Multivariate analysis of genetic data: exploring groups diversity

T. Jombart

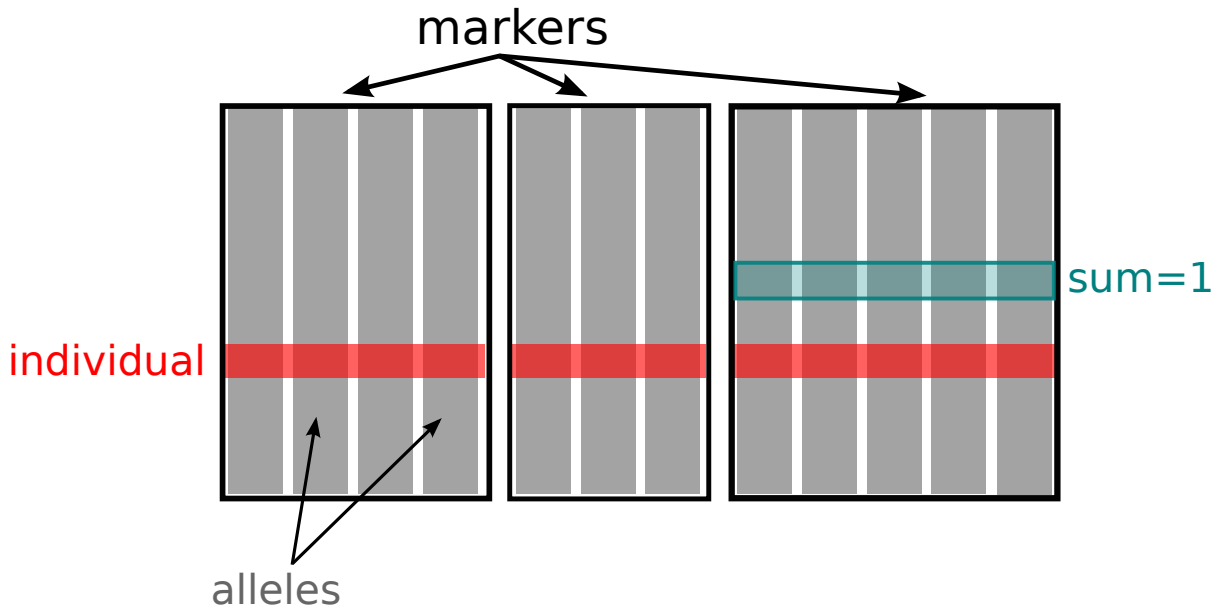
Imperial College London

Bogota
01-12-2010

Outline

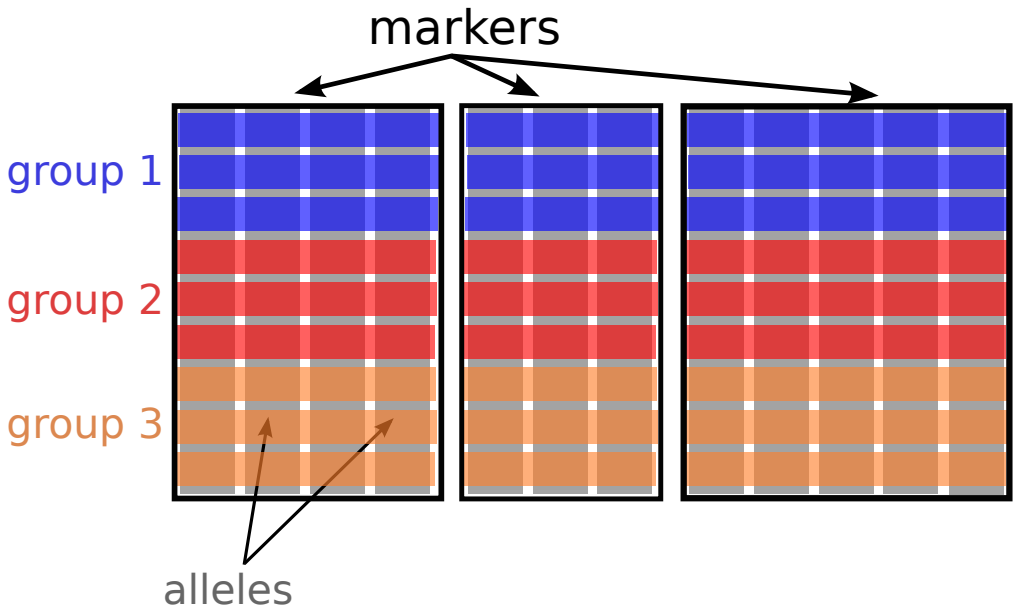
Outline

Genetic data: recall



- How to define groups?
- How to handle group information?

Genetic data: recall



- How to define groups?
- How to handle group information?

Finding and using group information

Finding groups:

- hierarchical clustering:
  - single linkage
  - complete linkage
  - UPGMA
- K-means

Using group information:

- multivariate analysis of group frequencies
- using groups as partitions
- discriminant analysis

Finding and using group information

Finding groups:

- hierarchical clustering:
  - single linkage
  - complete linkage
  - UPGMA
- K-means

Using group information:

- multivariate analysis of group frequencies
- using groups as partitions
- discriminant analysis

Outline

Outline

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

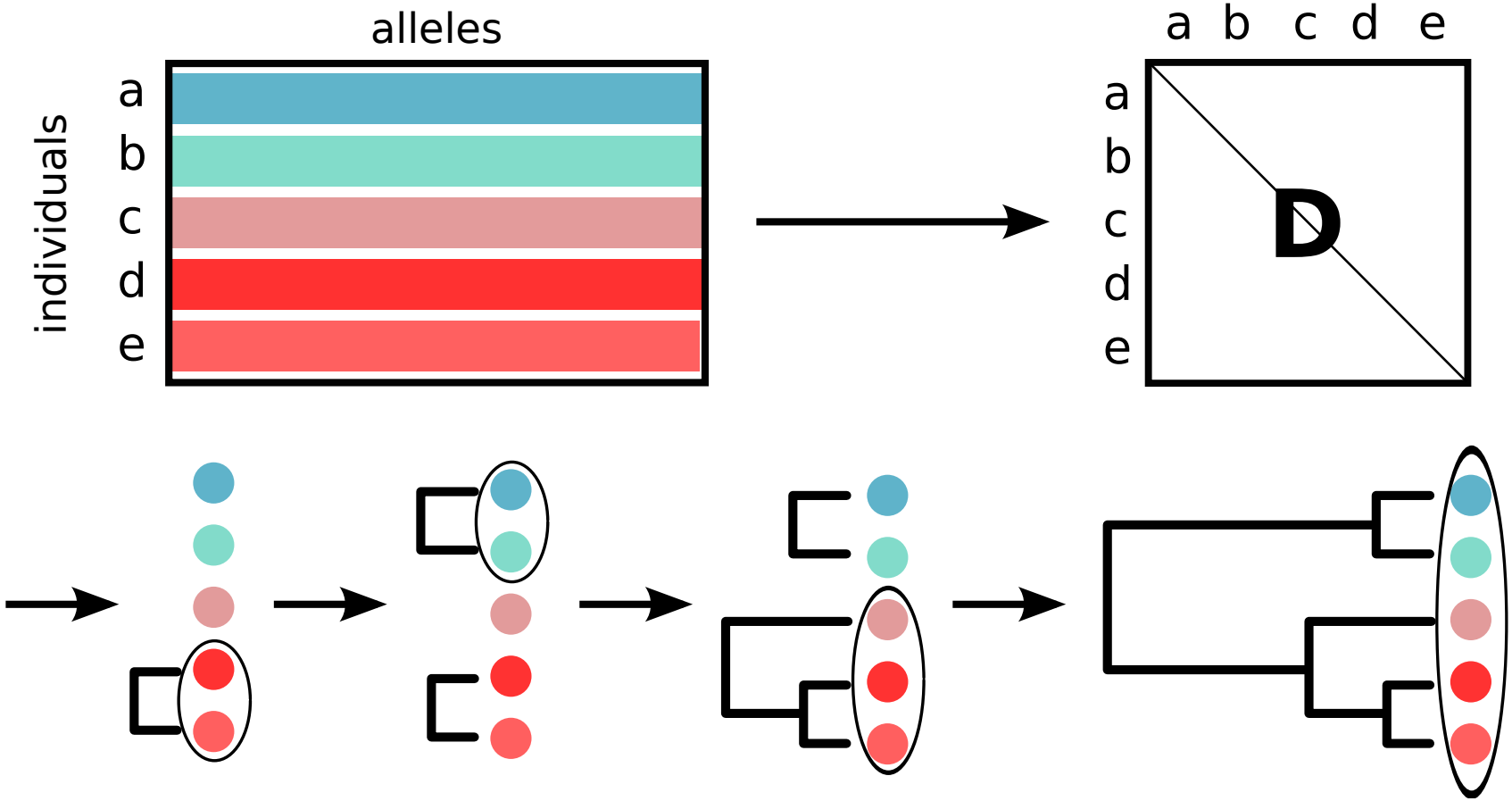
Rationale

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

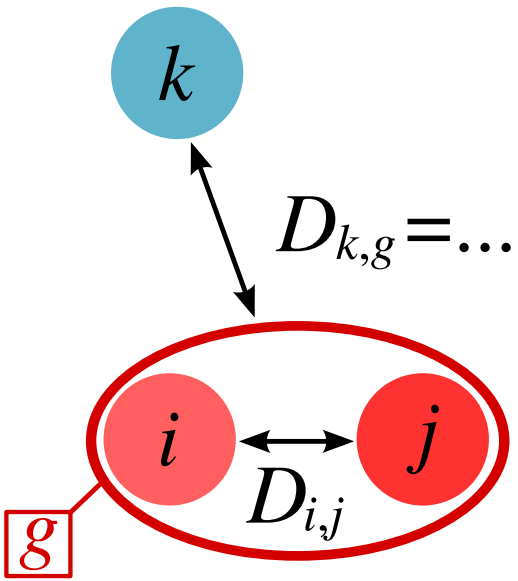
Rationale

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

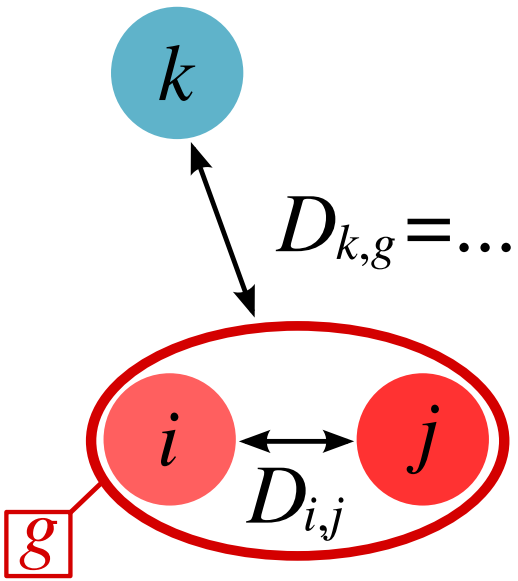


Differences between algorithms



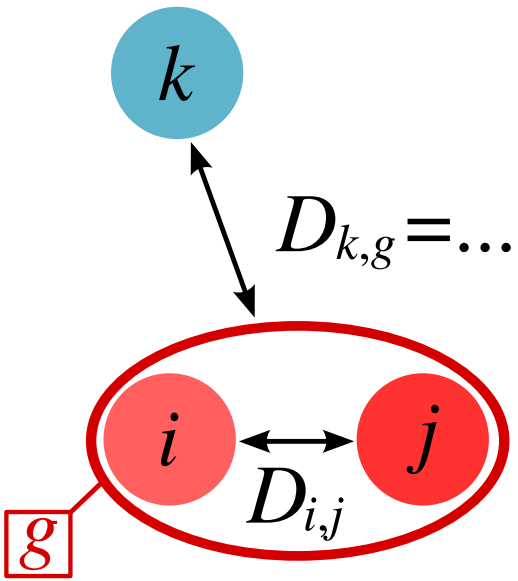
- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

Differences between algorithms



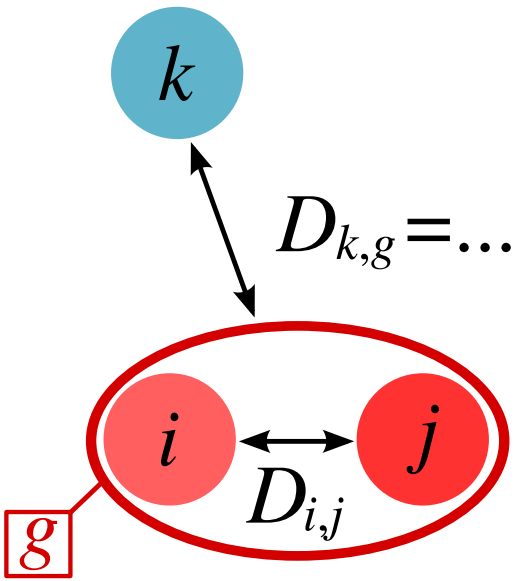
- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

Differences between algorithms



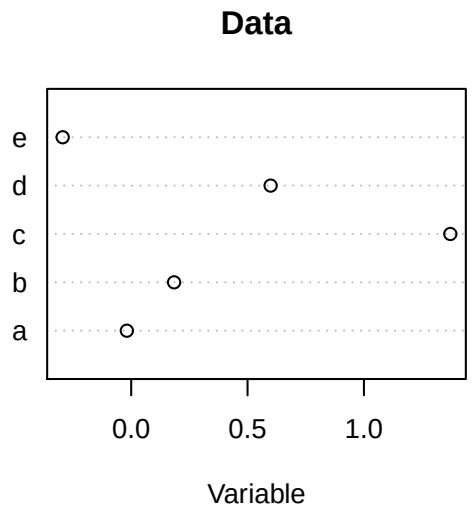
- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

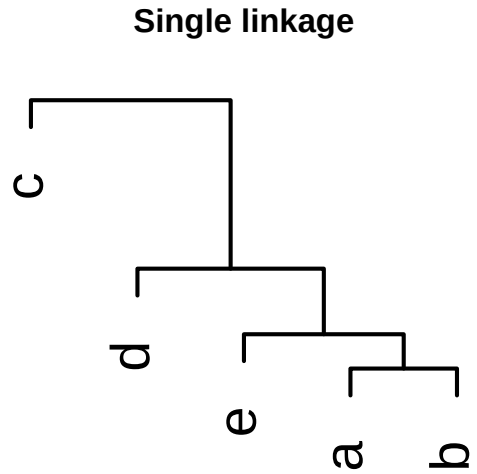
Differences between algorithms

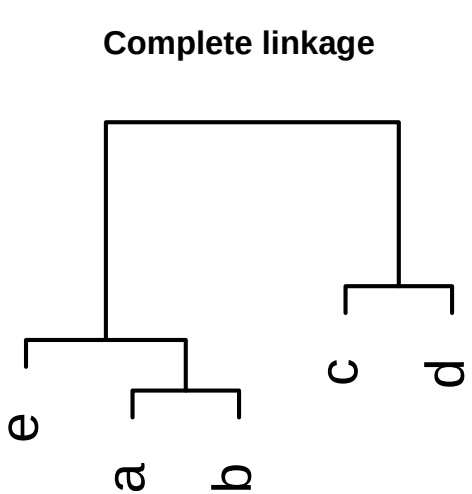


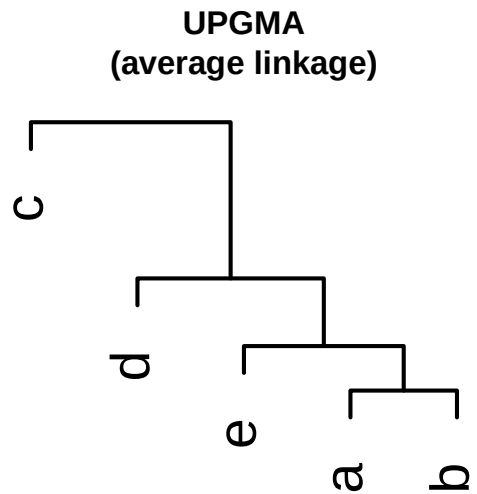
- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

Differences between algorithms



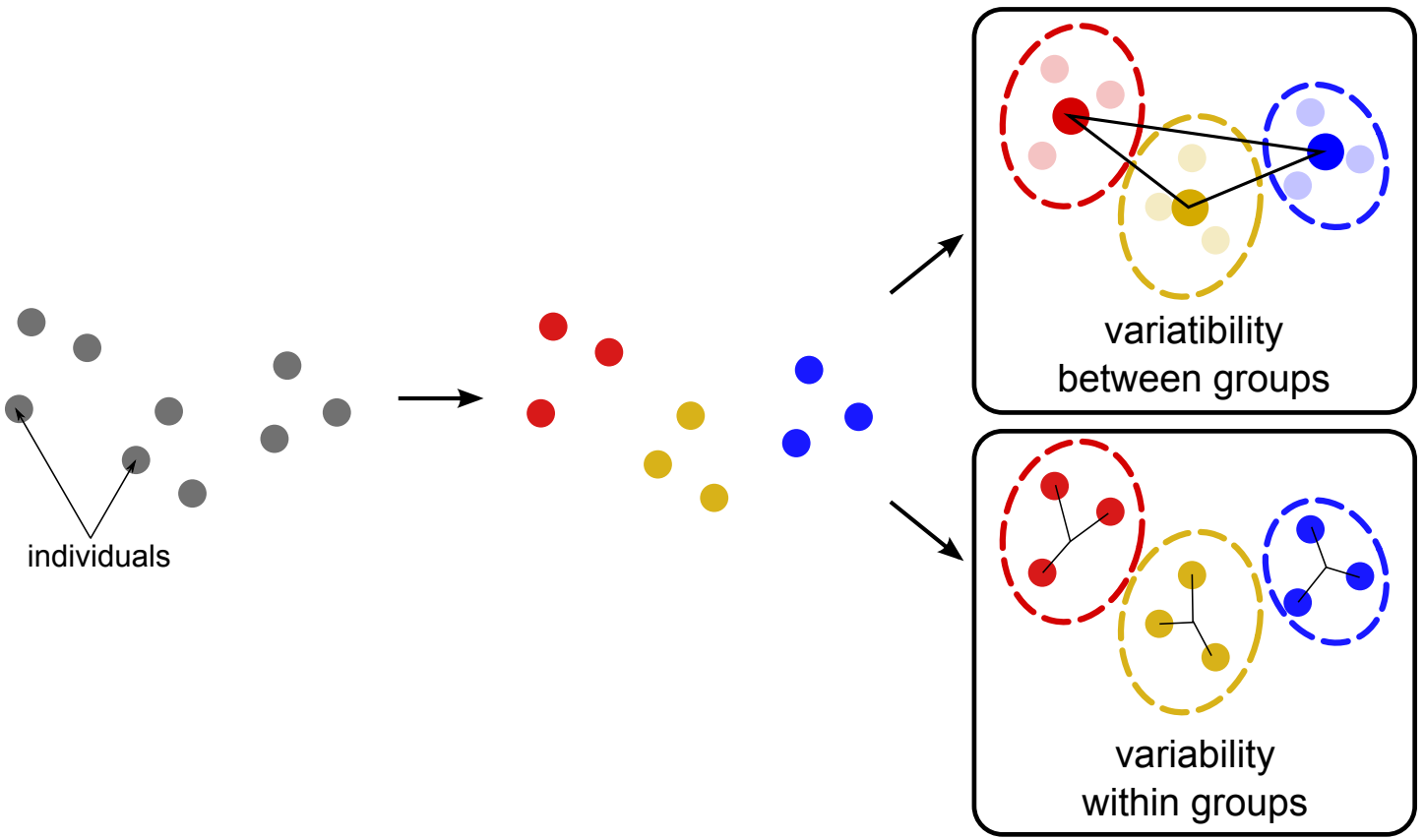






Outline

Variability between and within groups



K-means: underlying model

Univariate ANOVA model (ss: sum of squares):

$$\text{sst}(x) = \text{ssb}(x) + \text{ssw}(x)$$

with:

- $k = 1, \ldots, K$: number of groups
- μ_k: mean of group k; μ: mean of all data
- g_k: set of individuals in group k ($i \in g_k \Leftrightarrow i$ is in group k)
- $\text{sst}(\mathbf{x}) = \sum_i (x_i - \mu)^2$: total variation
- $\text{ssb}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (\mu_k - \mu)^2$: variation between groups
- $\text{ssw}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (x_i - \mu_k)^2$: (residual) variation within groups

K-means: underlying model

Univariate ANOVA model (ss: sum of squares):

$$\text{sst}(x) = \text{ssb}(x) + \text{ssw}(x)$$

with:

- $k = 1, \ldots, K$: number of groups
- μ_k: mean of group k; μ: mean of all data
- g_k: set of individuals in group k ($i \in g_k \Leftrightarrow i$ is in group k)
- $\text{sst}(\mathbf{x}) = \sum_i (x_i - \mu)^2$: total variation
- $\text{ssb}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (\mu_k - \mu)^2$: variation between groups
- $\text{ssw}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (x_i - \mu_k)^2$: (residual) variation within groups

K-means: underlying model

Univariate ANOVA model (ss: sum of squares):

$$\text{sst}(x) = \text{ssb}(x) + \text{ssw}(x)$$

with:

- $k = 1, \ldots, K$: number of groups
- μ_k: mean of group k; μ: mean of all data
- g_k: set of individuals in group k ($i \in g_k \Leftrightarrow i$ is in group k)
- $\text{sst}(\mathbf{x}) = \sum_i (x_i - \mu)^2$: total variation
- $\text{ssb}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (\mu_k - \mu)^2$: variation between groups
- $\text{ssw}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (x_i - \mu_k)^2$: (residual) variation within groups

K-means: underlying model

Univariate ANOVA model (ss: sum of squares):

$$\text{sst}(x) = \text{ssb}(x) + \text{ssw}(x)$$

with:

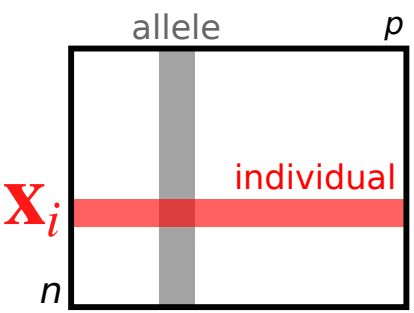
- $k = 1, \ldots, K$: number of groups
- μ_k: mean of group k; μ: mean of all data
- g_k: set of individuals in group k ($i \in g_k \Leftrightarrow i$ is in group k)
- $\text{sst}(\mathbf{x}) = \sum_i (x_i - \mu)^2$: total variation
- $\text{ssb}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (\mu_k - \mu)^2$: variation between groups
- $\text{ssw}(\mathbf{x}) = \sum_k \sum_{i \in g_k} (x_i - \mu_k)^2$: (residual) variation within groups

K-means: underlying model

Extension to multivariate data ($\mathbf{X} \in \mathbb{R}^{n \times p}$):

$$\mathrm{SST}(\mathbf{X}) = \mathrm{SSB}(\mathbf{X}) + \mathrm{SSW}(\mathbf{X})$$

with:



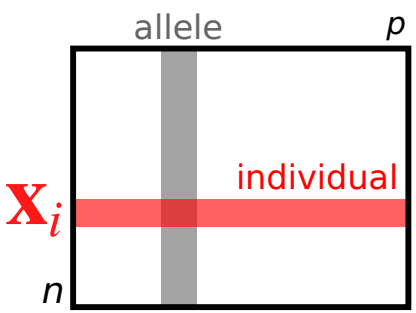
- $\boldsymbol{\mu}_k$: vector of means of group k; $\boldsymbol{\mu}$: vector of means of all data
- $\mathrm{SST}(\mathbf{X}) = \sum_i \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$: total variation
- $\mathrm{SSB}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|^2$: variation between groups
- $\mathrm{SSW}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$: (residual) variation within groups

K-means: underlying model

Extension to multivariate data ($\mathbf{X} \in \mathbb{R}^{n \times p}$):

$$\mathrm{SST}(\mathbf{X}) = \mathrm{SSB}(\mathbf{X}) + \mathrm{SSW}(\mathbf{X})$$

with:



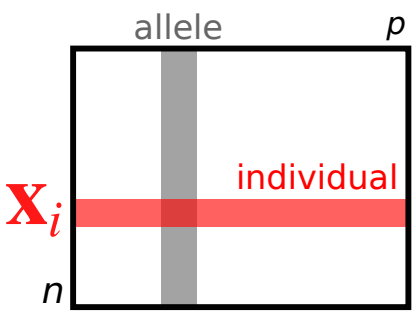
- $\boldsymbol{\mu}_k$: vector of means of group k; $\boldsymbol{\mu}$: vector of means of all data
- $\mathrm{SST}(\mathbf{X}) = \sum_i \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$: total variation
- $\mathrm{SSB}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|^2$: variation between groups
- $\mathrm{SSW}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$: (residual) variation within groups

K-means: underlying model

Extension to multivariate data ($\mathbf{X} \in \mathbb{R}^{n \times p}$):

$$\mathrm{SST}(\mathbf{X}) = \mathrm{SSB}(\mathbf{X}) + \mathrm{SSW}(\mathbf{X})$$

with:



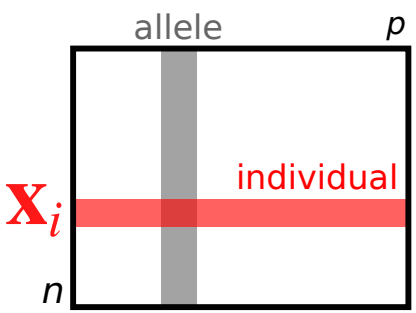
- $\boldsymbol{\mu}_k$: vector of means of group k; $\boldsymbol{\mu}$: vector of means of all data
- $\mathrm{SST}(\mathbf{X}) = \sum_i \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$: total variation
- $\mathrm{SSB}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|^2$: variation between groups
- $\mathrm{SSW}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$: (residual) variation within groups

K-means: underlying model

Extension to multivariate data ($\mathbf{X} \in \mathbb{R}^{n \times p}$):

$$\text{SST}(\mathbf{X}) = \text{SSB}(\mathbf{X}) + \text{SSW}(\mathbf{X})$$

with:



- $\boldsymbol{\mu}_k$: vector of means of group k; $\boldsymbol{\mu}$: vector of means of all data
- $\text{SST}(\mathbf{X}) = \sum_i \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$: total variation
- $\text{SSB}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|^2$: variation between groups
- $\text{SSW}(\mathbf{X}) = \sum_k \sum_{i \in g_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$: (residual) variation within groups

K-means rationale

Find K groups $\mathcal{G} = \{g_1, \ldots, g_k\}$ minimizing the sum of squares within-groups (SSW):

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

Note: this equates to finding groups maximizing the between-groups sum of squares.

K-means rationale

Find K groups $\mathcal{G} = \{g_1, \ldots, g_k\}$ minimizing the sum of squares within-groups (SSW):

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

Note: this equates to finding groups maximizing the between-groups sum of squares.

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. affect 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. affect 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:

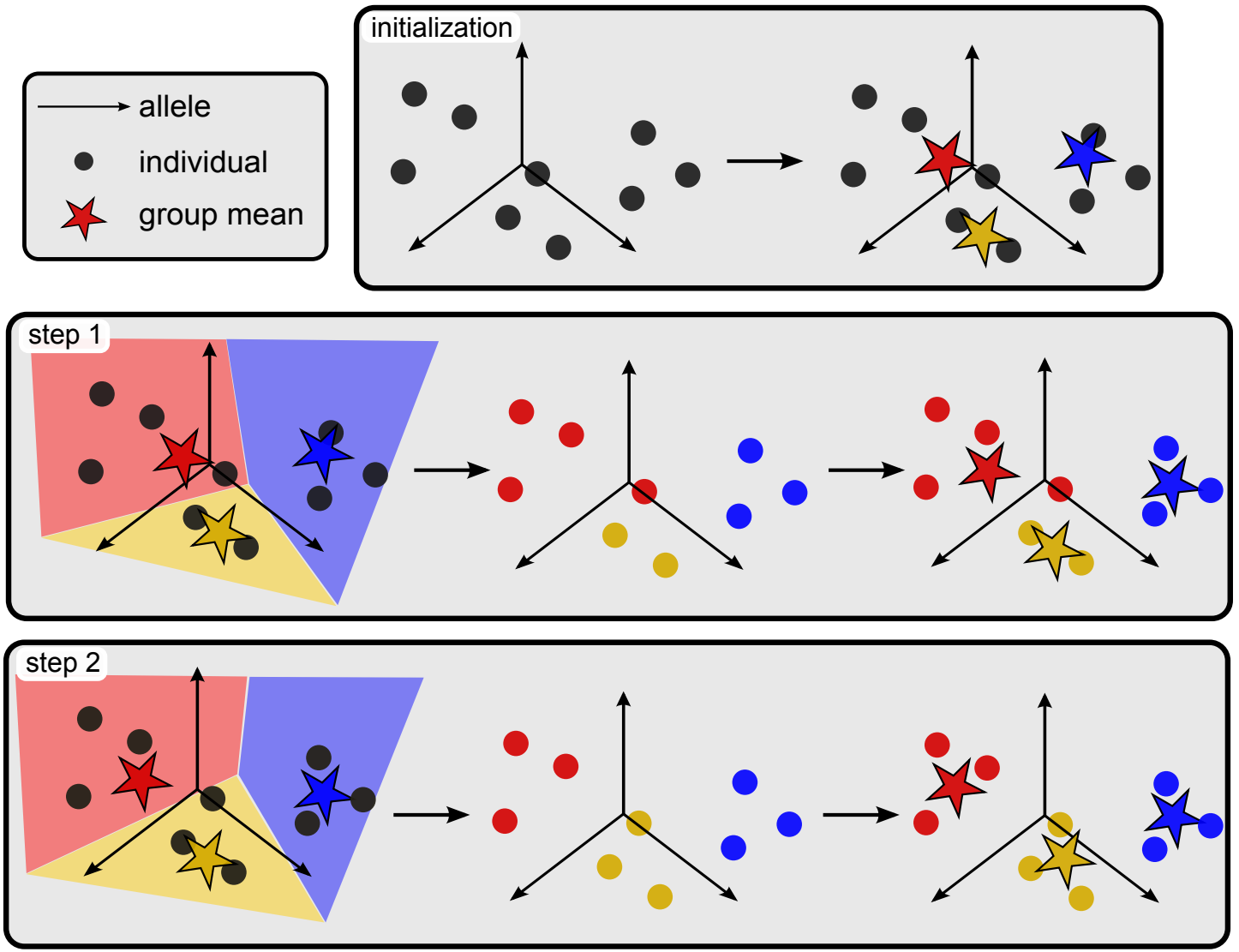
1. select random group means ($\boldsymbol{\mu}_k$, $k = 1, \ldots, K$)
2. affect each individual $\mathbf{x}_i$ to the closest group $\longrightarrow g_k$
3. update group means $\boldsymbol{\mu}_k$
4. go back to 2) until convergence (groups no longer change)

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. affect 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



Which K?

- K-means does not identify the number of clusters (K)
- each K-mean solution is a model with a likelihood
- model selection can be used to select K

Which K?

- K-means does not identify the number of clusters (K)
- each K-mean solution is a model with a likelihood
- model selection can be used to select K

Which K?

- K-means does not identify the number of clusters (K)
- each K-mean solution is a model with a likelihood
- model selection can be used to select K

Using Bayesian Information Criterion (BIC)

Defined as:

$$\text{BIC} = -2\log(\mathcal{L}) + k\log(n)$$

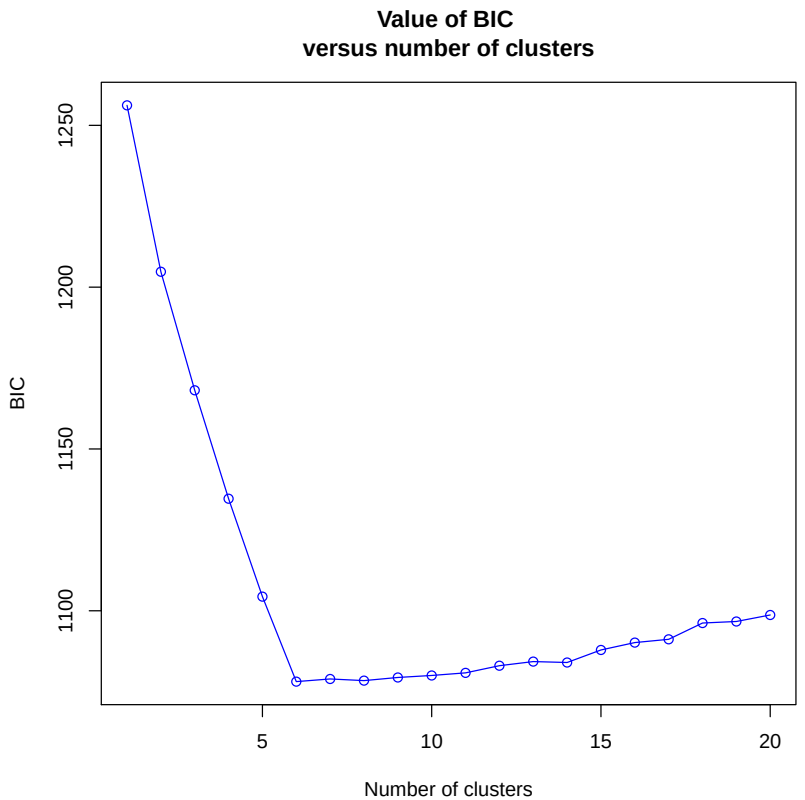
with:

- $\mathcal{L}$: likelihood
- k: number of parameters
- n: number of observations (individuals)

Smallest BIC = best model

K-means and BIC: example

Simulated data: 6 populations in island model

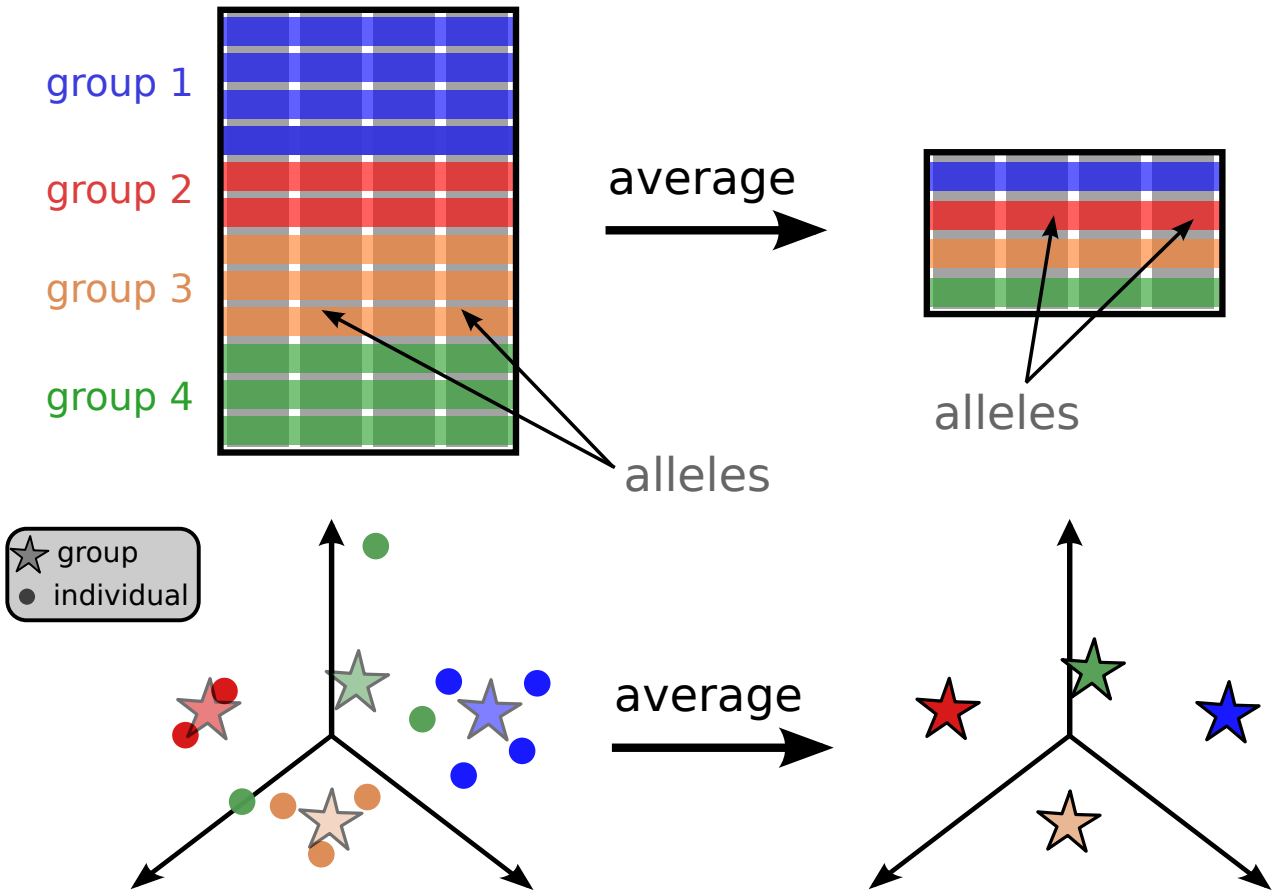


(Jombart et al. 2010, *BMC Genetics*)

Outline

Outline

Aggregating data by groups



$\longrightarrow$ multivariate analysis of group allele frequencies.

Analysing group data

Available methods:

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

Criticism:

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

Analysing group data

Available methods:

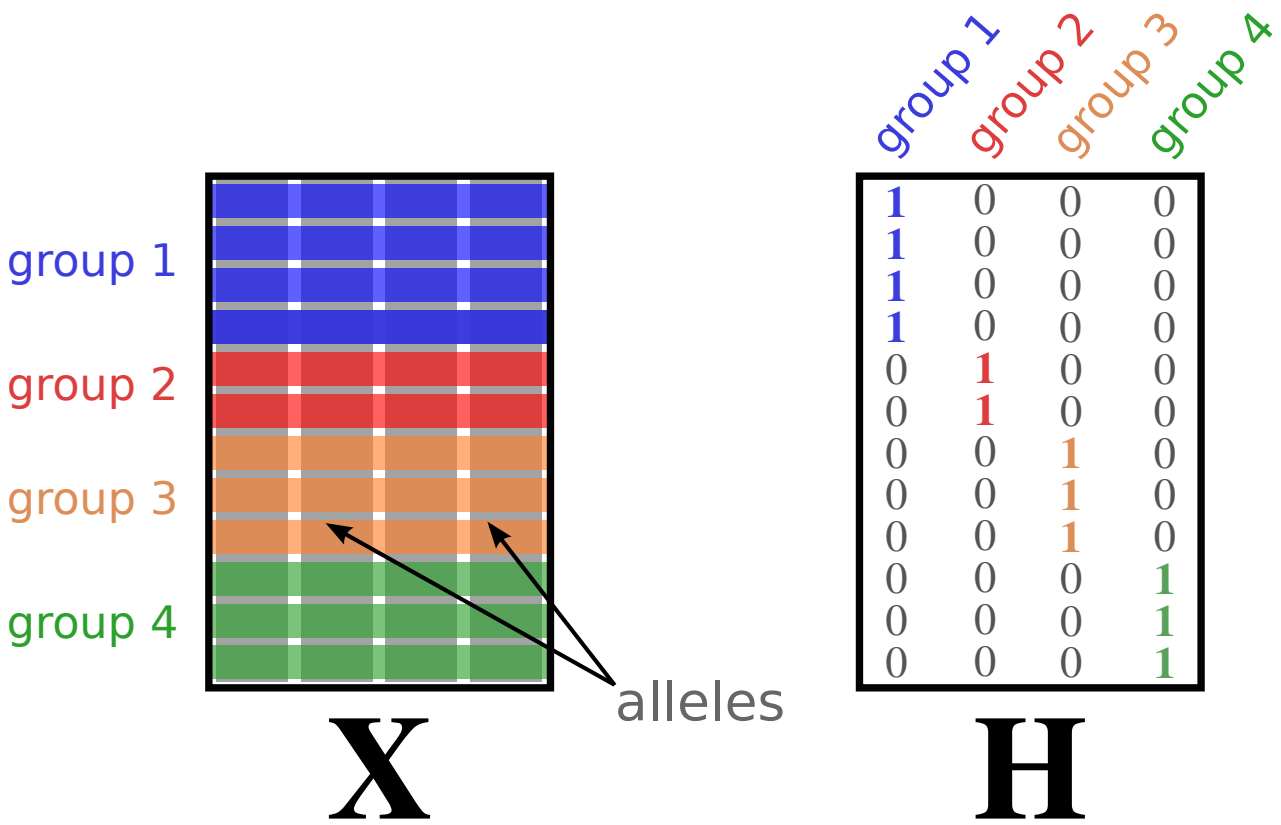
- Principal Component Analysis (PCA) of allele frequency table
- Genetic distance between populations $\longrightarrow$ Principal Coordinates Analysis (PCoA)
- Correspondance Analysis (CA) of allele counts

Criticism:

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

Outline

Using groups as partitions



Groups are coded as dummy vectors in $\mathbf{H}$.

The multivariate ANOVA model

The data matrix $\mathbf{X}$ can be decomposed as:

$$\mathbf{X} = \mathbf{PX} + (\mathbf{I} - \mathbf{P})\mathbf{X}$$

where:

- $\mathbf{P}$ is the projector onto $\mathbf{H}$: $\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{DH})^{-1})\mathbf{H}^T\mathbf{D}$, $\mathbf{D}$ being a metric in $\mathbb{R}^n$
- $\mathbf{PX}$ is a $n \times p$ matrix where each observation is replaced by the group average

The multivariate ANOVA model

Variation partition (same as K-means):

$$\text{VAR}(\mathbf{X}) = \text{B}(\mathbf{X}) + \text{W}(\mathbf{X})$$

where:

- $\text{VAR}(\mathbf{X}) = \text{trace}(\mathbf{X}^T\mathbf{D}\mathbf{X})$
- $\text{B}(\mathbf{X}) = \text{trace}(\mathbf{X}^T\mathbf{P}^T\mathbf{D}\mathbf{P}\mathbf{X})$
- $\text{W}(\mathbf{X}) = \text{trace}(\mathbf{X}^T(\mathbf{I}-\mathbf{P})^T\mathbf{D}(\mathbf{I}-\mathbf{P})\mathbf{X})$

Between-group analysis

$$\mathrm{VAR}(\mathbf{X}) = \mathrm{B}(\mathbf{X}) + \mathrm{W}(\mathbf{X})$$

Classical PCA:

- decompose $\mathrm{VAR}(\mathbf{X})$
- find $\mathbf{u}$ so that $\mathrm{var}(\mathbf{Xu})$ is maximum

Between-group PCA:

- decompose $\mathrm{B}(\mathbf{X})$
- find $\mathbf{u}$ so that $\mathrm{b}(\mathbf{Xu})$ is maximum

Between-group analysis

$$\text{VAR}(\mathbf{X}) = \text{B}(\mathbf{X}) + \text{W}(\mathbf{X})$$

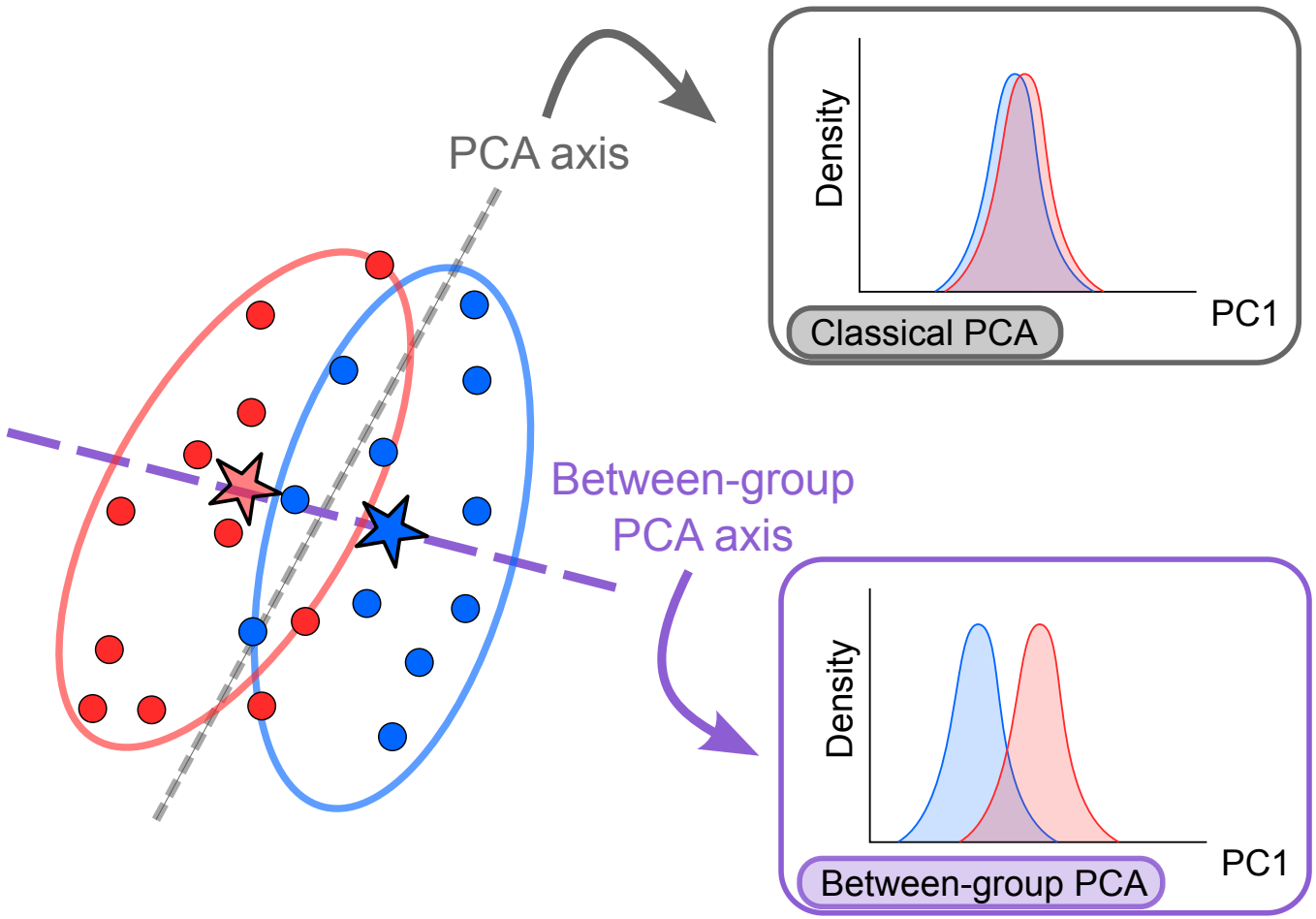
Classical PCA:

- decompose $\text{VAR}(\mathbf{X})$
- find $\mathbf{u}$ so that $\text{var}(\mathbf{Xu})$ is maximum

Between-group PCA:

- decompose $\text{B}(\mathbf{X})$
- find $\mathbf{u}$ so that $\text{b}(\mathbf{Xu})$ is maximum

Between-group PCA



Between-group PCA looks at between-group variability.

Outline

PCA, between-group PCA, and Discriminant Analysis

$$\text{VAR}(\mathbf{X}) = \text{B}(\mathbf{X}) + \text{W}(\mathbf{X})$$

Maximising different quantities:

- *PCA*: maximizes overall diversity (max var($\mathbf{Xu}$))
- *Between-group PCA*: maximizes group diversity (max b($\mathbf{Xu}$))
- *Discriminant Analysis*: maximizes group separation (max b($\mathbf{Xu}$), min w($\mathbf{Xu}$))

PCA, between-group PCA, and Discriminant Analysis

$$\text{VAR}(\mathbf{X}) = \text{B}(\mathbf{X}) + \text{W}(\mathbf{X})$$

Maximising different quantities:

- *PCA*: maximizes overall diversity (max var($\mathbf{Xu}$))
- *Between-group PCA*: maximizes group diversity (max b($\mathbf{Xu}$))
- *Discriminant Analysis*: maximizes group separation (max b($\mathbf{Xu}$), min w($\mathbf{Xu}$))

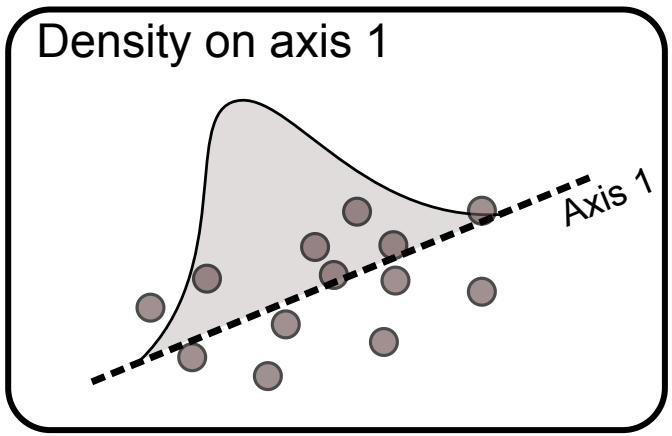
PCA, between-group PCA, and Discriminant Analysis

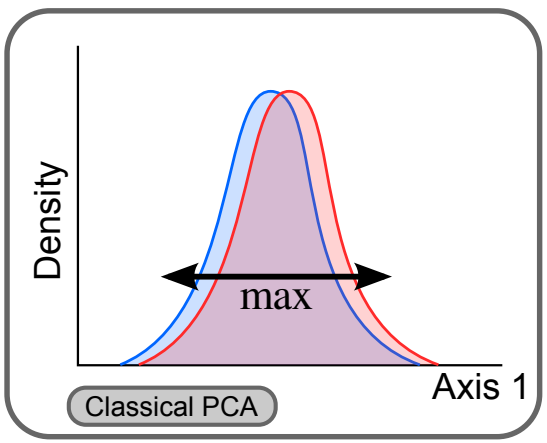
$$\text{VAR}(\mathbf{X}) = \text{B}(\mathbf{X}) + \text{W}(\mathbf{X})$$

Maximising different quantities:

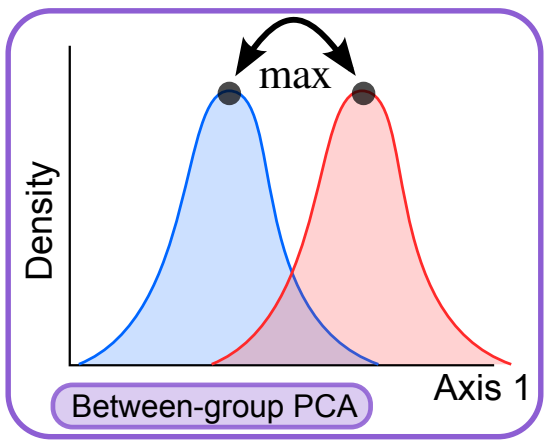
- *PCA*: maximizes overall diversity (max var$(\mathbf{Xu})$)
- *Between-group PCA*: maximizes group diversity (max b$(\mathbf{Xu})$)
- *Discriminant Analysis*: maximizes group separation (max b$(\mathbf{Xu})$, min w$(\mathbf{Xu})$)

Discriminant Analysis

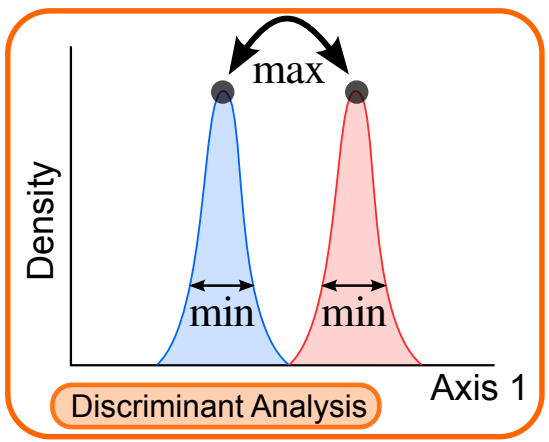




Max. total diversity



Max. diversity between groups



Max. separation of groups

Technical issues

Discriminant Analysis requires:

- $\mathbf{X}^T\mathbf{D}\mathbf{X}$ to be invertible $\Rightarrow$ less variables than observations
- $\mathbf{X}^T\mathbf{D}\mathbf{X}$ to be invertible $\Rightarrow$ uncorrelated variables

Genetic data:

- (almost) always (many) more alleles than individuals
- allele frequencies are by definition correlated ($\sum = 1$)
- linkage disequilibrium $\rightarrow$ correlated alleles

Technical issues

Discriminant Analysis requires:

- $\mathbf{X}^T\mathbf{D}\mathbf{X}$ to be invertible $\Rightarrow$ less variables than observations
- $\mathbf{X}^T\mathbf{D}\mathbf{X}$ to be invertible $\Rightarrow$ uncorrelated variables

Genetic data:

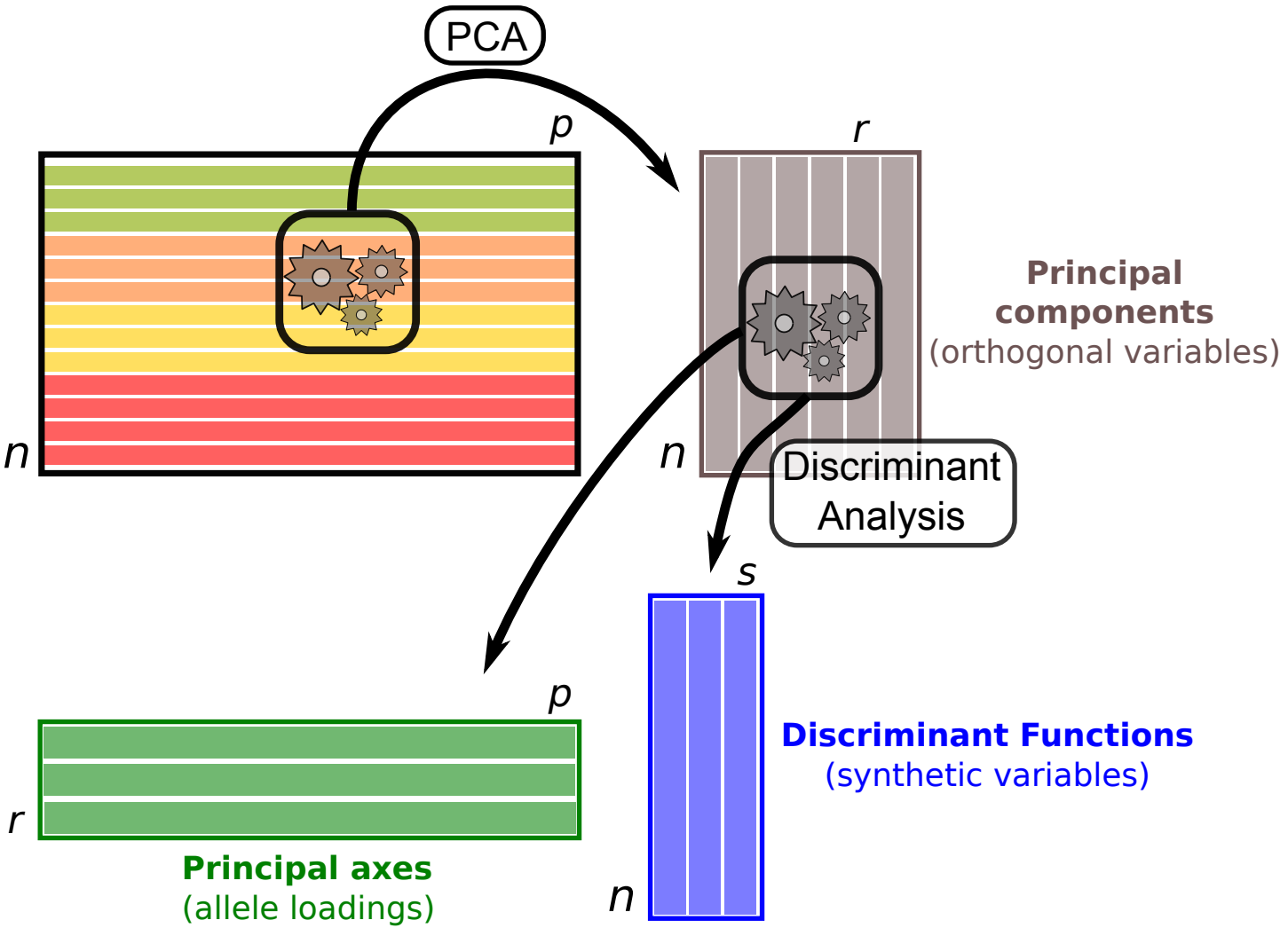
- (almost) always (many) more alleles than individuals
- allele frequencies are by definition correlated ($\sum = 1$)
- linkage disequilibrium $\rightarrow$ correlated alleles

Outline

Discriminant Analysis of Principal Components (DAPC)

- new method (Jombart et al. 2010, *BMC Genetics*)
- aim: modify DA for genetic data
- relies on data orthogonalisation/reduction using PCA

Rationale of DAPC



DAPC: summary

Discriminant Analysis requires:

- less variables than observations
- uncorrelated variables

Advantages of DAPC:

- always less PCs than observations
- PCs are uncorrelated
- still possible to compute allele contributions

DAPC: summary

Discriminant Analysis requires:

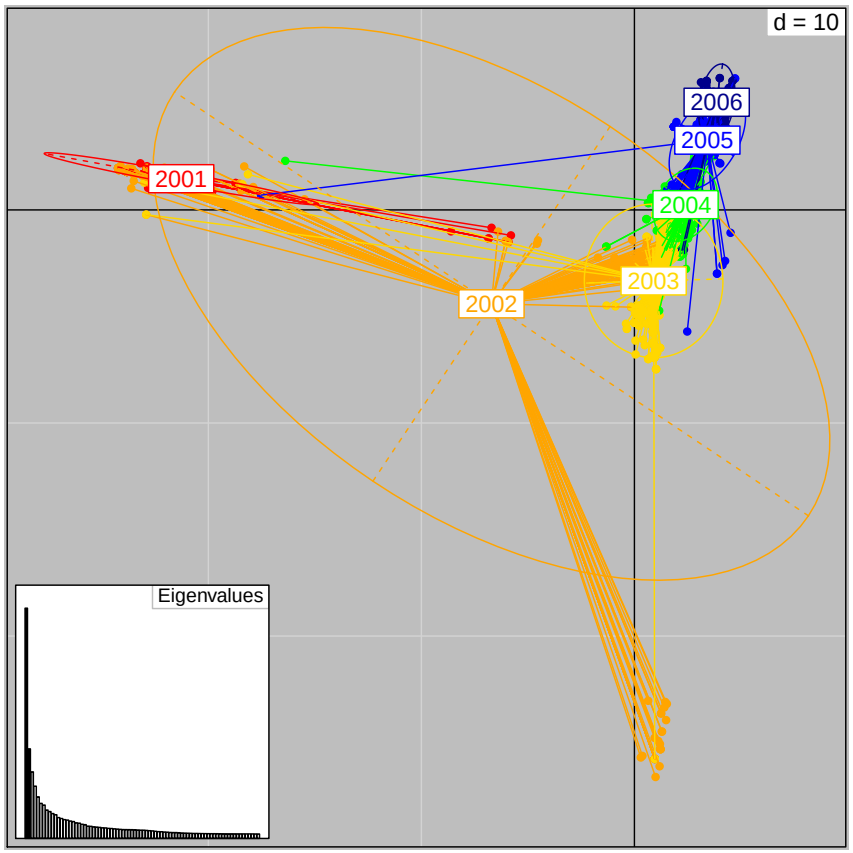
- less variables than observations
- uncorrelated variables

Advantages of DAPC:

- always less PCs than observations
- PCs are uncorrelated
- still possible to compute allele contributions

DAPC: example

Seasonal influenza (A/H3N2) data, PCA:



DAPC: example

Seasonal influenza (A/H3N2) data, DAPC: