Project for STK4150/9150 - Spring 2013

Geir Storvik

May 8, 2013

This is the problem set for the project part of the finals in STK4150/9150. The reports shall be individually written. The deadline for turning in the reports is Wednesday May 22 at 2 pm. Two copies marked with your candidate number must be submitted to the reception office at the Department of Mathematics at the seventh floor in N. H. Abels house. The report should only contain your answers/results and perhaps relevant plots. It should not extend more than 10 pages. The report must include a 1 page summary of your main findings. Computer code must be included, but as attachments and not part of the actual report (and need not be within the 10 page limit). Handwritten reports are acceptable. When you refer to material in attachments, be careful to indicate explicitly where.

Importantly, each student needs to submit a special extra page with her or his report. This page is the "erklæring" (self-declaration form), properly signed, and with the appropriate course form STK 4150 (master level) or STK 9150 (PhD level) clearly marked; it is available at the webpage as Exam Project, declaration form.

The problem set contains three exercises. For the two first, the sub-questions in total describe a minimal analysis that is required. You may score higher by considering additional aspects. Note however your space-limits, making it important to be clever in your choices of possible supplements.

Exercise 1 (Air quality data)
The file *Piemonte_data_byday.csv* contains data on PM_{10} (particulate matter with an aerodynamic diameter of less than 10 μm) daily observed over a period from 01/10/05 to 31/03/06 at 24 stations in the North-Italian region Piemonte. This is a dataset taken from Cameletti et al. [2011] which we will analyze using spatio-temporal models later. Here we will concentrate on models where the spatial correlation is ignored. In order to make the response approximately Gaussian distributed, we will consider $\log(PM_{10})$.
8 covariates are available:

`WS` daily mean wind speed (m/s),

`HMIX` daily maximum mixing height (m),

`P` daily precipitation (mm),

`TEMP` daily mean temperature ($^\circ K$),

`EMI` daily emissions (g/s),

`A` altitude (m)

`UTMX/UTMY` spatial geographic coordinates (km).

The dataset can be read into R by the commands

```
d <-read.table("Piemonte_data_byday.csv",header=TRUE,sep=",")
ndata = nrow(d)
nstations=length(unique(d$Station.ID))
ndays = as.integer(ndata/nstations)
d$Time = rep(1:ndays,each=nstations)
d$logPM10 = log(d$PM10)
```

(a). Consider first a simple regression model

$$y_t(s_i) = \mu + \mathbf{x}_t(s_i)^T \boldsymbol{\beta} + \varepsilon_t(s_i)$$

where $\mathbf{x}_t(s_i)$ is the vector of covariates at time t and station i, *Time* excluded.

Fit such a model using ordinary least squares and use different tools for checking whether the residuals are uncorrelated.

Hint: Possible commands are

```
fit = lm(logPM10~A+UTMX+UTMY+WS+TEMP+HMIX+PREC+EMI,data=d)
res = rep(NA,nrow(d))
res[!is.na(d$logPM10)] = resid(fit)
acf(res[d$Station==1],na.action=na.pass)
```

(b). Describe a possible permutation test that can be performed on the residuals to check for independence in time. Perform the test and summarize your results.

(c). Include now *Time* as a covariate. Look at the acf curve of the residuals for the first station again. Does it look like some of the autocorrelation is removed?

(d). Include now *Time* as a categorical covariate by the command

```
d$Time = as.factor(d$Time)
options(contrasts=c("contr.sum","contr.sum"))
fit3=lm(logPM10~A+UTMX+UTMY+WS+TEMP+HMIX+PREC+EMI+Time,data=d)
```

Again look at the autocorrelation function for the residuals. How do the autocorrelation function look in this case?

Also apply your permutation test from (b) to test for independence.

(e). The model in (d) can be written as

$$y_t(s_i) = \mu + \mathbf{x}_t(s_i)^T\boldsymbol{\beta} + \alpha_t + \varepsilon_t(s_i)$$

where $\{\alpha_t\}$ are parameters, one for each time point. The command

```
options(contrasts=c("contr.sum","contr.sum"))
```

makes a constraint $\sum_t \alpha_t = 0$ in order to make all the parameters identifiable.

Find the estimates of α_t for all t (note that you have to calculate the last one using the constraint above) and plot this as a function of time. Also look at the autocorrelation function of the estimates.

Also apply your permutation test from (b) to test for independence.

(f). An alternative to consider the α_t's to be fixed parameters, is to consider them as random variables. A particular choice is to assume $\alpha_t \overset{iid}{\sim} N(0, \sigma_\alpha^2)$. Fit such a model with INLA using the commands

```
library(INLA)
formula = logPM10~A+UTMX+UTMY+WS+TEMP+HMIX+PREC+EMI+f(Time,model="iid")
fit4=inla(formula,data=d,control.inla=list(int.strategy="eb"))
```

The "estimates" of the α's are in this case available in the object

```
fit4$summary.random$Time[,2]
```

Why do I write "estimates" in this case?

Compare these estimates with the one obtained by considering the α's as fixed parameters. In particular, look at the square sum of the two types of estimates. Why do you think the square sum of the estimates based on the α's being treated as random are smaller?

(g). For the model where the α_t's are treated as random, calculate the covariance structure for $\{Y_t(\mathbf{s}_i)\}$.

(h). Now consider an extension where the α_t's are considered to follow an AR(1) process. Such a model can be fitted into INLA through the commands

```
formula2 = logPM10~A+UTMX+UTMY+WS+TEMP+HMIX+PREC+EMI+f(Time,model="ar1")
fit5=inla(formula2,data=d,control.inla=list(int.strategy="eb"))
```

What is the estimate of the autoregressive parameter in this case?

Compare the "estimates" of the α_t's for this model with the ones obtained using the iid assumption.

(i). Residuals from the previous model can be obtained by the commands

```
formula2 = logPM10~A+UTMX+UTMY+WS+TEMP+HMIX+PREC+EMI+f(Time,model="ar1")
fit5=inla(formula2,data=d,control.inla=list(int.strategy="eb"),
          control.predictor=list(compute=TRUE))
d$resid = d$logPM10-fit5$summary.fitted.values$mean
```

Use your permutation test on the residuals from this model fit. Summarize your findings.

(j). Calculate the covariance structure for $\{Y_t(\mathbf{s}_i)\}$ for the extended model.

Exercise 2 (Air quality data (cont))
Consider again the data from exercise 1. We will now extend the model by taking spatial dependence into account as well. In that case we need to read in the spatial coordinates for the data, which can be done with the following commands (including plotting the spatial points as well as the border of the region:

```
coordinates <-read.table("coordinates.csv",header=TRUE,sep=",")
borders = read.table("Piemonte_borders.csv",header=TRUE,sep=",")
plot(borders, lwd=3,type="l")
points(coordinates$UTMX, coordinates$UTMY,pch=20,col=2)
text(coordinates$UTMX, coordinates$UTMY,1:24)
```

We will first explore some of the spatial structure in the data.

(a). From the last model considered in the previous exercise, pick out the residuals corresponding to the first day of data and explore the spatial structure using different tools you have learned in the course.

Repeat for some other days. Also perform a permutation test for spatial independence on the whole dataset.

Summarize your findings.

In the following we will now try to build up a spatio-temporal model. In order to avoid some numerical problems, we will also reduce the dataset to the first 50 days:

```
d = d[d$Time<=50,]
```

Consider a model of the form

$$y_t(s) = \mu + \mathbf{x}_t(s)\boldsymbol{\beta} + \alpha_t + \delta(\mathbf{s}) + \varepsilon_t(\mathbf{s})$$

where now $\{\delta(\mathbf{s})\}$ is a zero-mean Gaussian spatial process. Further, we will in the beginning only use A as a covariate.

(b). Assume first that $\{\delta(\mathbf{s})\}$ has an independence structure. Such a model can be fitted by INLA with the commands

```
ind = rep(1:nstations,ndays)
formula = logPM10 ~  A + f(Time,model="ar1") +f(ind,model="iid")
res = inla(formula,data=d,control.inla=list(int.strategy="eb"))
```

Fit this model and extract the predictions of the δ's.

Use different methods to explore whether there are any spatial structure in the δ's.

(c). Assume now that $\{\delta(\mathbf{s})\}$ has an exponential covariance structure, that is

$$\text{Cov}[\delta(s), \delta(s+\mathbf{h})] = \sigma_\beta^2 \exp(-||\mathbf{h}||/\theta)$$

INLA does not have this covariance function available directly. Assume however that θ is known. Show that the precision matrix for $\boldsymbol{\delta} = (\delta(\mathbf{s}_1), ..., \delta(\mathbf{s}_n))$ can be written as $\tau\mathbf{C}$ where $\mathbf{C}$ is a known matrix. This situation is covered by the *generic0* model in INLA. Such a model can be fitted by the commands

```
dis = as.matrix(dist(coordinates[,c("UTMX","UTMY")],upper=TRUE,diag=TRUE))
theta = 8
Sigma.delta = exp(-dis/theta)
C0 = solve(Sigma.delta)
formula2=logPM10~A+f(Time,model="ar1")+f(ind,model="generic0",Cmatrix=C0)
res2 = inla(formula2,data=d,control.inla=list(int.strategy="eb"))
```

for $\theta = 8$. Perform these commands as well.

Also try out different θ. Specify some criterion to choose θ. Use the best value in the following.

Compare the model you get with the best θ value with the independence model. Which model gives the best fit?

(d). A further extension of the model is to assume

$$y_t(s) = \mu + \mathbf{x}_t(s)\boldsymbol{\beta} + \delta_t(\mathbf{s}) + \varepsilon_t(\mathbf{s})$$

where now $\{\delta_t(\mathbf{s})\}$ is a spatio-temporal process following the dynamic model

$$\delta_t(\mathbf{s}) = \rho\delta_{t-1}(\mathbf{s}) + \eta_t(\mathbf{s})$$

Such a model can be fitted in INLA using the commands

```
formula3 = logPM10 ~  A + f(Time,model="ar1") +
  f(ind,model="generic0",Cmatrix=C0,group=Time,control.group=list(model="ar1"))
res3 = inla(formula3,data=d,control.inla=list(int.strategy="eb"))
res3$mlik
```

Compare this model with the ones you have obtained earlier.

(e). Now repeat the points above including the covariates *WS,HMIX,P,TEMP,EMI*.

Why is it not necessary to include *UTMX,UTMY*?

Why does it seem like the spatial structure becomes less important now?

Also try out some alternative models. Which models do you prefer?

(f). Summarize your findings.

Exercise 3
In Cameletti et al. [2011] an alternative approach based on CAR models is used. We will not do any analysis on this. However, write a small essay about possible approaches based on CAR models that could be used in this case. Also include a discussion of pros and cons on such approaches compared to the geostatistical approach considered in exercise 2. (The paper cited above use a quite complicated approach, so you probably do not need to look at that, but rather do your own thinking).

References

M. Cameletti, F. Lindgren, DP Simpson, and H. Rue. Spatio-temporal modelling of particulate matter concentration through the spde approach. AStA Adv Stat Anal, Submitted, 2011.