

Fuzzy linear discriminant analysis

Linear discriminant analysis (Fisher, 1936) is usually performed to investigate differences among multivariate classes, to determine which attributes discriminate between the classes, and to determine the most parsimonious way to distinguish among classes. Similar to analysis of variance for single attribute, we could compute the within-class variance to evaluate the dispersion within class, and between-class variance to examine the differences between the classes.

Discriminant analysis for hard classifications has been applied in soil science for more than 60 years. The first of such application is by Cox and Martin (1937) where they use discriminant analysis to find out whether some chemical properties give significant information on the presence of Azotobacter in soil. Hughes and Lindley (1955) illustrated the application in classification of some Snowdonian soil. Oertel (1961), Norris and Loveday (1971), Webster and Burrough (1974) used this method to allocate soil profiles into existing soil classes.

Bell et al. (1992) related eight landscape parameters to soil drainage classes, and use discriminant analysis to allocate new observations to the classes. Henderson and Ragg (1980) employed a multivariate logistic method of discriminant analysis to assess the usefulness of some soil morphological properties for distinguishing between four taxonomic units.

The theory is readily accessible in Webster and Oliver (1990). We generalize the theory to a fuzzy linear discriminant analysis, which considers the membership of each individual to each of the classes.

Consider the data matrix \mathbf{X} (elements x_{il} , $i = 1, \dots, n$; $l = 1, \dots, p$) classified by fuzzy k means to give membership matrix \mathbf{M} (elements m_{ij} , $i = 1, \dots, n$; $j = 1, \dots, k$) with the centroid matrix \mathbf{C} (elements c_{jl} , $j = 1, \dots, k$; $l = 1, \dots, p$), where n is the number of sample, p is the number of attributes, and k is the number of classes.

The sums of squares and products (SSP) within-classes matrix \mathbf{W}_f , or also called within-classes fuzzy scatter matrix (Bezdek, 1981):

$$\mathbf{W}_{fk} = \sum_{i=1}^n m_{ij}^{\phi} (\mathbf{x}_i - \mathbf{c}_k)(\mathbf{x}_i - \mathbf{c}_k)^T \quad (1)$$

The elements are:

$$w_{lm} = \sum_{j=1}^k \sum_{i=1}^n m_{ij}^{\phi} (x_{il} - c_{jl})(x_{im} - c_{jm}), \quad \forall l, l = 1, \dots, p; \forall m, m = 1, \dots, p \quad (2)$$

The fuzzy covariance matrix of class k is defined as:

$$\mathbf{C}_{fk} = \frac{\sum_{i=1}^n m_{ij}^{\phi} (\mathbf{x}_i - \mathbf{c}_k)(\mathbf{x}_i - \mathbf{c}_k)^T}{\sum_{i=1}^n m_{ij}^{\phi}} \quad (3)$$

The sum of squares product (SSP) between-classes matrix \mathbf{B}_f

$$\mathbf{B}_{fk} = \left(\sum_{i=1}^n m_i^{\phi} \right) (\mathbf{c}_k - \bar{\mathbf{x}})(\mathbf{c}_k - \bar{\mathbf{x}})^T \quad (4)$$

which has elements:

$$b_{lm} = \left(\sum_{j=1}^k \sum_{i=1}^n m_{ij}^{\phi} \right) (c_{jl} - \bar{x}_l)(c_{jm} - \bar{x}_m), \quad \forall l, l = 1, \dots, p; \forall m, m = 1, \dots, p \quad (5)$$

where \bar{x}_l and \bar{x}_m are the overall means of the l th and m th variates.

When the membership is hard (m equals 0 or 1) Equations (1) and (5) reduce to the conventional \mathbf{W} , and \mathbf{B} matrices (Webster and Oliver, 1990).

The total SSP matrix \mathbf{T}_f is calculated as

$$\mathbf{T}_f = \mathbf{B}_f + \mathbf{W}_f \quad (6)$$

The ratio of the determinant of the within-classes to the total SSP matrix is the Wilks' Λ (Wilks, 1932; Webster, 1970), similarly we have fuzzy Wilks' Λ :

$$\Lambda = \frac{|\mathbf{W}_f|}{|\mathbf{T}_f|} \quad (7)$$

The fuzzy Wilks' Λ is a measure of the difference between classes. The value Λ varies from 0 to 1, with 0 suggesting class means differ (and the attributes differentiate the classes more), and 1 suggesting all class means are the same. Mariott (1971) suggested that when plotting k^2L versus k , k^2L will drop steadily from 1 when $k = 1$ class. Identifying the class where there is a sharp decline in k^2L value indicates the optimum number of classes. However, since there is a fuzzy membership embedded within matrices \mathbf{W}_f and \mathbf{T}_f , the value k^2L could be greater than 1

Similar to principal component analysis, in discriminant analysis one computes $k-1$ or p discriminant (canonical) functions. The first function maximizes the differences between the classes. The successive functions will be orthogonal or independent to other functions, hence their contributions to the discrimination between classes will not overlap. These functions or canonical variates are calculated

from the eigenvalues and eigenvectors of matrix $\mathbf{W}_f^{-1} \mathbf{B}_f$. The solution is to find the eigenvalues (latent roots) λ :

$$|\mathbf{W}_f^{-1} \mathbf{B}_f - \lambda \mathbf{I}| = \mathbf{0} \quad (8)$$

which can be solved by:

$$(\mathbf{W}_f^{-1} \mathbf{B}_f - \lambda \mathbf{I}) \mathbf{e}_i = \mathbf{0} \quad (9)$$

where \mathbf{e}_i is the i columns of eigenvectors.

The projection of a data vector \mathbf{x} on the i th canonical axis is computed as:

$$\mathbf{z}_i = \mathbf{x}^T \mathbf{e}_i. \quad (10)$$

Similarly the class means centroids of the j th class are projected as:

$$\mathbf{z}_i = \mathbf{c}_j^T \mathbf{e}_i. \quad (11)$$

References

- Bell, J.C., Cunningham, R.L., Havens, M.W., 1992. Calibration and validation of a soil-landscape model for predicting soil drainage class. *Soil Sci. Soc. Am. J.* 56, 1860-1866.
- Bezdek, J.C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Ilenum press, New York.
- Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179-188.
- Henderson, R., Ragg, J.M., 1980. A reappraisal of soil mapping in an area of Southern Scotland. part II The Usefulness of some morphological properties and of a discriminant analysis in distinguishing between the dominant taxa of four mapping units. *J. Soil Sci.* 31, 573-580.
- Hughes, R.E., Lindley, D.V. 1955. Application of biometric methods to problems classification in ecology. *Nature, London* 175, 806-807.
- Mariott, F.H.C., 1971. Practical problems in a method of cluster analysis. *Biometrics* 27, 501-514.
- Oertel, A.C., 1961. Chemical discrimination of terra rossas and redzinas. *J. soil Sci.* 12, 111-118.
- Webster, R., 1971. Wilks' Criterion: a measure for comparing the value of general purpose soil classifications. *J. Soil Sci.* 22, 254-260.
- Webster, R., Burrough, P.A., 1974. Multiple discriminant analysis in soil survey. *J. Soil Sci.* 25, 120-134.
- Webster, R., Oliver, M.A., 1990. *Statistical Methods in Soil and Land Resource Survey*. Oxford University Press, Oxford.