Multivariate analysis of genetic data — exploring group diversity —

Thibaut Jombart

MRC Centre for Outbreak Analysis and Modelling
Imperial College London

Genetic data analysis with R
PR∼Statistics, Glasgow
05-08-2015

Outline

Introduction

Identifying groups

- Hierarchical clustering
- K-means

Exploring group diversity

- Aggregating data
- Optimizing group differences
- Discriminant Analysis of Principal Components

Outline

Introduction

Identifying groups

Hierarchical clustering

K-means

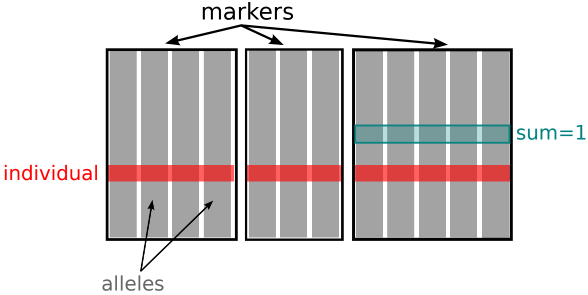
Exploring group diversity

Aggregating data

Optimizing group differences

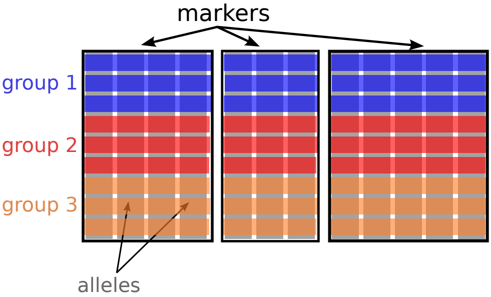
Discriminant Analysis of Principal Components

Genetic data: introducing group data



- How to identify groups?
- How to explore group diversity?

Genetic data: introducing group data



- How to identify groups?
- How to explore group diversity?

Outline

Introduction

Identifying groups

- Hierarchical clustering
- K-means

Exploring group diversity

- Aggregating data
- Optimizing group differences
- Discriminant Analysis of Principal Components

Hierarchical clustering: a variety of algorithms

- single linkage
- complete linkage
- UPGMA
- Ward
- ...

Rationale

1. compute pairwise genetic distances $\mathbf{D}$ (or similarities)
2. group the closest pair(s) together
3. (optional) update $\mathbf{D}$
4. return to 2) until no new group can be made

Rationale

1. compute pairwise genetic distances $\mathbf{D}$ (or similarities)
2. group the closest pair(s) together
3. (optional) update $\mathbf{D}$
4. return to 2) until no new group can be made

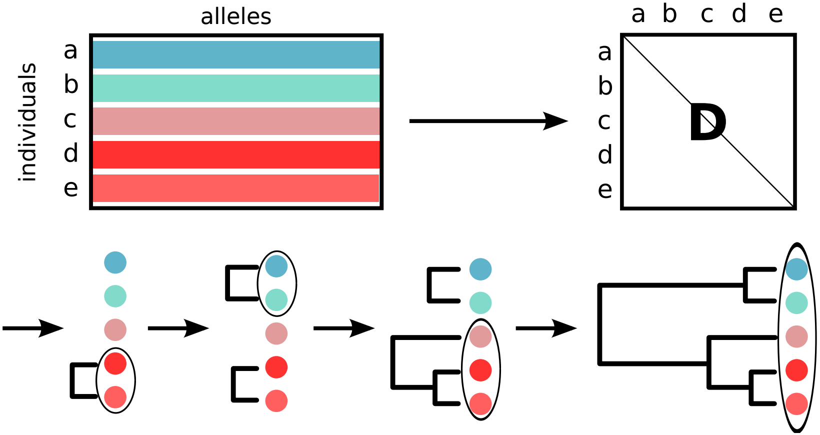
Rationale

1. compute pairwise genetic distances $\mathbf{D}$ (or similarities)
2. group the closest pair(s) together
3. (optional) update $\mathbf{D}$
4. return to 2) until no new group can be made

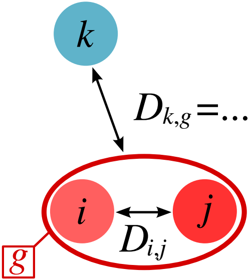
Rationale

1. compute pairwise genetic distances $\mathbf{D}$ (or similarities)
2. group the closest pair(s) together
3. (optional) update $\mathbf{D}$
4. return to 2) until no new group can be made

Rationale

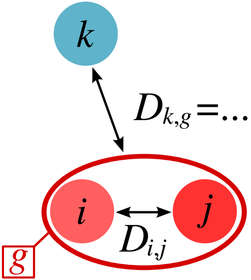


Differences between algorithms



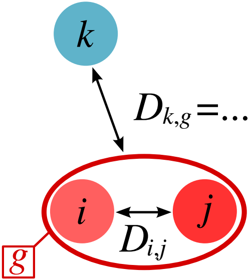
- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

Differences between algorithms



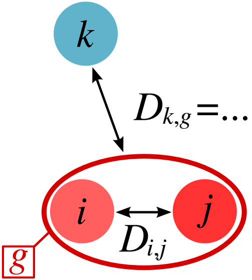
- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

Differences between algorithms



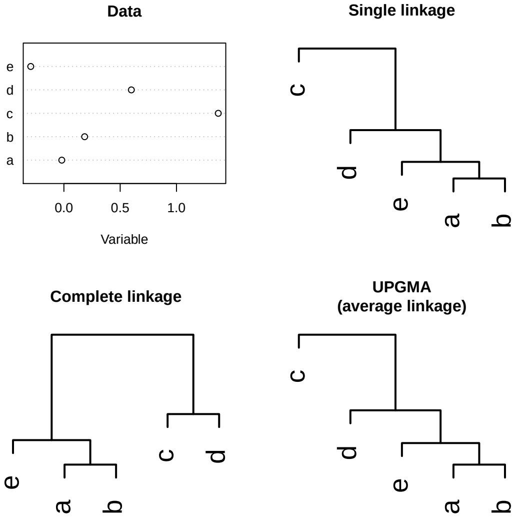
- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

Differences between algorithms



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

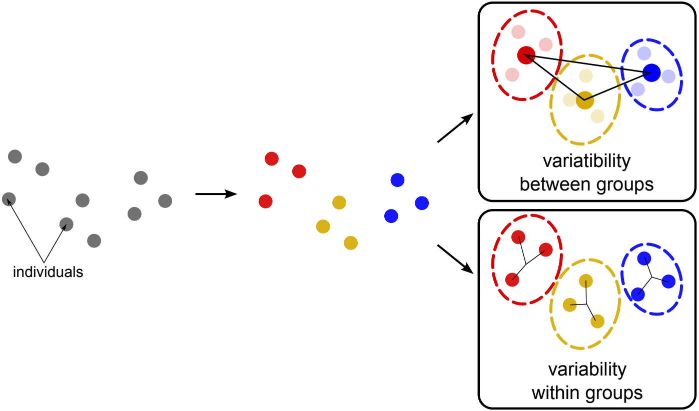
Differences between algorithms



K-means underlying model

ANOVA model:

$$\text{total var.} = (\text{var. between groups}) + (\text{var. within groups})$$



K-means rationale

Find groups which minimize *within group var.* (equally: maximize *between group var.*).

In other words:
Identify a partition $\mathcal{G} = \{g_1, \ldots, g_k\}$ solving:

$$\arg \min_{\mathcal{G}=\{g_1,\ldots,g_k\}} \sum_k \sum_{i \in g_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

with:

- $\mathbf{x}_i \in \mathbb{R}^p$: vector of allele frequencies of individual i
- $\boldsymbol{\mu}_k \in \mathbb{R}^p$: vector of means allele frequencies of group k

K-means rationale

Find groups which minimize *within group var.* (equally: maximize *between group var.*).

In other words:
Identify a partition $\mathcal{G} = \{g_1, \ldots, g_k\}$ solving:

$$\arg \min_{\mathcal{G}=\{g_1,\ldots,g_k\}} \sum_k \sum_{i \in g_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

with:

- $\mathbf{x}_i \in \mathbb{R}^p$: vector of allele frequencies of individual i
- $\boldsymbol{\mu}_k \in \mathbb{R}^p$: vector of means allele frequencies of group k

K-means rationale

Find groups which minimize *within group var.* (equally: maximize *between group var.*).

In other words:
Identify a partition $\mathcal{G} = \{g_1, \dots, g_k\}$ solving:

$$\arg \min_{\mathcal{G}=\{g_1,\dots,g_k\}} \sum_k \sum_{i \in g_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

with:

- $\mathbf{x}_i \in \mathbb{R}^p$: vector of allele frequencies of individual i
- $\boldsymbol{\mu}_k \in \mathbb{R}^p$: vector of means allele frequencies of group k

K-means algorithm

The K-mean problem is solved by the following algorithm:

1. select random group means ($\boldsymbol{\mu}_k$, $k = 1, \ldots, K$)
2. assign each individual $\mathbf{x}_i$ to the closest group $\longrightarrow g_k$
3. update group means $\boldsymbol{\mu}_k$
4. go back to 2) until convergence (groups no longer change)

K-means algorithm

The K-mean problem is solved by the following algorithm:

1. select random group means ($\boldsymbol{\mu}_k$, $k = 1, \dots, K$)
2. assign each individual $\mathbf{x}_i$ to the closest group $\longrightarrow g_k$
3. update group means $\boldsymbol{\mu}_k$
4. go back to 2) until convergence (groups no longer change)

K-means algorithm

The K-mean problem is solved by the following algorithm:

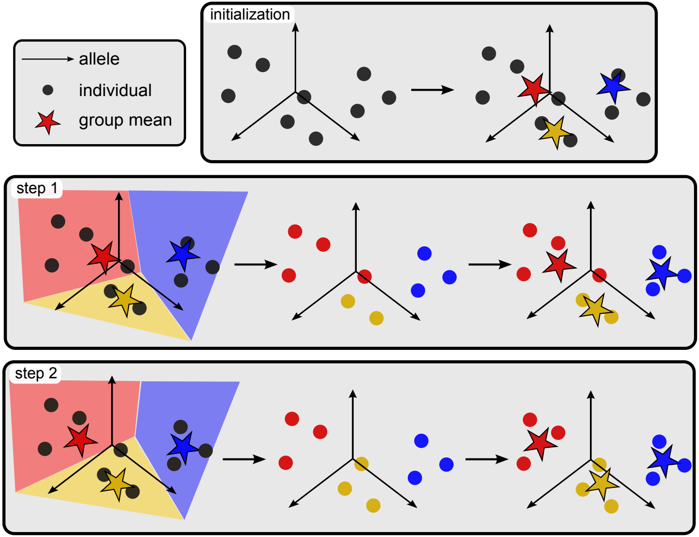
1. select random group means ($\boldsymbol{\mu}_k$, $k = 1, \ldots, K$)
2. assign each individual $\mathbf{x}_i$ to the closest group $\longrightarrow g_k$
3. update group means $\boldsymbol{\mu}_k$
4. go back to 2) until convergence (groups no longer change)

K-means algorithm

The K-mean problem is solved by the following algorithm:

1. select random group means ($\boldsymbol{\mu}_k$, $k = 1, \ldots, K$)
2. assign each individual $\mathbf{x}_i$ to the closest group $\longrightarrow g_k$
3. update group means $\boldsymbol{\mu}_k$
4. go back to 2) until convergence (groups no longer change)

K-means algorithm



K-means: limitations and extensions

Limitations

- slower for large numbers of alleles (e.g. 100,000)
- K-means does not identify the number of clusters (K)

Extension

- run K-means after dimension reduction using PCA
- try increasing values of K
- use Bayesian Information Criterion (BIC) for model selection

K-means: limitations and extensions

Limitations

- slower for large numbers of alleles (e.g. 100,000)
- K-means does not identify the number of clusters (K)

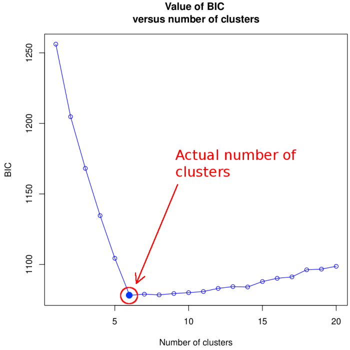
Extension

- run K-means after dimension reduction using PCA
- try increasing values of K
- use Bayesian Information Criterion (BIC) for model selection

Genetic clustering using K-means & BIC

(Jombart *et al.* 2010, *BMC Genetics*)

Simulated data: island model with 6 populations



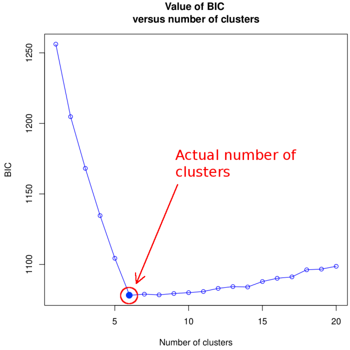
Performances:

- K-means $\geq$ STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

Genetic clustering using K-means & BIC

(Jombart *et al.* 2010, *BMC Genetics*)

Simulated data: island model with 6 populations



Performances:

- K-means $\geq$ STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

Outline

Introduction

Identifying groups

- Hierarchical clustering
- K-means

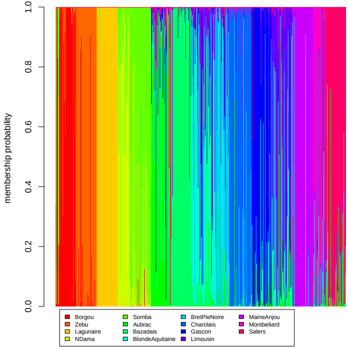
Exploring group diversity

- Aggregating data
- Optimizing group differences
- Discriminant Analysis of Principal Components

Why identifying clusters is not the whole story

Example of cattle breeds diversity (30 microsatellites, 704 individuals).

Group membership probabilities:

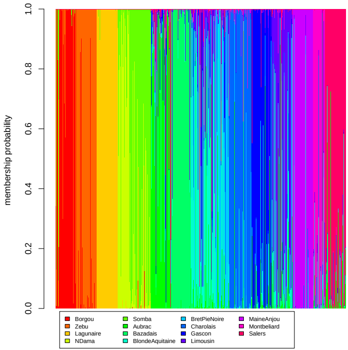


Important to assess the relationships between clusters.

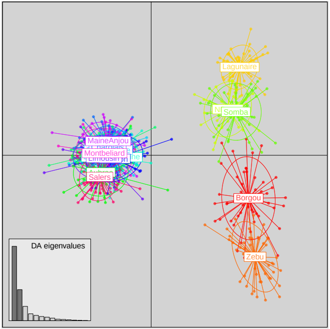
Why identifying clusters is not the whole story

Example of cattle breeds diversity (30 microsatellites, 704 individuals).

Group membership probabilities:



Multivariate analysis:

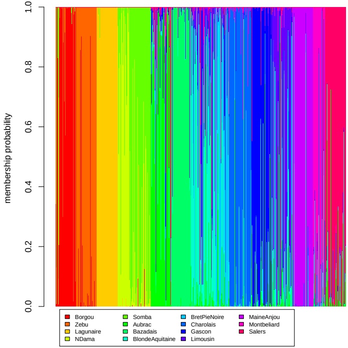


Important to assess the relationships between clusters.

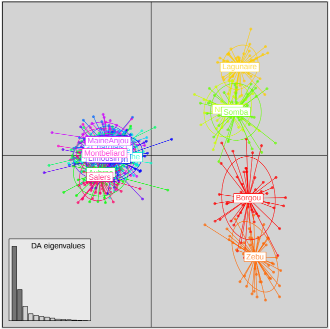
Why identifying clusters is not the whole story

Example of cattle breeds diversity (30 microsatellites, 704 individuals).

Group membership probabilities:

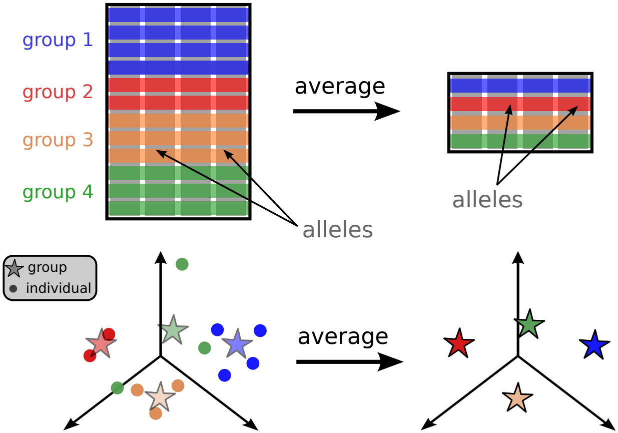


Multivariate analysis:



Important to assess the relationships between clusters.

Aggregating data by groups



⟶ multivariate analysis of group allele frequencies.

Analysing group data

Available methods:

- Principal Component Analysis (PCA) of allele frequency table
- Genetic distance between populations ⟶ Principal Coordinates Analysis (PCoA)
- Correspondance Analysis (CA) of allele counts

Criticism:

- Lose individual information
- Neglect within-group diversity
- CA: possible artefactual outliers

Analysing group data

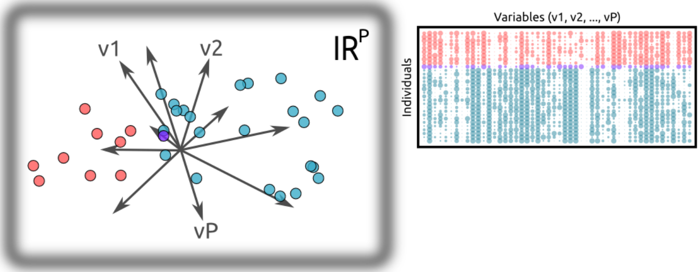
Available methods:

- Principal Component Analysis (PCA) of allele frequency table
- Genetic distance between populations ⟶ Principal Coordinates Analysis (PCoA)
- Correspondance Analysis (CA) of allele counts

Criticism:

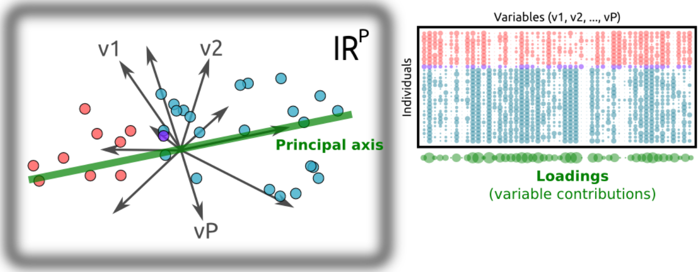
- Lose individual information
- Neglect within-group diversity
- CA: possible artefactual outliers

Multivariate analysis: reminder



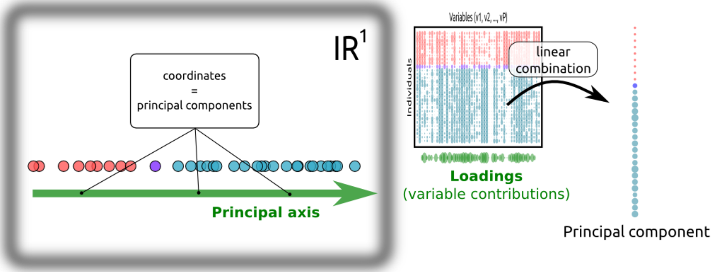
Find principal components with *maximum total variance*.

Multivariate analysis: reminder



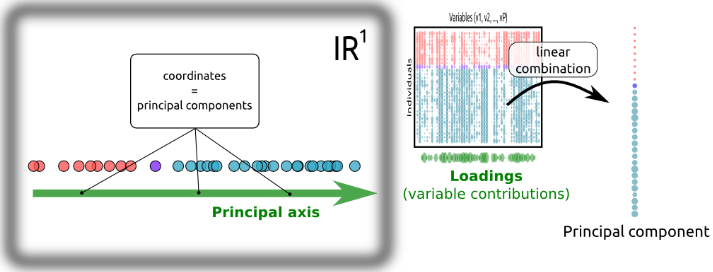
Find principal components with *maximum total variance*.

Multivariate analysis: reminder



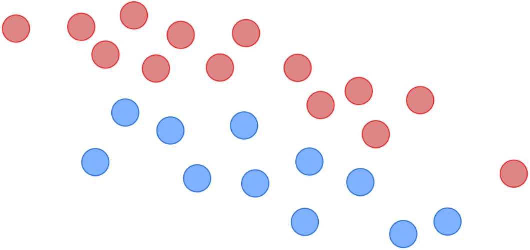
Find principal components with *maximum total variance*.

Multivariate analysis: reminder



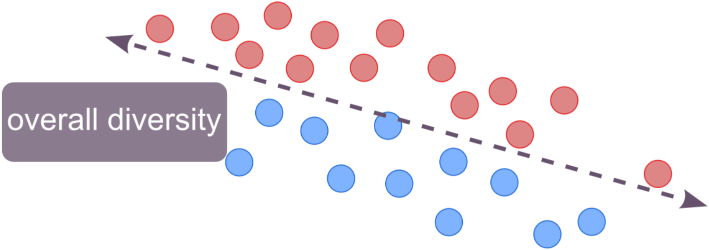
Find principal components with *maximum total variance*.

But total variance may not reflect group differences



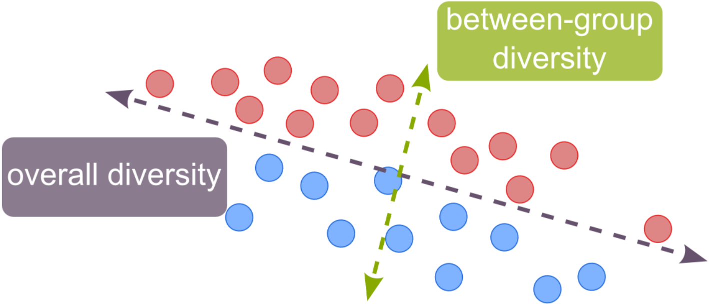
Need to optimize different criteria.

But total variance may not reflect group differences



Need to optimize different criteria.

But total variance may not reflect group differences



Need to optimize different criteria.

Optimizing different criteria

Similar approaches to PCA can be used to optimize different quantities:

- **PCA**: *total* variance
- **Between-group PCA**: variance *between* groups
- **Within-group PCA**: variance *within* groups
- **Discriminant Analysis**: variance *between* groups / variance *within* groups

Optimizing different criteria

Similar approaches to PCA can be used to optimize different quantities:

- **PCA**: *total* variance
- **Between-group PCA**: variance *between* groups
- **Within-group PCA**: variance *within* groups
- **Discriminant Analysis**: variance *between* groups / variance *within* groups

Optimizing different criteria

Similar approaches to PCA can be used to optimize different quantities:

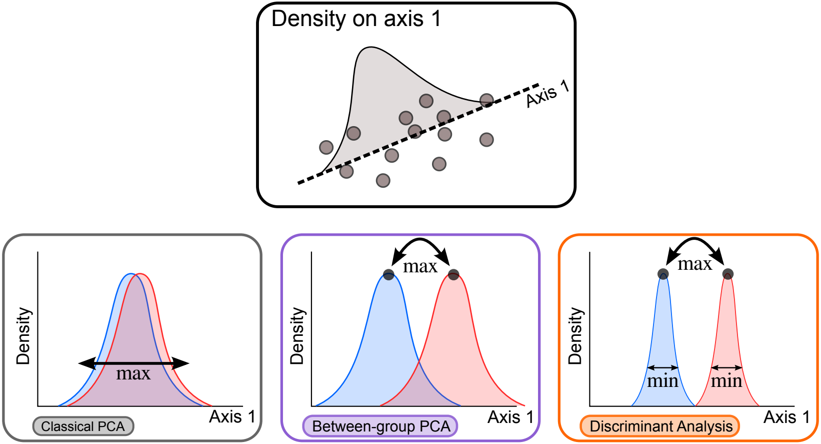
- **PCA**: *total* variance
- **Between-group PCA**: variance *between* groups
- **Within-group PCA**: variance *within* groups
- **Discriminant Analysis**: variance *between* groups / variance *within* groups

Optimizing different criteria

Similar approaches to PCA can be used to optimize different quantities:

- **PCA**: *total* variance
- **Between-group PCA**: variance *between* groups
- **Within-group PCA**: variance *within* groups
- **Discriminant Analysis**: variance *between* groups / variance *within* groups

From PCA to DA: increasing group differentiation



Max. total diversity

Max. diversity between groups

Max. separation of groups

Discriminant Analysis: limitations and extensions

Limitations:

- DA requires less variables (alleles) than observations (individuals)
- DA requires uncorrelated variables (no frequencies, no linkage disequilibrium)

Discriminant Analysis of Principal Components (DAPC)[1]:

- data orthogonalisation/reduction using PCA before DA
- overcomes limitations of DA
- group membership probabilities, group prediction

[1] Jombart et al. 2010, *BMC Genetics*

Discriminant Analysis: limitations and extensions

Limitations:

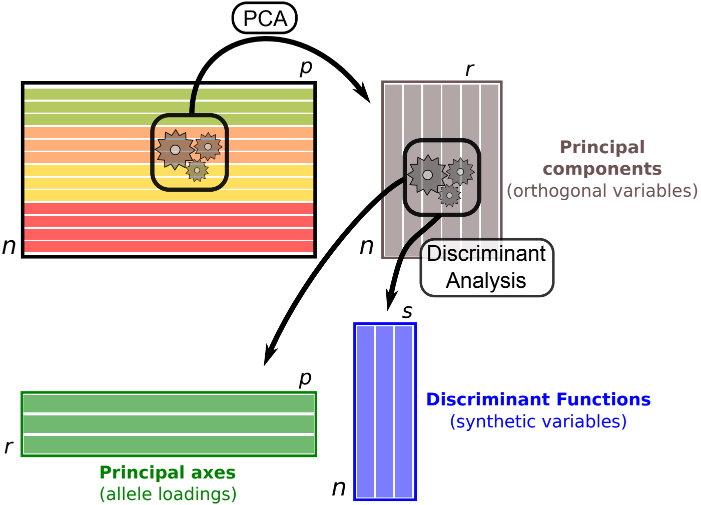
- DA requires less variables (alleles) than observations (individuals)
- DA requires uncorrelated variables (no frequencies, no linkage disequilibrium)

Discriminant Analysis of Principal Components (DAPC)[1]:

- data orthogonalisation/reduction using PCA before DA
- overcomes limitations of DA
- group membership probabilities, group prediction

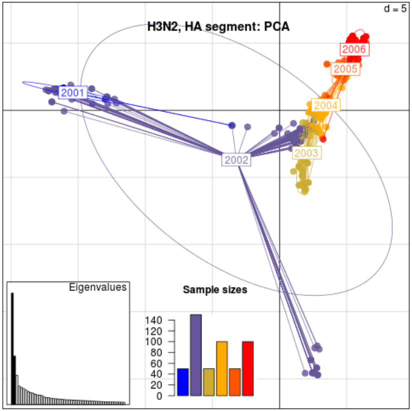
[1] Jombart et al. 2010, *BMC Genetics*

Rationale of DAPC



PCA of seasonal influenza (A/H3N2) data

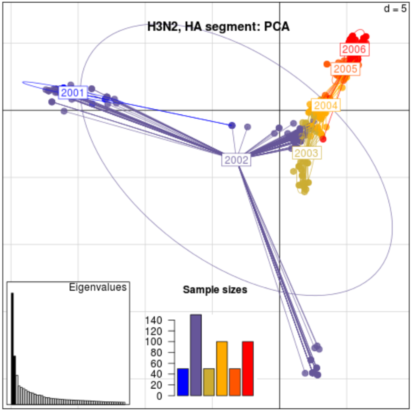
Data: seasonal influenza (A/H3N2), 500 HA segments.



Little temporal evolution, burst of diversity in 2002??

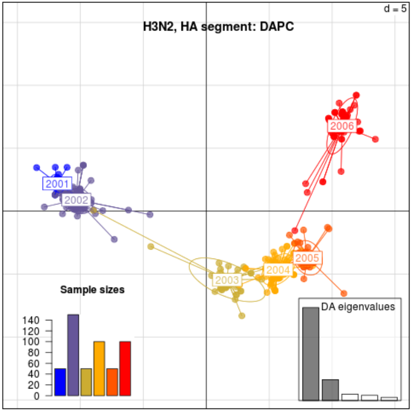
PCA of seasonal influenza (A/H3N2) data

Data: seasonal influenza (A/H3N2), 500 HA segments.



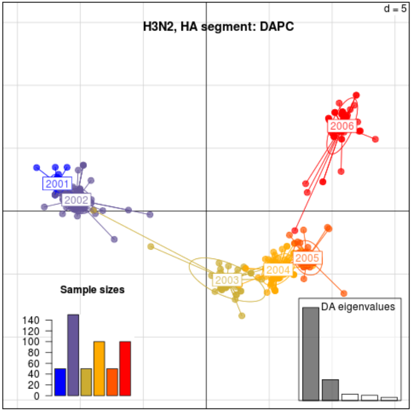
Little temporal evolution, burst of diversity in 2002??

DAPC of seasonal influenza (A/H3N2) data



Strong temporal signal, originality of 2006 isolates (new alleles).

DAPC of seasonal influenza (A/H3N2) data



Strong temporal signal, originality of 2006 isolates (new alleles).

Other features

DAPC can be used to:

- provides group assignment probabilities
- can use supplementary individuals
- can predict group membership of new data
- can be used for variable selection

Time to get your hands dirty (again)!

The pdf of the practical is online:

`http://adegenet.r-forge.r-project.org/`

or

Google → adegenet → documents → "Workshop Glasgow, August 2015"