Module 18 Multivariate Analysis for Genetic data
Session 12 Discriminant Analysis II

Jan Graffelman

jan.graffelman@upc.edu

[1] Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2] Department of Biostatistics
University of Washington
Seattle, WA, USA

26th Summer Institute in Statistical Genetics (SISG 2021)

Contents

Error rates and Confusion matrix

- It is of interest to evaluate the performance of a classification rule.
- There are several criteria to do so.
- Actual error rate (AER, density dependent)

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x}) d\mathbf{x}$$

- Apparent error rate (APER, not density dependent) based on the confusion matrix

		Predicted class	
		π_1	π_2
True Class	π_1	n_{11}	n_{12}
	π_2	n_{21}	n_{22}

- APER obtained as

$$\text{APER} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

- APER underestimates the AER.

Jackknife or hold-one-out

Procedure:

- Take the data from group π_1. Omit the ith observation, build the classifier with $n_1 - 1 + n_2$ observations.
- Classify the ith observation using the classifier.
- Repeat for all observations in π_1.
- Calculate n_{1M}^H, the number of observations that were held out and misclassified.
- Do the same for group π_2 and calculate n_{2M}^H.
- Obtain an estimate of the expected actual error rate

$$E\left(\text{AER}\right) = \frac{n_{1M}^H + n_{2M}^H}{n_1 + n_2}$$

Allele intensities revisited

LDA

	non T carrier	T carrier
non T carrier	46	0
T carrier	0	52

$$\text{APER} = \frac{0+0}{46+52} = 0$$

With cross-validation

$$\text{E(AER)} = 0$$

QDA

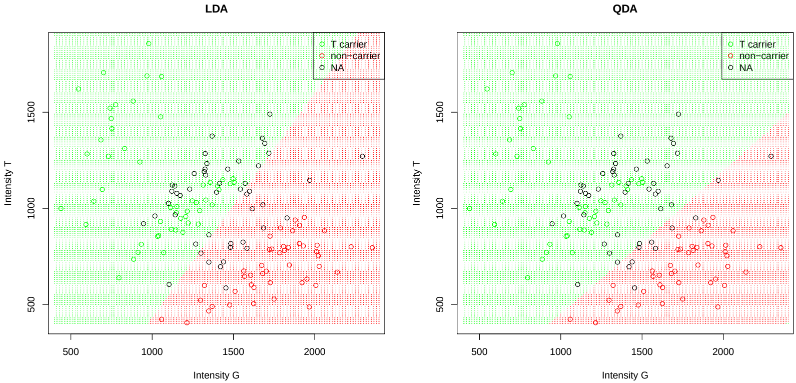
	non T carrier	T carrier
non T carrier	46	0
T carrier	0	52

$$\text{APER} = \frac{0+0}{46+52} = 0$$

With cross-validation

$$\text{E(AER)} = 0.0102$$

Visualisation



LDA with multiple groups

- The ECM rule can be extended to k groups
- Fisher's discriminant analysis

ECM rule

ECM rule with k groups (equal costs)

Assign $\mathbf{x}$ to π_k if

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \forall \quad i \neq k$$

Fisher's linear discriminant analysis

- Searches for an optimal linear combination:

$$Z_1 = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$

- Maximizes the ratio of variability between groups to variability within groups
- Objective function

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

- $\mathbf{W}$ is the matrix with within-group sums-of-squares
- For a single group *i*

$$\mathbf{W}_i = (\mathbf{X}_i - \mathbf{1}\mathbf{m}_i')'(\mathbf{X}_i - \mathbf{1}\mathbf{m}_i')$$

- $\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_i$
- $\mathbf{B}$ is the matrix with between-group sums-of-squares
- $\mathbf{T}$ is the matrix with total sums-of-squares

$$\mathbf{T} = (\mathbf{X} - \mathbf{1}\mathbf{m}')'(\mathbf{X} - \mathbf{1}\mathbf{m}') \qquad \mathbf{T} = \mathbf{W} + \mathbf{B}$$

Solution

- The optimal weights are found by solving an eigenvector-eigenvalue problem

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$$

- The number of dimensions d in the solution is given by $\min(k-1, p)$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{A} = \mathbf{A}\mathbf{D}_{\lambda}$$

- Eigenvectors scaled to satisfy $\mathbf{A}'\mathbf{S}_p\mathbf{A} = \mathbf{I}$
- Selecting the first two eigenvalues and eigenvectors allows for dimension reduction

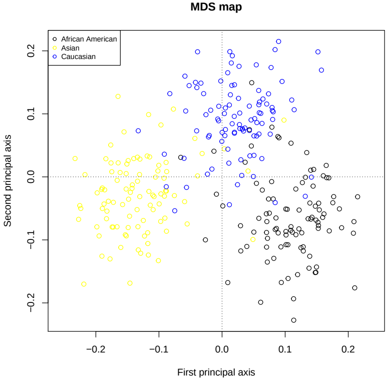
NIST autosomal STR data revisited

The data:

- 29 autosomal STRs
- Consider individuals with African-American, Asian and Caucasian ancestry
- Sample sizes balanced by subsampling

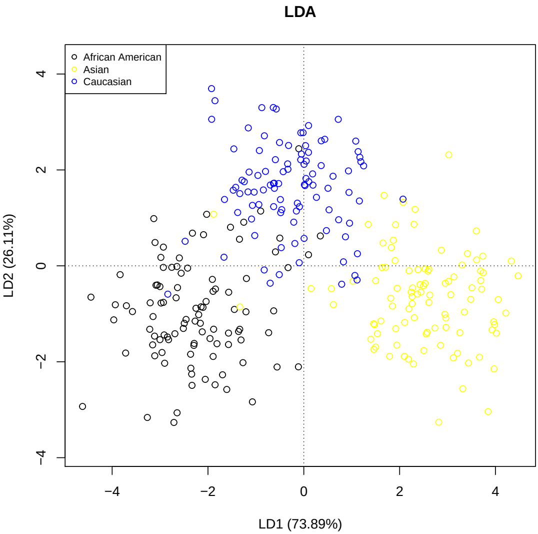
Prior to discriminant analysis:

- STRs coded as binary variables
- Quantification of the data by MDS based on Jaccard metric



Can we predict ancestry from an STR profile?

STR data in discriminant space



Numerical output

	1	2
Eigenvalue	550.38	194.45
Fraction	0.74	0.26
Cumulative	0.74	1.00

		Principal axis									
	prior	1	2	3	4	5	6	7	8	9	10
Afr. Ame.	0.333	0.108	-0.063	-0.001	-0.001	0.002	0.001	0.002	0.010	0.010	0.009
Asian	0.333	-0.132	-0.029	0.007	-0.002	0.010	0.005	-0.007	-0.002	-0.011	0.003
Caucasian	0.333	0.024	0.092	-0.006	0.003	-0.012	-0.006	0.006	-0.009	0.002	-0.012

Confusion matrix

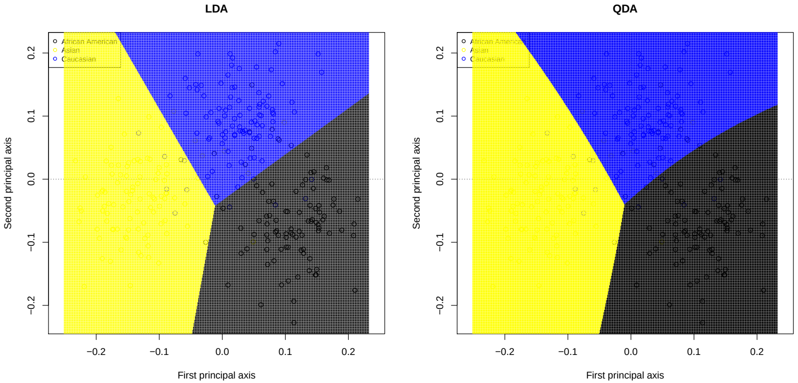
LDA			
	Afr. Ame.	Asian	Caucasian
Afr. Ame.	86	0	11
Asian	1	92	4
Caucasian	5	6	86

APER $= 0.093$

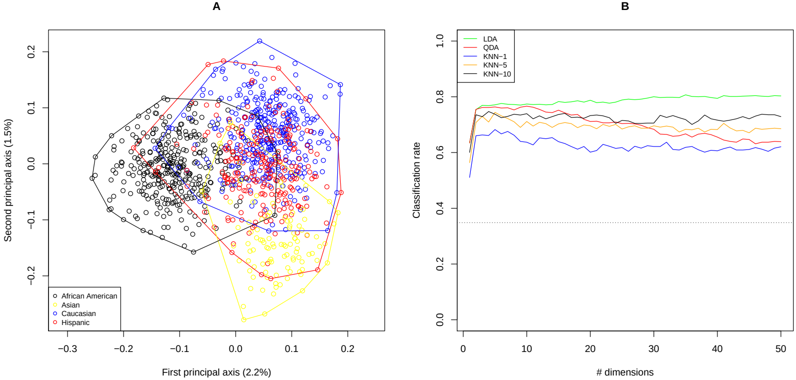
QDA			
	Afr. Ame.	Asian	Caucasian
Afr. Ame.	91	0	6
Asian	2	93	2
Caucasian	5	3	89

APER $= 0.062$

NIST STR data revisited



More complex...



Alternative statistical techniques

- An alternative technique for two-group DA is logistic regression
- An alternative technique for multi-group DA is the multinomial logit model

References

- Hand, D.J. (1981) Discrimination and Classification. Wiley, New York.
- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall, Chapter 11.
- Lachenbruch, P.A. (1975) Discriminant Analysis. Hafner Press, New York.
- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) An Introduction to Statistical Learning. Springer, New York.