4	74.6
Rd	36.4	35.2	37.6	37.0

Table 7. Misclassified Observations (Phase 2).

Observation	True Group	Pred. Group	Group	Square Distance	Probability
11	M3	M4	M1	88.210	0.000
			M2	62.920	0.000
			M3	13.260	0.404
			M4	12.480	0.596
27	M3	M4	M1	82.822	0.000
			M2	63.392	0.000
			M3	10.161	0.151
			M4	6.700	0.849

Table 8. Summary of Classification (Final Phase).

GROUP	M1	M2	M3	M4
M1	3	0	0	0
M2	0	8	0	0
M3	0	0	11	0
M4	0	0	0	7
TOTAL	3	8	11	7
N Correct	3	8	11	7
Proportion	1.000	1.000	1.000	1.000

N = 29, N Correct = 29, Prop. Correct = 1.000

Table 9. Linear Discriminant Function for Group (Final Phase).

	M1	M2	M3	M4
Constant	- 14551	- 15102	- 15905	- 16558
Mic	- 1100	- 1120	- 1167	- 1198
SL 2.5	171	175	169	169
ST	103	106	110	113
UR	452	461	480	493
+b	266	278	274	285
Rd	31	30	32	32