University of Helsinki
Department of Mathematics and Statistics
Spatial analysis of area data

Final exam, January 2009

Select **four (4)** out of the five assignments.

1. Explain briefly the meaning of the following concepts

   a) spatial autocorrelation
   b) spatial regression
   c) spatial smoothing
   d) correlogram
   e) pseudolikelihood
   f) Gaussian Markov random field

2. Brook's lemma can be presented as

$$\frac{p_{\boldsymbol{Y}}(\boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y}_0)} = \prod_{i=1}^{n} \frac{p_i(y_i \mid Y_j = y_{j0}, j < i; Y_j = y_j, j > i)}{p_i(y_{i0} \mid Y_j = y_{j0}, j < i; Y_j = y_j, j > i)},$$

   where $\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$ and $\boldsymbol{y}_0 = \begin{bmatrix} y_{10} & y_{20} & \dots & y_{n0} \end{bmatrix}^T$.

   a) What does $p_{\boldsymbol{Y}}(\boldsymbol{y})$ mean in this formula?
   b) How about $p_i(y \mid Y_j = y_j, j \neq i)$?
   c) Give a general (verbal) definition of a CAR model.
   d) What is the connection of Brook's lemma to CAR models?

3. Suppose we are simulating some distribution $p(\boldsymbol{x})$ using a Metropolis–Hastings -method, and the current value of $\boldsymbol{x}$ is $\boldsymbol{x}^{(t)}$.

   a) How is the next value $\boldsymbol{x}^{(t+1)}$ determined (describe the phases of this update; the equations are not necessary)?
   b) Why should the correlation between $x^{(t)}$ and $\boldsymbol{x}^{(t+1)}$ be as low as possible?
   c) What can one do to control this correlation?

4. Values (x_i, y_i), $i = 1, 2, \ldots, n$, of variables x and y were observed from n regions, and model

$$y_i = a + bx_i + u_i$$

was fitted to the observed data

i) assuming the residuals u_i to be identically distributed and mutually independent with expected values $Eu_i = 0$, $i = 1, \ldots, n$, and

ii) assuming that

$$u_i = \rho \sum_{j=1}^{n} w_{ij} u_j + e_i,$$

where the e_i's are identically distributed and mutually independent with expected values $Ee_i = 0$, $i = 1, \ldots, n$,

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a common border,} \\ 0, & \text{if not,} \end{cases}$$

and ρ is a parameter, whose value is estimated from the data.

The obtained estimates of parameters a, b, and ρ, their standard errors and p-values in the tests of null hypotheses $H_0 : a = 0$, $H_0 : b = 0$ $H_0 : \rho = 0$ were

i)

Parameter	estimate	std.error	p-value
a	-2.5328	0.4104	1.16e-06
b	0.3018	0.1368	0.0358

ii)

Parameter	estimate	std.error	p-value
a	-0.65610	0.92594	0.4786
b	0.25252	0.15959	0.1136
ρ	0.44226	0.026412	< 2.22e-16

a) Interpret results obtained with assumptions i).

b) Interpret results obtained with assumptions ii).

c) What do you think is the reason for the difference in results?

d) Which model would you choose? Why?

e) What is your conclusion on the relationship between x and y?

f) By which name is the model based on assumtions ii) known?

5. Suppose that random vector $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T$ follows the three-dimensional normal distribution with mean vector $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$ and precision matrix

$$\boldsymbol{Q} = \begin{bmatrix} 1 & -\rho & 0 \\ -\rho & 2 & -\rho \\ 0 & -\rho & 1 \end{bmatrix},$$

where $0 < \rho < 1$, so that the probability density function of the distribution of $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) = C(\rho) \exp(-\tfrac{1}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}),$$

where $C(\rho)$ is a constant, whose value does not depend on $\boldsymbol{x}$. Derive the density function $p(x_2 \mid x_1, x_3)$ of the conditional distribution of x_2 given x_1 and x_3.