Discriminant Analysis

Janette Walde

janette.walde@uibk.ac.at

Department of Statistics
University of Innsbruck

Outline I

Outline II

Basics

- Discriminant Analysis (DA) is used to predict group membership from a set of metric predictors (independent variables X).
- How can the variables be *linearly* combined to best classify a subject into a group?
- DA is concerned with testing how well (or how poorly) the observation units are classified.
- DA is interested in exactly how the groups are differentiated not just that they are significantly different (as in MANOVA).

Problems

- Using multi-temporal satellite imagery to characterize forest wildlife habitat: The case of ruffed grouse.(Forest Ecology and Management, Vol. 260 (9), 2010, pp 1539-1547).
- Potential isotopic and chemical markers for characterizing organic fruits: (Food Chemistry, Vol. 125 (3), 2011, pp 1072-1082). Combining isotopic and chemical−physical markers (pH, fruit weight, etc.) a good discrimination between organic and conventional fruits of different species (oranges, clementines, strawberries, peaches produced in Italy between 2006 and 2008) was achieved.

Questions I

The primary goal is to find a dimension(s) that groups differ on and create classification functions.

- Can group membership be accurately predicted by a set of independent variables?
- Along how many dimensions do groups differ reliably?
  - DA creates discriminant functions and each is assessed for significance.
  - Usually the first one or two discriminant functions are sufficient.

Questions II

  * Each DA function is orthogonal to the previous and the number of dimensions (discriminant functions) is equal to either the $G - 1$ or J, which ever is smaller.

- Are the discriminant functions interpretable or meaningful?
  * Does a DA function differentiate between groups in some meaningful way?
  * How do the DA functions correlate with each predictor?
- Can we classify new (unclassified) subjects into groups? Given the classification functions how accurate are we? And when we are inaccurate is there some pattern to the misclassification?

Questions III

- What is the strength of association between group membership and the predictors?
- Which predictors are most important in predicting group membership?
- Can we predict group membership after removing the effects of one or more covariates?
- Can we use DA to estimate population parameters?

Modeling approach

DA involves deriving a variate, the linear combination of two (or more) independent variables that will discriminate best between a-priori defined groups.
Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance relative to the within-group variance. The linear combination for a discriminant analysis, also known as the discriminant function, is derived from an equation that takes the following form:

$$Z_{ik} = b_{0i} + b_{1i}X_{1k} + ... + b_{Ji}X_{Jk} \quad (1)$$

Z_{ik} ... discriminant score of discriminant function i for object k

Modeling approach, cont.

$$Z_{ik} = b_{0i} + b_{1i}X_{1k} + ... + b_{Ji}X_{Jk}$$

Z_{ik} ... discriminant score of discriminant function i for object k, $i = 1, ..., G - 1$

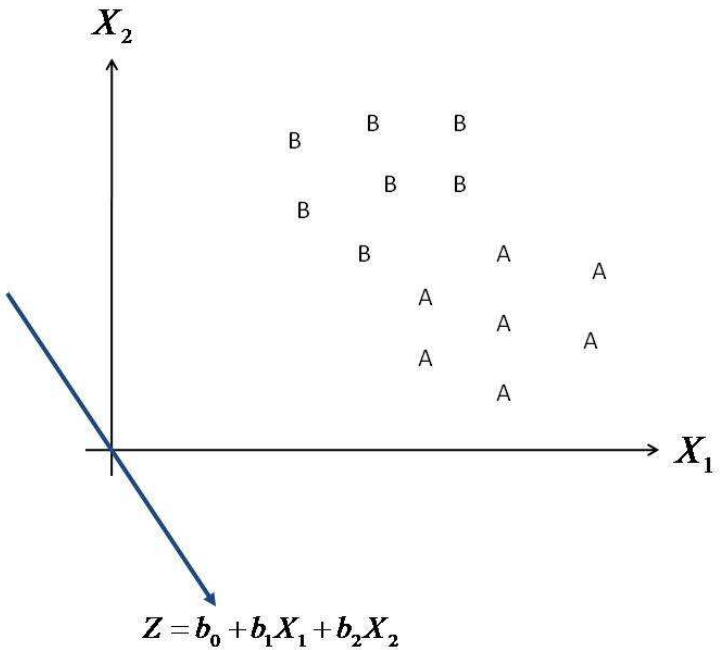
X_{jk} ... independent variable j for object $k, j = 1, 2, ..., J$

b_{ji} ... discriminant weight for independent variable j and discriminant function i

b_{0i} ... constant of discriminant function i

Note: Different kinds of specifications for DA functions are available.

Geometric representation of the two-group discriminant function



Estimation of the DA function(s)

$$\max_b \gamma = \frac{\sum_{g=1}^{G} I_g \left(\overline{Z}_g - \overline{Z}\right)^2}{\sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(\overline{Z}_{gi} - \overline{Z}_g\right)^2} = \frac{SS_b}{SS_w} \tag{2}$$

γ ... discriminant criteria (eigen value)

I_g ... size of group g

$\overline{Y}_g$ ... mean of group g

Z_{gi} ... ith discriminant value of group g

SS_b ... sum of squared deviations between groups, explained deviation

SS_w ... sum of squared deviations within groups, remaining/unexplained deviations

Computational method

- **Simultaneous estimation** involves computing the discriminant function so that all of the independent variables are considered concurrently.
- **Stepwise estimation** involves entering the independent variables into the discriminant function one at a time on the basis of their discriminating power.
  The Mahalanobis D^2 measure is appropriate for stepwise procedure. It is computed in the in the original space of the predictor variables. The selection rule is to maximize Mahalanobis D^2 between groups. The Mahalanobis D^2 procedure performs a stepwise DA similar to a stepwise regression analysis, designed to develop the best one-variable model, followed by the best two-variable model, and so forth.

More than one discriminant function

The number K of discriminant functions must be both less than the number of groups G and less than the number of independent variables J: $K \leq \min\{G-1, J\}$. For each discriminant function we get an optimal eigenvalue γ_k. The relative eigenvalue of discriminant function k is a measure of the importance of the $k-$th discriminant function:

$$EP_k = \frac{\gamma_k}{\gamma_1 + \gamma_2 + \cdots + \gamma_K} \tag{3}$$

In general two discriminant functions suffice.

Statistical Significance I

- After computing the discriminant functions the level of significance must be assessed. A number of different statistical criteria are available. We discuss the measure of Wilks' lambda that evaluates the statistical significance of the discriminatory power of the discriminant function.

Statistical Significance II

- If **one or more functions** are deemed **not statistically significant**, the discriminant model should be **re-estimated** with the number of functions to be derived limited to the number of significant functions. In this manner, the assessment of predictive accuracy and the interpretation of the discriminant functions will be based only on significant functions.

Statistical Significance, cont.

As γ gives the maximal value of the discriminant criteria, a high value of γ indicates high quality. However, γ has no upper limit. Therefore appropriate transformations of γ are used:

$$\frac{\gamma}{1+\gamma} = \frac{SS_b}{SS_b + SS_w} = \frac{\text{explained variation}}{\text{total variation}} \quad (4)$$

$$\frac{1}{1+\gamma} = \frac{SS_w}{SS_b + SS_w} = \frac{\text{unexplained variation}}{\text{total variation}} \quad (5)$$

$\sqrt{(4)}$... **canonical correlation coefficient** c.
(5) is called **Wilks' Lambda** Λ.

Wilks' Lambda

Wilks' Lambda (Λ) is an inverse quality criterium.
If K DA functions are computed the characteristics γ, c, and Λ are computed separately for each DA function. In order to analyze the dissimilarity of the groups **multivariate Wilks' Lambda** is calculated:

$$\Lambda = \prod_{k=1}^{K} \frac{1}{1+\gamma_k} \tag{6}$$

γ_k denotes the eigen value of the $k-$th DA function.

Wilks' Lambda, Cont.

A suitable transformation of Λ allows a significance test regarding the DA function:

$$\chi^2_B = -\left(N - \frac{J+G}{2} - 1\right) \ln(\Lambda) \tag{7}$$

N ... number of observation units

J ... number of variables

G ... number of groups

Λ ... multivariate Wilks' Lambda

Wilks' Lambda, Cont.

χ^2_B is approximately χ^2-distributed with $J \cdot (G-1)$ degrees of freedom.

Hypothesis:

H_0: The groups are not different from each other.

H_1: There are different groups.

Assumptions:

- Multivariate normal distribution of the X-variables in the groups.
- Identical covariance matrices in the groups.

Wilks' Lambda, Cont.

In order to test whether an additional DA function is necessary given k DA functions are already estimated Wilks' Lambda for the residual discriminant value can be used:

$$\Lambda_k = \prod_{q=k+1}^{K} \frac{1}{1+\gamma_q} \tag{8}$$

χ^2_B (cf. (7)) is again χ^2-distributed with $(J-k)\cdot(G-k-1)$ degrees of freedom. It can be useful not to use all significant DA functions. Generally, 2-3 DA functions are sufficient.

Assumptions of Discriminant Analysis

- Homogeneous within-group variances.
- Multivariate normality within groups.
- Linearity among all pairs of variables.
- No multi-collinearity.
- Prior probabilities.

Assessing Group Membership Prediction Accuracy - Hit Ratio

In order to establish the quality of the DA the hit ratio (HR) is computed. The hit ratio gives the correctly classified observation units divided by the number of observation units. The results are summarized in a classification matrix:

true class membership	predicted class membership	
	Group A	*Group B*
Group A	c_{11}	c_{12}
Group B	c_{21}	c_{22}

$$HR = \frac{c_{11}+c_{22}}{c_{11}+c_{12}+c_{21}+c_{22}}$$

Measuring predictive accuracy relative to chance

- The hit ratio must be compared with the a-priori hit ratio or a random assignment.
- Including a-priori probabilities in the computation.
- The data set is employed to compute the DA functions as well as the classification $\rightarrow$ over-estimation of the hit ratio. Therefore, a **validation set** should be used in order to estimate the hit ratio appropriately.

Comment: As a result one obtains the probability of belonging to a specific class g and therefore also the probability for second best class.

Importance of the Independent Variables

If the discriminant function is statistically significant and the classification accuracy is acceptable, the focus lies on making substantive interpretations of the findings. This process involves examining the discriminant functions to determine the relative importance of each independent variable in discriminating between the groups. Three methods of determining the relative importance have been proposed:

1. Standardized discriminant weights.
2. Discriminant loadings (structure correlation).
3. Partial F-values.

Discriminant weights

This approach examines the sign and magnitude of the **standardized discriminant weight** (discriminant coefficient) assigned to each variable in computing the discriminant functions.
When the sign is ignored, each weight represents the relative contribution of its associated variable to that function. Independent variables with relatively larger weights contribute more to the discriminating power of the function than do variables with smaller weights.

$$b_j^* = b_j \cdot s_j \tag{9}$$

b_j ... discriminant coefficient of variable j

s_j^2 ... within group variance of variable j

Importance of independent variable with more than one DA function

If K DA functions are computed, K standardized discriminant weights are obtained for each variable. Thus, the **mean standardized discriminant weight** for each variable is calculated:

$$\overline{b}_j = \sum_{k=1}^{K} \left| b^*_{jk} \right| \cdot EP_k \tag{10}$$

b^*_{jk} ... standardized discriminant weight for variable j and discriminant function k

EP_k ... Eigenvalue proportion of DA function k

Discriminant Loadings

Discriminant loadings (structure correlations) measure the simple linear correlation between each independent variable and the discriminant function.

The discriminant loadings reflect the variance that the independent variables share with the DA function and can be interpreted as assessing the relative contribution of each independent variable to the DA function.

Partial F values

The relative discriminating power of the independent variables can be interpreted through the use of partial F values. This is accomplished by examining the absolute sizes of the significant F values and ranking them. Large F values indicate greater discriminatory power .

In practice, rankings using the F-values approach are the same as the ranking derived from using discriminant weights, but the F values indicate the associated level of significance for each variable.

Classification functions of R.A. Fisher

$$
\begin{aligned}
F_1 &= f_{01} + f_{11}X_1 + f_{21}X_2 + \ldots + f_{J1}X_J \\
F_2 &= f_{02} + f_{12}X_1 + f_{22}X_2 + \ldots + f_{J2}X_J \\
&\vdots \quad \vdots \quad \vdots \\
F_G &= f_{0G} + f_{1G}X_1 + f_{2G}X_2 + \ldots + f_{JG}X_J
\end{aligned}
$$

Rule: The observation units belong to the group for which the value F_g is maximal.

Note: Do not mix up Fisher's classification functions with the (canonical) DA functions.