Doing one dimentional point process with INLA

Elias T. Krainski

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação

University of Glasgow, 3-5, July 2019

Burkitt dataset

- The `burkitt` dataset from the `splancs` package (Rowlingson and Diggle 1993)
- Records cases of Burkitt's lymphoma in the Western Nile district of Uganda during the period 1960-1975 (see, Bailey and Gatrell 1995, Chapter 3).
- This dataset contains the five columns:

Variable	Description
`x`	Easting
`y`	Northing
`t`	Day, starting at 1/1/1960 of onset
`age`	age of child patient
`dates`	Day, as string yy-mm-dd

The Burkitt dataset

```
data('burkitt', package = 'splancs')
burkitt$dates <- as.Date(paste0("19",burkitt$dates),"%Y-%m-%d'
burkitt$year <- burkitt$t/365.25
head(burkitt, 3)
```

```
##       x   y   t age      dates  year
## 1 300 302 413  22 1961-02-17 1.131
## 2 291 270 472   5 1961-04-17 1.292
## 3 326 263 511  12 1961-05-26 1.399
```

```
tail(burkitt, 3)
```

```
##         x   y    t age      dates  year
## 186 279 339 5753   5 1975-10-02 15.75
## 187 277 379 5755   7 1975-10-04 15.76
## 188 258 350 5775   5 1975-10-24 15.81
```

Burkitt data summary

```
##         x               y               t              age       
##  Min.   :255   Min.   :247   Min.   : 413   Min.   : 2.00  
##  1st Qu.:269   1st Qu.:327   1st Qu.:2412   1st Qu.: 5.00  
##  Median :282   Median :344   Median :3704   Median : 6.00  
##  Mean   :286   Mean   :339   Mean   :3530   Mean   : 7.11  
##  3rd Qu.:300   3rd Qu.:362   3rd Qu.:4700   3rd Qu.: 8.00  
##  Max.   :335   Max.   :399   Max.   :5775   Max.   :36.00  
##      dates                 year      
##  Min.   :1961-02-17   Min.   : 1.13  
##  1st Qu.:1966-08-08   1st Qu.: 6.60  
##  Median :1970-02-21   Median :10.14  
##  Mean   :1969-08-30   Mean   : 9.66  
##  3rd Qu.:1972-11-13   3rd Qu.:12.87  
##  Max.   :1975-10-24   Max.   :15.81  
```

Time at which cases observed and temporal knots

- Need to fit a SPDE model

```
rt <- range(burkitt$year)
w0 <- diff(rt)/50
tknots <- seq(min(burkitt$year)-w0, max(burkitt$year)+w0, w0)
```

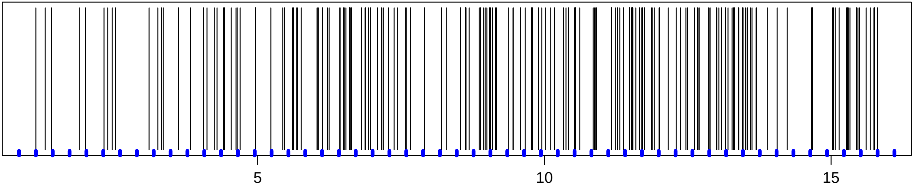


Figure 1: Time when each event occurred (black) and knots used for inference (blue).

The data and an approach

- The observed data, $\mathbf{y}$, is the set of n events at the following time points, $\mathbf{y} : t_1, t_2, ..., t_n$
  - in the burkitt example it is 413, , 472, , . . . , 5775.
  - and time domain is "1960-01-01" to "1975-12-31"

```
datelims <- ISOdate(c(1960, 1975), c(1, 12), c(1, 31))
tdomain <- difftime(datelims, datelims[1], units='days')
tdomain.y <- as.numeric(tdomain/365.25)
tdomain
```

```
## Time differences in days
## [1]    0 5843
```

```
tdomain.y
```

```
## [1]  0 16
```

Modeling

- We consider an intensity function varying over time

$$\lambda(t)$$

- This describes how likely a case is to happen at time t

Modeling

- We consider an intensity function varying over time

$$\lambda(t)$$

- This describes how likely a case is to happen at time t

Common approach

- Aggregate the data into discrete grid
- Consider the number of cases in each grid cell t as Poisson(λ_t)
- Model the $\log(\lambda_t)$

The log-Likelihood for a set of points

- Given the time domain $[0, L]$ and these data points, the log likelihood of the intesity function $\lambda(t)$ is

$$l(\lambda(.)|\mathbf{y}) = \boxed{L} - \log\left(\int_0^L \boxed{\lambda(t)} \partial t\right) + \sum_{i=1}^{n} \log(\lambda \boxed{(t_i)})$$

  - L is the size of the time domain
  - $\lambda(t)$ is the intensity at time t
  - (t_i) are time coordinates of the cases

Direct likelihood approximation, D. P. Simpson et al. (2016)

The log likelihood

$$l(\lambda(.)|\mathbf{y}) = L - \log\left(\int_0^L \lambda(t)\partial t\right) + \sum_{i=1}^{n} \log(\lambda(t_i))$$

is approximated by

$$l(\mathbf{y}) \approx L - \sum_{j=1}^{m} w_j \log \lambda(t) + \sum_{i=1}^{n} \log(\lambda(t_i))$$

- On a set of m integration points $t_1, ..., t_m$
- With the weights $w_1, ..., w_m$
    - if the grid is equaly spaced, $w_j = w_0$

The log-Gaussian Cox point process model

- The lGCpp assumes that the log of λ follows a Gaussian process

The log-Gaussian Cox point process model

- The lGCpp assumes that the log of λ follows a Gaussian process

Figure 2: A latent Gaussian model !!!

The idea into practice

In order to implement the idea of evaluating

$$l(\mathbf{y}) \approx L - \sum_{j=1}^{m} w_j \log \lambda(t) + \sum_{i=1}^{n} \log(\lambda(t_i))$$

- Consider a set of m time knots as the integration points
- Compute the weights $w_1, ..., w_m$
  - if the grid is equaly spaced, $w_j = w_0$
- Consider a Poisson likelihood with observations as
  - 0 (zero) at the event locations
  - 1 (one) at the integration weights and expected cases as
  - w_j at the integration points
  - 0 (zero) at the event locations
- use the SPDE setup as usual

Temporal mesh and Projection matrix

Considering the time knots set before

```
mesh.t <- inla.mesh.1d(tknots)
```

Considering the time locations and the temporal mesh:

```
At <- inla.spde.make.A(mesh = mesh.t, loc = burkitt$year)
dim(At)
rowSums(At)
colSums(At)
```

Data stack definition

- Use approach in D. P. Simpson et al. (2016)
- At the data locations
  - y: 1 (one)
  - E: 0 (zero), expected value at a point
- At the integration points (mesh nodes)
  - y: 0 (zero)
  - E: *width* of each time knot

The data stack built

```
stk <- inla.stack(
  data = list(y = rep(0:1, c(mesh.t$n, n)),
              expected = c(rep(w0, mesh.t$n), rep(0,n))),
  A = list(rbind(Diagonal(n = mesh.t$n), At),
          1),
  effects = list(idx=1:mesh.t$n,
                 a0 = rep(1, mesh.t$n + n)))
```

Priors

- PC-priors derived in Fuglstad et al. (2018) for the range and the marginal standard deviation

```
spde1 <- inla.spde2.pcmatern(mesh = mesh.t,
  prior.range = c(0.2, 0.01), # P(practic.range < 0.2) = 0.01
  prior.sigma = c(1, 0.01)) # P(sigma > 1) = 0.01
```

Model fitting

```r
form <- y ~ 0 + a0 + f(idx, model = spde1)

burk.res1 <- inla(form, family = 'poisson',
  data = inla.stack.data(stk), E = expected,
  control.predictor = list(A = inla.stack.A(stk),
                           compute=TRUE))
```

Parameter results

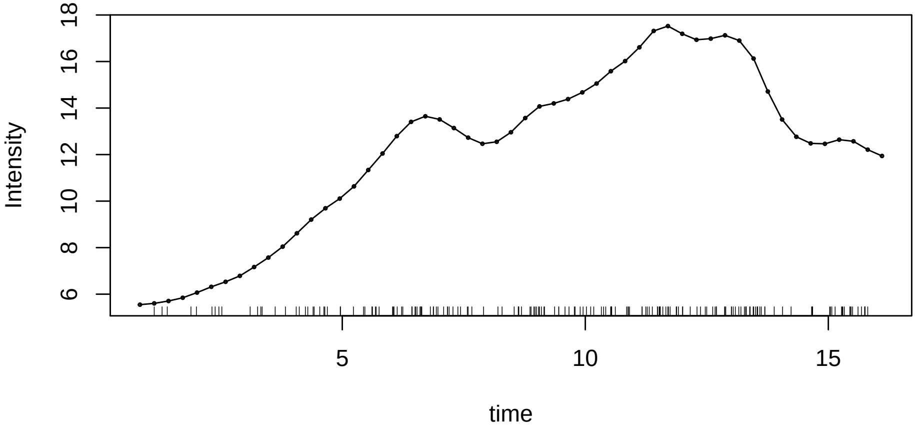
```
burk.res1$summary.fix
```

```
##      mean    sd 0.025quant 0.5quant 0.975quant  mode       kl
## a0 2.424 0.457      1.484    2.428      3.355 2.436 0.0019(
```

```
burk.res1$summary.hyper
```

```
##                    mean     sd 0.025quant 0.5quant 0.975quant
## Range for idx 8.6872 6.4016     1.9682    6.979    25.5883
## Stdev for idx 0.4253 0.1751     0.1766    0.395     0.8531
```

Predicted



- Expected number of cases

```
## [1] 189.5
```

Posterior marginal distributions

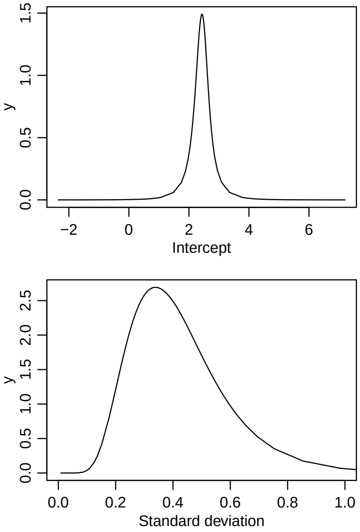


Figure 3: Intercept and random field parameters posterior marginal distributions.

References

Bailey, T. C., and A. C. Gatrell. 1995. *Interactive Spatial Data Analysis*. Harlow, UK: Longman Scientific & Technical.

Fuglstad, G-A., D. Simpson, F. Lindgren, and H. Rue. 2018. "Constructing Priors That Penalize the Complexity of Gaussian Random Fields." *Journal of the American Statistical Association* 114 (525). Taylor & Francis: 445–52. doi:10.1080/01621459.2017.1415907.

Rowlingson, B. S., and P. J. Diggle. 1993. "Splancs: Spatial Point Pattern Analysis Code in S-Plus." *Computers & Geosciences* 19 (5): 627–55. doi:https://doi.org/10.1016/0098-3004(93)90099-Q.

Simpson, D. P., J. B. Illian, F. Lindren, S. H Sørbye, and H. Rue. 2016. "Going Off Grid: Computationally Efficient Inference for Log-Gaussian Cox Processes." *Biometrika* 103 (1): 49–70.