UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in STK4150 — Environmental and spatial statistics

Day of examination: Friday, 8 June, 2012.

Examination hours: 09.00 – 13.00.

This problem set consists of 4 pages.

Appendices: None

Permitted aids: Approved calculator, Cressie and Wikle: Statistics for Spatio-temporal data

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

Consider a random process $\{Y(\boldsymbol{s}); \boldsymbol{s} \in \mathcal{D} \subset \mathcal{R}^2\}$.

(a) Specify the exact requirements for $\{Y(\boldsymbol{s}); \boldsymbol{s} \in \mathcal{D}\}$ to be a Gaussian random field.

What are the main advantages in using the Gaussian random fields for specifying joint distributions for spatially dependent data?

(b) Assume now that $Z(\boldsymbol{s}_i), i = 1, ..., n$ are count observations at the observation locations $\boldsymbol{s}_i, i = 1, ..., n$. Assume more specifically that

$$\begin{aligned} E[Y(\boldsymbol{s})] =& \mu(\boldsymbol{s}) \\ \text{Cov}[Y(\boldsymbol{s}), Y(\boldsymbol{s}')] =& C_y(\boldsymbol{s}, \boldsymbol{s}') \\ Z(\boldsymbol{s}_i)|\boldsymbol{Y} \overset{ind}{\sim}& \text{Poisson}(\exp(Y(\boldsymbol{s}_i))) \end{aligned}$$

What are the main advantages in using a hierarchical model compared to direct modelling of a joint distribution for $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n))^T$.

(c) Under the assumptions above and for simplicity now also assuming $C_y(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}) = C_y^0(\boldsymbol{h})$, find $E[Z(\boldsymbol{s}_i)]$ and $\text{Cov}[Z(\boldsymbol{s}), Z(\boldsymbol{s} + \boldsymbol{h})]$.

Hint: If $x \sim N(\mu, \sigma^2)$, then

$$E[\exp(x)] = \exp(\mu + 0.5\sigma^2)$$

Note: In particular finding the covariance is a bit difficult, perhaps you should go further to the next points before spending too much time on this.

(Continued on page 2.)

The rates of lip cancer in 56 counties in Scotland have been analyzed by Clayton and Kaldor (1987) and Breslow and Clayton (1993). The form of the data includes the observed and expected cases (expected numbers based on the population and its age and sex distribution in the county), a covariate measuring the percentage of the population engaged in agriculture, fishing, or forestry, and the "position" of each county expressed as a list of adjacent counties.
A possible model for these data are

$$Z_i \sim \text{Poisson}(\exp(Y_i))$$
$$Y_i \sim \log(E_i) + \beta_0 + \beta_1 x_i/10 + b_i$$

where β_0 is an intercept term representing the baseline (log) relative risk of disease across the study region, x_i is the covariate "percentage of the population engaged in agriculture, fishing, or forestry" in district i, with associated regression coefficient β_1 and b_i is an area-specific random effect capturing the residual or unexplained (log) relative risk of disease in area i. We often think of b_i as representing the effect of latent (unobserved) risk factors.
We will consider three different models for the b_i's:

Model 1 The b_i's are iid and $N(0,\tau_1^{-1})$.

Model 2 The b_i's follow a CAR model with $b_i|\boldsymbol{b}_{-i} \sim N(\frac{1}{n_i}\sum_{j\sim i} b_j, 1/(n_i\tau_2))$

Model 3 Each $b_i = b_i^{iid} + b_i^{car}$ where the b_i^{iid}'s are iid as for model 1 and the b_i^{car}'s follow a CAR model as for model 2

In all cases, the b_i's are restricted to sum to zero. Note that the τ's here are *precisions.*

(d) Is it possible to put these models into the general model discussed in (*b*)?

What is the benefit in including two random effects in Model 3?

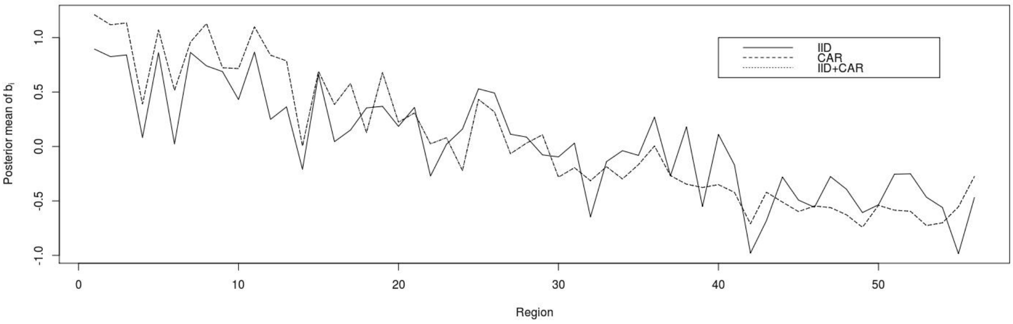
(e) The table below show estimates and standard errors for the parameters involved for the three models based on a Bayesian approach with vague priors (actually based on the INLA program)

	Model 1		Model 2		Model 3	
Param	Estimate	SE	Estimate	SE	Estimate	SE
β_0	-0.4893	0.1560	-0.2999	0.1273	-0.3001	0.1283
β_1	0.6832	0.1394	0.4056	0.1379	0.4055	0.1397
τ_1	3.0863	0.8988	NA	NA	19227	19201
τ_2	NA	NA	2.0529	0.7499	2.051	0.7513

(Continued on page 3.)

Furthermore, the figure below shows predictions of the b_i's for the 56 regions (the predictions for models 2 and 3 are not possible to distinguish in this plot).

Based on these results, which model would you prefer?



Problem 2

Consider the pure time series model

$$Y_t = aY_{t-1} + \varepsilon_t, \quad t = 1, 2, ..., T \tag{1}$$

where $\{\varepsilon_t\}$ are iid and $N(0, \sigma_\varepsilon^2)$.

(a) Assume $\text{var}(Y_t) = \sigma_Y^2$ for all t. Find σ_Y^2. What requirements on a are needed?

(b) Show that (1) is a Markov Random Field (MRF) in one dimension. What is the neighborhood system in this case.

Now extend the model to a spatio-temporal process $\{Y_t(\boldsymbol{s}_i), i = 1, ..., m, t = 0, 1, ...\}$, discrete both in time and space, where

$$\boldsymbol{Y}_t = \boldsymbol{M}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t \tag{2}$$

where $\boldsymbol{Y}_t = (Y_t(\boldsymbol{s}_1), ..., Y_t(\boldsymbol{s}_m))^T$ and $\boldsymbol{\varepsilon}_t = (\varepsilon_t(\boldsymbol{s}_1), ..., \varepsilon_t(\boldsymbol{s}_m))^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$ and where the ε's are independent in time.

(c) Assume $\mathsf{var}(\boldsymbol{Y}_t) = \boldsymbol{\Sigma}_Y$ for all t. What requirements are needed for $\boldsymbol{M}$ in order to make this possible? You do not need to derive these requirements but specify them and argue why such requirements are reasonable.

(d) Assume $\boldsymbol{Y}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0$ is specified such that the stationary assumption from (c) is fulfilled. Show that

$$\begin{aligned}\log p(\boldsymbol{Y}) = \text{Const} &- \tfrac{1}{2}(\boldsymbol{Y}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}((\boldsymbol{Y}_1 - \boldsymbol{\mu}_0) \\ &- \tfrac{1}{2}\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{M}\boldsymbol{Y}_{t-1})^T\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{Y}_t - \boldsymbol{M}\boldsymbol{Y}_{t-1})\end{aligned}$$

where $p(\boldsymbol{Y})$ is the joint density of all $Y_t(\boldsymbol{s}_i)$'s.

(Continued on page 4.)

Assume now that $\boldsymbol{Q}_\varepsilon = \boldsymbol{\Sigma}_\varepsilon^{-1}$ is sparse such that $Q_{i,j} = 0$ if $j \notin \mathcal{Q}_i$ while $\boldsymbol{M}$ is diagonal. Show that (2) is a Markov Random Field and find the neighborhood structure in this case.

(e) Consider in this point a general $\underline{\mathrm{M}}$ (i.e. it does not have to be diagonal anymore). Find $\mathsf{cov}[Y_t(\boldsymbol{s}_i), Y_{t+\tau}(\boldsymbol{s}_j)]$ for all $t, \tau, \boldsymbol{s}_i, \boldsymbol{s}_j$.

What kind of simplifying structure do these covariances have. Discuss the benefits of these simplifying structures.

END