Big Data in Environmental Research

Pre-World Congress Meeting of New Researchers in Statistics and Probability

Problem sheet - SOLUTIONS

1.

2.

3.

4. *Zeroth model* We get the PIT plot using the following code:

```
hist(result0$cpo$pit)
```

and the histogram for Y and the expected counts:

```
hist(Y)
expected = Germany$E*exp(result0$summary.linear.predictor$mean)
hist(expected)
```

First model We use the same code as before, except replace `result0` by `result1`.

Second model We fit the model, look at the `PIT` plot and `DIC` using the following code:

```
result2 = inla(formula2, family="poisson", data=Germany, E=E,
        control.compute=list(dic=TRUE,cpo=TRUE))
hist(result2$cpo$pit)
result$dic$dic
hist(Y)
expected=Germany$E*exp(result2$summary.linear.predictor$mean)
hist(expected)
```

Third model

```
result2 = inla(formula3, family="poisson", data=Germany, E=E,
        control.compute=list(dic=TRUE,cpo=TRUE))
hist(result3$cpo$pit)
result$dic$dic
hist(Y)
expected=Germany$E*exp(result3$summary.linear.predictor$mean)
hist(expected)
```

5. `CARBayes` performs MCMC simulation. We need the CARBayes, spdep and shapefiles libraries.

```
library(CARBayes)
library(spdep)
library(shapefiles)
```

First we need to read in the shape file and the details of the areas:

```
dbf <- read.dbf(dbf.name="englandlocalauthority.dbf")
dbf$dbf <- dbf$dbf[ ,c(2,1,3:7)]
shp <- read.shp(shp.name="englandlocalauthority.shp")
shp$shp[[23]]$points <- shp$shp[[23]]$points[-c(460, 461, 462, 463, 464, 465), ]
```

Next we calculate the SMRs and combine them with the spatial information:

```
observed <- read.csv(file="copdmortalityobserved.csv", row.names=1)
expected <- read.csv(file="copdmortalityexpected.csv", row.names=1)
SMR <- observed[ ,-1]/expected
colnames(SMR) <- c("SMR2001", "SMR2002", "SMR2003", "SMR2004", "SMR2005", "SMR2006",
   "SMR2007", "SMR2008", "SMR2009", "SMR2010")
```

We can plot the SMR using the following code:

```
SMRspatial <- combine.data.shapefile(SMR, shp, dbf)
range <- seq(min(SMR$SMR2010)-0.01, max(SMR$SMR2010)+0.01, length.out=11)
n.col <- length(range)-1
spplot(SMRspatial, zcol=c("SMR2010"),scales=list(draw=TRUE),
     xlab="Easting", ylab="Northing", at=range, col="transparent",
     col.regions=hsv(0.6, seq(0.2, 1, length.out=n.col), 1))
```

The adjacency matrix is created by using the following code:

```
W.nb <- poly2nb(SMRspatial, row.names = rownames(SMR))
W.list <- nb2listw(W.nb, style="B")
W.mat <- nb2mat(W.nb, style="B")
```

Now we can fit the CAR smoothing model, and compute the risk:

```
formula <- observed$Y2010~offset(log(expected$E2010))
model <- S.CARleroux(formula=formula, family="poisson", W=W.mat,
     burnin=20000, n.sample=100000, thin=10, fix.rho=TRUE, rho=1)
print(model)
plot(model$samples$beta)
plot(model$samples$tau2)

risk <- model$fitted.values / expected$E2010
```

6. Fitting the model requires the R-INLA package.

```
library(INLA)
```

The adjacency matrix created in the previous question can be converted into the INLA format using the following code:

```
nb2INLA("UK.adj",W.nb)
```

Next we create areas IDs to match the values in UK.adj.

```
data = as.data.frame(cbind(observed,expected))
data$ID <- 1:324
```

And finally we fit the model:

```
m1 <- inla(observed$Y2010~f(ID, model="besag", graph="UK.adj"), family="poisson",
           E=expected$Y2010, data=data, control.predictor=list(compute=TRUE),
           control.compute= list(dic=TRUE, cpo = TRUE))
print(m1)
```

The risks can be estimated using the following code:

```
risk <- m1$summary.fitted.values / expected$E2010
```

7. We use the code from Question 5 to estimate the risks for 2001. Then we can use the following code to plot the estimated smoothed risks:

```
SMRspatial@data$risk <- risk
range <- seq(min(SMR$SMR2001)-0.01, max(SMR$SMR2001)+0.01, length.out=11)
n.col <- length(range)-1
spplot(SMRspatial, zcol=c("SMR2001"),scales=list(draw=TRUE),
     xlab="Easting", ylab="Northing", at=range, col="transparent",
     col.regions=hsv(0.6, seq(0.2, 1, length.out=n.col), 1))
```

The same code can be used for the INLA model, and then the plotting can be done using

```
SMRspatial@data$risk <- risk
range <- seq(min(SMR$SMR2010)-0.01, max(SMR$SMR2010)+0.01, length.out=11)
n.col <- length(range)-1
spplot(SMRspatial, zcol=c("SMR2010"),scales=list(draw=TRUE),
       xlab="Easting", ylab="Northing", at=range, col="transparent",
       col.regions=hsv(0.6, seq(0.2, 1, length.out=n.col), 1))
```

We can compare the two models e.g. using DIC, which can be extracted by using the following code

```
model$modelfit      # for CARBayes
m1$dic$dic        # for INLA
```

To investigate the sensitivity we need to add an extra argument to the `CARBayes` model

```
model <- S.CARleroux(formula=formula, family="poisson", W=W.mat,
     burnin=20000, n.sample=100000, thin=10, fix.rho=TRUE, rho=1,prior.var.beta=99)
```

and then use the following code

```
plot(model$samples$beta)
```

For the `INLA` model, we modify the code the following way

```
hyper.list.pc <- list(prec=list(prior="pc.prec",param=c(0.6,0.01)))
m1 <- inla(observed$Y2010~f(ID, model="besag", graph="UK.adj",hyper=hyper.list.pc),
           family="poisson",
           E=expected$Y2010, data=data, control.predictor=list(compute=TRUE),
           control.compute= list(dic=TRUE, cpo = TRUE))
```

and analyse the results after changing the list in `param` using e.g.

```
plot(m1)
```

8. Use code from previous questions, change year in the definition of the model.

9.
```
library(rgdal)
library(gcmr)
library(maptools)
library(classInt)
library(spdep)
library(shapefiles)
library(CARBayes)

# read in the shape file and convert it to the British National Grid
dbf <- read.dbf(dbf.name="scot.dbf")
shp <- read.shp(shp.name="scot.shp")


scot_LL <- readOGR(".", "scot")

proj4string(scot_LL) = CRS("+proj=longlat +ellps=WGS84")
EPSG <- make_EPSG()
EPSG[grep("British National Grid", EPSG$note), 1:2]

scot_BNG0 <- spTransform(scot_LL, CRS("+init=epsg:27700"))

# read in the Scottish health data and merge with the names of the areas
scot_dat <- read.table("scotland.dat.txt", skip = 1)

names(scot_dat) <- c("District", "Observed", "Expected", "PcAFF",
```

```
"Latitude", "Longitude")


scot_dat = merge(x=scot_dat, y=dbf$dbf, by.x = "District", by.y="ID", all.x = T, all.y=F)


# create a set of smooth risks and add it to scot_dat as an extra column named SMR
scot_dat$SMR = scot_dat$Observed/scot_dat$Expected


# combine the estimate of risks with the shapefile
smr.matrix = as.matrix(scot_dat$SMR, ncol=1)[,1,drop=FALSE]
rownames(smr.matrix) = scot_dat$NAME

SMRspatial <- combine.data.shapefile(smr.matrix, shp, dbf)

# convert the geography of scot_dat to match the British National Grid
row.names(scot_dat) <- formatC(scot_dat$District, width = 2,
flag = "0")
ID <- formatC(scot_BNG0$ID, width = 2, flag = "0")
scot_BNG1 <- spChFIDs(scot_BNG0, ID)
scot_BNG <- spCbind(scot_BNG1, scot_dat[match(ID, row.names(scot_dat)),
])


# create a map of the risks, with diffetent colours for different ranges
colors <- colorRampPalette(c('mistyrose2', 'pink4'))(256)
spplot(scot_BNG, zcol="SMR",col.regions = colors)
```