Discriminant Analysis, Classification and Regression Trees, and Indicator Species Analysis

Jacob Weverka (with major help from the UMASS Landscape Ecology Lab)

Discriminant Analysis

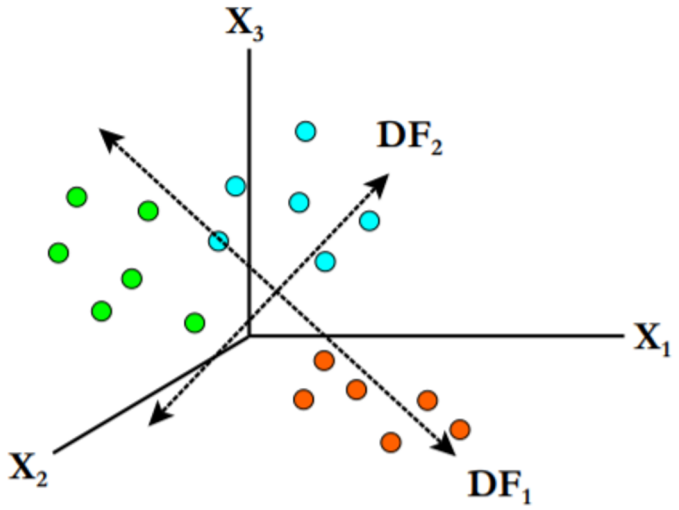
- Identify and describe relationship between multivariate data and a grouping variable
- Use relationship to make predictions
- Can go in either direction
  - $y_{categorical} \sim x_1 + x_2 + \cdots x_n$ (analogous to multiple regression)
  - OR
  - $y_1 + y_2 + \cdots y_n \sim x_{categorical}$ (analogous to ANOVA)

Can I use DA?

- 1 categorical variable and 2 or more discriminating variables (continuous, categorical, or count)
- Groups must be mutually exclusive
- No missing data
- Don’t need equal sample size, but similar size is preferable
- At least 2 samples per group, and at least 2 more samples than number of variables

Canonical Analysis of Discriminance

- Objective: test and describe difference between 2 or more pre-defined groups using 2 or more discriminating variables
- Method: Create linear combinations of discriminating variables that best distinguish between groups (canonical functions)
  - Each sample gets a ‘canonical score’ for each
  - Each group has a centroid of canonical scores
  - Look for separation between groups
- Ideally look for ecological interpretation



Classification

- Use canonical functions to classify samples
- Several parametric and non-parametric methods
- Parametric
  - Linear discriminant functions (LDA) – assumes equal multivariate normal distributions for all groups
  - Quadratic discriminant functions – assumes unequal multivariate normal distribution for all groups
- Non-parametric
  - Kernal – estimate group-specific densities and classify using highest local density
  - K nearest neighbor – classifies by using groups based on which has the nearest neighbor

DA Assumptions

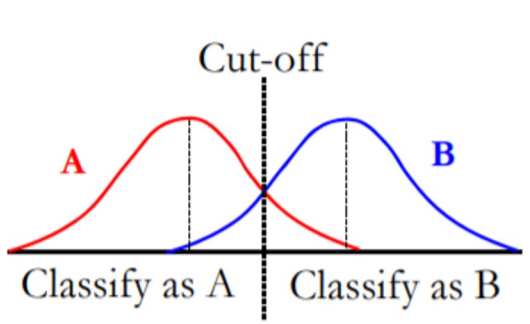
- Inferential use of DA requires assumptions
- Equal variance-covariance matrices between groups
  - Look for univariate variance homogeneity for discriminating variables (e.g. Fligner-Killeen nonparametric), visually inspect
- Multivariate normality within groups
  - Look for univariate normality in groups, try a multivariate normality test (e.g Energy test)
- Collinearity & Multicollinearity: No 2 variables are perfectly correlated, no variable can be calculated by combination of other variables
  - If assumption broken, drop one of the variables from the analysis
- Samples are independent

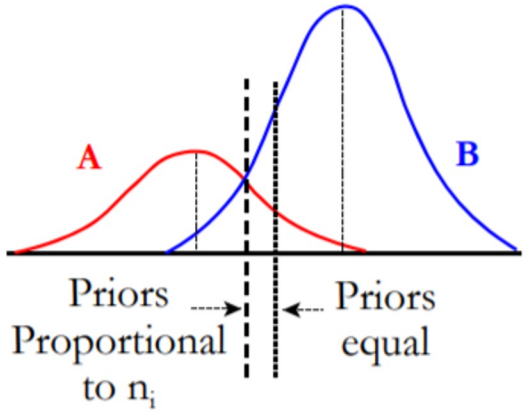
DA Assumptions Continued

- Prior probabilities of group membership identifiable
  - Not necessarily equal
- Linearity – variables change linearly along gradients

N	Priors	Actual	Predicted: Nest	Predicted: Not nest	#errors
7	0.5	Nest	0.83 (5.81)	0.17 (1.19)	1.19
93	0.5	Not nest	0.17 (15.81)	0.83 (77.19)	15.81
				total errors	17.00
				error rate	17.0%

Priors	Actual	Predicted: Nest	Predicted: Not nest	#errors
0.07	Nest	0.48 (3.36)	0.52 (3.64)	3.64
0.93	Not nest	0.02 (1.86)	0.98 (91.14)	1.86
			total errors	5.50
			error rate	5.5%





If Assumptions are Violated

- Procedure is moderately robust to violations
- Calculate canonical functions – do they have ecological meaning?
- Try nonparametric methods or CART (see later)

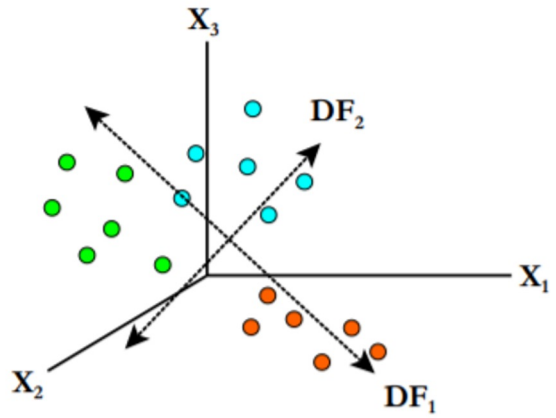
Variable Selection for DA

- If you have to many discriminators or are trying to identify the most useful from a large set of discriminators
- Forward variable selection using variable that minimizes Wilk's lambda statistic

How it works

- Solve canonical functions
- $|A - \lambda W| = 0$
  - A = among groups sums of squares and cross products matrix
  - W = within groups sums of squares and cross products matrix
  - λ = Vector of eigenvalue solutions
- Canonical function coefficiencts come from eigenvalue vector

$DF1 = .8x_1 + .3x_2 - .2x_3$
$DF2 = .4x_1 - .8x_2 + .2x_3$



Assessing Canonical Functions Importance

- Is a canonical function significant, and how many should be used?
  - Multiple methods
- Relative percent variance
  - Relative value of coefficients
  - Shows how much variation is associated with each axis

$$\Phi_i = \frac{\lambda_i}{\sum_{i=1}^{Q} \lambda_i}$$

- Canonical Correlation
  - Correlate canonical functions to groups
  - Significance test
- Graph and evaluate
- Test classification accuracy (chance-corrected)
- Are they ecologically interpretable?

Interpreting Canonical Functions

- Multiple methods
- Standardized canonical scores
  - Weights of standardized variables, use to describe variable importance
- Total Structure coefficients
  - Correlation between variable and canonical functions
- Covariance controlled partial F
  - Ranks partial F ratios of all variable contribution to model
- Potency index
  - % of discriminating power of retained canonical functions associated with each variable

Canonical function validation

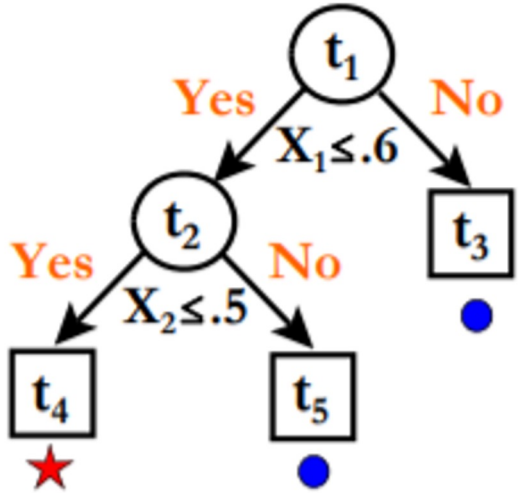
- Results only reliable if mean and dispersion estimates are also reliable (larger sample size is preferable)
- Split-sample validation
  - Training and testing datasets
  - How well does model from training dataset fit results from testing dataset?

Other DA considerations

- Different philosophies on interpreting canonical functions – performance measures vs. ecological meaning
- Validation of model is important
- May be limited by parametric assumptions

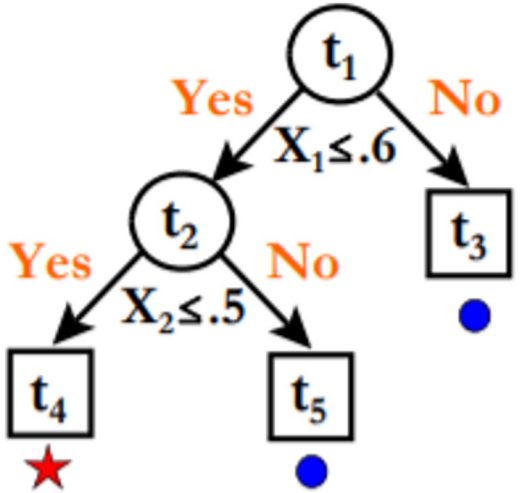
Classification and Regression Trees (CART)

- Nonparametric procedure for description of data
- Recursive partitioning of data into smaller groups
- Can take numeric and categorical variable
- Can handle missing data, robust to outliers
- Good for non-homogenous relationships, complex data, mixed data
- Relatively interpretable
- Automatic variable selection



CART Elements

- A set of questions: is $x_m > c$
  - Or linear combinations of continuous variable/Boolean combinations of categorical variables
- A rule for selecting best split at a node
- A rule for the size of the tree (when to stop splitting)



How it works

- At each node, calculate the best split point for each potential variable
- Choose the variable with the best split
- Continue doing this for each node until every node is terminal

Splitting Criteria

- Node impurity: mixing of classes within a node
- At each node, select split that reduces impurity the most
  - If regression tree, minimize variance
- Can also specify prior probabilities for each class, which affect the tree's classification
  - Or misclassification cost

■ Information Index:

$$i(t) = -\sum p(j/t) \ln p(j/t)$$

■ Gini Index:

$$i(t) = 1 - \sum p^2(j/t)$$

$p(j/t)$ = probability that a case is in class j given that it falls into node t; equal to the proportion of cases at node t in class j if priors are equal to class sizes.

■ Twoing Index:

At every node, select the conglomeration of classes into two superclasses so that considered as a two-class problem, the greatest decrease in node impurity is realized.

How big to make a tree?

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

$R(T)$ = Overall missclassification cost

α = Complexity parameter (complexity cost per terminal node)

$|\tilde{T}|$ = Complexity = # terminal nodes

- More splits = better fit
- More splits = greater risk of overfitting
- Method – grow tree to largest extent possible, then prune back to an appropriate size
- V-fold cross validation
  - Divide data into V subsets, drop each subset and make a tree on remaining data, validate on dropped subset
  - Tree with minimum estimated error rate chosen
  - Smallest tree within 1 SE of minimum error rate

CART Variable Importance

- Measured by sum of decrease in impurity cause by variable at each node
- Mainly reported as relative magnitude

Random Forest

- Machine Learning method based on CART
- Bootstrap approach containing several CARTs
- Each tree built on 2/3 of data
- Random subset of explanatory variables tried at each node
- No pruning
- Classification error rate for each tree determined by making prediction for “out-of-bag” sample
- RF prediction determined by prediction of majority of trees

Random Forest – variable importance

- Determined by permuting each variable and measuring increase in prediction error
- Higher increase in error = more important variable
- Or determine by mean decrease in impurity for each variable
- Otherwise RF is a bit of a black box

CART Limitations and Considerations

- Could be computationally complex
- May bias toward variables with more possible splits
- Results not based on underlying theoretical model
  - Blessing and a curse

Indicator Species Analysis

- Nonparametric procedure for distinguishing groups based on species compositional data
- Compute index of relative abundance of a species within each group
- Compute relative frequency of a species within each group
- Indicator value is product of these two numbers
- Indicator value of 1 = perfect indicator, 0 = not an indicator