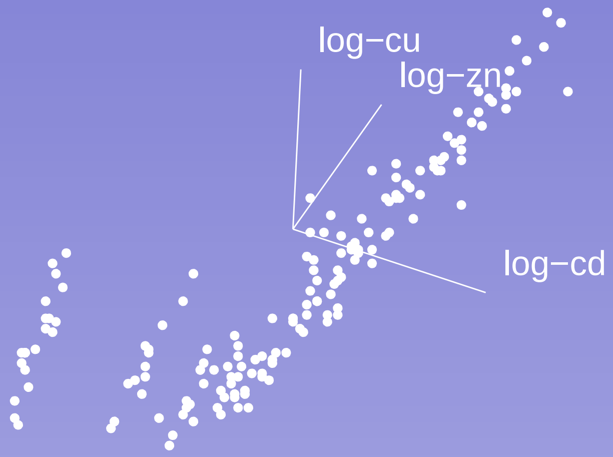
Spatial Analysis and GIS 2

Edzer J. Pebesma

November 2005



Technical details

- SAGIS2 has two main aspects: (i) Multivariate Analysis, (ii) Geostatistics
- assignment: 10%; tests: (i) 40%, (ii) 50%, all compulsory
- study guide: `http://www.geog.uu.nl/~pebesma/sagis2/`
- computer classes: `http://webct.uu.nl/`
- reader: *verkoopruimte*
- teaching assistant: Hanneke Schuurmans

What is multivariate analysis?

joint analysis of multiple variables, in relation to (i) a dependent variable (ii) each other.

supervised :

- *prediction* of a single variable from a set of predictor variables
- one dependent, multiple independent
- simple regression analysis → multiple regression analysis
- statistical learning

unsupervised :

- *simultaneous analysis* of multiple variables
- what is the (common) story
- what is their (cor)relation, interaction

What is geostatistics

- prediction, not (only) under a given condition, but at a specific spatial location
- spatial correlation plays a (lead) role
- naturally extends (multiple) regression models
- univariable; multivariable extends unsupervised multivariate analysis

Multivariate analysis

- matrix algebra
- multiple regression
- ordination techniques:
  - ⋆ principal component analysis
  - ⋆ (factor analysis)
  - ⋆ (correspondence analysis)
- clustering and classification:
  - ⋆ discriminant and canonical analysis
  - ⋆ cluster analysis

Goals of multivariate analysis

unsupervised : data reduction, finding groups

supervised : predicting values; predicting class membership

general : finding patterns, stories, exploring hypothesis

Why matrix algebra?

1. multivariate data are easily expressed as matrices

2. dimension “disappears”

3. geometric interpretation

Vectors and spaces

a vector has a length and a direction (coordinates).

length: $|\vec{a}| = \sqrt{a_1^2 + a_2^2 + ... + a_n^2}$

addition: $\vec{c} = \vec{a} + \vec{b}$:

scalar multiplication: $\vec{c} = 5\vec{a} = (5a_1, ..., 5a_n)$,

inner product: $\vec{a} \cdot \vec{b} = a_1b_1 + a_2b_2 + ... + a_nb_n$ (dot product, scalar product): scalar; length of one vector projected on the other, times the length of the other:

$$\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos\phi.$$

angle ϕ between $\vec{a}$ and $\vec{b}$: $\phi = cos^{-1}\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}$

outer product: $\vec{a}\times\vec{b}$ (vector product): vector; perpendicular to the surface formed by $\vec{a}$ and $\vec{b}$, such that ($\vec{a}$, $\vec{b}$ and $\vec{a}\times\vec{b}$) for a right-handed set Length: $|\vec{a}||\vec{b}|\sin\phi$.

$\vec{a}$ and $\vec{b}$ in the same direction: if $\vec{a}\cdot\vec{b} = |\vec{a}||\vec{b}|$ (Vectors exactly in the same or opposite direction are called *dependent*.)

$\vec{a}$ and $\vec{b}$ orthogonal (perpendicular) if $\vec{a}\cdot\vec{b} = 0$

$\vec{a}$ and $\vec{b}$ orthonormal: when $\vec{a}\cdot\vec{b} = 0$, and $|\vec{a}| = |\vec{b}| = 1$.

How do we define a space?

Minimum requirement for n dimensions: n independent, non-zero vectors

Say, a basis is formed by n vectors $\{a_1, a_2, ..., a_n\}$, then any point in the space spanned by these vectors can be expressed as $(\lambda_1 a_1, \lambda_2 a_2, ..., \lambda_n a_n)$

- orthonormal basis 2D: $(1, 0), (0, 1)$
- non-orthonormal basis 2D: $(1, 0), (1, 1)$
- invalid basis 2D: $(1, 0), (2, 0)$

What is a matrix?

Square pattern (table) of numbers:

$$\begin{bmatrix} 2 & 1 & 3 \\ 0 & -1 & 9 \end{bmatrix}$$

Rows: $m = 2$, columns: $n = 3$

First row: *top*; first column: *left*

an $m \times n$ (2×3) matrix

Special cases:

row-vector $m = 1$:

$$a_{1,.} = \begin{bmatrix} 2 & 1 & 3 \end{bmatrix}$$

column-vector $n = 1$:

$$a_{.,2} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

scalar $m = n = 1$: $[3]$ or just: $a_{1,3} = 3$

transposed : $a_{i,j} \rightarrow a_{j,i}$ (rows and columns exchanged). Symbol: A^T or A'.

$$\begin{bmatrix} 2 & 1 & 3 \\ 0 & -1 & 9 \end{bmatrix}' = \begin{bmatrix} 2 & 0 \\ 1 & -1 \\ 3 & 9 \end{bmatrix}$$

square matrix $m = n$:

$$\begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix}$$

symmetric matrix $m = n$, $a_{i,j} = a_{j,i}$ (square, and $A^T = A$):

$$\begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix}$$

diagonal matrix $a_{i,j} = 0$ for each i, j where $i \neq j$ (always square and symmetric)

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

identity matrix I $a_{i,j} = 0$ for each i, j where $i \neq j$, diagonal elements have value

1 (special case of diagonal matrix)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

null matrix $a_{i,j} = 0$, for each i, j

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Why matrix algebra?

Notation compact, structured

Abstraction *structure* of calculations arises, independent of dimensions

Data matrix question forms, soil samples, “boorformulieren” etc.

Geometry volumes, distances, etc.

Practice computer languages (matlab, octave, mathematica, S (S-Plus, R), ... ; programming libraries

```
A = B * C
```

Element-wise matrix operations

addition $A = B + C$:

$$\begin{bmatrix} 1 & 3 & 3 \\ 2 & 3 & 0 \\ 5 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ 2 & 2 & 0 \\ 5 & 3 & 1 \end{bmatrix}$$

subtraction $A = B - C$:

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 \\ 2 & 3 & 0 \\ 5 & 3 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ 2 & 2 & 0 \\ 5 & 3 & 1 \end{bmatrix}$$

scalar multiplication $A = cB$:

$$2\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Multiplication: the matrix product

$$\begin{array}{rcrcr} 2x_1 & + & 3x_2 & = & 6 \\ 4x_1 & + & x_2 & = & 12 \end{array}$$

$$\begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 12 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \times x_1 + 3 \times x_2 = 6$$

$$\begin{bmatrix} 2 & 3 \\ \overrightarrow{4 \quad 1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\big\downarrow = \begin{bmatrix} 6 \\ 12 \end{bmatrix}$$

$BC \neq CB$:

$$\begin{bmatrix} 0 & 1 & 3 \\ 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 5 & 5 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 3 \\ 2 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 4 & 3 & 1 \\ 2 & 3 & 5 \\ 0 & -1 & -3 \end{bmatrix}$$

Rules for matrix multiplication

1. $B = C \Rightarrow AB = AC$ en $BA = CA$

2. $(A + B)C = AC + BC$ en $C(A + B) = CA + CB$

3. $(AB)C = A(BC)$

4. $AI = IA = A$

5. $(AB)' = B'A'$ [from which follows: $A'A = (A'A)'$]

Systems of equations $Ax = b$

$$\begin{array}{rcrcrcr} 2x_1 & + & x_2 & + & 3x_3 & = & 6 \\ x_1 & + & 3x_2 & + & 3x_3 & = & 12 \\ 2x_1 & - & x_2 & & & = & -3 \end{array}$$

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 12 \\ -3 \end{bmatrix}$$

or: $Ax = b$

Approach:

1. zero lower left triangle

2. back substition

zero element $(3, 1)$, subtract line 1 from line 3:

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 0 & -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 12 \\ -9 \end{bmatrix}$$

zero element $(2, 1)$, subtract line 1 from (2 × line 2):

$$\begin{bmatrix} 2 & 1 & 3 \\ 0 & 5 & 3 \\ 0 & -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 18 \\ -9 \end{bmatrix}$$

zero element (3,2), add (2 × line 2) to (5 × line 3):

$$\begin{bmatrix} 2 & 1 & 3 \\ 0 & 5 & 3 \\ 0 & 0 & -9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 18 \\ -9 \end{bmatrix}$$

$-9x_3 = -9 : \ x_3 = 1.$

Substitution in 2 yields: $5x_2 + 3 = 18 : \ x_2 = 3$

Substitution in 1 yields: $2x_1 + 3 + 3 = 6 : \ x_1 = 0.$

Multiple systems of equations

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} 6 & 2 \\ 12 & 1 \\ -3 & 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 0 & -2 & -3 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} 6 & 2 \\ 12 & 1 \\ -9 & 2 \end{bmatrix}$$

Matrix inversion: $AA^{-1} = I$

definition matrix inversion:

$$AX = I \Leftrightarrow X = A^{-1}$$

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 2 & -1 & 0 \end{bmatrix} X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\left[\begin{array}{ccc|ccc} 2 & 1 & 3 & 1 & 0 & 0 \\ 1 & 3 & 3 & 0 & 1 & 0 \\ 2 & -1 & 0 & 0 & 0 & 1 \end{array}\right]$$

subtract line 1 from line 3:

$$\left[\begin{array}{rrr|rrr} 2 & 1 & 3 & 1 & 0 & 0 \\ 1 & 3 & 3 & 0 & 1 & 0 \\ 0 & -2 & -3 & -1 & 0 & 1 \end{array}\right]$$

multiply line 2 with 2:

$$\left[\begin{array}{rrr|rrr} 2 & 1 & 3 & 1 & 0 & 0 \\ 2 & 6 & 6 & 0 & 2 & 0 \\ 0 & -2 & -3 & -1 & 0 & 1 \end{array}\right]$$

subtract line 1 from line 2:

$$\left[\begin{array}{rrr|rrr} 2 & 1 & 3 & 1 & 0 & 0 \\ 0 & 5 & 3 & -1 & 2 & 0 \\ 0 & -2 & -3 & -1 & 0 & 1 \end{array}\right]$$

multiply row 2 by 2, multiply row 3 by 5:

$$\left[\begin{array}{rrr|rrr} 2 & 1 & 3 & 1 & 0 & 0 \\ 0 & 10 & 6 & -2 & 4 & 0 \\ 0 & -10 & -15 & -5 & 0 & 5 \end{array}\right]$$

add line 2 to line 3:

$$\left[\begin{array}{rrr|rrr} 2 & 1 & 3 & 1 & 0 & 0 \\ 0 & 10 & 6 & -2 & 4 & 0 \\ 0 & 0 & -9 & -7 & 4 & 5 \end{array}\right]$$

multiply line 1 and 2 by 3:

$$\left[\begin{array}{rrr|rrr} 6 & 3 & 9 & 3 & 0 & 0 \\ 0 & 30 & 18 & -6 & 12 & 0 \\ 0 & 0 & -9 & -7 & 4 & 5 \end{array}\right]$$

add line 3 to line 1; add 2 times line 3 to line 2:

$$\left[\begin{array}{ccc|ccc} 6 & 3 & 0 & -4 & 4 & 5 \\ 0 & 30 & 0 & -20 & 20 & 10 \\ 0 & 0 & -9 & -7 & 4 & 5 \end{array}\right]$$

divide line 2 by 10:

$$\left[\begin{array}{ccc|ccc} 6 & 3 & 0 & -4 & 4 & 5 \\ 0 & 3 & 0 & -2 & 2 & 1 \\ 0 & 0 & -9 & -7 & 4 & 5 \end{array}\right]$$

subtract line 2 from line 1; multiply line 3 by -1:

$$\left[\begin{array}{ccc|ccc} 6 & 0 & 0 & -2 & 2 & 4 \\ 0 & 3 & 0 & -2 & 2 & 1 \\ 0 & 0 & 9 & 7 & -4 & -5 \end{array}\right]$$

multiply line 1 by $1\frac{1}{2}$, multiply line 2 by 3:

$$\left[\begin{array}{ccc|rrr} 9 & 0 & 0 & -3 & 3 & 6 \\ 0 & 9 & 0 & -6 & 6 & 3 \\ 0 & 0 & 9 & 7 & -4 & -5 \end{array}\right]$$

this yields:

$$9IA^{-1} = \begin{bmatrix} -3 & 3 & 6 \\ -6 & 6 & 3 \\ 7 & -4 & -5 \end{bmatrix}$$

or:

$$A^{-1} = \frac{1}{9}\begin{bmatrix} -3 & 3 & 6 \\ -6 & 6 & 3 \\ 7 & -4 & -5 \end{bmatrix} = \begin{bmatrix} -1/3 & 1/3 & 2/3 \\ -2/3 & 2/3 & 1/3 \\ 7/9 & -4/9 & -5/9 \end{bmatrix}$$

Solving systems with the inverse

$$A^{-1}A = I$$

$$AX = B$$

(given A^{-1} exist!)

$$A^{-1}AX = A^{-1}B$$

$$X = A^{-1}B.$$

Singular matrix

$$[0], \quad \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$A = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 4 & 2 & 6 \end{bmatrix}$$

when we subtract 2 times line 1 from line 3 aftrekken, we have:

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

$$|A| = 0$$

Application: linear regression

observation i:

$$y_i = \beta_0 1 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + e_i = \sum_{j=0}^{p} X_{i,j}\beta_j + e_i$$

observation i, matrix notation:

$$y_i = [1 \; X_{i,1} \; \ldots \; X_{i,p}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + e_i = X_{i,.}\beta + e$$

all observations, matrix notation, $X_{i,.}$ the i-th row in X:

$$y = X\beta + e$$

Least squares solution

find β for which the sum of squared residuals is minimal.

$$R = \sum_{i=1}^{n} e_i^2 = e'e = (y - X\beta)'(y - X\beta) =$$

$$y'y - 2\beta'X'y + \beta'X'X\beta$$

The derivative to β is:

$$\frac{\delta R}{\delta \beta} = -2X'y + 2X'X\beta$$

(this is a $p \times 1$ vector) The least squares estimate $\hat{\beta}$ of β is obtained by solving $\delta R/\delta\beta = 0$:

$$-2X'y + 2X'X\hat{\beta} = 0$$
$$X'X\hat{\beta} = X'y$$

$$X'X\hat{\beta} = X'y$$

compare with

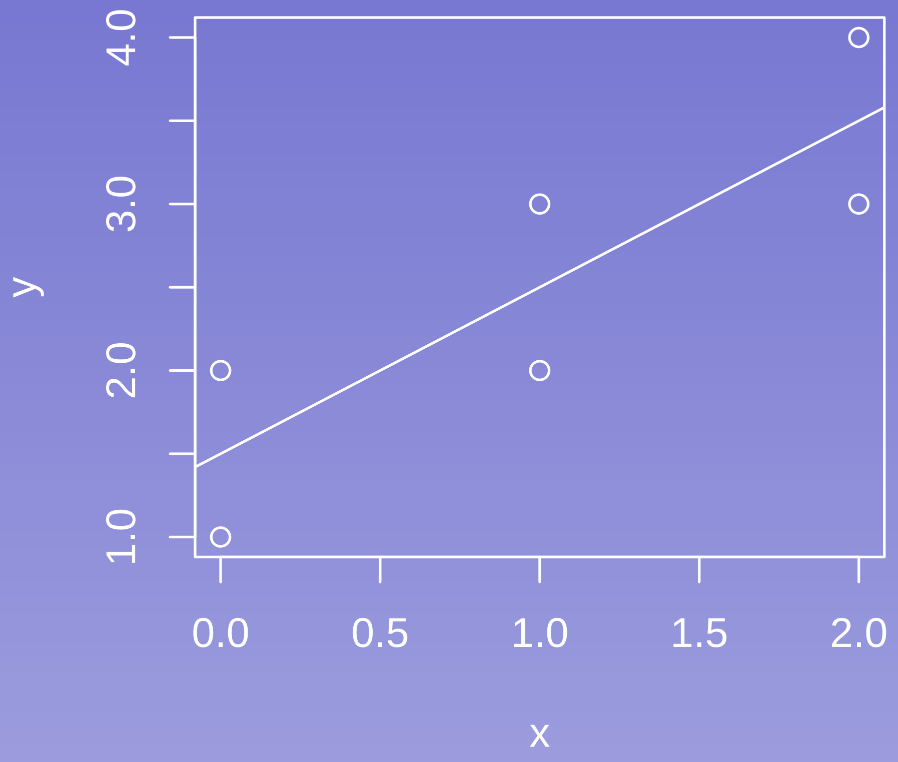
$$Ax = b$$

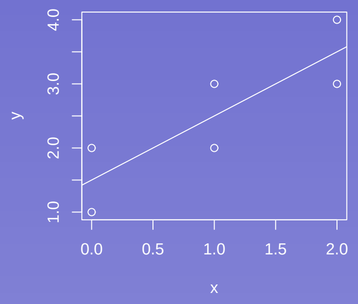
$Ax = b$, with $A = X'X$, $x = \beta$ and $b = X'y$

Solution:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Example simple linear regression


y
1.0
2.0
3.0
4.0
0.0
0.5
1.0
1.5
2.0
x



$$y = \beta_0 \times 1 + \beta_1 \times x = X\beta + e$$

$$\begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 6 & 6 \\ 6 & 10 \end{bmatrix}, \quad \hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$$

$$y = 1.5 + 1X + e$$

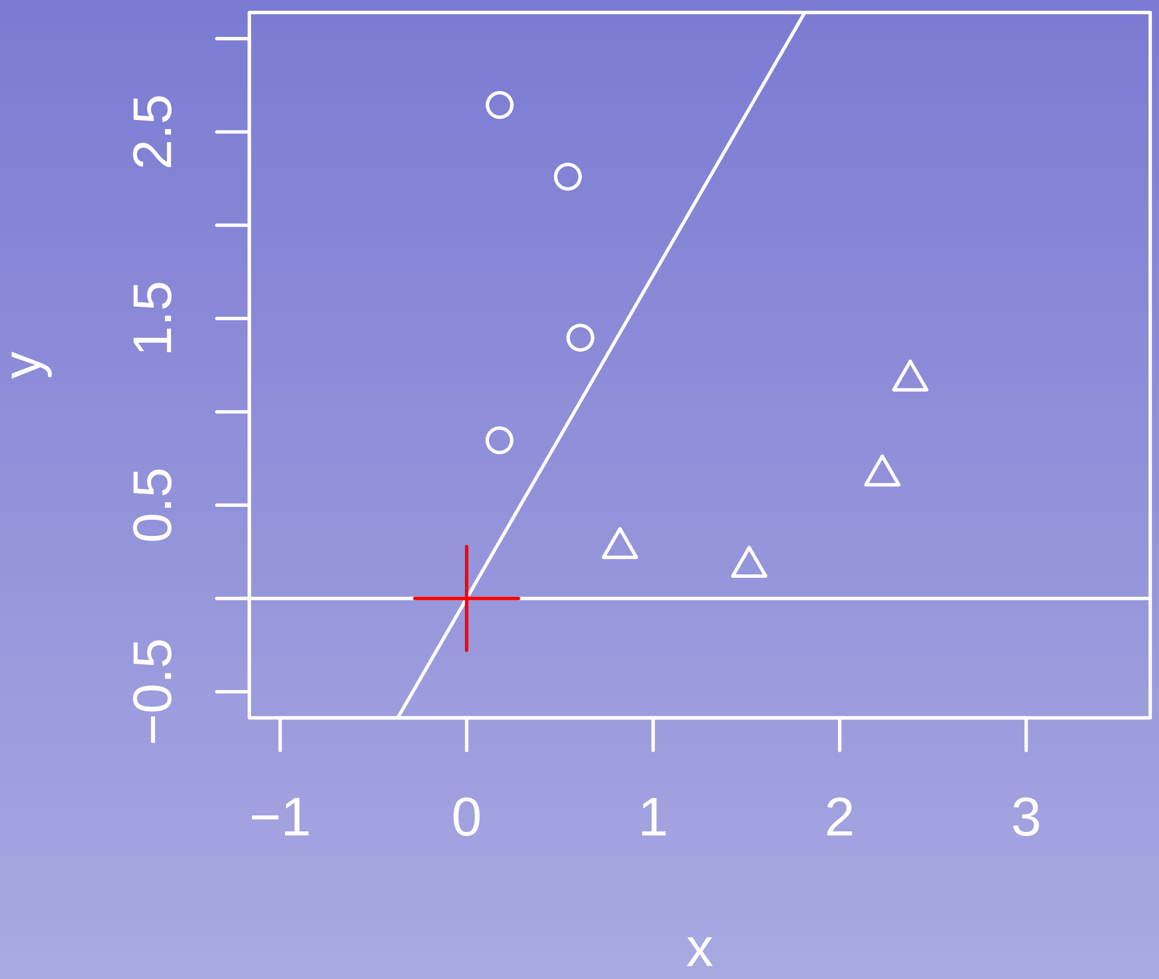
Application: projection

Projection matrices P satisfy:

- $P'P = PP' = I$
- vectors are normalized and orthogonal (orthonormal)

Projection of points in X is done by XP

Projected points have new coordinates, but their relative postions are not disturbed (angles, distances)


−0.5
0.5
1.5
2.5
y
−1
0
1
2
3
x

$$X = \begin{bmatrix} 0.18 & 2.64 \\ 0.61 & 1.40 \\ 0.18 & 0.85 \\ 0.54 & 2.26 \end{bmatrix}, P = \begin{bmatrix} 0.50 & -0.87 \\ 0.87 & 0.50 \end{bmatrix}, XP = \begin{bmatrix} 2.38 & 1.17 \\ 1.52 & 0.17 \\ 0.82 & 0.27 \\ 2.23 & 0.66 \end{bmatrix}$$

in 2D, for angle ϕ (counter clockwise from x),

$$P = \begin{bmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{bmatrix}$$

$$\phi = \pi/2 : P = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$\phi = \pi : P = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Determinant

$$\begin{array}{rcrcr} x_1 & + & 2x_2 & = & 3 \\ x_1 & + & 3x_2 & = & 5 \end{array}$$

We can eliminate x_2 by multiplying eq. 1 by 3 and eq. 2 by 2:

$$\begin{array}{rcrcr} 3x_1 & + & 6x_2 & = & 9 \\ 2x_1 & + & 6x_2 & = & 10 \end{array}$$

next, we subtract eq. 2 from eq. 1, yielding $x_1 = -1$; substitution yields $x_2 = 2$.

The general case

$$\begin{array}{rcrcl} a_1 x_1 & + & b_1 x_2 & = & c_1 \\ a_2 x_1 & + & b_2 x_2 & = & c_2 \end{array}$$

multiply eq. 1 by b_2 and eq. 2 by b_1:

$$\begin{array}{rcrcl} a_1 b_2 x_1 & + & b_1 b_2 x_2 & = & b_2 c_1 \\ b_1 a_2 x_1 & + & b_1 b_2 x_2 & = & b_1 c_2 \end{array}$$

subtracting eq. 2 from eq. 1 yields:

$$x_1 = \frac{b_2 c_1 - b_1 c_2}{a_1 b_2 - a_2 b_1}, \text{and } x_2 = \frac{a_2 c_1 - a_1 c_2}{a_1 b_2 - a_2 b_1}$$

only a finite solution when $a_1 b_2 - a_2 b_1 \neq 0$.

$$a_1b_2 - a_2b_1 = \left| \begin{array}{cc} a_1 & b_1 \\ a_2 & b_2 \end{array} \right|.$$

Calculation of determinant using sub-determinants

$$\begin{vmatrix} 1 & 2 & 3 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{vmatrix} = 1\begin{vmatrix} 2 & 2 \\ 1 & 3 \end{vmatrix} - 2\begin{vmatrix} 0 & 2 \\ 1 & 3 \end{vmatrix} + 3\begin{vmatrix} 0 & 2 \\ 1 & 1 \end{vmatrix} =$$

$$1 \times 4 - 2 \times -2 + 3 \times -2 = 2$$

tekens:

$$\begin{array}{cccc} + & - & + & - \\ - & + & - & + \\ + & - & + & - \\ - & + & - & \ldots \\ \ldots & & & \end{array}$$

Save calculations by simplifying the system:

$$\begin{vmatrix} 1 & 2 & 3 \\ 0 & 2 & 2 \\ 1 & 1 & 3 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & 2 & 2 \\ 0 & -1 & 0 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{vmatrix}$$

$$= 1 \times 2 \times 1 = 2$$

If, in $Ax = b$ the determinant of A, $|A| = 0$, then A is *singular* and $Ax = b$ cannot be solved for x (if $b \neq 0$).

Eigenvectors, eigenvalues

Given a square matrix A, suppose that a vector $x \neq 0$ exists, such that

$$Ax = \lambda x$$

with λ a constant (scalair), then x is an *eigenvector* of A, en λ is the corresponding *eigenvalue*.

A square matrix A has as many eigenvectors as rows (columns), and the complete set of eigenvectoren satisfies:

$$AX = X\Lambda$$

with eigenvectors the columns of X, en with Λ a diagonal matrix with diagonal elements the eigenvalues of the corresponding eigenvectors.

Calculation of eigenvectors and -values

We can write

$$Ax = \lambda x$$

as

$$Ax - \lambda x = 0$$

of

$$(A - \lambda I)x = 0.$$

Solution:

1. solve

   $$|(A - \lambda I)| = 0$$

   for the eigenvalues λ;

2. substitute these values in $Ax = \lambda x$ and solve for the eigenvectors x

Properties

1. Symmetric matrices have orthogonal eigenvectors

2. Eigenvalues of 0 correspond to eigenvectors in the directions (dimensions) that are not present in the matrix.

Example: eigenvalues/vectors

Suppose $A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$. Solve
$\begin{vmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{vmatrix} = (4-\lambda)(3-\lambda) - 2 = \lambda^2 - 7\lambda + 10 = 0$. This can be decomposed into $(\lambda - 2)(\lambda - 5) = 0$ and the eigenvalues are $\lambda_1 = 2$ and $\lambda_2 = 5$. The eigenvectors are found by solving

$$\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \text{ and } \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 5 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The first equation leads to the solution $2x_1 + x_2 = 0$, for which any (scalar) multiple of $[-1 \;\; 2]'$ is a solution. The second eigenvalue leads to multiples of $[1 \;\; 1]'$ as solution. Computer programs normalize the eigenvectors; signs are arbitrary.

Covariance, correlation

- variance: measures variability

$$\mathrm{Var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

- covariance: measures linear depenence, non-normalized

$$\mathrm{Cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})$$

- $\mathrm{Cov}(x,x) = \mathrm{Var}(x)$

- $|Cov(x,y)| \leq \sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}$

- correlation: a normalized measure $[-1,1]$ of *linear* dependency between x and y:

$$\mathrm{Corr}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}$$

- symmetric: $\mathrm{Cov}(x,y) = \mathrm{Cov}(y,x)$, $\mathrm{Corr}(x,y) = \mathrm{Corr}(y,x)$

- if x and y are normalized (mean zero, unit variance), then $\mathrm{Corr}(x,y) = \mathrm{Cov}(x,y)$

covariance/correlation matrix

Given a data matrix X (m rows with records, n columns with variables x_j),

- the covariance matrix C is the $n \times n$ matrix with elements $C_{i,j} = \mathrm{Cov}(x_i, x_j)$
- the correlation matrix R is the $n \times n$ matrix with elements $R_{i,j} = \mathrm{Corr}(x_i, x_j)$
- covariance/correlation matrices are square and symmetric
- the diagonal of covariance matrices: $C_{i,i} = \mathrm{Var}(x_i)$
- the diagonal of correlation matrices: $R_{i,i} = 1$

Data reduction

Main goal in observational studies:

> How can we reduce the research findings to a few relevant and clearcut conclusions, *unambiguously supported by the observations*

Data reduction

Main goal in observational studies:

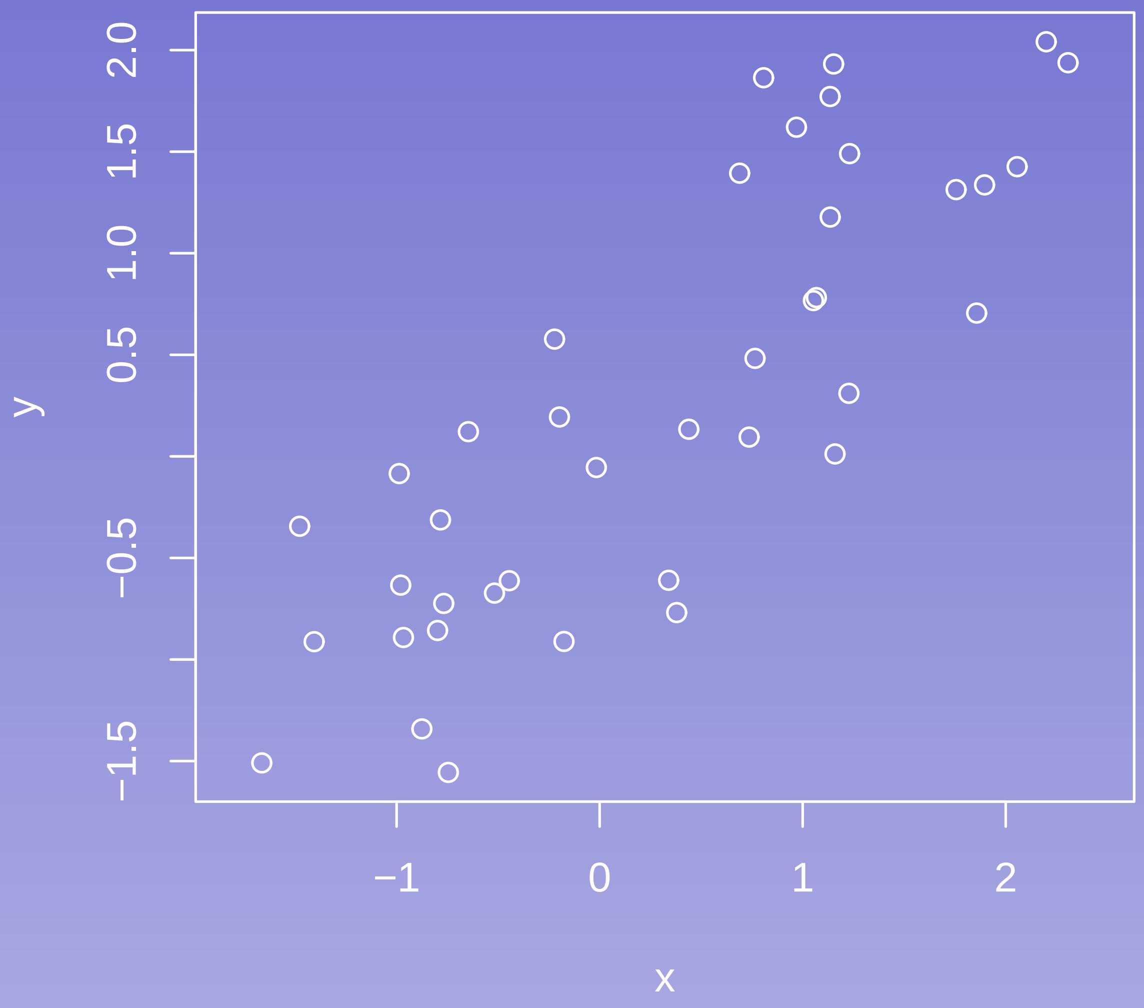
> How can we reduce the research findings to a few relevant and clearcut conclusions, *unambiguously supported by the observations*

Main approaches in multivariate analysis:

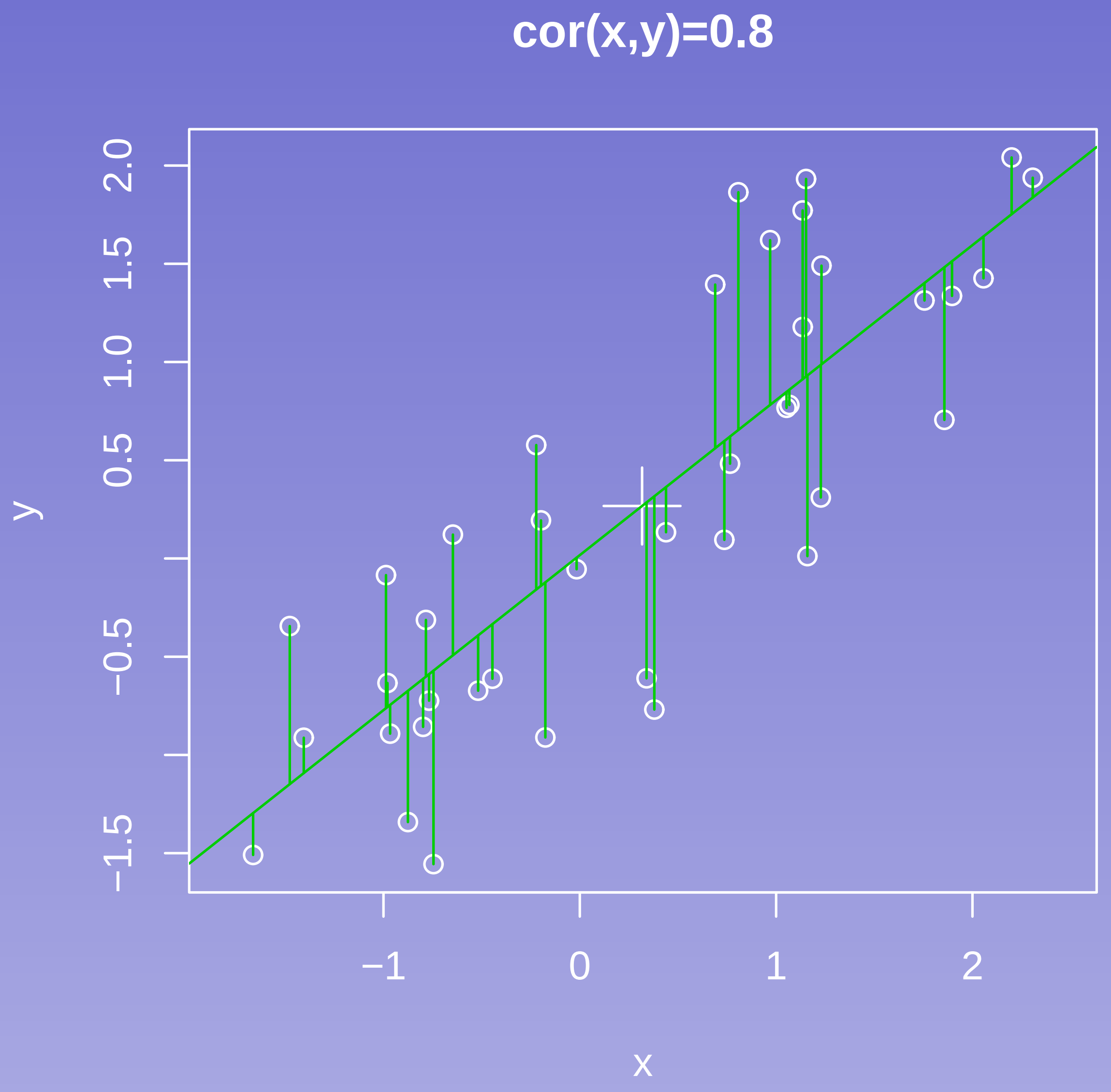
1. **variables** find one or a few variables that summarize all variability

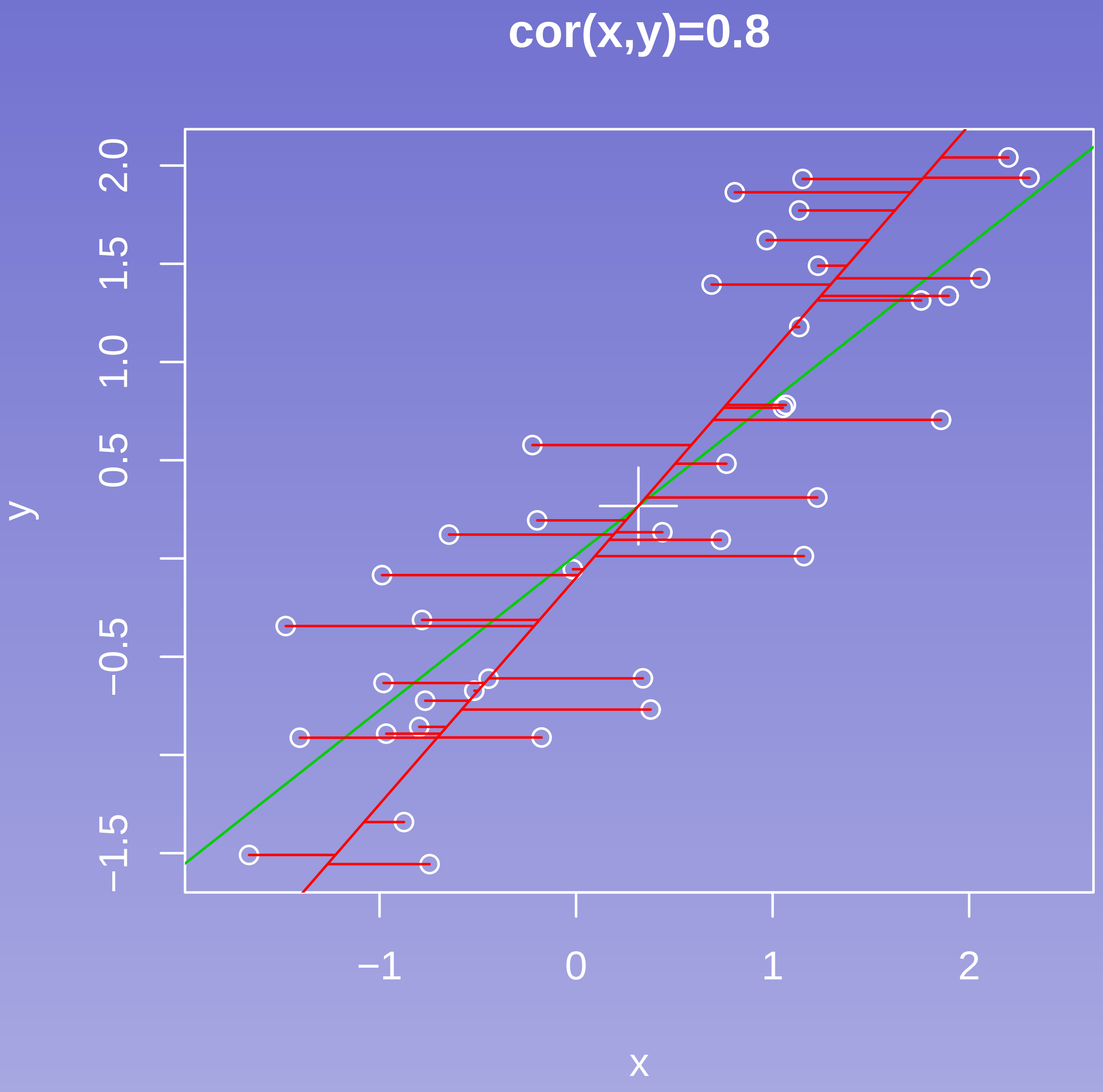
2. **cases** find (or test) a grouping variable that summarize much of the case-to-case variability

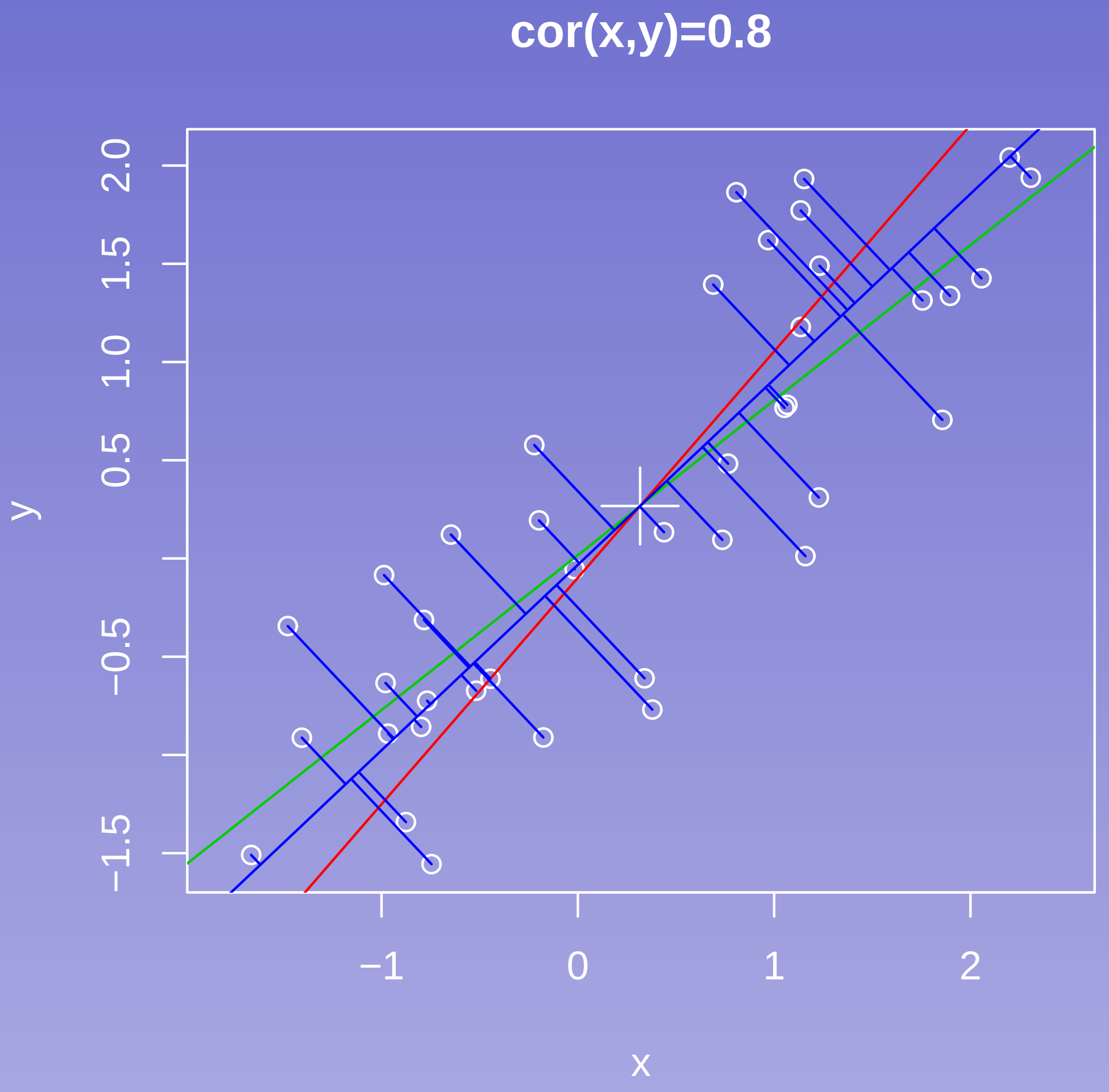
1. ordination methods; 2. clustering, discrimination

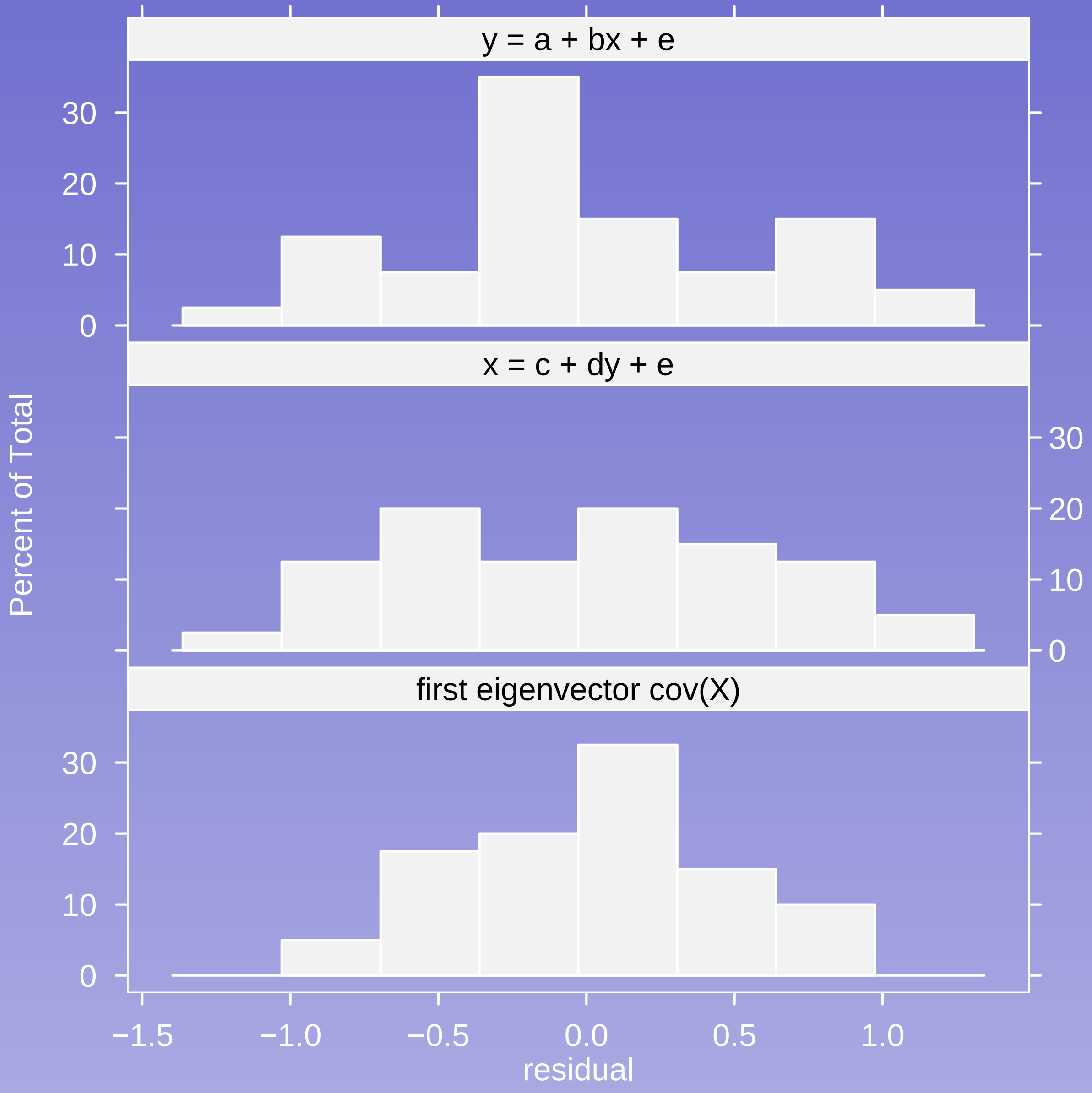


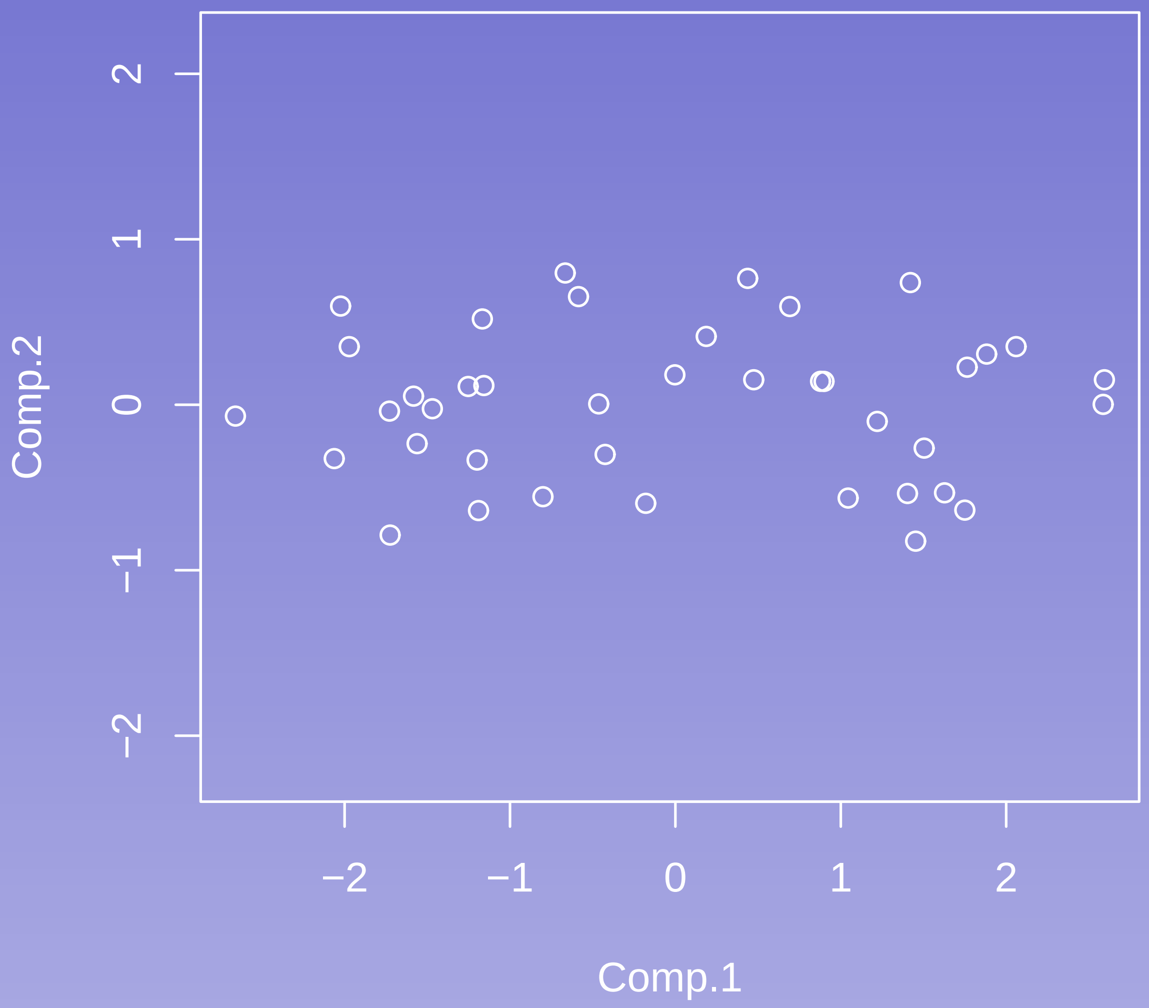
y
−1.5
−0.5
0.5
1.0
1.5
2.0
−1
0
1
2
x


cor(x,y)=0.8
y
−1.5
−0.5
0.5
1.0
1.5
2.0
−1
0
1
2
x


cor(x,y)=0.8
y
x
−1.5
−0.5
0.5
1.0
1.5
2.0
−1
0
1
2


cor(x,y)=0.8
y
−1.5
−0.5
0.5
1.0
1.5
2.0
−1
0
1
2
x


y = a + bx + e
x = c + dy + e
first eigenvector cov(X)
Percent of Total
0
10
20
30
−1.5
−1.0
−0.5
0.0
0.5
1.0
residual


Comp.2
2
1
0
−1
−2
−2
−1
0
1
2
Comp.1

Eigenvalue/vector properties

$$AX = X\Lambda$$

- if A is symmetric, X is orthonormal
- if A is orthonormal, λ_i are all equal
- the more A deviates from orthonormal, the large the difference between λ_1 and λ_n
- if A is singular, one or more of the λ_i are zero; the number of positive λ_i's equals the number of dimension spanned by the columns (rows) of A
- the sum of λ_i equals the sum of the diagonal elements in A, $A_{i,i}$

Centering and normalizing data variables

Suppose data are stored in a matrix X;

centering means that each column (variable) x_j is replaced by $x_j - \bar{x}_j \Rightarrow$ zero mean

normalizing means that each column (variable) x_j is replaced by $\frac{x_j - \bar{x}_j}{\sigma_j} \Rightarrow$ zero mean, unit variance

centerend data: $\mathrm{Cov}(X) = \frac{1}{m-1} X'X \Rightarrow$ column inproducts
normalized data: $\mathrm{Corr}(X) = \frac{1}{m-1} X'X$

Eigenvalues/vectors: properties

- eigenvectors of a symmetric matrix are orthonormal
- eigenvectors are *ordered* by their corresponding eigenvalue; the first eigenvector has (by definition) the largest eigenvalue
- sum of eigenvalues equals sum of diagonal elements A
- given a value of λ, how to solve for x?
  - ⋆ substitute λ in $(A - \lambda I)x = 0$
  - ⋆ now $A - \lambda I$ is known, and 0 is known
  - ⋆ try values for e.g. $x_{1,1}$
  - ⋆ make sure that you don't end up with a 0 vector

Singular value decomposition (SVD)

$$X_{n\times m} = R_{n\times r}\Lambda_{r\times r}K'_{r\times m}$$
$$r \leq m \leq n$$

X: *centered* data matrix
R: columns: eigenvectors of XX'
Λ: singular values of X (square root of pos. eigenvalues $X'X$ of XX')
K: columns: eigenvectors of $X'X$

$X'X$ symmetric $\Rightarrow$ K orthogonal
XX' symmetric $\Rightarrow$ R orthogonal
$K'K = KK' = I$ (K orthonormal)
$R'R = RR' = I$ (R orthonormal)

Consequences SVD

$$X_{n\times m} = R_{n\times r}\Lambda_{r\times r}K'_{r\times m}$$

project X on eigenvectors of $X'X$: post-multiply with K:

$$Y_{n\times r} = X_{n\times m}K_{m\times r} = R\Lambda K'K = R\Lambda$$

$X'X \Rightarrow$ columns of K are *independent*:

$Y'Y = (XK)'XK$
svd: $X = R\Lambda K' \Rightarrow XK = R\Lambda K'K = R\Lambda I = R\Lambda$
$Y'Y = (R\Lambda)'R\Lambda = \Lambda'R'R\Lambda = \Lambda'I\Lambda = \Lambda'\Lambda = \Lambda^2$ (diag.)

singular values of X are the square root of singular values (eigenvalues) of $X'X$.

PCA by SVD

- centered (possibly normalized) data matrix X
- $Y = XK$, K the eigenvectors of $X'X$ $(\mathrm{Cov}(X)$ or $\mathrm{Corr}(X))$
- $X'X$ is symmetric $\Rightarrow$ K is a projection matrix
- $Y'Y = \Lambda^2$:
  - ⋆ the variables Y are independent
  - ⋆ the variance of the Y is Λ^2

Principal components – what are they?

Principal components (PC's) are *directions* (new axes);

- the first PC explains maximum variability in a data set
- the second PC explains, independent from the first, maximum (remaining) variability
- subsequent PC's are independent

Principal components: loadings and scores

Principal components are formed by the eigenvectors of the covariance or correlation matrix; if X_j is the j-th centered column in data matrix X,

$$PC_1 = \alpha_{1,1}X_1 + \alpha_{2,1}X_2 + ... + \alpha_{n,1}X_m$$

with α the first eigenvector (column) of $X'X$. We call the coefficients α the *loadings* of a PC. They tell the direction. Each PC has as much loadings as X has variables (columns).

The projected (new) values along the new axis (PC) are called the *scores*. The number of scores for a PC is equal to the number of cases (rows) in X.

The eigenvalues equals the variance taken into account by a PC. The sum of the eigenvalue equals the sum of the variances (diagonal elements of $\mathrm{Cov}(X) = \frac{1}{m-1}X'X$

Rationale behind principal components

- hopefully, a few PC's summarize the *essence* of the data:
- retain the first few PC's, and abandon the rest.
- *always* a good first "shot" at correlated data (exploration)
- unfortunately, essential messages may be "hidden" in later PC's, or distributed over many PC's

The “size and shape” effect

Often, the more interesting information is in the *second* (or later) component; examples:

fossil data the first component measures size, the second shape (width/height); size tells something about age, shape about species

spectral curves first component measures brightness (exposed vs. shaded areas), the second differences in spectral curve shape (amount of vegetation, water etc.)

sediment chemistry first component measures clay vs sand (i.e., sedimentation environment dynamics), the second (and further) the specific composition characteristics of the clay, maybe related to origin of sediment

pollution the first component may measure degree of pollution, the second the composition (relative ratios) of the pollution components, maybe connected to the origin of pollution

Use covariance or correlation?

User choice: IT MATTERS

- if variables should be given equal *weight* (importance) in the analysis, use correlations.
- if differences in variances reflect the difference in *importance* of variables, use covariances (e.g., grain size distrib.?)
- if in doubt, use correlations.

Curve data in Physical Geography

Example of curves:

- grain size distribution
- hyperspectral data (wavelength)
- depth: e.g. moisture depth profile, variables $\theta(z_i)$
- spatial series: space replicates are the variables, moments in time the observations
- time series: time replicates are variables, spatial locations the observations

The more densely sampled the curve, the more correlated the variables.

Factor Analysis

Goal:

What is the relation (correlatin) of m observed variables with p $(p < m)$ underlying, unobserved factors?

- Factor analysis: seeks from n original variables p underlying, unknown (and not directly observable) variables, called common factors
- p is known, prior to analysis

- statistical model:

$$X_j = \sum_{r=1}^{p} a_{jr} f_r + \epsilon_j$$

X_j j-th variable
a_{jr} loading of the j-th variable on the r-th factor
f_r th r-th factor
ϵ_j random variable, unique to X_j

- the set of m ϵ_j's is called the unique factor
- Difference from regression analysis: the f_r are unobservable
- Differences from PCA are *subtle*.
- PCA: from n original to n new axes: explorative, geometric
- FA: statistical model: observation $=$ structure $+$ noise.

- if $p = m$: FA ≈ PCA

How do we determine p?

- theory (*not* statistical theory, and neither physics!!)
- 2 or 3, ..., 7? (never more)
- experimenation ... which is not prior knowledge!
- number of factors for which eigenvalue > 1 ... ?

Factor rotation

- general idea: if p ($p \geq 2$) factors explain 80% of the variance, then any p orthogonal factors in this p-dimensional subspace explain this 80% of the variability.

- PCA: first PC explains maximum variability

- rotated factors: first factor does not explain maximum variability

- why then rotate? Interpretability. Factors with loadings close to either 0 or $+1$, -1 have the advantage that they are associated with certain (groups of) variables, and not with others (varimax).

Nominal variables and cross tables

- Data: nominal $(0,1)$ or ordinal $(1,2,3,...,n)$
- binary, e.g. present (1) or absent (0)
- nominal, e.g. sand (0), clay (1), peat (3)
- ordinal: low, intermediate, high
- Linear combination of variables: meaningless

Question: how do two nominal (or two sets of binary) variables relate to each other?

Dune data set

- How do plant species relate to each other?
- How do plant species relate to environmental conditions?
- 30 species, 20 quadrats

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
Belper	3	0	2	0	0	0	0	2	0	0	2	0	0	2	2	0	0	0	0	0
Empnig	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
Junbuf	0	3	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	4	2
Junart	0	0	0	3	0	0	4	0	0	3	0	0	4	0	0	4	0	0	0	0
Airpra	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	3	0	0
Elepal	0	0	0	8	0	0	4	0	0	5	0	0	0	0	0	4	4	0	0	0

```
Rumace 0 0 0 0 6 0 0 5 0 0 0 0 2 0 0 0 0 0 2 3
Viclat 0 0 0 0 0 0 0 0 0 0 1 2 0 1 0 0 0 0 0 0
Brarut 0 0 2 4 6 0 2 2 0 4 2 4 2 6 2 4 0 3 4 2
Ranfla 0 2 0 2 0 0 2 0 0 2 0 0 0 0 0 4 2 0 0 0
Cirarv 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Hyprad 0 0 0 0 0 0 0 0 2 0 0 2 0 0 0 0 0 5 0 0
Leoaut 5 2 2 0 3 0 3 3 2 2 3 5 2 5 2 2 2 6 2 3
Potpal 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 2 0 0 0
Poapra 4 2 4 0 3 4 4 2 1 0 4 4 4 3 5 0 0 0 0 4
Calcus 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 3 4 0 0 0
Tripra 0 0 0 0 5 0 0 2 0 0 0 0 0 0 0 0 0 0 0 2
Trirep 5 2 1 0 5 0 2 2 0 1 6 3 3 2 2 0 6 2 3 2
Antodo 0 0 0 0 3 0 0 4 4 0 4 0 0 0 0 0 0 4 0 2
Salrep 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 5 0 3 0 0
Achmil 3 0 0 0 2 1 0 2 2 0 4 0 0 0 0 0 0 0 0 2
Poatri 7 9 5 2 4 2 4 6 0 0 4 0 5 0 6 0 0 0 4 5
Chealb 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Elyrep 4 0 4 0 0 4 0 4 0 0 0 0 6 0 4 0 0 0 0 0
Sagpro 0 2 5 0 0 0 2 0 0 0 0 2 2 0 0 0 0 3 4 0
Plalan 0 0 0 0 5 0 0 5 2 0 3 3 0 3 0 0 0 0 0 5
```

```
Agrsto 0 5 8 7 0 0 4 0 0 4 0 0 3 0 4 5 4 0 4 0
Lolper 5 0 5 0 6 7 4 2 0 0 6 7 2 2 6 0 0 0 0 6
Alogen 2 5 2 4 0 0 5 0 0 0 0 0 3 0 7 0 0 0 8 0
Brohor 4 0 3 0 0 0 0 2 0 0 4 0 0 0 0 0 0 0 0 2
```

Data reduction by groups

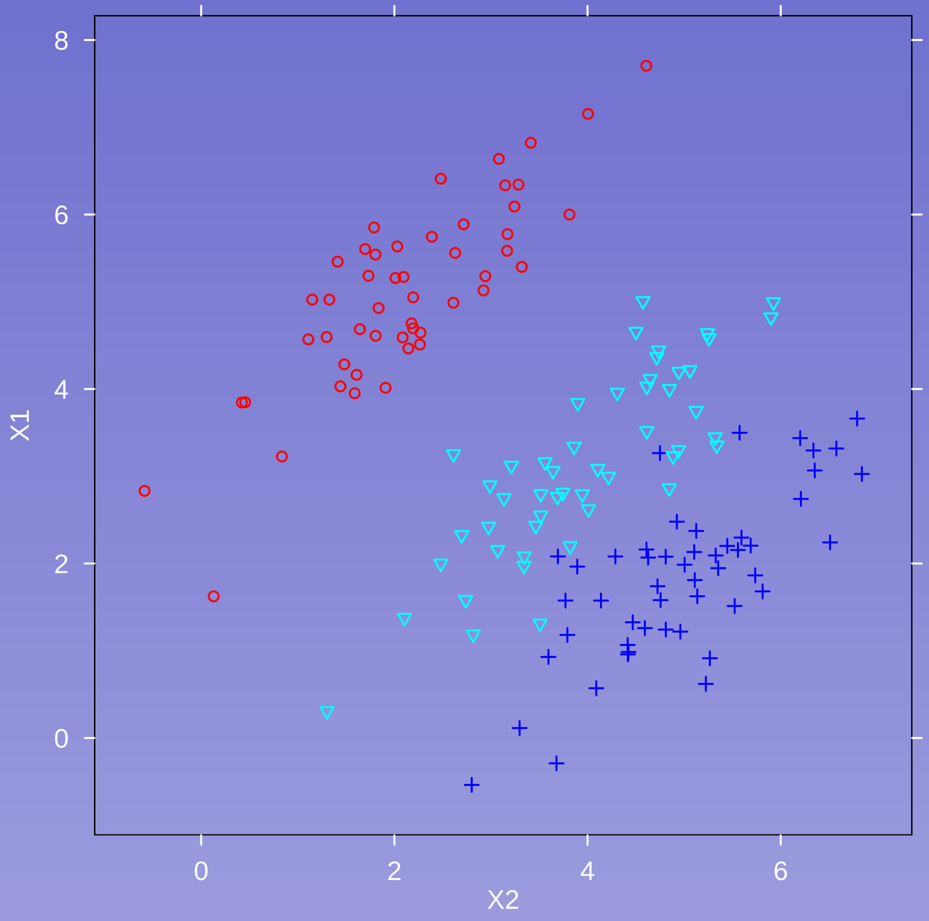
Idea:

a (single) grouping variable (nominal variable) may reflect a simple but adequate structure, and may summarize the multivariate variability in a (large part of the) data set. We may seek such a grouping variable (clustering), or measure its strength or predict group membership given from all other variables (discriminant analysis).

Discrimination and Clustering

- discriminant analysis is concerned with how well a set of variables can predict a *given* grouping variable, *given the grouping variable is known*. ⇒ supervised: grouping variable dependent, other variables independent

- cluster analysis is concerned with *finding groups* from an, a prior, ungrouped data set. ⇒ unsupervised: no distinction between dependent/independent.


X1
X2
8
6
4
2
0
0
2
4
6

Discriminant functions

Discriminant function:

$$R = \lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_m X_m = \lambda' X$$

(axis) with:
λ_i : loadings
R : scores (R_j score of observation j on Discr.fn.)

mean of A: $\bar{A} = [\bar{A}_1, \bar{A}_2, ..., \bar{A}_m]'$

$$R_A = \lambda_1 \bar{A}_1 + ... + \lambda_m \bar{A}_m = \lambda' \bar{A}$$

$$R_B = \lambda_1 \bar{B}_1 + ... + \lambda_m \bar{B}_m = \lambda' \bar{B}$$

group means projected on the discriminant axis

Criteria for a good discriminant axis:

- $R_A - R_B$ as large as possible
- Var$(\lambda' A)$ and Var$(\lambda' B)$ as small as possible

How to find the function?

$$\mathsf{Var}(A) = \Sigma \Rightarrow \mathsf{Var}(\lambda' A) = \lambda' \Sigma \lambda$$

assume homoscedastic within-group covariances:

$$\mathsf{Var}(A) = \mathsf{Var}(B) = \ldots = \mathsf{Var}(Z)$$

Problem: find λ such that

$$\frac{|R_A - R_B|}{\lambda' \Sigma \lambda}$$

is maximized, given $\lambda' \lambda = 1$ (maximize not *only* $|R_A - R_B|$).

Solution (cmp. multiple regression):

$$\lambda = \Sigma^{-1}(\bar{A} - \bar{B})$$

Testing multivariate differences

Significance testing of the multivariate difference $\bar{A} - \bar{B}$

compare to two-sample t-test:

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_p}$$

$|t| > t(\alpha, \text{dF})$: significant difference
$|t| < t(\alpha, \text{dF})$: non-significant difference

Multivariate:

$$D^2 = (\bar{A} - \bar{B})\Sigma^{-1}(\bar{A} - \bar{B})$$

If $\Sigma = I$: $(\bar{A} - \bar{B})'(\bar{A} - \bar{B})$
If Σ diagonal: scaling axis
Other cases: scaling + rotation

Assumptions for the test

(Davis:)

1. observations were taken at random from the population
2. probability of being in group A or B is equal
3. within-group distribution: mulitivariate normal
4. within-group covariances: identical (Σ)
5. no mis-classifications

Wilk’s lambda

$$\Lambda = \frac{|W|}{|T|}$$

W : within-class covariantie matrix
T : total covariantie matrix

Multivariate analogue of $1 - R^2$

Λ: measures effectivity of the division in groups
0 : effective
1 : not effective

Suppose 1 row (or column) in W is 0 $\Rightarrow |W| = 0 \Rightarrow$ perfect distinction $|T| > 0$

Canonical analysis

- Problem: more than two groups (A,B,C) or (A,B,...,Q)
- multiple axes are needed

Search for p axes that

- are independent (orthogonal) and
- as good as possible distinguish the groups

$p \leq m$ en $p \leq k - 1$

eigenvector technique in the space of standardized differences.

Alternative approaches

- SVM, support vector machines
- ANN, artificial neural networks
- logistic regression

Cluster analysis

Clustering: search for a *good* division into groups, based on measured values.

- EDA
- data reduction
- allocation

Why not?

- prediction, mapping
- hypothesis testing

Clustering approaches

1. partitioning methods

2. arbitrary origin

3. hierarchical agglomerative

Problems

- when is a clustering a good clustering?
- how many groups should we distinguish?

Hierarchical methods

Find a measure of similarity (distance) between:

- objects and objects
- objects and clusters
- clusters and clusters

Where is the cluster?

- single linkage
- complete linkage

- weighted pair group
- centroid

Penalizing merges

How do we weight clusters

- relative to each other
- relative to other objects

Ward's method: minimize error sum of squares: merge cluster A with size n_A and cluster B with size n_B when

$$\frac{n_A n_B}{n_A + n_B} d^2_{AB}$$

is at minimum $\Rightarrow$ prefers merging of object and/or small clusters

$$\frac{n_A n_B}{n_A + n_B} d^2_{AB}$$

given $d^2 = 1$,

$$\frac{n_A n_B}{n_A + n_B}$$

$n_A,\ n_B$	1	2	5	10	20
1	0.50	0.67	0.83	0.91	0.95
2	0.67	1.00	1.43	1.67	1.82
5	0.83	1.43	2.50	3.33	4.00
10	0.91	1.67	3.33	5.00	6.67
20	0.95	1.82	4.00	6.67	10.00

Choices, choices, choices...

- similarity measure (distances)
- fusion criterium (when to merge)

Questions:

theory is there a theoretical foundation for the choices made?

statistical properties to what extent would another sample lead to the same clustering? ⇒ sampling variability

optimality in which sense is the clustering found the best?

K-means clustering

Arbitrary origin methods –

Idea: start with k arbitrary origins and repeat:

1. calculate distances of each object to the k centres
2. assign each object to its nearest centre
3. shift the cluster centres to the mean of the objects assigned to it

until convergence (no change).

Risk: local minimum

k arbitrary origins: take the group means of a random partitioning of the objects into k groups.

Integral criteria

When, instead of distances an integral criterion (e.g. Wilk's Λ) is used: start with an arbitrary partition, and repeat:

1. exchange n objects at random

2. accept the change if the criterion improved

(simulated annealing)

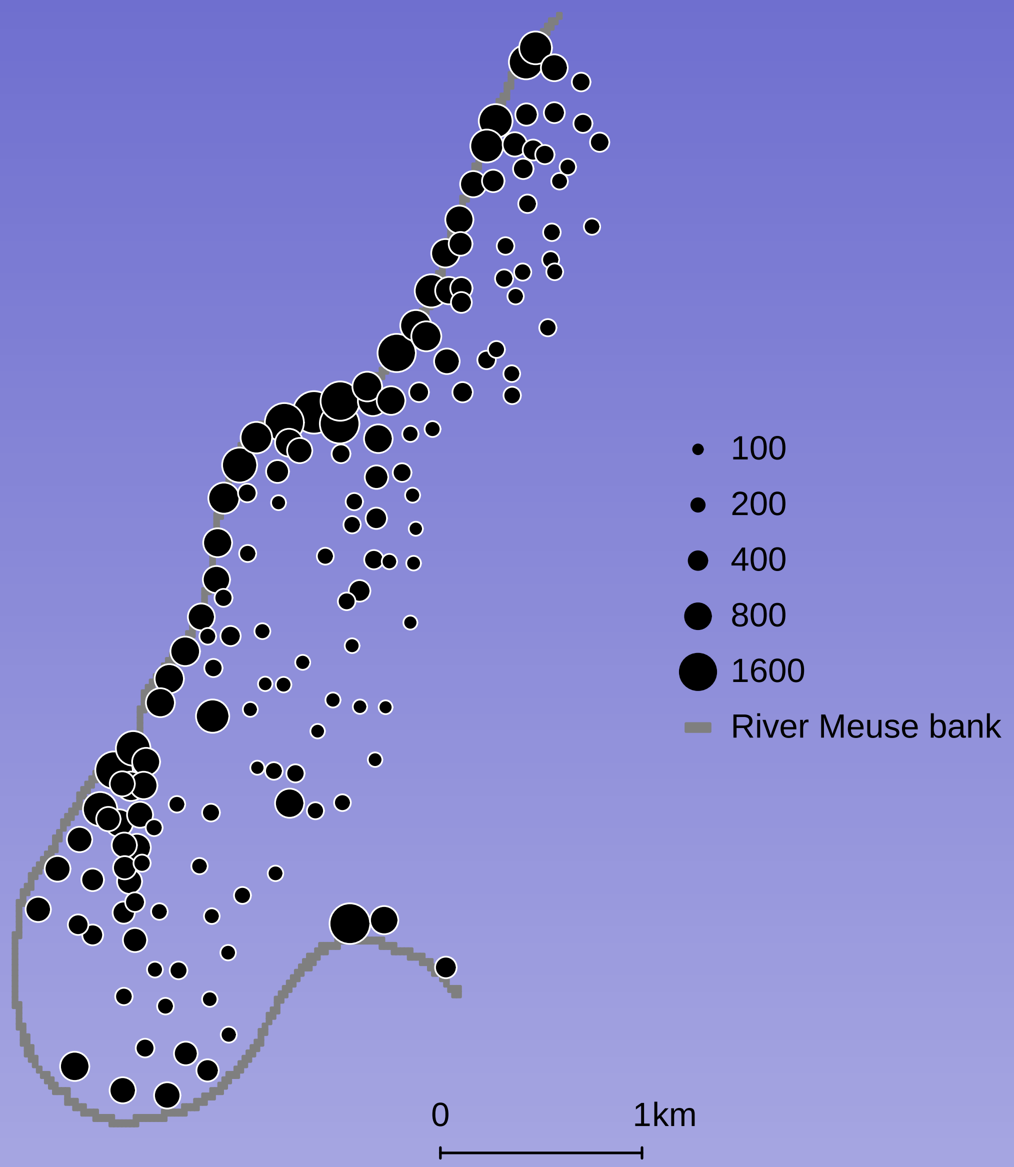
Risk: minimum is a local minimum

How many clusters

- theory (compare FA)
- clusters found can be interpreted
- 7 (like legend units on a map)

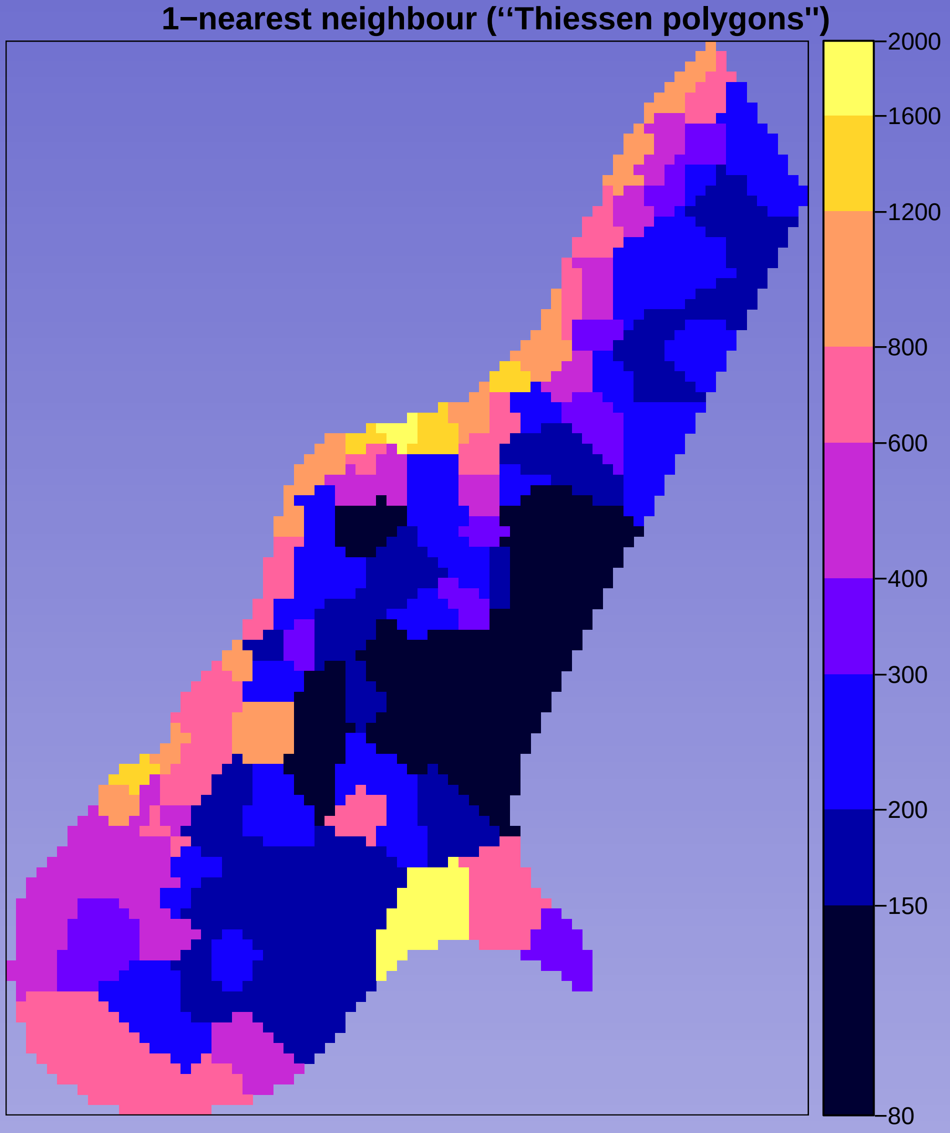
drawbacks:

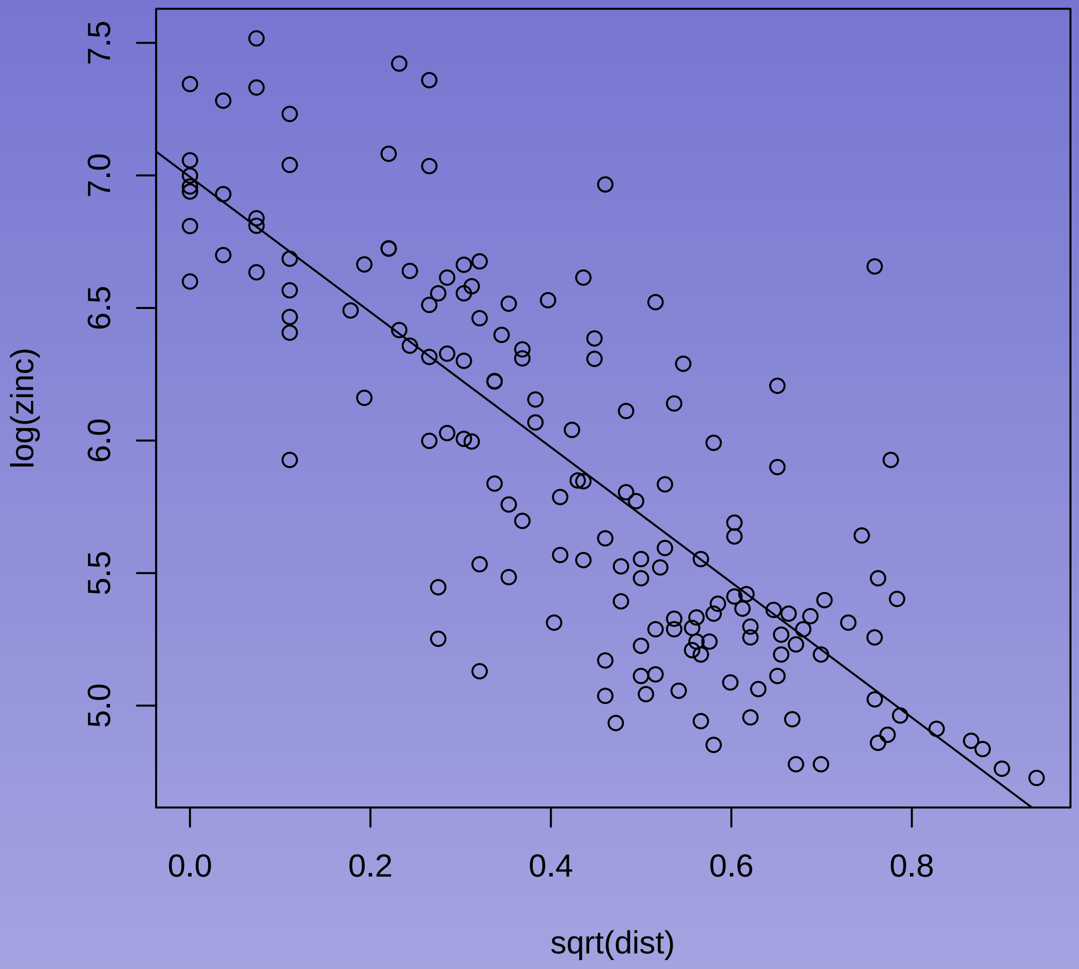
- freedom of choice, no theory
- local optima
- no repercussions on degrees of freedom lost

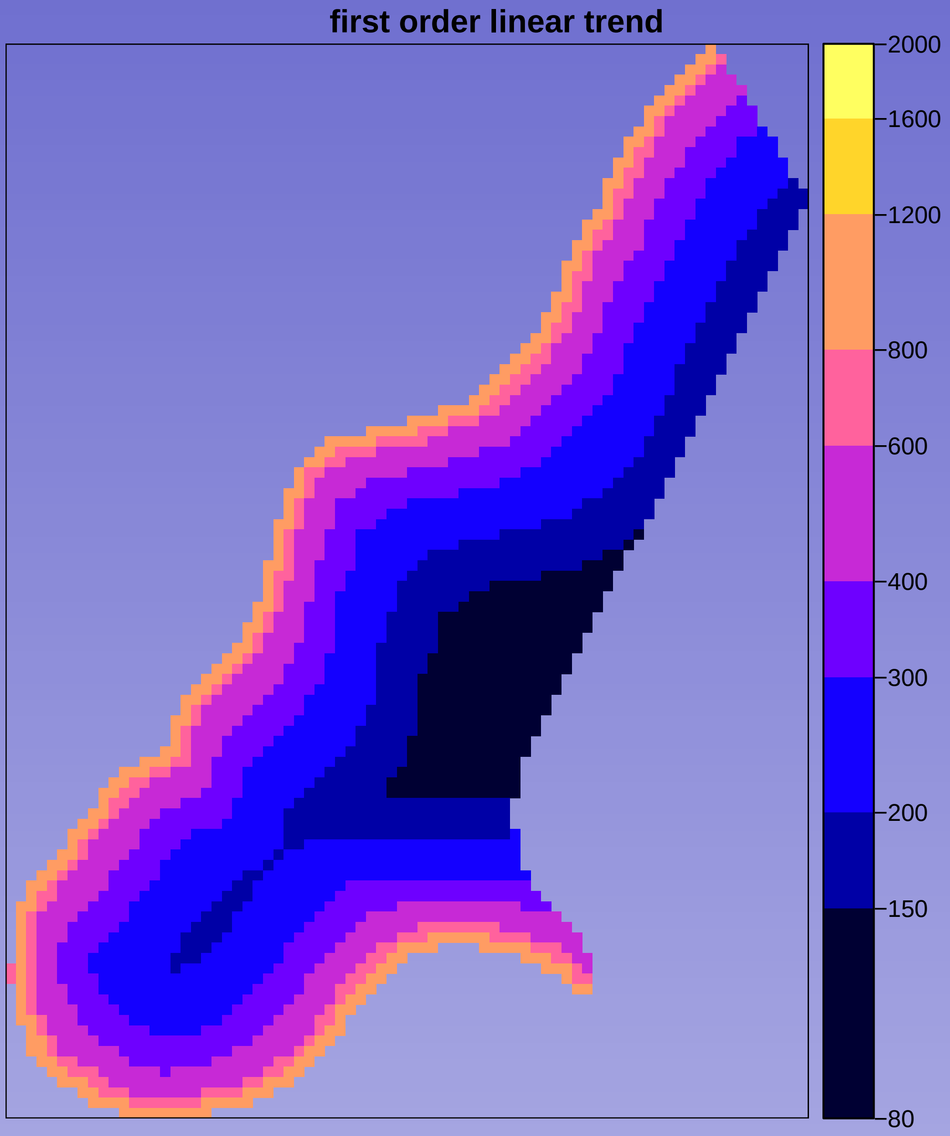


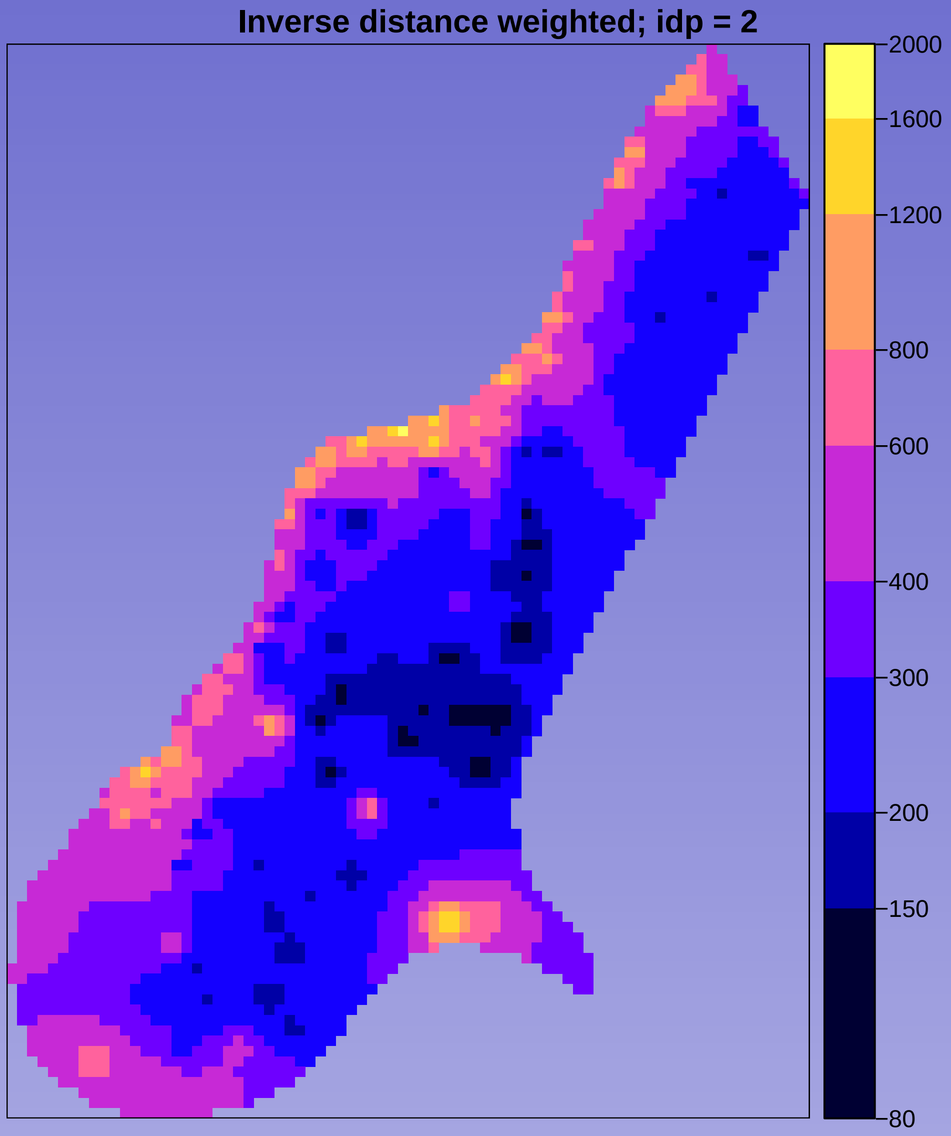
Spatial interpolation:

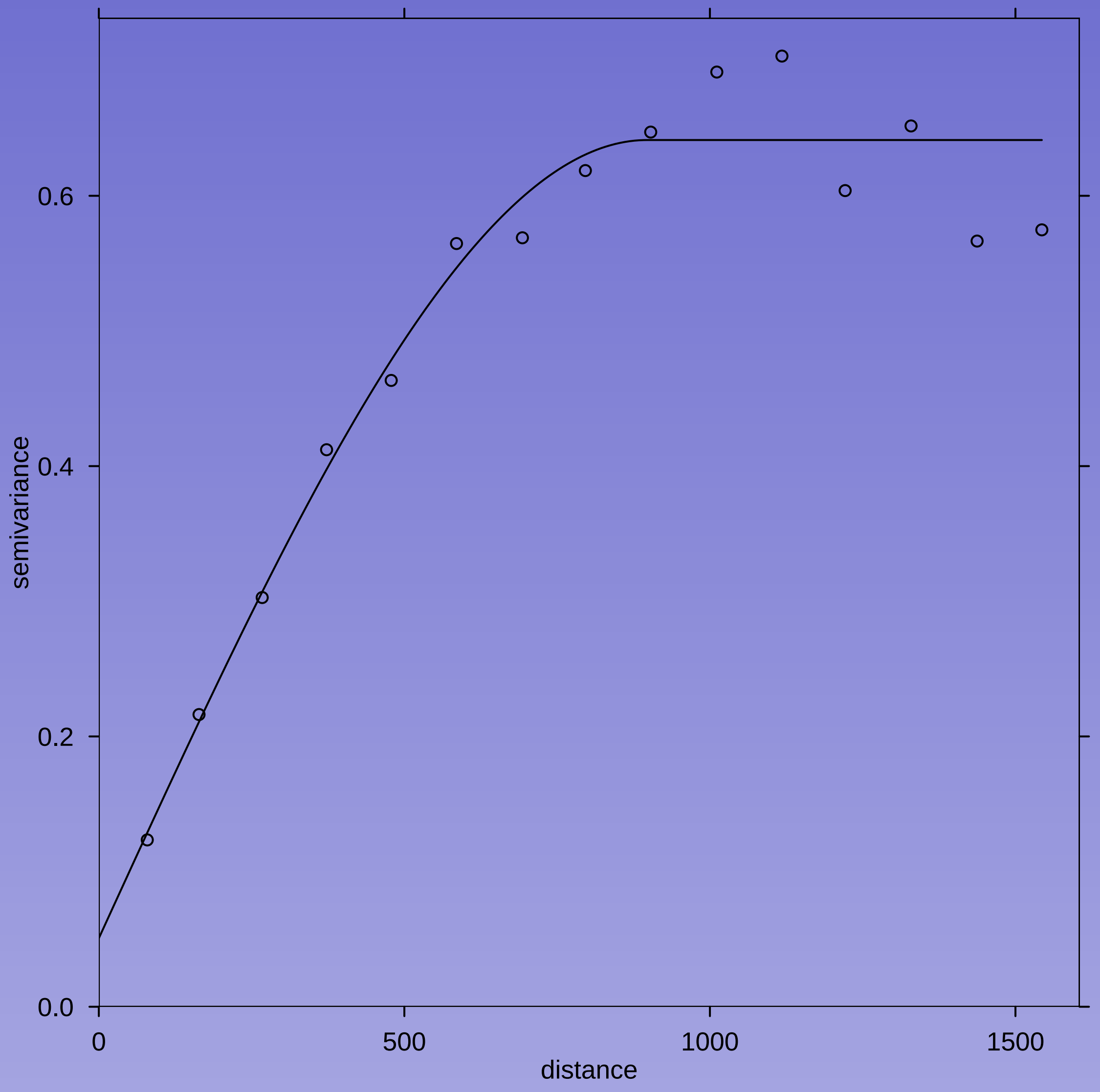
SAGIS2, 2005

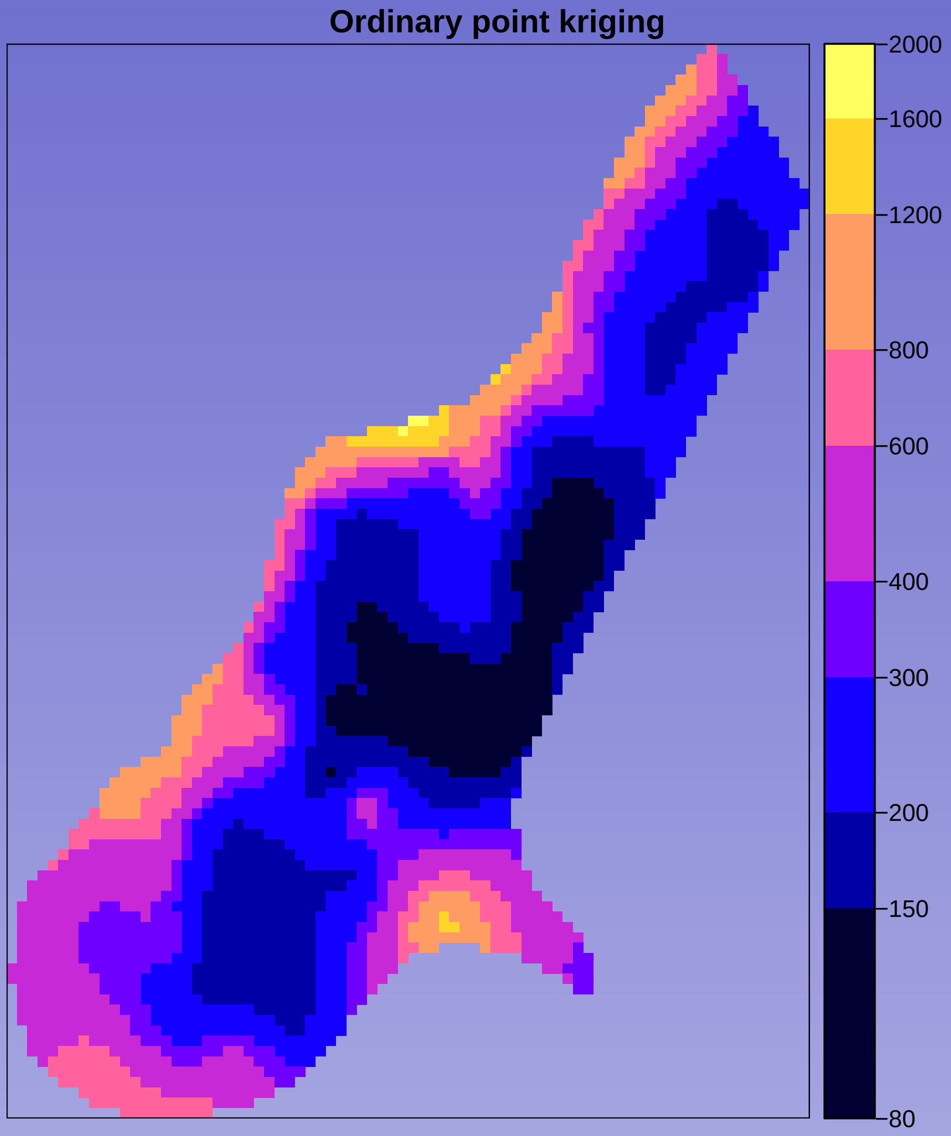

1−nearest neighbour (‘‘Thiessen polygons'')
2000
1600
1200
800
600
400
300
200
150
80

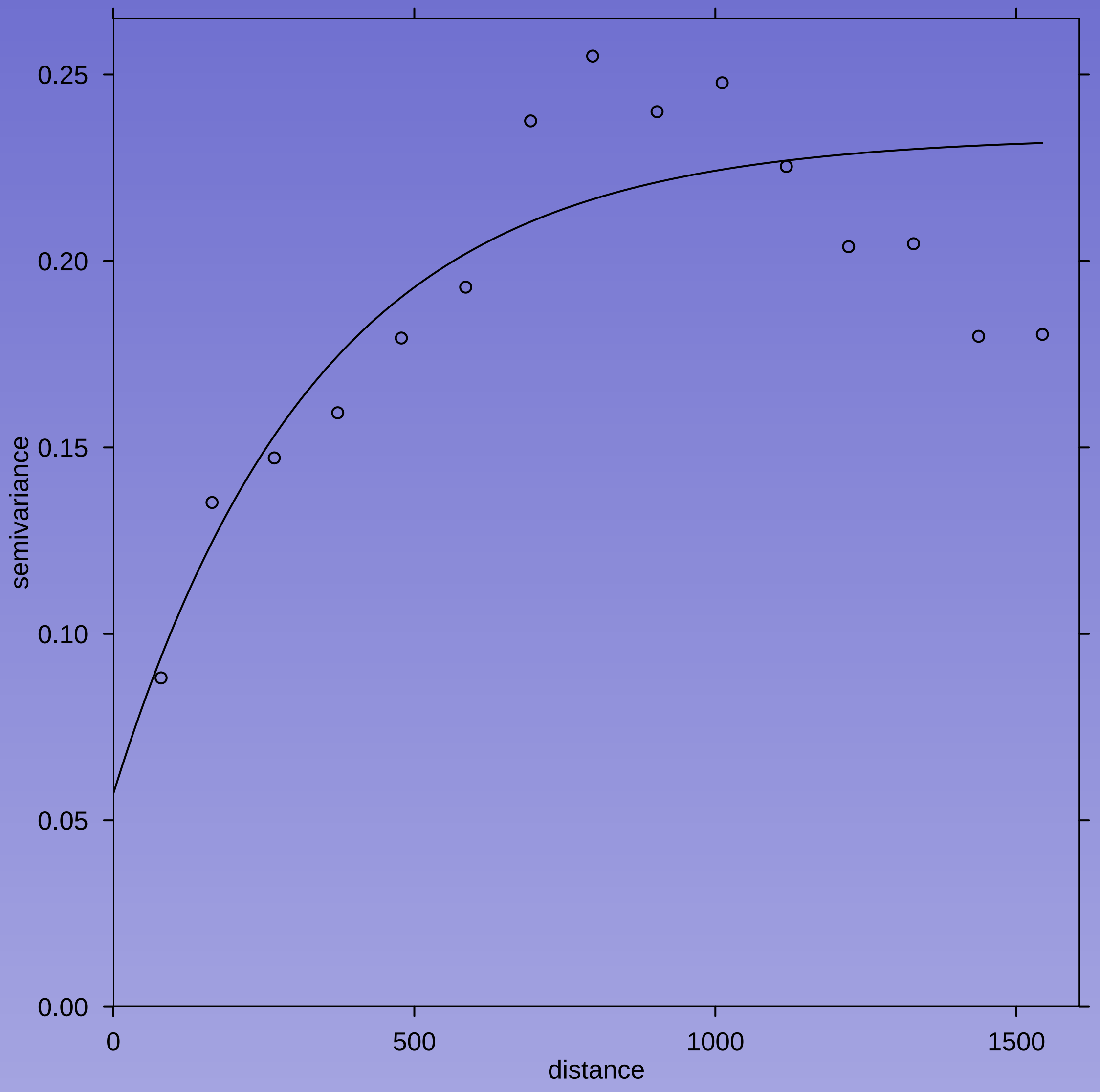

log(zinc)
7.5
7.0
6.5
6.0
5.5
5.0
0.0
0.2
0.4
0.6
0.8
sqrt(dist)

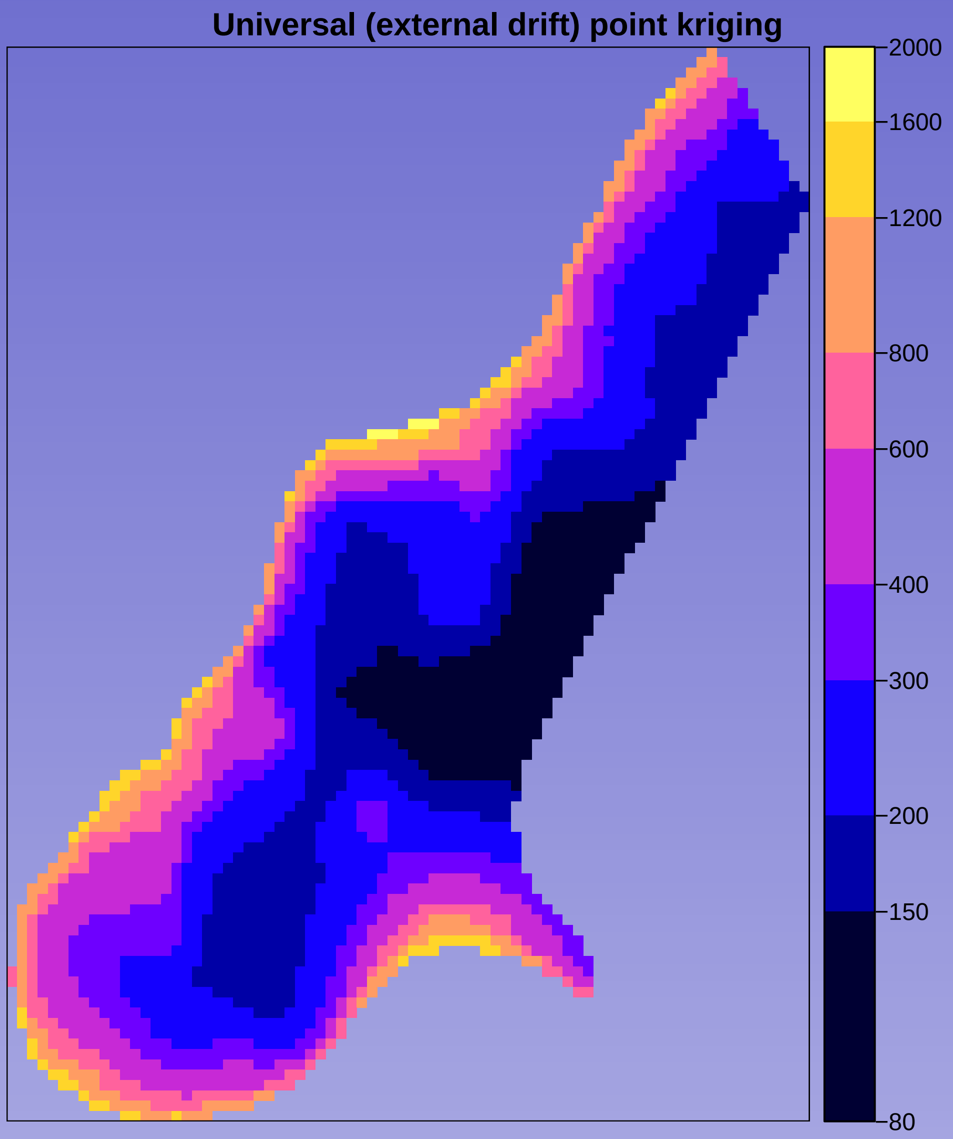

first order linear trend
2000
1600
1200
800
600
400
300
200
150
80


Inverse distance weighted; idp = 2
2000
1600
1200
800
600
400
300
200
150
80


semivariance
0.0
0.2
0.4
0.6
0
500
1000
1500
distance


Ordinary point kriging
2000
1600
1200
800
600
400
300
200
150
80


0.25
0.20
0.15
0.10
0.05
0.00
semivariance
0
500
1000
1500
distance



Universal (external drift) point kriging

Spatial statistics

“Statistics for spatial data” (Noel Cressie, 1993):

- point pattern data: a pattern where the actual spatial locations are of interest (e.g. are they random or clustered – diseases, crime scenes)
- lattice data: attributes are measured on regions that collectively form the study area, e.g. postal code regions, NUTS regions, image pixels
- geostatistical data: a variable has been sampled on some set of locations; the interest is the value of that variable on any set of locations (e.g. pollution, gold concentration)

Primary data

- measured attribute
- spatial location (x, y; z?), locations projected
- other attributes ...

GIS data base

- x and y coordinates of prediction locations
- land use, soil type
- elevation (DEM)
- distance to key features (pollution/diffusion source (point/line) or sink; breeding colony, ...)
- remotely sensed images

Simple approaches to spatial prediction

- linear regression:
    - ⋆ using an "external" predictor
    - ⋆ using coordinates as predictors
    - ⋆ global, or local?
    - ⋆ weighted, using distance?
- categorical predictors: spatial ANOVA
- inverse distance weighted

Linear regression as spatial predictor

Examples:

- `> lm(log(zinc)~sqrt(dist), meuse)`
- rainfall and orography (altitude)
- temperature and altitude, or latitude (scale!)
- log(pollution) and distance to source, measured along flow path

Trend surface interpolation

Polynomials in x, y (or x, y, z):

$$Z(x,y) = \beta_0 + \beta_1 x + \beta_2 y + e(x,y)$$

$$Z(x,y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy + e(x,y)$$

coefficient vector β is globally constant:

- allows testing, etc.
- testing assumes... independence

Local trend surface interpolation

For predicting $Z(x_0, y_0)$, given the model

$$Z(x, y) = \beta_0 + \beta_1 x + \beta_2 y + e(x, y)$$

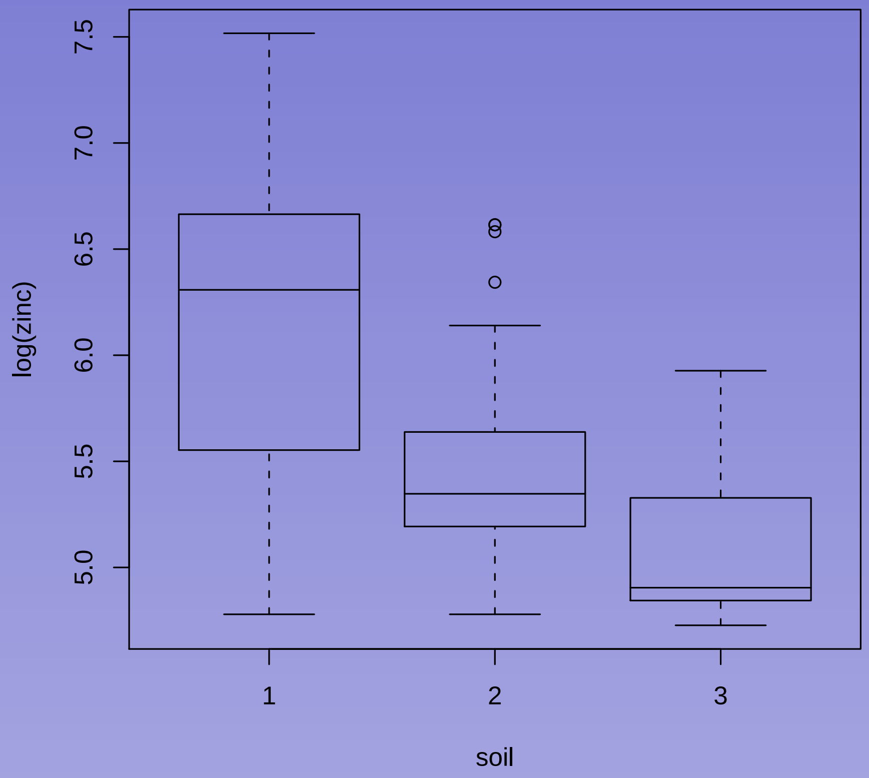
pick only data in a local neighbourhood around (x_0, y_0)

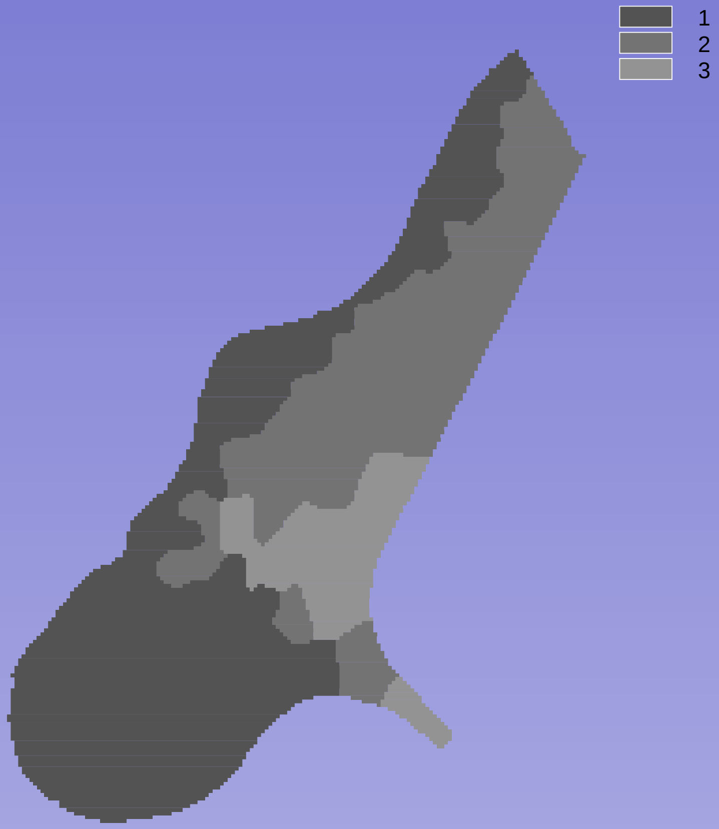
How to define a neighbourhood?

- distance
- number of nearest observations (n must be larger than p [$=$ nrows(β)]!)
- combined criteria

Problem: surface is discontinuous; solution: use loess, splines

Categorical predictors: ANOVA





Inverse distance interpolation

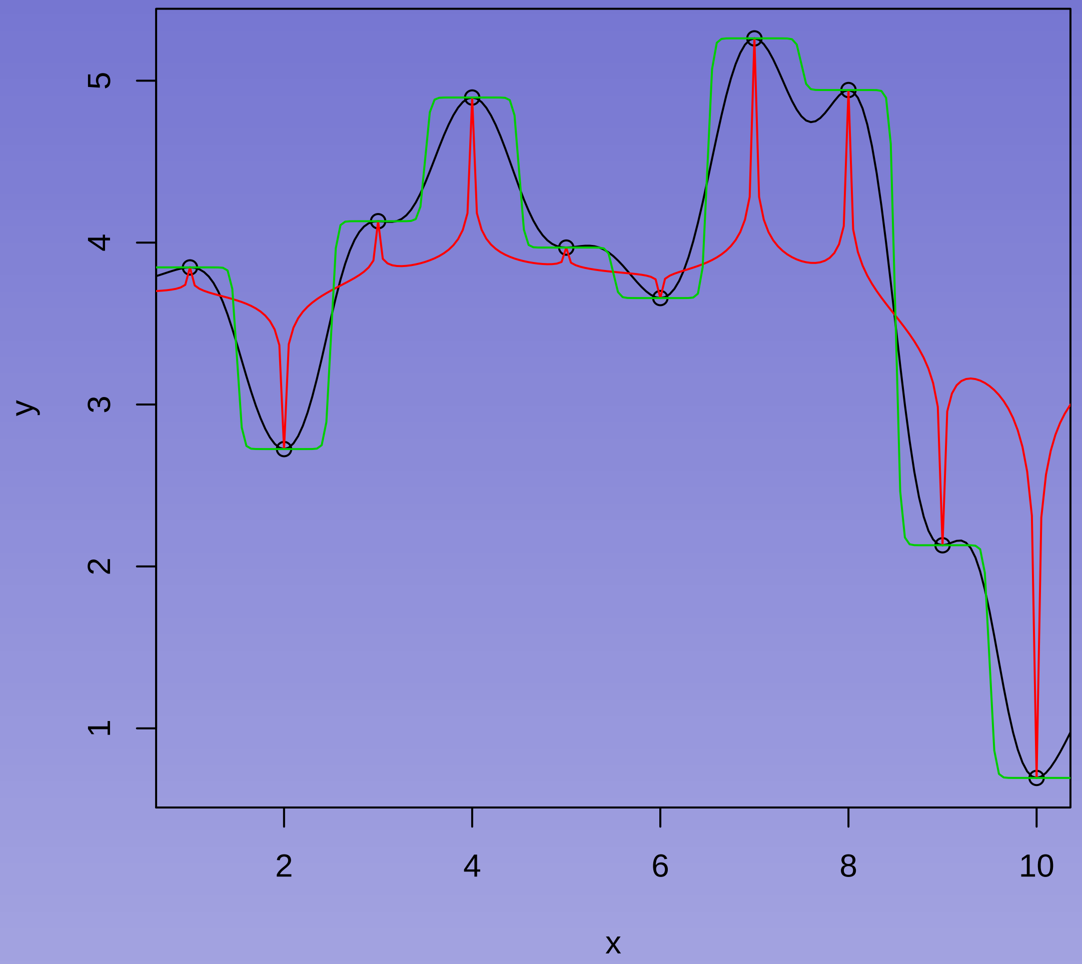
Use a weighted average:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

with $s_0 = \{x_0, y_0\}$, or $s_0 = \{x_0, y_0, \text{depth}_0\}$ weights inverse proportional to power p of distance:

$$\lambda_i = \frac{|s_i - s_0|^{-p}}{\sum_{i=1}^{n} |s_i - s_0|^{-p}}$$

- power p: tuning parameter
- if for some i, $|s_i - s_0| = 0$, then $\lambda_i = 1$ and other weights become zero
- $\Rightarrow$ exact interpolator


5
4
3
2
1
y
2
4
6
8
10
x

inverse distance power: 2, .5, 10

GWR - geographically weighted regression

e.g. `log(zinc) ~ sqrt(dist)`; prediction of $Z(s_0)$

- apply in a local neighbourhood
- apply weights, inverse proportional to $|s_0 - s_i|$; functions:
  - ⋆ Gaussian
  - ⋆ Bi-cubic
  - ⋆ span?
- compare to loess:
  - ⋆ loess: regression *and* weighting in covariates,
  - ⋆ GWR: regression in covariates, weighting in geographic distances

- books by R.S. Fotheringham and co-workers

Random variables and random functions

A random variable (RV) Z is a variable whose outcome is subject to chance. A continuous RV Z has a distribution function: $F_Z(x) = Pr(Z \leq x)$ which can be written as

$$F_Z(x) = \int_{-\infty}^{x} f_Z(u)du$$

with $f_Z(x) \geq 0$, and $f_Z(x)$ (defined as) the probability density function.

Expectation: $\mathrm{E}(Z) = \int_{-\infty}^{\infty} xf(x)dx$

Variance: $\mathrm{Var}(Z) = \mathrm{E}[(Z - \mathrm{E}(Z))^2]$

Covariance: $\mathrm{Cov}(Y, Z) = \mathrm{E}[(Y - \mathrm{E}(Y))(Z - \mathrm{E}(Z))]$

A set of random variables, $Z(s)$, $s \in \{s_1, s_2, ..., s_n\}$ is called a *random function*

$\mathrm{Var}(\lambda_1 Z_1 + \lambda_2 Z_2) = \lambda_1^2 \mathrm{Var}(Z_1) + \lambda_2^2 \mathrm{Var}(Z_2) + 2\lambda_1\lambda_2 \mathrm{Cov}(Z_1, Z_2)$

For vector $\mathbf{Z} = [Z_1 \; Z_2 \; ... \; Z_n]'$: $\mathrm{Var}(\lambda'\mathbf{Z}) = \lambda'\mathrm{Var}(\mathbf{Z})\lambda$

Best linear prediction (*a.k.a.* simple kriging)

Suppose we know μ, and $Z(s) = \mu + e(s)$. The linear predictor

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) = \lambda' Z$$

has variance

$$\mathrm{Var}(Z(s_0) - \hat{Z}(s_0)) = \mathrm{Var}(Z(s_0) - \lambda' Z)$$

which can be written as

$$\mathrm{Var}(Z(s_0) - \lambda' Z) = \mathrm{Var}(Z(s_0)) + \lambda' \mathrm{Var}(Z) \lambda - 2\lambda' \mathrm{Cov}(Z(s_0), Z)$$

so we need all variances of $Z(s_0)$ (scalar), of Z (matrix) and covariances of $Z(s_0)$ and Z (vector).

Next, find weights such that $\mathrm{Var}(Z(s_0) - \hat{Z}(s_0))$ is minimized, and we have the best (minimum variance) linear predictor.

Best linear prediction weights

Let $V = \mathrm{Var}(Z) \quad (n \times n)$ and $v = \mathrm{Cov}(Z(s_0), Z) \quad (n \times 1)$, and scalar $\mathrm{Var}(Z(s_0)) = \sigma_0^2$.

Expected squared prediction error $\mathrm{E}(Z(s_0) - \hat{Z}(s_0))^2 = \sigma^2(s_0)$

Replace Z with $Z - \mu$ (assume $\mu = 0$)

$\sigma^2(s_0) = \mathrm{E}(Z(s_0) - \lambda' Z)^2 = \mathrm{E}(Z(s_0))^2 - 2\lambda'\mathrm{E}(Z(s_0)Z) + \lambda'\mathrm{E}(ZZ')\lambda$

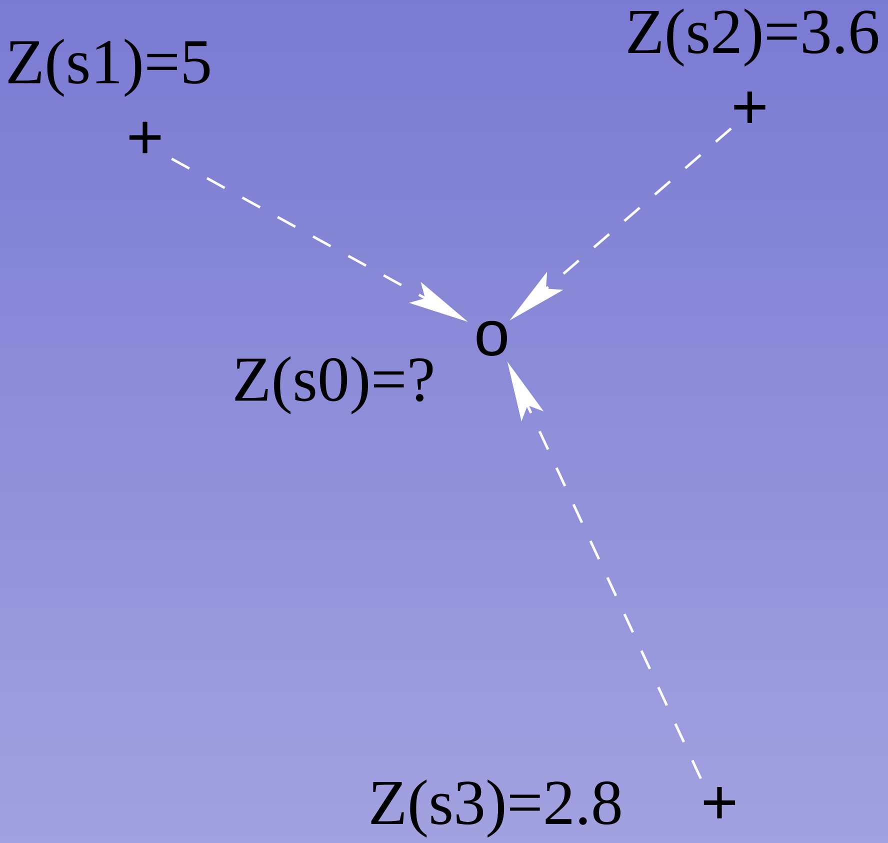
$= \mathrm{Var}(Z(s_0)) - 2\lambda'\mathrm{Cov}(Z(s_0), Z) + \lambda'\mathrm{Var}(Z)\lambda = \sigma_0^2 - 2\lambda' v + \lambda' V \lambda$

Choose λ such that $\frac{\delta\sigma^2(s_0)}{\delta\lambda} = -2v + 2\lambda' V = 0$

$\lambda' = vV^{-1}$

BLP: $\hat{Z}(s_0) = \mu + v'V^{-1}(Z - \mu) \qquad \sigma^2(s_0) = \sigma_0^2 - v'V^{-1}v$

Spatial Prediction



Stationarity 1

Given prediction location s_0, and data locations s_1 and s_2, we need: $\mathrm{Var}(Z(s_0))$, $\mathrm{Var}(Z(s_1))$, $\mathrm{Var}(Z(s_2))$, $\mathrm{Cov}(Z(s_0), Z(s_1))$, $\mathrm{Cov}(Z(s_0), Z(s_2))$, $\mathrm{Cov}(Z(s_1), Z(s_2))$.

How to get these covariances?

- given a single measurement $z(s_1)$, we can not infer $\mathrm{Var}(Z(s_1))$
- given two measurements $z(s_1)$ and $z(s_2)$, we can *never* infer $\mathrm{Cov}(Z(s_1), Z(s_2))$
- geven a time series at s_1 and s_2, we can infer $\mathrm{Cov}(Z(s_1), Z(s_2))$, but how to infer $\mathrm{Cov}(Z(s_0), Z(s_1))$ and $\mathrm{Cov}(Z(s_0), Z(s_2))$?

Solution: assume stationarity.

Stationarity 2

Stationarity of the

mean $\mathrm{E}(Z(s_1)) = \mathrm{E}(Z(s_2)) = \ldots = m$

variance $\mathrm{Var}(Z(s_1)) = \mathrm{Var}(Z(s_2)) = \ldots = \sigma_0^2$

covariance $\mathrm{Cov}(Z(s_1), Z(s_2)) = \mathrm{Cov}(Z(s_3), Z(s_4))$ if $s_1 - s_2 = s_3 - s_4$: distance/direction dependence

Second order stationarity: $\mathrm{Cov}(Z(s), Z(s+h)) = C(h)$

which implies: $\mathrm{Cov}(Z(s), Z(s)) = \mathrm{Var}(Z(s)) = C(0)$

The function $C(h)$ is the covariogram of the random function $Z(s)$

From covariance to semivariance

Covariance: $\mathrm{Cov}(Z(s), Z(s+h)) = C(h) = \mathrm{E}[(Z(s)-m)(Z(s+h)-m)]$

Semivariance: $\gamma(h) = \frac{1}{2}\mathrm{E}[(Z(s)-Z(s+h))^2]$

$\mathrm{E}[(Z(s)-Z(s+h))^2] = \mathrm{E}[(Z(s))^2 + (Z(s+h))^2 - 2Z(s)Z(s+h)]$

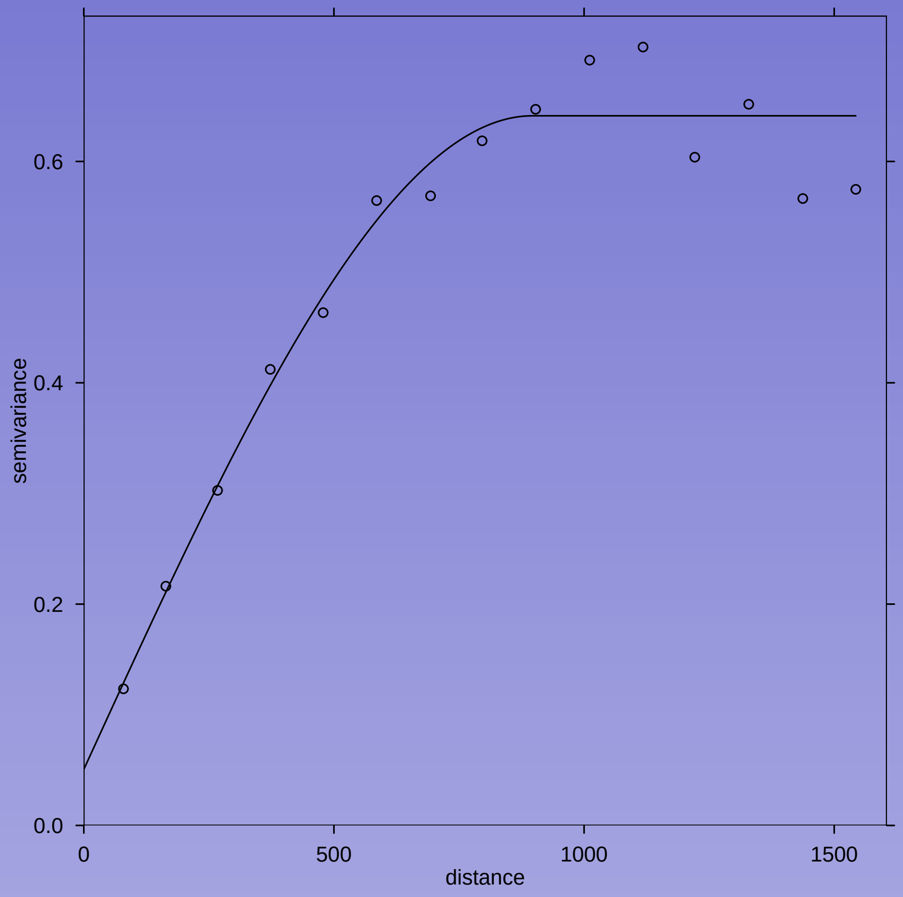
[Assume $m=0$]:

$\mathrm{E}[(Z(s)-Z(s+h))^2] = \mathrm{E}[(Z(s))^2] + \mathrm{E}[(Z(s_h))^2] - 2\mathrm{E}[Z(s)Z(s+h)] =$
$2\mathrm{Var}(Z(s)) - 2\mathrm{Cov}(Z(s), Z(s+h)) = 2C(0) - 2C(h)$

$\gamma(h) = C(0) - C(h)$

$\gamma(h)$ is the semivariogram of $Z(s)$.

The *Variogram*



The *Variogram*

- the central tool to geostatistics
- like a mean squares (variance) in analysis of variance, like a t to a t-test
- measures spatial correlation
- subject to debate: it involves *modelling*
- synonymous to *semivariogram*, but
- semivariance is *not* synonymous to variance

Variogram: how to compute

average squared differences:

$$\gamma(\tilde{h}) = \frac{1}{2N_h} \sum_{i=1}^{N_h} (Z(s_i) - Z(s_i + h))^2 \quad h \in \tilde{h}$$

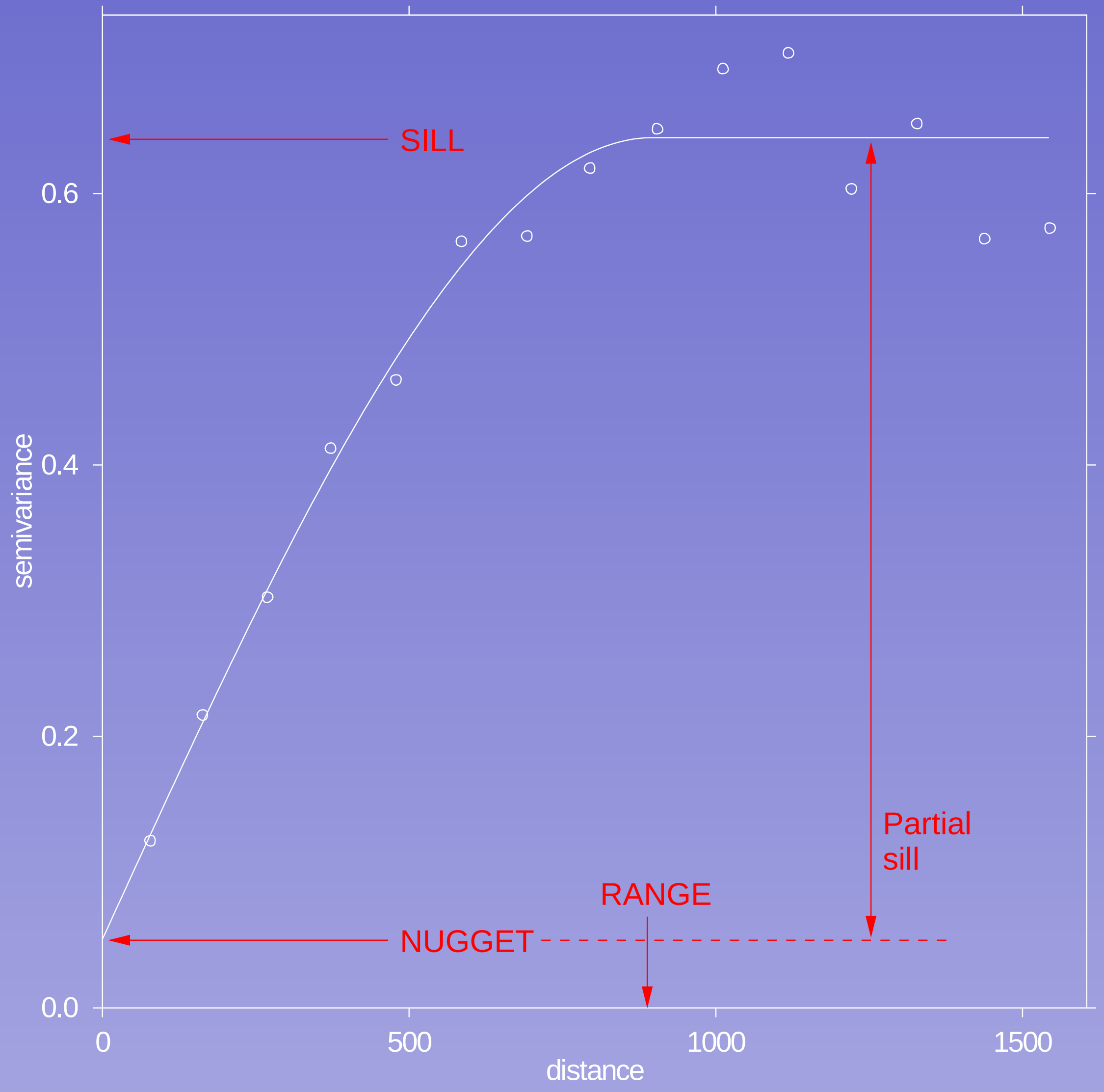
- divide by $2N_h$:
  - ⋆ if finite, $\gamma(\infty) = \sigma^2$
  - ⋆ *semi*variance

- if data are not gridded, group N_h pairs $s_i, s_i + h$ for which $h \in \tilde{h}$, $\tilde{h} = [h_1, h_2]$

- choose about 10-25 distance intervals $\tilde{h}$, from length 0 to about on third of the area size

- “plot” $\tilde{h}$ at the average value of all $h \in \tilde{h}$

Variogram: terminology


SILL
Partial
sill
RANGE
NUGGET
semivariance
0.0
0.2
0.4
0.6
distance
0
500
1000
1500

gstat coding (R):

```
> vgm(psill = 0.6, model = "Sph", range = 900, nugget = 0.06)
  model psill range
1   Nug  0.06     0
2   Sph  0.60   900
> vgm(0.6, "Sph", 900, 0.06)
  model psill range
1   Nug  0.06     0
2   Sph  0.60   900
```

Why prefer the variogram over the covariogram

Covariance: $\mathrm{Cov}(Z(s), Z(s+h)) = C(h) = \mathrm{E}[(Z(s)-m)(Z(s+h)-m)]$

Semivariance: $\gamma(h) = \frac{1}{2}\mathrm{E}[(Z(s)-Z(s+h))^2]$

$\gamma(h) = C(0) - C(h)$

- tradition
- $C(h)$ needs (an estimate of) m, $\gamma(h)$ does not
- $C(0)$ may not exist (∞!), when $\gamma(h)$ does (e.g., Brownian motion)
- *software* wants $\gamma(h)$.

Ordinary kriging

- Simple kriging: $Z(s) = \mu + e(s)$, μ known
- Ordinary kriging: $Z(s) = m + e(s)$, m unknown
- SK: linear predictor $\lambda' Z$ with λ such that $\sigma^2(s_0) = \mathrm{E}(Z(s_0) - \lambda' Z)^2$ is minimized
- OK: linear predictor $\lambda' Z$ with λ such that it

  1. has minimum variance $\sigma^2(s_0) = \mathrm{E}(Z(s_0) - \lambda' Z)^2$, and
  2. is unbiased $\mathrm{E}(\lambda' Z) = m$

- second constraint: $\sum_{i=1}^{n} \lambda_i = 1$, weights sum to one.

- BLUP: $\hat{Z}(s_0) = \hat{m} + vV^{-1}(Z - \hat{m})$
  with $\hat{m} = (\mathbf{1}'V^{-1}\mathbf{1})^{-1}\mathbf{1}'V^{-1}Z$, and
  $\sigma^2(s_0) = \sigma_0^2 - v'V^{-1}v + (1 - \mathbf{1}'V^{-1}v)'(\mathbf{1}'V^{-1}\mathbf{1})^{-1}(1 - \mathbf{1}'V^{-1}v)$

gstat: status of project

- open source (GPL) project, `http://www.gstat.org/`, started in 1992
- `gstat` (or `gstat.exe`) is a stand-alone binary that
  - ⋆ input: e.g. data (e.g. through GDAL) and uses gnuplot to show variograms
  - ⋆ reads (ascii) data and (usually) writes maps.
- `gstatw.exe`: gstat+GUI; stand-alone; *very* limited functionality, little development expected
- `gstat` library: an S (S-PLUS or R) library that manipulates data in an S data environment; developmented; successor or gstat stand-alone
- 30.000 lines of ANSI-C code, 1500 lines of S code

- gstat and gstat S library are fully documented
- recent: gstat S library depends on `sp`
- upcoming: “Applied Spatial Data Analysis with R”, by R Bivand & E Pebesma

Kriging in a local neighbourhood

OK: $Z(s) = m + e(s)$

- instead of assuming m globally constant, we can assume it is (only) constant in a local neighbourhood around s_0 (expressed in distance, or number of nearest points)
- *local* stationarity of mean
- for each neighbourhood, m is re-estimated
- the smaller the neigbhourhood, the more it costs
- OK, neighbourhood size 1: 1-nearest neighbour predictor

- OK, neighbourhood size 0: missing value
- SK, neighbourhood size 1: prediction between nearest neighbour and μ
- OK, neighbourhood size 0: prediction is μ
- large neighbourhood ($n \gg 50$): prediction is *practically* identical to kriging in global neighbourhood (SK, OK; not UK)
- if we have many data (e.g. $n \gg 1000$), kriging in global neighbourhood becomes cumbersome because of computation of V^{-1} $\Rightarrow$ kriging in a large, local neighbourhood may be much faster

Support: point kriging

- measurements have a certain *support*: the physical (spatial, temporal) “size” of the sample that was measured.

- we call this the *point support*, although strictly speaking, unlike points, point support is larger than zero.

- the larger the support, the lower the variability
  - compare hourly, daily averaged, and yearly averaged temperatures
  - compare gauge rainfall, rainfall averaged over 1 km^2, or rainfall averaged over 100 km^2.
  - not an easy concept when using bulk sampling, soil mixture samples etc.

- predictions usually refer to estimates for quantities that would have been measured on the same support as that of the measurements (point support prediction; point kriging)

Support: point or block kriging

- predictions of mean values for areas, larger than the point support is called *block kriging*; predict $Z(B_0) = |B_0|^{-1} \int_{B_0} Z(u)du$

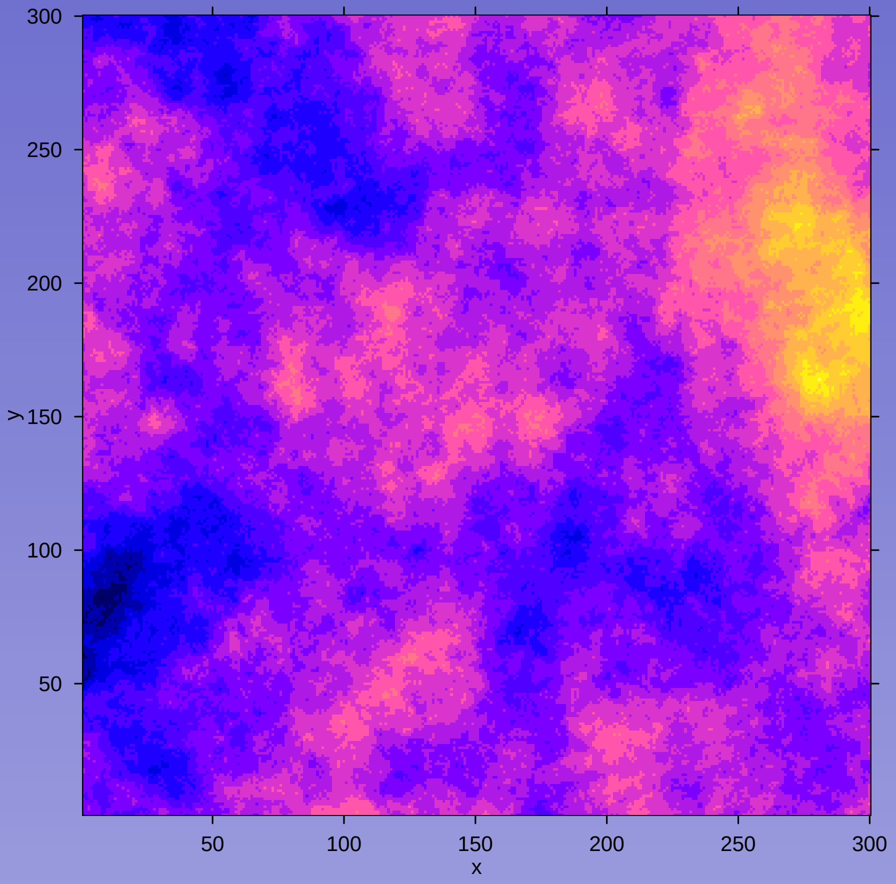
- the larger the support of the block, the smaller the prediction errors that come with it

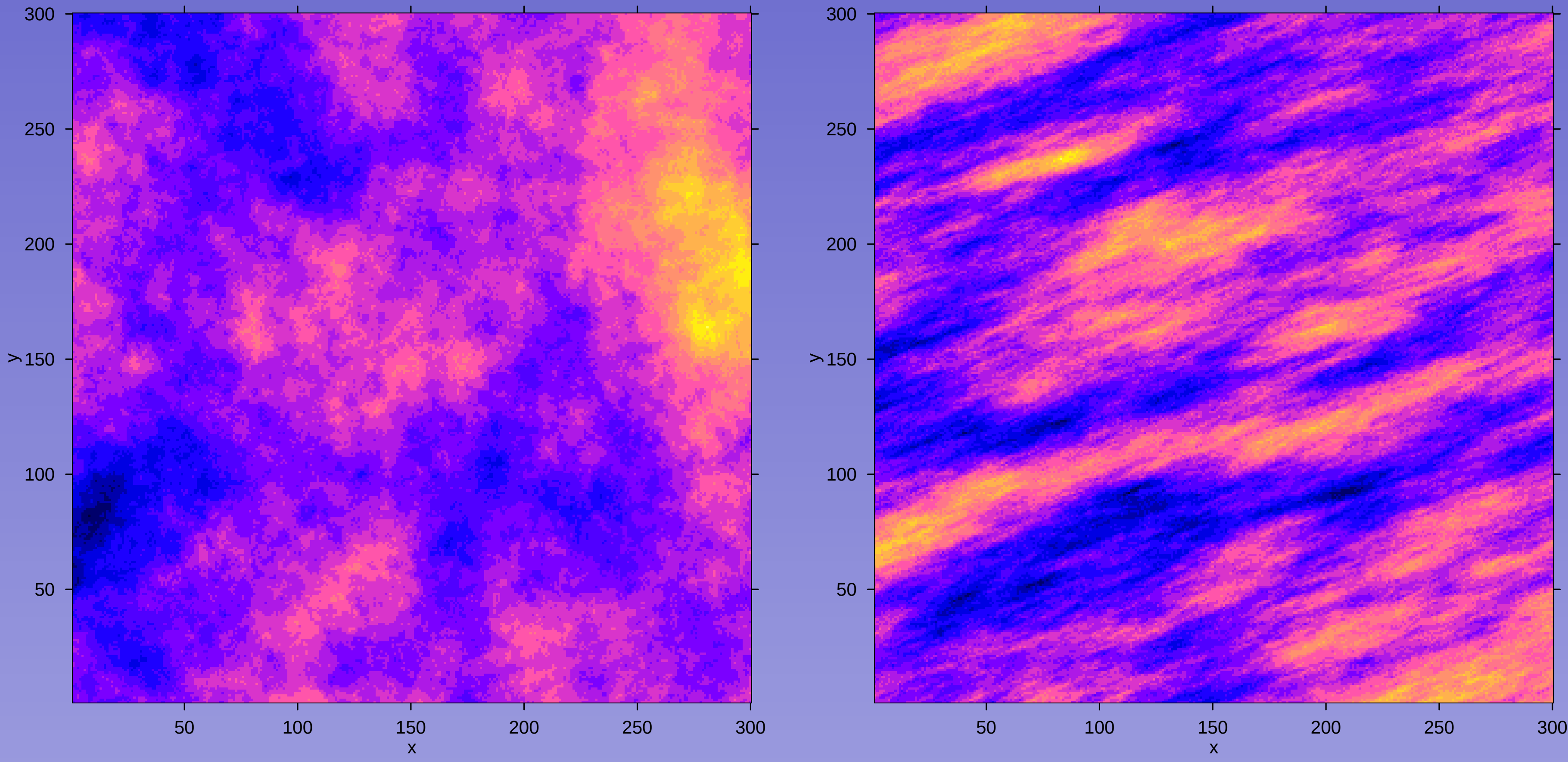
- how large blocks should we choose? Some ideas:
  - ⋆ trade-off: larger blocks have smaller prediction errors, but less spatial resolution (in the end, the block covers the study area)
  - ⋆ is legislation related to a target support?
  - ⋆ the size of model grid cells
  - ⋆ the size of units that can e.g. be mined (ore) or excavated (polluted soil)

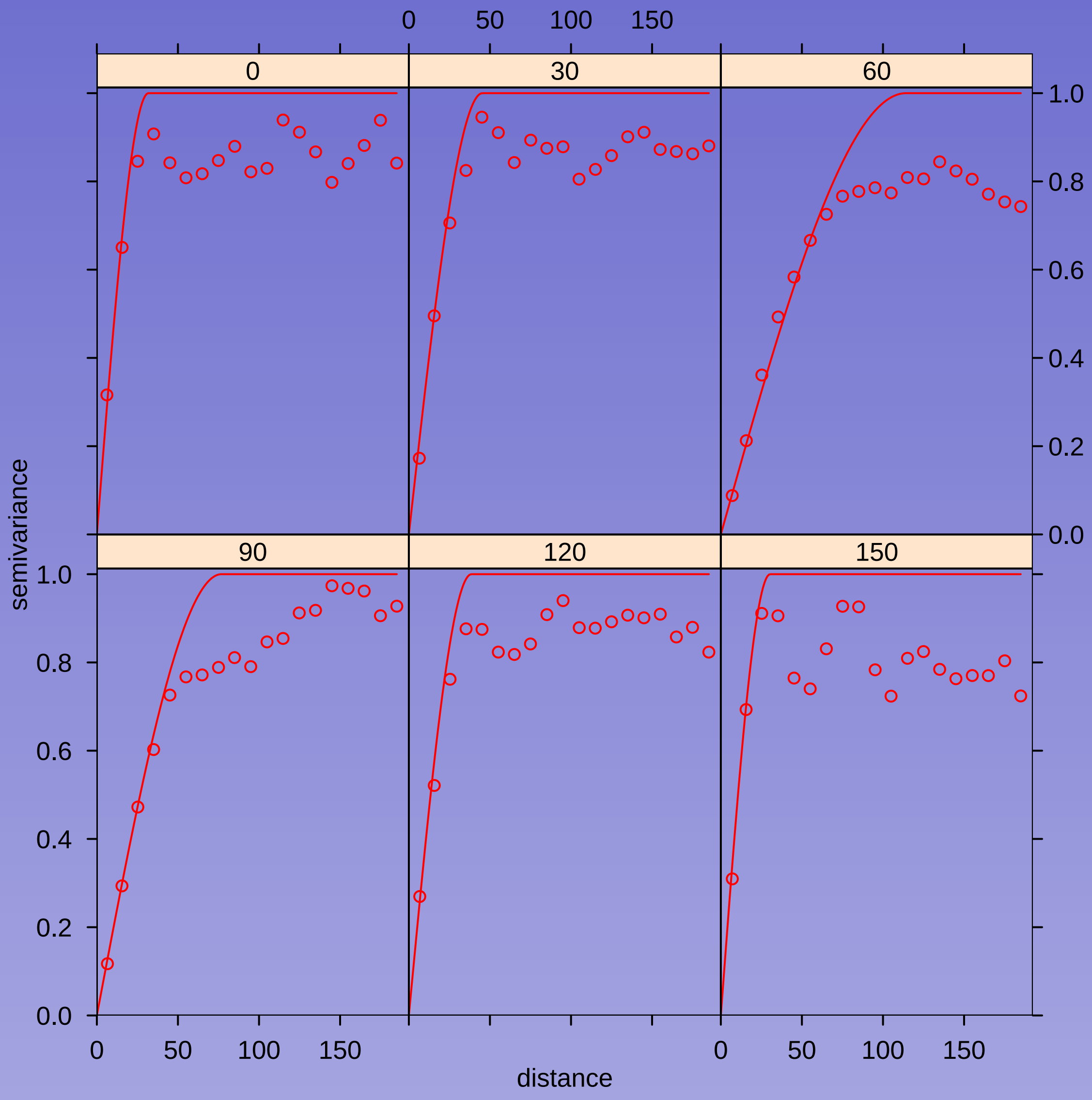
- ⋆ related to monitoring network density: how much of the spatial pattern is *lost*?

Isotropy and anisotropy

- spatial correlation may depend on direction
- usually it will, but to what extent?
- large samples are needed to explore this


300
250
200
y 150
100
50
50
100
150
200
250
300
x


300
250
200
y 150
100
50
50
100
150
200
250
300
x
300
250
200
y 150
100
50
50
100
150
200
250
300
x


0
30
60
90
120
150
semivariance
distance
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150

Box 6.2. Computing kriging weights

Box 6.2 in Burrough and McDonnell contains errors.

- Use the excel sheet `kriging_graphics.xls` instead (available from http://webct.uu.nl).

- authors: Hans Zuuring (Univ of Montana) and Derek Karssenberg

- lets you interact with:
  - ⋆ sample locations
  - ⋆ sample values
  - ⋆ variogram parameters: nugget, sill, range

Kriging standard deviation (or variance)

The kriging standard deviation $\sigma(s_0)$ (or kriging variance, $\sigma^2(s_0)$) is (or should be) a measure for the *expected* error $Z(s_0) - \hat{Z}(s_0)$, or prediction *accuracy*.

The kriging standard deviation:

- depends on data configuration (closeness to s_0, degree of clustering)
- does not depend on data values (!)
- is zero at observation locations
- is smaller for blocks than for points
- is different for different variogram models

- is smaller when the variogram is smaller (i.e., lower)

“Wrong” variograms will yield invalid prediction standard deviations. ⇒ cross validation.

Prediction intervals

- (kriging) prediction yields $\hat{Z}(s_0)$, not $Z(s_0)$.
- we don't know $Z(s_0)$, but we know the average magnitude of $Z(s_0) - \hat{Z}(s_0)$ (zero), and its variance: $\sigma^2(s_0)$
- if the prediction error is normally distributed, the *prediction interval*

$$[\hat{Z}(s_0) - 2\sigma(s_0), \quad \hat{Z}(s_0) + 2\sigma(s_0)]$$

covers, with approximately 95% probability, the true value $Z(s_0)$

- why *approximately*?

- ⋆ kriging assumes we *know* the variogram, $\gamma(h)$, whereas we can only estimate it (cf. $Z \Rightarrow t$)
- ⋆ variables never follow a normal distribution

Stretching stationarity

- (anisotropy: direction dependent variograms)
- transform $Z(s)$ non-linearly (e.g. log, sqrt, Box-Cox)
- transform the geographic space non-linearly
- re-fit variograms in each local neighbourhood
- stratify the area $\Rightarrow$ stratified kriging
- modify (extend) trend function $\Rightarrow$ universal kriging

Stratified kriging

Instead of considering the whole study area as a single variable, split it into "homogeneous" sub-areas, that are different with respect to

- mean levels
- variability, or spatial correlation structure (variogram)
- both

Within each sub-area (stratum), model the variogram using within-stratum data, and interpolate.

Sub-areas should be known a priory and never derived from the interpolation data. Examples: soil maps, land use coverages, hydrological sub-systems, geomorphological units, ...

Each stratum should have sufficient observations for variogram modelling.

Universal kriging/external drift

Universal kriging extends the ordinary kriging model

$$Z(s) = m + e(s)$$

to the more general models

$$Z(s) = \beta_0 + \beta_1 f_1(s) + \beta_2 f_2(s) + ... + \beta_p f_p(s) + e(s) = F(s)\beta + e(s)$$

- kriging involves for each location (i) estimation of the trend, (ii) prediction of the residual
- universal kriging is still “exact”, reproducing observations
- the variogram needed is the *residual* variogram of $e(s)$

- if the regressors $f_j(s)$ explain a considerable part of the variation in $Z(s)$, the residual variability in $e(s)$ is much smaller than the variation in $Z(s)$, the variogram sill is much lower, and *prediction will be more accurate* (prediction errors will be smaller)

Cross validation

(Point) kriging yields the observed value at an observation location, so comparison of $\hat{Z}(s_0)$ and $Z(s_0)$ is not informative about the quality of the spatial interpolation. Residuals are zero.

When we want to evaluate (compare) different interpolation methods, different kriging variaties, or a kriging variaty with different variogram models, we use cross validation. One version of cross validation is leave-one-out cross validation.

Leave-one-out cross validation: for each observation $Z(s_i)$

- take $Z(s_i)$ out of the data set
- interpolate, given the remaining data, $\hat{Z}_{[i]}(s_i)$

- calculate the residual $Z(s_i) - \hat{Z}_{[i]}(s_i)$ and the normalized residuals, or z-score,

$$\frac{Z(s_i) - \hat{Z}_{[i]}(s_i)}{\sigma(s_i)}$$

Cross validation statistics

Ideally

- the correlation between $Z(s_i)$ and $\hat{Z}_{[i]}(s_i)$ should be close to 1
- the variability of $\hat{Z}_{[i]}(s_i)$ should be close to that of $Z(s_i)$ (but will always be smaller: the *smoothing effect*)
- residuals should have zero mean, and small variance (small range, etc.)
- z-scores should contain no outliers (values outside, say $[-3, 3]$)
- z-scores should have unit standard deviation (only this "validates" the prediction error standard deviation).

Indicator kriging of nominal variables

Suppose we want to interpolate a binary, nominal variable, e.g. sand or clay. We can code this into a 0/1 variable: 0 if an observation is sand, 1 if it is clay. A 0/1 variable is also called an indicator variable.

We can interpolate 0/1 values, as any values, after modelling its variogram. Note that if the fraction of 1 is p, the variance (sill) of the variable is $p(1-p)$.

The interpolated map shows values mostly between 0 and 1, some are outside this range.

These interpolated values can be seen as estimated "probabilities" that the interpolated value is 1 (i.e., sand).

Block kriging indicator values yields estimates of the *fraction* of 1-values within the block, *not* the probability that the block *mean* is 1.

Indicator kriging of continuous variables

For continuous variables with a very weird distribution (e.g. counts with many zeros), it sometimes is a good idea to work with indicator transforms of the variable:

$$I(Z(s),c) = \begin{cases} 1 & \text{if } Z(s) \leq c, \\ 0 & \text{otherwise} \end{cases}$$

for one or more cutoffs c. Indicator kriging results can be seen as *estimates* of $Pr(Z(s) \leq c)$, the probability that $Z(s)$ is less than c.

- Order relation violations:
  - ⋆ $\hat{I}(Z(s),c) < 0$
  - ⋆ $\hat{I}(Z(s),c) > 1$
  - ⋆ $\hat{I}(Z(s),c_1) > \hat{I}(Z(s),c_2)$ when $c_1 < c_2$

- For indicator *block* kriging estimates, the result should be regarded as estimates of the *fraction* of the block where $Z(s) \leq c$, and not as an estimate of the probability $Pr(Z(B_0) \leq c)$:

$$I(\int_{B_0} Z(u)du, c) \neq \int_{B_0} I(Z(u), c)du$$

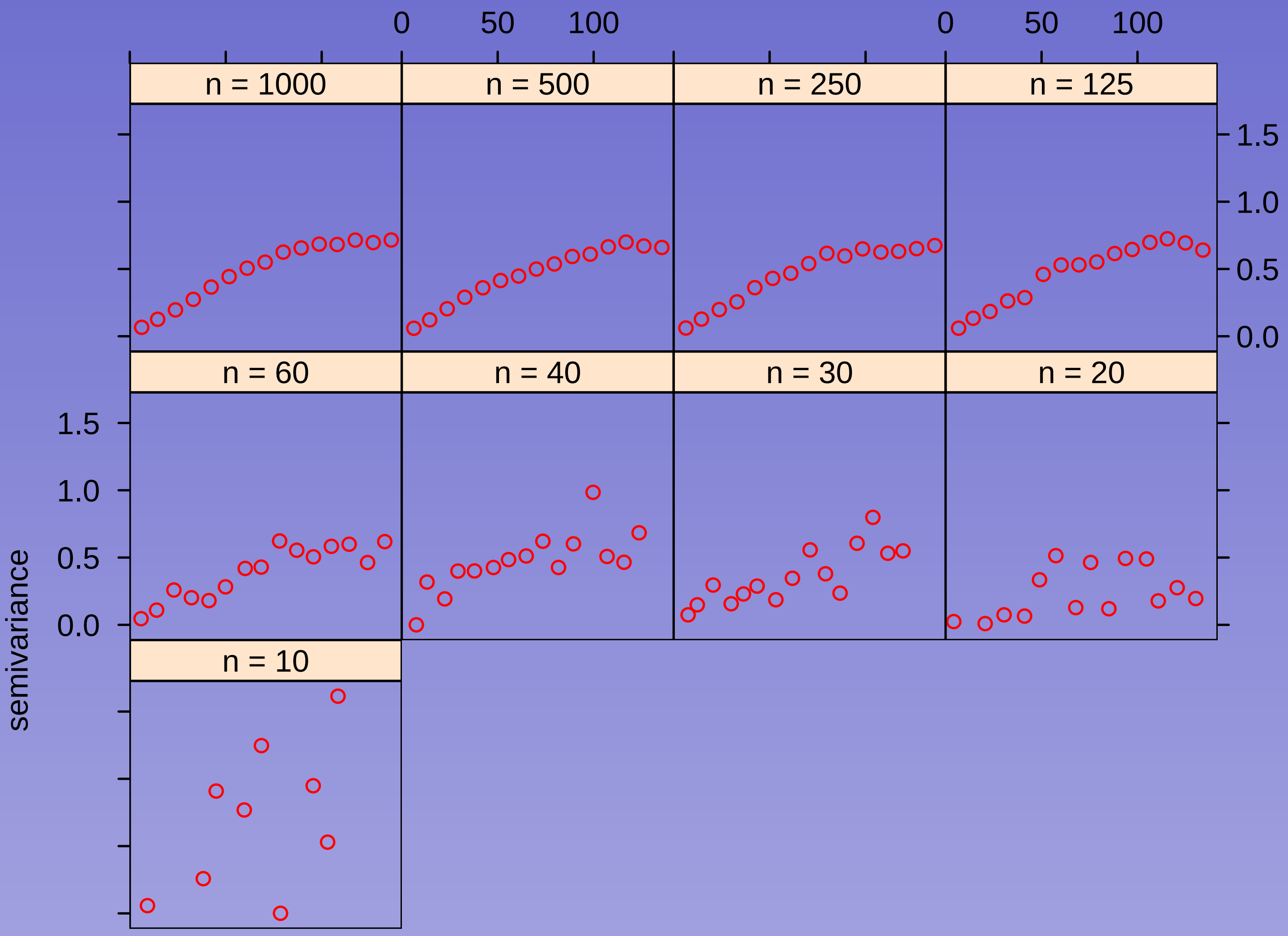
- For the case in the computer course, I would not recommend this but rather use logarithms, and a Gaussian distribution: assuming that $Z(s_0)$ follows a Gaussian distribution with mean $\hat{Z}(s_0)$ and standard deviation $\sigma(s_0)$, $Pr(Z(s_0) \leq c$ can be derived for any c.

Data requirements for variogram estimation

Geostatistical interpolation requires that the interpolated data provide sufficient information to estimate the variogram. At least three factors play a role for this: sample size, sample configuration, and data distribution.

Variogram estimation becomes harder:

- the smaller the sample size gets
- the more the data are restricted to a few clusters, or the more they are (too much) regularly spaced, lacking short distances
- the more the distribution of the data is dominated by a few extreme values (large absolute skewness; extreme outliers)


0
50
100
0
50
100
n = 1000
n = 500
n = 250
n = 125
n = 60
n = 40
n = 30
n = 20
n = 10
1.5
1.0
0.5
0.0
1.5
1.0
0.5
0.0
semivariance

Sample configuration and variogram estimation

There is no agreement on a “universally” optimal sampling configuration for geostatistical research (i.e., variogram modelling, followed by spatial prediction), but:

- for spatial prediction, regular (lattice, or triangular) sampling is optimal (in case of isotropy; otherwise stretched lattices);
- for variogram modelling, all distances should be present, *including sufficient information about short distances* (which are not present when sampling regularly)
- cross validation on a regular sampling grid will not reveal deficiencies in modelled short distance behaviour of the variogram; interpolated maps will be dominated by this short distance variogram behaviour.

- compromise: most effort put to regular spread, sufficient effort to short distance replicates.

- related questions: adding sampling points to an existing design, or reducing ("optimizing") an existing monitoring network.

What you should know about kriging

- what do sill, nugget, range, and anisotropy tell about spatial variability of an observed variable?
- what happens if we predict a value at an observation location? and what if we do a block kriging at an observation location?
- what does the prediction variance measure?
- what is the difference between point and block kriging, in terms of predictions and prediction variances?
- why is the interpolator discontinuous at observation locations when the nugget is positive?

- why is the prediction variance pattern independent on data, but only dependent on data configuration?
- what are the causes for positive nugget effect?
- how to interpret cross validation statistics, how to choose between interpolation methods based on cross validation
- what is meant by the smoothing effect?

Geostatistical software

Free:

- gstat: stand alone; S-PLUS library / R package
- other R packages: geoR, geoRglm, sgeostats, vardiag: http://cran.r-project.org
- gslib: FORTRAN library (Stanford; http://www.gslib.com)
- GsTL: C++ template library (Stanford; use google)

Commercial:

- S-PLUS: S+SpatialStats module (http://www.insightful.com)

- ArcGIS geostatistical analysist ($2500) (http://www.esri.com)
- Isatis (http://www.geovariance.fr); GoCAD (google)

What's left?

- conditional simulation
- multivariable geostatistics: cokriging
- case studies: (i) groundwater quality in the Netherlands, (ii) spatio-temporal trends in sediment pollution from NCP sea floor sediment sample data

Conditional simulation

Idea: generate a large set of fields (realizations) that

- honour the data (are conditioned to the data)
- on average, reflect the kriging prediction and variance
- each have a spatial variability, equal to that of the data (in contrast to the kriging prediction map, which is much smoother than the data)
- when to use it? When Z is input to a non-linear model, e.g. some transport or flow model.

Sequential Gaussian simulation algorithm

For a set of prediction locations (the “mask” map), repeat

1. pick a random, unvisited prediction location, call it s_0
2. given the observed and simulated data, calculate the (simple) kriging mean $\hat{Z}(s_0)$ and kriging standard deviation $\sigma(s_0)$
3. draw a random variate from the Gaussian distribution with mean $\hat{Z}(s_0)$ and standard deviation $\sigma(s_0)$
4. add the value to the data set

until all prediction locations have been visited

Cokriging

- instead of a single variable $Z(s)$, we have multiple variables $Z_1(s), ..., Z_m(s)$ that are (spatially) cross correlated

- spatial cross correlation: cross variograms

  ⋆ $\gamma_{a,b}(h) = \mathrm{E}((Z_a(x) - Z_a(x+h))(Z_b(x) - Z_b(x+h)))$
  ⋆ $\tilde{\gamma}_{a,b}(h) = \mathrm{E}((Z_a(x) - m_a)(Z_b(x+h) - m_b))$

- $Z_2(s), ...Z_m(s)$ help for the estimation of $Z_1(s_0)$ (compare universal kriging)

- result is prediction vector $(\hat{Z}_1(s_0), ..., \hat{Z}_m(s_0))$ and prediction error covariance matrix!

Contrasts

given $\mathbf{y}(s_0) = (y_{86}(s_0), y_{91}(s_0), y_{96}(s_0), y_{00}(s_0))'$, we can calculate *contrasts*

$$C(s_0) = \lambda' \mathbf{y}(s_0)$$

- four-year mean: $\lambda' = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
- difference '86+'91 vs. 96+00: $\lambda' = (-\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$
- yearly increase: $\lambda' = (-0.065, -0.02, 0.025, 0.061)$
- SE: λ'Cov$(\mathbf{y}(s_0))\lambda$ is available!