Multivariate analysis of genetic data — uncovering spatial structures —

Thibaut Jombart

MRC Centre for Outbreak Analysis and Modelling
Imperial College London

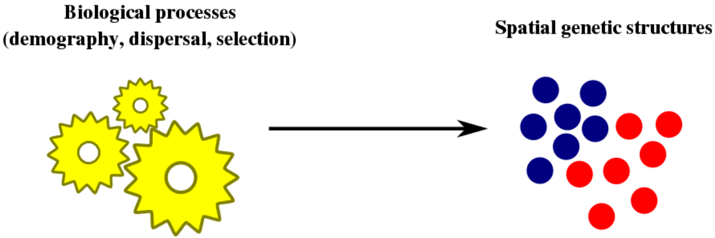
Genetic data analysis with R
PR∼Statistics, Glasgow
07-08-2015

Outline

Outline

From processes to structures

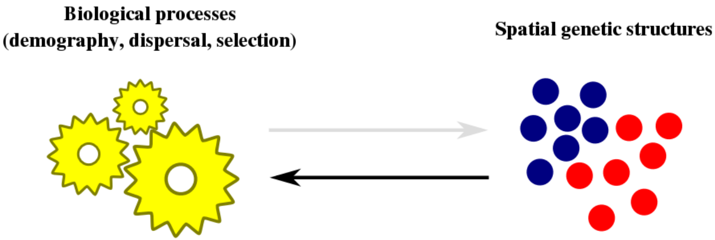
Genetic structure: non-random distribution of genetic diversity.



Identify structures to infer processes.

From processes to structures

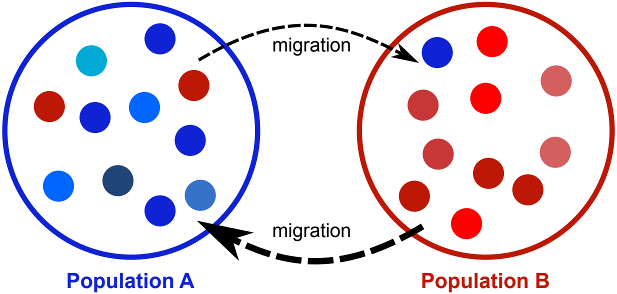
Genetic structure: non-random distribution of genetic diversity.



Identify structures to infer processes.

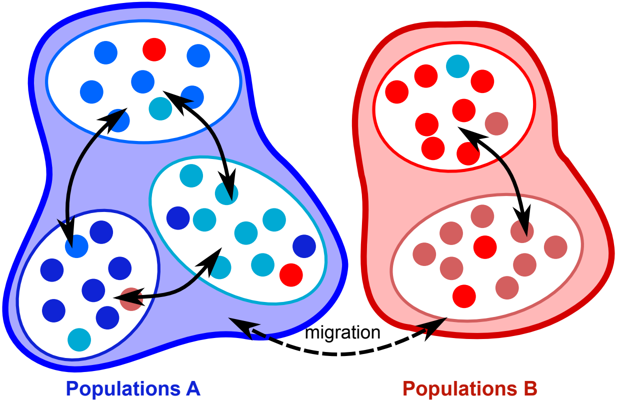
Island model

Reproduction within populations + migration.



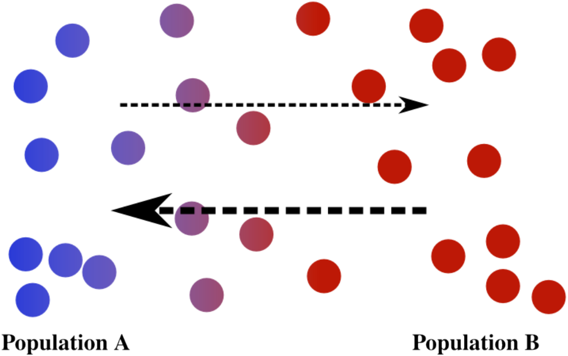
Hierarchical island model

Reproduction within subpopulations + stratified migration.



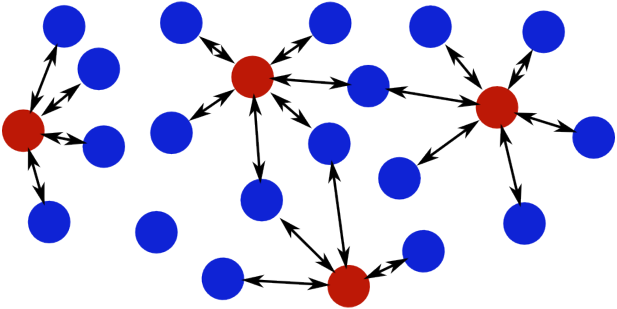
Isolation by distance (IBD)

Reproduction between neighbours → '*diffusion*' of genes



Inbreeding avoidance

Mating with individuals from another population → '*repulsion*' structure



Genetic models and spatial structures

- *island / hierarchical island model*: patches of related genotypes
- *isolation by distance (IBD)*: clines of genetic differentiation
- *inbreeding avoidance*: repulsion structure

⇒ Genetic processes often create spatial structures.
How can we reveal them?

Genetic models and spatial structures

- *island / hierarchical island model*: patches of related genotypes
- *isolation by distance (IBD)*: clines of genetic differentiation
- *inbreeding avoidance*: repulsion structure

⇒ Genetic processes often create spatial structures.
How can we reveal them?

Genetic models and spatial structures

- *island / hierarchical island model*: patches of related genotypes
- *isolation by distance (IBD)*: clines of genetic differentiation
- *inbreeding avoidance*: repulsion structure

⇒ Genetic processes often create spatial structures.
How can we reveal them?

Genetic models and spatial structures

- *island / hierarchical island model*: patches of related genotypes
- *isolation by distance (IBD)*: clines of genetic differentiation
- *inbreeding avoidance*: repulsion structure

$\Rightarrow$ Genetic processes often create spatial structures.
How can we reveal them?

Outline

Introduction

Testing spatial structures
- Moran's Index
- Mantel's correlation

Multivariate analysis of spatial patterns

Spatial autocorrelation

Definitions:

- *in general*: values of a variable non independent from the corresponding spatial locations
- *in genetics*: genetic distance is correlated to spatial distance

Two types of spatial autocorrelation:

- **positive**: closer individuals are more similar than at random
- **negative**: closer individuals are more dissimilar than at random

Spatial autocorrelation

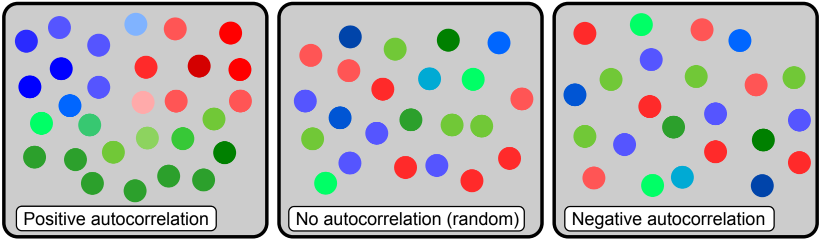
Definitions:

- *in general*: values of a variable non independent from the corresponding spatial locations
- *in genetics*: genetic distance is correlated to spatial distance

Two types of spatial autocorrelation:

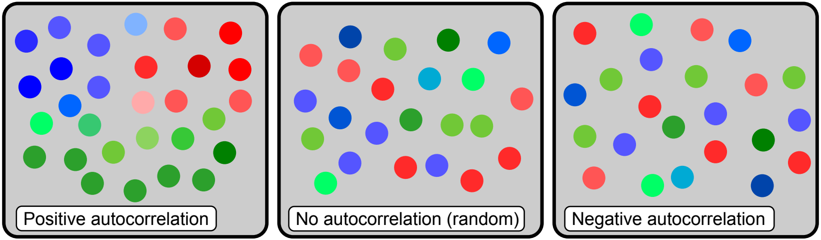
- **positive**: closer individuals are more similar than at random
- **negative**: closer individuals are more dissimilar than at random

Spatial autocorrelation: illustration



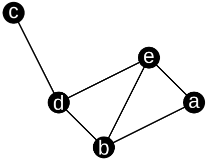
How do we measure spatial autocorrelation?

Spatial autocorrelation: illustration



How do we measure spatial autocorrelation?

From spatial coordinates to spatial weights



Matrix of spatial weights $\mathbf{L}$

Row i : uniform weights for neighbours of i.

	a	b	c	d	e
a	0.000	0.500	0.000	0.000	0.500
b	0.333	0.000	0.000	0.333	0.333
c	0.000	0.000	0.000	1.000	0.000
d	0.000	0.333	0.333	0.000	0.333
e	0.333	0.333	0.000	0.333	0.000

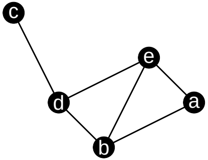
Let $\mathbf{x}$ be a variable with one value at each location.

The lag vector $\mathbf{Lx}$ computes mean values of neighbours.

From spatial coordinates to spatial weights

Matrix of spatial weights $\mathbf{L}$

Row i : uniform weights for neighbours of i.

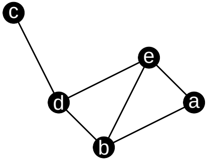


	a	b	c	d	e
a	0.000	0.500	0.000	0.000	0.500
b	0.333	0.000	0.000	0.333	0.333
c	0.000	0.000	0.000	1.000	0.000
d	0.000	0.333	0.333	0.000	0.333
e	0.333	0.333	0.000	0.333	0.000

Let $\mathbf{x}$ be a variable with one value at each location.

The lag vector $\mathbf{Lx}$ computes mean values of neighbours.

From spatial coordinates to spatial weights



Matrix of spatial weights $\mathbf{L}$

Row i : uniform weights for neighbours of i.

	a	b	c	d	e
a	0.000	0.500	0.000	0.000	0.500
b	0.333	0.000	0.000	0.333	0.333
c	0.000	0.000	0.000	1.000	0.000
d	0.000	0.333	0.333	0.000	0.333
e	0.333	0.333	0.000	0.333	0.000

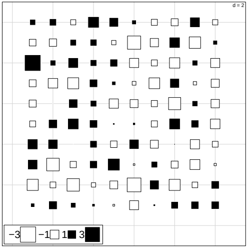
Let $\mathbf{x}$ be a variable with one value at each location.

The lag vector $\mathbf{Lx}$ computes mean values of neighbours.

A variable and its lag-vector

Lag vector :

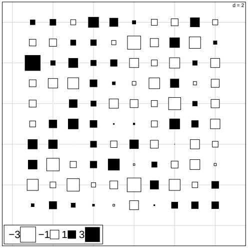
Random:



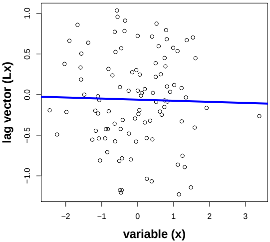
Regression of $\mathbf{Lx}$ onto $\mathbf{x}$:

A variable and its lag-vector

Random:



Lag vector :



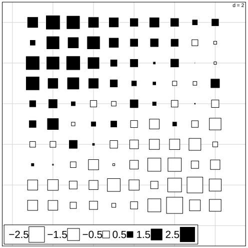
Regression of $\mathbf{Lx}$ onto $\mathbf{x}$:

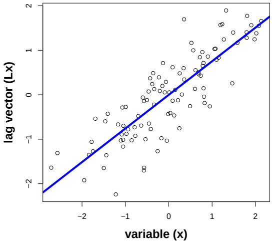
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	0.02	0.02	0.06	0.8081
Residuals	98	31.53	0.32		

A variable and its lag-vector

Lag vector :

Positive autocorrelation:





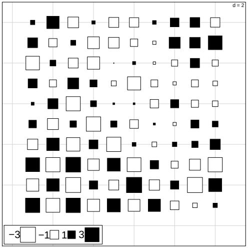
Regression of $\mathbf{Lx}$ onto $\mathbf{x}$:

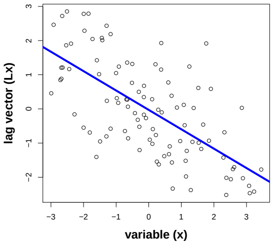
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
xG	1	65.91	65.91	245.69	0.0000
Residuals	98	26.29	0.27		

A variable and its lag-vector

Lag vector :

Negative autocorrelation:





Regression of $\mathbf{Lx}$ onto $\mathbf{x}$:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
xL	1	87.56	87.56	77.80	0.0000
Residuals	98	110.29	1.13		

Moran's index: definition

Moran's I:

$$I(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{L}\mathbf{x}}{n}\frac{1}{\text{var}(\mathbf{x})}$$

where:

- $\mathbf{x} \in \mathbb{R}^n$: a centred variable (e.g. allele frequency, PC)
- $\mathbf{L}$: matrix of spatial weights (nxn)
- $\mathbf{Lx}$: lag vector
- $I_0 = \frac{-1}{n-1} \approx 0$: null value (no autocorrelation, i.e. random spatial distribution)

$\Rightarrow$ Moran's I varies like $\langle \mathbf{x}, \mathbf{Lx} \rangle$.

Moran's index: definition

Moran's I:

$$I(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{L} \mathbf{x}}{n} \frac{1}{\text{var}(\mathbf{x})}$$

where:

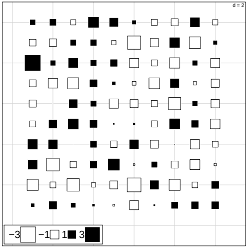
- $\mathbf{x} \in \mathbb{R}^n$: a centred variable (e.g. allele frequency, PC)
- $\mathbf{L}$: matrix of spatial weights (nxn)
- $\mathbf{Lx}$: lag vector
- $I_0 = \frac{-1}{n-1} \approx 0$: null value (no autocorrelation, i.e. random spatial distribution)

$\Rightarrow$ Moran's I varies like $\langle \mathbf{x}, \mathbf{Lx} \rangle$.

Variable, lag-vector, Moran's I

Lag vector :

Random:

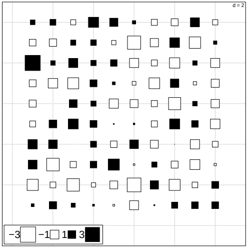


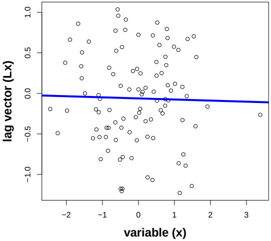
Moran's I:

Variable, lag-vector, Moran's I

Lag vector :

Random:



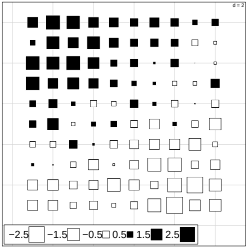


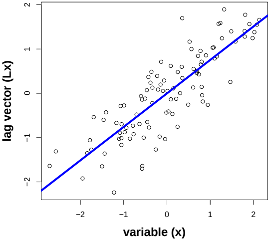
Moran's I:
$I(\mathbf{x}) \approx I_0$

Variable, lag-vector, Moran's I

Lag vector :

Positive autocorrelation:



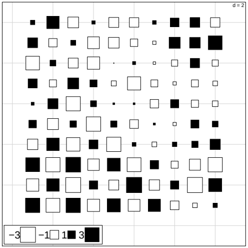


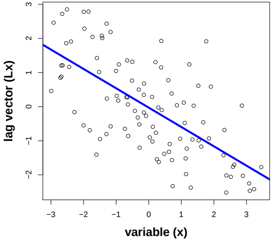
Moran's I:
$I(\mathbf{x}) > I_0$

Variable, lag-vector, Moran's I

Lag vector :

Negative autocorrelation:





Moran's I:

$I(\mathbf{x}) < I_0$

Testing Moran's I

Monte Carlo procedure:

- compute I from the data
- permute randomly the locations to get a value of I under H_0: "$\mathbf{x}$ is distributed at random across space."
- repeat this operation a large number of times to obtain a reference distribution of I under H_0
- compare initial value to the reference distribution to get a p-value.

Testing Moran's I

Monte Carlo procedure:

- compute I from the data
- permute randomly the locations to get a value of I under H_0: "$\mathbf{x}$ is distributed at random across space."
- repeat this operation a large number of times to obtain a reference distribution of I under H_0
- compare initial value to the reference distribution to get a p-value.

Testing Moran's I

Monte Carlo procedure:

- compute I from the data
- permute randomly the locations to get a value of I under H_0: "$\mathbf{x}$ is distributed at random across space."
- repeat this operation a large number of times to obtain a reference distribution of I under H_0
- compare initial value to the reference distribution to get a p-value.

Testing Moran's I

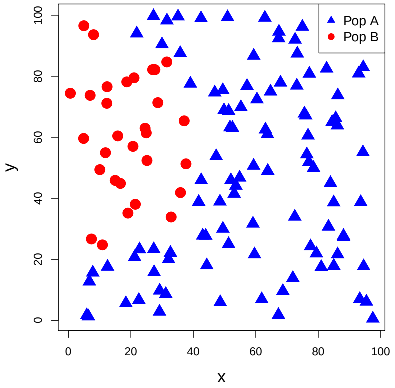
Monte Carlo procedure:

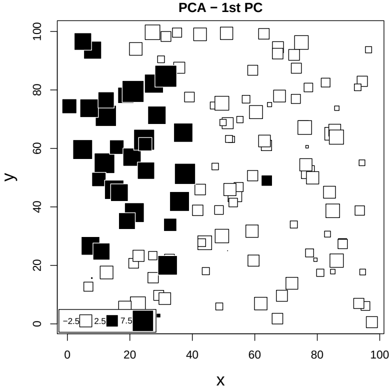
- compute I from the data
- permute randomly the locations to get a value of I under H_0: "$\mathbf{x}$ is distributed at random across space."
- repeat this operation a large number of times to obtain a reference distribution of I under H_0
- compare initial value to the reference distribution to get a p-value.

Application: testing spatial structures in principal components

Data (2 population, island model):

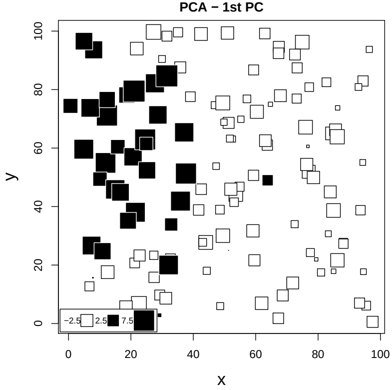
PCA results, PC 1:



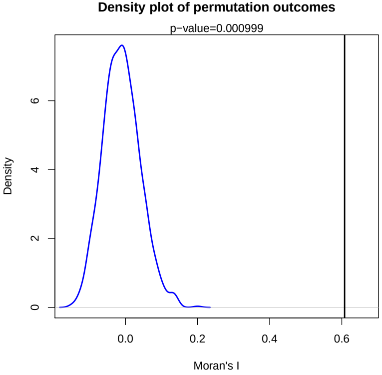


Application: testing spatial structures in principal components

PCA results, PC 1:



Moran's I test of PC1:



Univariate /vs/ multivariate correlation

- Moran's I is univariate
- *solution*: test a few principal components
- *problems:*
    - does not use all the genetic information
    - which PC to test?
    - correction for multiple testing

⇒ need for multivariate tests

Univariate /vs/ multivariate correlation

- Moran's I is univariate
- *solution*: test a few principal components
- *problems:*
    - does not use all the genetic information
    - which PC to test?
    - correction for multiple testing

⇒ need for multivariate tests

Univariate /vs/ multivariate correlation

- Moran's I is univariate
- *solution*: test a few principal components
- *problems:*
    - does not use all the genetic information
    - which PC to test?
    - correction for multiple testing

⇒ need for multivariate tests

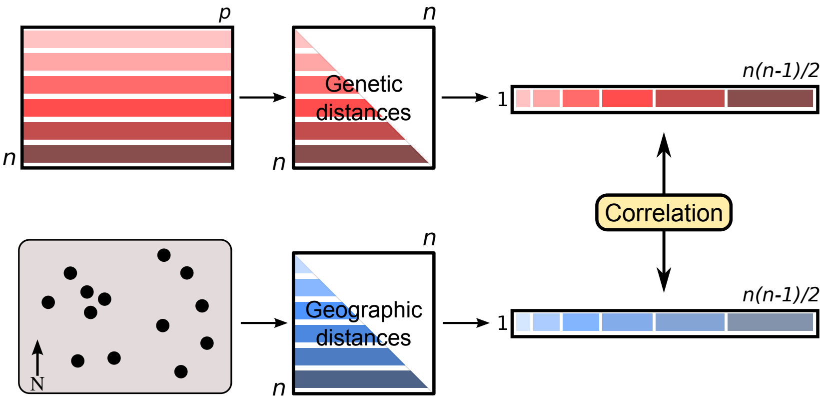
Univariate /vs/ multivariate correlation

- Moran's I is univariate
- *solution*: test a few principal components
- *problems:*
    - does not use all the genetic information
    - which PC to test?
    - correction for multiple testing

$\Rightarrow$ need for multivariate tests

Mantel's correlation: rationale

Correlation between two unfolded distance matrices.



Mantel's correlation: definition

Notations:

- $\mathbf{X} = [x_{ij}]$ ($\mathbf{X} \in \mathbb{R}^{n \times n}$): genetic distances
- $\mathbf{Y} = [y_{ij}]$ ($\mathbf{Y} \in \mathbb{R}^{n \times n}$): geographic distances
- $\bar{x}$, $\bar{y}$: means of x and y (excepting diagonals)
- s_x, s_y: standard deviation of x and y (excepting diagonals)

Original definition (unstandardized):

$$z_{\mathsf{M}} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_{ij} y_{ij}$$

Mantel's correlation: definition

Notations:

- $\mathbf{X} = [x_{ij}]$ ($\mathbf{X} \in \mathbb{R}^{n \times n}$): genetic distances
- $\mathbf{Y} = [y_{ij}]$ ($\mathbf{Y} \in \mathbb{R}^{n \times n}$): geographic distances
- $\bar{x}$, $\bar{y}$: means of x and y (excepting diagonals)
- s_x, s_y: standard deviation of x and y (excepting diagonals)

Standardized coefficient (true correlation):

$$r_{\mathsf{M}} = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (\frac{x_{ij} - \bar{x}}{s_x})(\frac{y_{ij} - \bar{y}}{s_y})$$

Testing Mantel correlation

Monte Carlo procedure:

- compute z_{M} or r_{M} from the data
- permute randomly the rows and columns of one matrix, recompute the test statistic (i.e., under H_0: "no correlation")
- repeat this operation many times to generate a reference distribution
- compare initial value to the reference distribution to get a p-value.

Testing Mantel correlation

Monte Carlo procedure:

- compute z_{M} or r_{M} from the data
- permute randomly the rows and columns of one matrix, recompute the test statistic (i.e., under H_0: "no correlation")
- repeat this operation many times to generate a reference distribution
- compare initial value to the reference distribution to get a p-value.

Testing Mantel correlation

Monte Carlo procedure:

- compute z_M or r_M from the data
- permute randomly the rows and columns of one matrix, recompute the test statistic (i.e., under H_0: "no correlation")
- repeat this operation many times to generate a reference distribution
- compare initial value to the reference distribution to get a p-value.

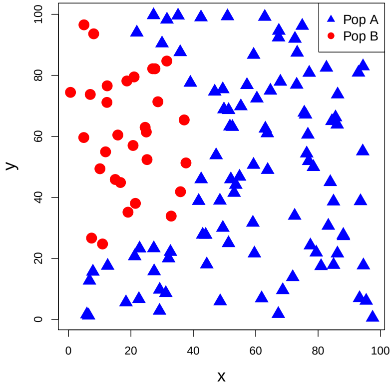
Testing Mantel correlation

Monte Carlo procedure:

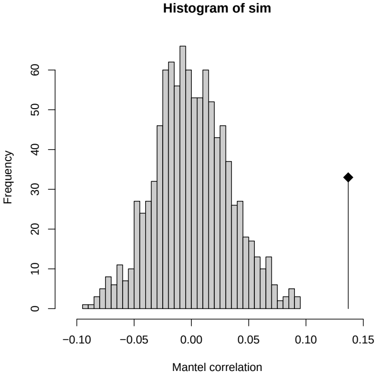
- compute z_{M} or r_{M} from the data
- permute randomly the rows and columns of one matrix, recompute the test statistic (i.e., under H_0: "no correlation")
- repeat this operation many times to generate a reference distribution
- compare initial value to the reference distribution to get a p-value.

Application: testing spatial structures

Data (2 population, island model):



Mantel test:



Outline

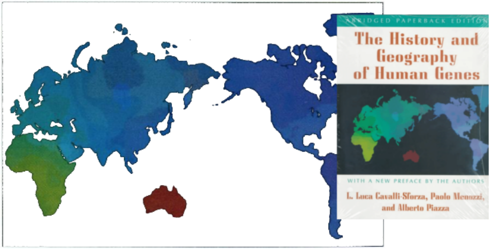
Introduction

Testing spatial structures
- Moran's Index
- Mantel's correlation

Multivariate analysis of spatial patterns

Mapping principal components

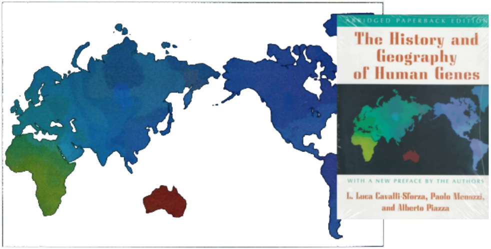
Maps of the three first principal components of PCA.



Are we actually looking for spatial patterns here?

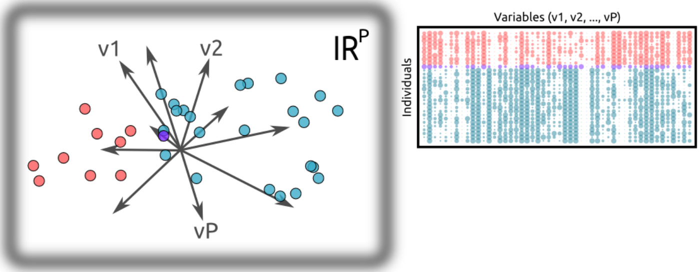
Mapping principal components

Maps of the three first principal components of PCA.



Are we actually looking for spatial patterns here?

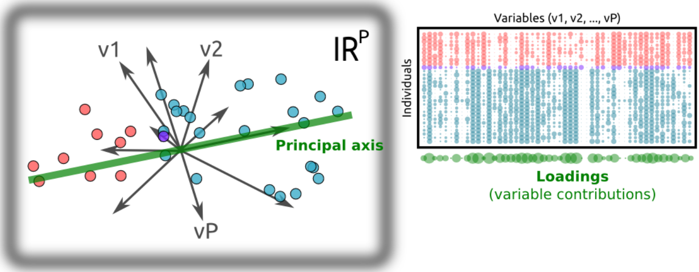
Multivariate analysis: reminder



Principal components with *maximum total variance*.

⇒ Spatial information is not taken into account.

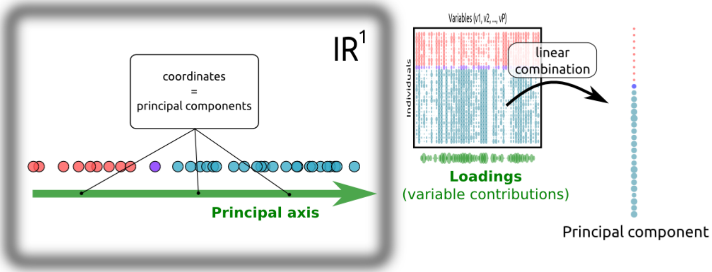
Multivariate analysis: reminder



Principal components with *maximum total variance*.

⇒ Spatial information is not taken into account.

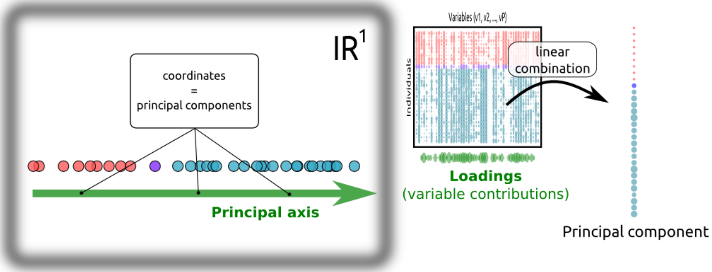
Multivariate analysis: reminder



Principal components with *maximum total variance*.

⇒ Spatial information is not taken into account.

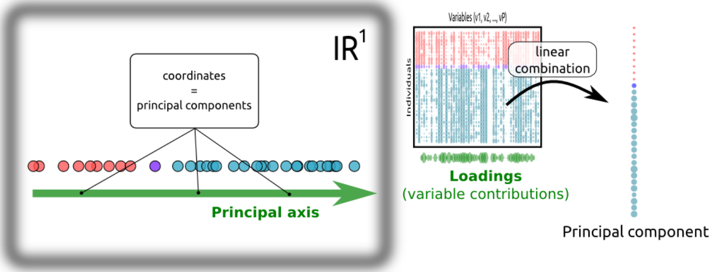
Multivariate analysis: reminder



Principal components with *maximum total variance*.

⇒ Spatial information is not taken into account.

Multivariate analysis: reminder



Principal components with *maximum total variance*.

⇒ Spatial information is not taken into account.

Using spatial information

- usual multivariate analyses ignore spatial information
- they may reveal obvious spatial structures, but overlook finer patterns

  ⇒ need for taking spatial information into account

Using spatial information

- usual multivariate analyses ignore spatial information
- they may reveal obvious spatial structures, but overlook finer patterns

$\Rightarrow$ need for taking spatial information into account

Using spatial information

- usual multivariate analyses ignore spatial information
- they may reveal obvious spatial structures, but overlook finer patterns

  $\Rightarrow$ need for taking spatial information into account

Spatial Principal Component Analysis (sPCA): rationale

Principal components should:

- display variability ⇒ optimize *total variance*
- display positive autocorrelation ⇒ *large Moran's I*
- (or) display negative autocorrelation ⇒ *low (negative) Moran's I*

sPCA decomposes: $(\textit{total variance}) \times (\textit{Moran's } I)$

Spatial Principal Component Analysis (sPCA): rationale

Principal components should:

- display variability ⇒ optimize *total variance*
- display positive autocorrelation ⇒ *large Moran's I*
- (or) display negative autocorrelation ⇒ *low (negative) Moran's I*

sPCA decomposes: $(\textit{total variance}) \times (\textit{Moran's } I)$

Spatial Principal Component Analysis (sPCA): rationale

Principal components should:

- display variability ⇒ optimize *total variance*
- display positive autocorrelation ⇒ *large Moran's I*
- (or) display negative autocorrelation ⇒ *low (negative) Moran's I*

sPCA decomposes: (*total variance*) × (*Moran's I*)

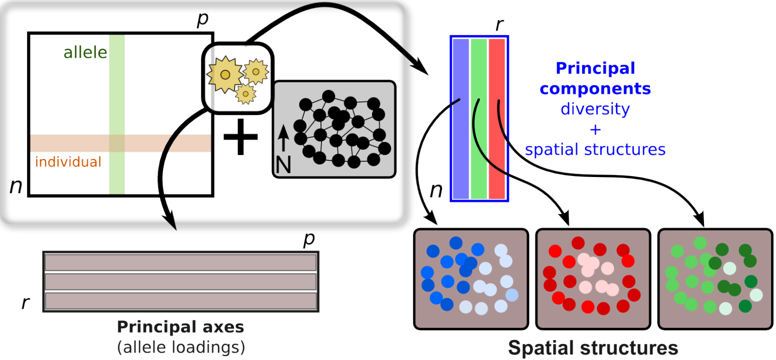
Spatial Principal Component Analysis (sPCA): rationale

Principal components should:

- display variability ⇒ optimize *total variance*
- display positive autocorrelation ⇒ *large Moran's I*
- (or) display negative autocorrelation ⇒ *low (negative) Moran's I*

sPCA decomposes: (*total variance*) × (*Moran's I*)

Spatial Principal Component Analysis (sPCA): outputs



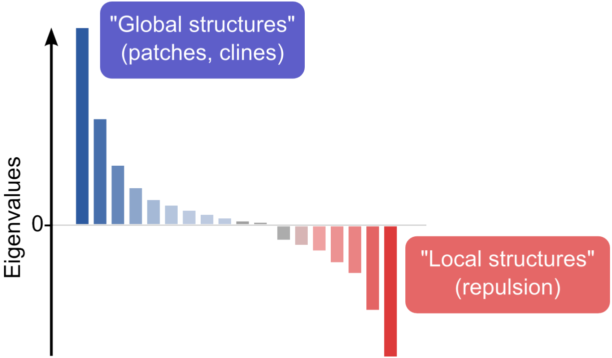
Principal components
diversity
+
spatial structures

Principal axes
(allele loadings)

Spatial structures

Global and local structures

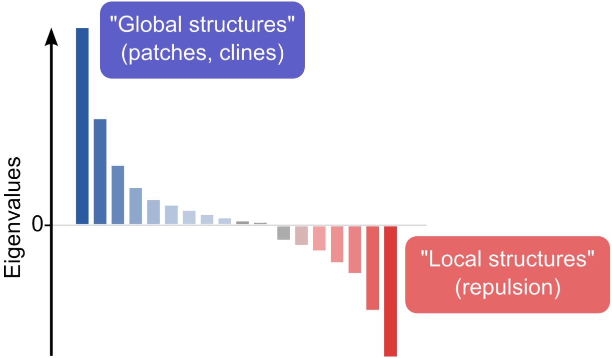
Unlike other multivariate methods, sPCA has **positive** and **negative** eigenvalues



How do we get these in practice?

Global and local structures

Unlike other multivariate methods, sPCA has **positive** and **negative** eigenvalues



How do we get these in practice?

Time to get your hands dirty (one last time)!

The pdf of the practical is online:

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

or

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