STAT40810 — Stochastic Models

Brendan Murphy

Week 9

Poisson Process

Big O, Little o

- A common notation in mathematics, physics, computer science and statistics is the Big O, little o notation.
- When we write $O(h)$, as $h \to 0$, we mean that something is approximately proportional to h when h tends to zero.
- When we write $o(h)$, as $h \to 0$, we mean that something is smaller than h as h tends to zero.
- More formally, ...

$$f(h) = O(h), \text{ if } \lim_{h \to 0} \frac{f(h)}{h} = C \neq 0$$

$$f(h) = o(h), \text{ if } \lim_{h \to 0} \frac{f(h)}{h} = 0$$

Counting processes

Counting processes are increasing processes $\{X_t, t \in \mathbb{R}_+\}$ with $X_t \in \mathbb{N}$. *They are an important class of processes in continuous time.*

Notation

The idea is to count something over time (eg, the arrivals of customers): X_t is the number of arrivals between time 0 and time t, so naturally we set $X_0 = 0$.
Let $N_{(s,t]}$ be the number of arrivals during the interval $(s, t]$:

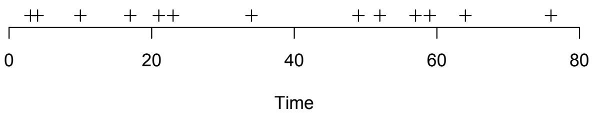
$$N_{(s,t]} = X_t - X_s.$$

Note that $X_t = N_{(0,t]}$.

Example: Ireland vs New Zealand

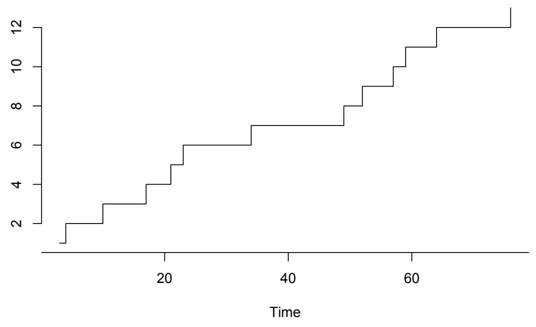
Let's consider the times where there were scores in the weekend game between Ireland and New Zealand.

The scores were at these times...



Example: Ireland vs New Zealand

The counting process looks like this...



Definition of a Poisson process

Poisson process

A Poisson process with *intensity* $\lambda > 0$ is a counting process $\{X_t\}$ with

1. independent increments;
2. $\mathbb{P}(X_{t+h} - X_t = 1) = \lambda h + o(h)$ when $h \to 0$;
3. $\mathbb{P}(X_{t+h} - X_t > 1) = o(h)$ when $h \to 0$.

Some remarks about (1)

By "independent increments" we mean that, for any $t_1 < t_1 \leq t_3 < t_4$,

$$X_{t_2} - X_{t_1} \text{ indep. of } X_{t_4} - X_{t_3}.$$

In other words:

$$N_{(t_1,t_2]} \text{ is indep. of } N_{(t_3,t_4]}.$$

This is sometimes referred as the *loss of memory* property.

Remarks on (2) and (3)

Recall (2) and (3):

$$\mathbb{P}(X_{t+h} - X_t = 1) = \lambda h + o(h),$$

$$\mathbb{P}(X_{t+h} - X_t > 1) = o(h).$$

We can write (3) in a different way:

$$\frac{\mathbb{P}(X_{t+h} - X_t > 1)}{h} \to 0, \text{ as } h \to 0.$$

Examples of Poisson Property

- In the rugby game, two scores will never happen *exactly* at the same time.
- In a shop/bank two customers will never arrive *exactly* at the same time, and the probability that two customers arrive at a very small period of time is negligible.
- Remark that (2) and (3) imply that:

$$\mathbb{P}(X_{t+h} - X_t = 0) = 1 - \lambda h + o(h).$$

About Poisson

Siméon Denis Poisson (1781-1840), the French mathematician and physicist, who received the Copley medal from the Royal Society of London in 1832.

Fundamental theorem

A reminder: the Poisson distribution

$$\mathbb{P}\{X = k\} = \frac{\gamma^k}{k!} e^{-\gamma}.$$

Then:

- $\mathbb{E}(X) = \mathbb{V}\mathrm{ar}(X) = \gamma.$

Theorem

Let $\{X_t\}$ be a Poisson process with intensity λ and $N_{(s,t]} = X_t - X_s$ for $s < t$.
Then:

$$N_{(s,t]} \sim \mathcal{P}(\lambda(t-s)).$$

An immediate consequence

The arrival rate

The arrival rate between times s and t is:

$$\frac{X_t - X_s}{t - s}.$$

Consequence of the theorem

For a Poisson process, the expected arrival rate is constant, equal to λ:

$$\mathbb{E}\left[\frac{X_t - X_s}{t - s}\right] = \lambda.$$

Arrival times

Definition

Let S_1, S_2, ... be the arrival times:

$$S_k = \min\{t \geq 0 : \quad X_t = k\}$$

(and by convention $S_0 = 0$).

Note that:

$$X_t = \sum_{i=1}^{\infty} \mathbf{1}_{S_i \leq t}.$$

Distribution of the arrival times

Theorem

For any $k \geq 1$, $S_k \sim \Gamma(k, \lambda)$ where we recall that the $\Gamma(k, \lambda)$ distribution has density:

$$f(x) = \frac{\lambda^k}{\Gamma(k)} e^{-\lambda x} x^{k-1}.$$

Gap between arrival times

Definition

For any $k \geq 1$, $T_k = S_k - S_{k-1}$.

Theorem

The T_k are iid with distribution $\mathcal{E}(\lambda)$, where we recall that the exponential distribution $\mathcal{E}(\lambda)$ has density

$$f(x) = \lambda e^{-\lambda x}.$$

In particular, $\mathbb{E}(T_k) = 1/\lambda$ and $\mathrm{Var}(T_k) = 1/\lambda^2$.

Marked Poisson process

Defintion - marked Poisson process

A marked Poisson process is defined as:

- a Poisson process $\{X_t, t \geq 0\}$ with arrival times S_1, S_2, ...
- a collection of iid random variables M_1, M_2, ... indep. of $\{X_t\}$.

For example:

- S_k is the time of the k-th score in a rugby game.
- M_k is the points awarded for the k-th score.

Or:

- S_k is the arrival time of the k-th customer.
- M_k is the money spent by the k-th customer.

Or:

- S_k is the time of occurence of the k-th earthquake;
- M_k is its magnitude.

...

Thinned Poisson process

Definition - thinned Poisson process

Let $\{X_t\}$ be a marked Poisson process with arrival times S_1, S_2, ... and marks M_1, M_2, ... $\sim \mathcal{B}e(p)$.
We define the thinned Poisson process $\{Y_t\}$ by

$$Y_t = \sum_{i=1}^{\infty} \mathbf{1}_{S_i \leq t, M_i = 1}.$$

In other words, $\{Y_t\}$ is obtained from $\{X_t\}$ by erasing the arrivals S_i with $M_i = 0$.

Theorem

The process $\{Y_t\}$ is actually a Poisson process with parameter $p\lambda$.

Superposition of Poisson process

Theorem

Let $\{X_t\}$ and $\{Y_t\}$ be two Poisson processes independent of each other, with intensity given by λ and μ, respectively. Let us put $Z_t = X_t + Y_t$. Then $\{Z_t\}$ is a Poisson process with intensity $\lambda + \mu$.

Poisson Process: Estimation

- Estimating the intensity of a Poisson process is straightforward.
- Suppose we observe a process for the time interval $(0, T]$.
- We know that the number of events has a Poisson(λT) distribution.
- Thus,

$$\hat{\lambda} = \frac{\{\#\text{Events observed}\}}{T}.$$

- For the rugby game we get $\hat{\lambda} = 0.1625$.
- That is, we would expect to see 0.1625 scores per minute.
- The time between scores should be exponentially distributed with rate 0.1625 (mean=6 minutes and 9 seconds)

STAT40810 — Stochastic Models

Brendan Murphy

Week 9

Inhomogeneous Poisson Process

Inhomogeneous Poisson process: definition

In some cases, the assumption that the expected arrival rate is constant is not realistic.

Definition

Let $\lambda(t) > 0$ be a function of the time t. An inhomogeneous Poisson process with intensity $\lambda(t)$ is a counting process $\{X_t\}$ with

1. independent increments;
2. $\mathbb{P}(X_{t+h} - X_t = 1) = \lambda(t)h + o(h)$ when $h \to 0$;
3. $\mathbb{P}(X_{t+h} - X_t > 1) = o(h)$ when $h \to 0$.

Inhomogeneous Poisson process: properties

Let $\{X_t\}$ be an inhomogeneous Poisson process with intensity $\lambda(t)$. Then we can prove that

$$N_{(s,t]} = X_t - X_s \sim \mathcal{P}\left(\int_s^t \lambda(x)\mathrm{d}x\right).$$

In particular, when $\lambda(t) = \lambda$ is constant, $\{X_t\}$ is a Poisson process and we have:

$$N_{(s,t]} = X_t - X_s \sim \mathcal{P}\left((t-s)\lambda\right).$$

Example Code: Homogeneous

- We can fit a Poisson process to the Ireland vs New Zealand data using the `NHPoisson` package.
- Let's start with a homogeneous Poisson process.

```
# Load the NHPoisson package
library(NHPoisson)

# Read in the scoring time data
x <- c(3,4,10,17,21,23,34,49,52,57,59,64,76)

# Fit a homogeneous Poisson process
fit0 <- fitPP.fun(n=80,posE=x,start=list(b0=0))
summary(fit0)
```

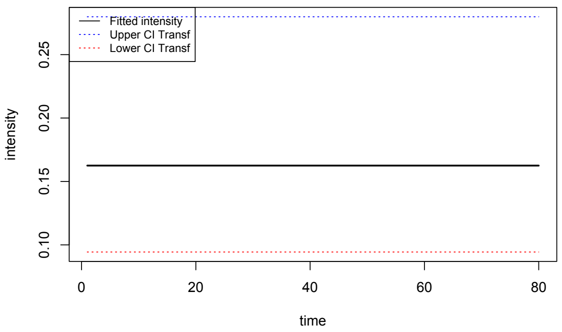

Fit

- The output is:

```
Maximum likelihood estimation

Call:
fitPP.fun(start = list(b0 = 0), posE = x, nobs = 80)

Coefficients:
    Estimate Std. Error
b0 -1.817077  0.2773501

-2 log L: 73.24401
```

Estimated Intensity

- The estimated intensity (plotted) is:



Example Code: Linear

- The NHPoisson package allows the intensity to depend on covariates (in this case we use time).

```
# Allow for a linear intensity function
# Set up the time covariate (I have made this 1 to 80 in 1 minute intervals)
timerange<-seq(1,80,by=1)
covariates<-timerange

#Fit the inhomogeneous Poisson process
fit1 <- fitPP.fun(covariates=covariates,posE=x,start=list(b0=0,b1=0))
summary(fit1)
```


Fit

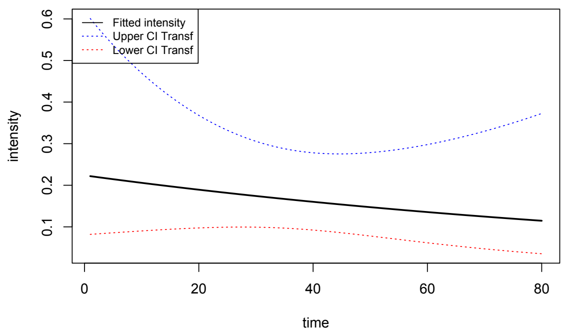
- The output is:

```
Maximum likelihood estimation

Call:
fitPP.fun(covariates = covariates, start = list(b0 = 0, b1 = 0),
    posE = x)

Coefficients:
         Estimate Std. Error
b0 -1.497201255 0.51854630
b1 -0.008356157 0.01214462

-2 log L: 72.76531
```

- We haven't gained much by allowing for inhomogeneity using a linear intensity.

Estimated Intensity

- The estimated intensity (plotted) is:



Example Code: Quadratic

- We also investigate if a quadratic intensity is worthwhile. That is,

$$\log \lambda(t) = b_0 + b_1 t + b_2 t^2.$$

```
# Allow for a quadratic intensity function
timerange<-seq(1,80,by=1)
covariates<-cbind(timerange,timerange^2)

#Fit the inhomogeneous Poisson process
fit2 <- fitPP.fun(covariates=covariates,posE=x,start=list(b0=0,b1=0,b2=0))
summary(fit2)
```


Fit

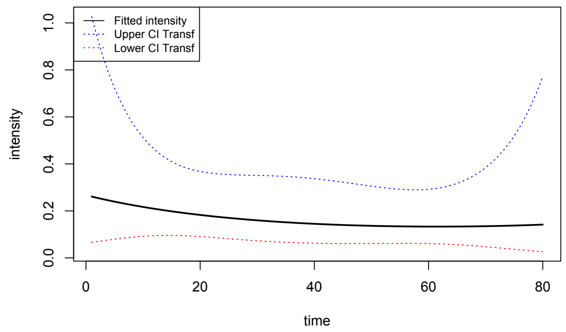
- The output is:

```
Maximum likelihood estimation

Call:
fitPP.fun(covariates = covariates, start = list(b0 = 0, b1 = 0,
    b2 = 0), posE = x)

Coefficients:
        Estimate   Std. Error
b0 -1.3217711615 0.7377627733
b1 -0.0225982871 0.0459999286
b2  0.0001835105 0.0005719938

-2 log L: 72.66407
```

- We haven't gained much by allowing for inhomogeneity using a quadratic intensity either.

Estimated Intensity

- The estimated intensity (plotted) is:



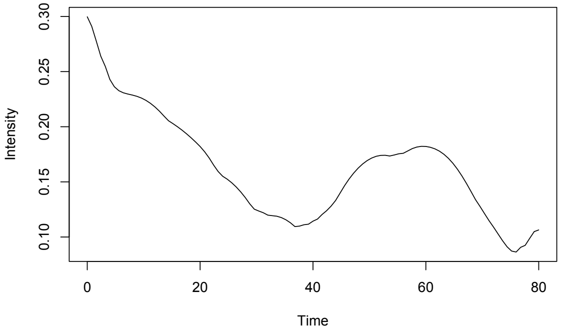
Non parametric Intensity

- The `lpint` package in R allows for the intensity to be estimated non-parametrically.
  It estimates the intensity in an analogous manner to the kernel smoothing regression methods we have already seen.

```
#Load the lpint package
library(lpint)

# Read in the scoring time data
x <- c(3,4,10,17,21,23,34,49,52,57,59,64,76)

# Set up the number of scores at each time point
#Set to be 1 in this case
y<-rep(1,length(x))

#Fit the inhomogeneous Poisson process
fit <- lpint(jmptimes=x,jmpsizes=y,Tau=80)
plot(fit,type="l",xlab="Time",ylab="Intensity")
```

Estimated Intensity

- The estimated intensity (plotted) is:



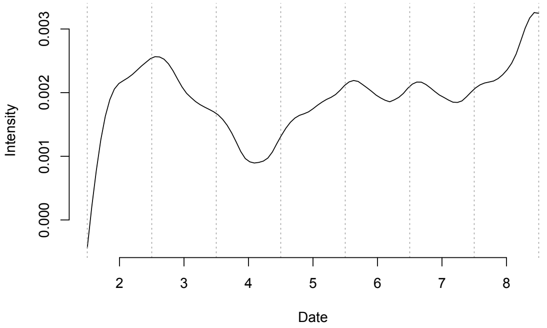
Example: Washington DC Bikeshare Scheme

- Data were collected from the Capital Bikeshare scheme in Washington DC.

- The times at which bikes were collected from the "Massachusetts Ave & Dupont Circle NW" bike station were recorded from July 2nd to July 8th, 2016.
- A total of 1166 collection events were observed in this seven day period.

Estimated Intensity: Bikeshare Scheme

- The intensity of the process was estimated using the `lpint` package:



- Notice the dip in intensity on the 4th of July!

Code

```
# Read in data (file downloaded from the Capital bikeshare scheme website)
dat0 <- read.csv("~/Downloads/2016-Q3-cabi-trips-history-data/2016-Q3-Trips-History-Data-1.csv")

# Extract data on collections from station 31200
dat <- dat0[dat0$Start.station.number==31200,]

# Extract the relevant events (found manually)
dat<-dat[10510:9345,]

# Extract the event times
x<- as.character(dat$Start.date)

# Put into numeric format (seconds from start at midnight 7/2/2016)
library(lubridate)
x <- mdy_hm(x)
x <- as.numeric(x)
x <- x-as.numeric(mdy_hm("7/2/2016 0:00"))

# Set up the number of scores at each time point
# Set to be 1 in this case
y<-rep(1,length(x))

#Fit the inhomogeneous Poisson process
fit <- lpint(jmptimes=x,jmpsizes=y,Tau=7*24*60*60)

# Plot the fit
plot(fit,type="l",xlab="Date",ylab="Intensity",axes=FALSE)
abline(v=(0:7)*(24*60*60),col="darkgray",lty=3)
axis(1,at=((0:6)+0.5)*(24*60*60),labels=c(2:8))
axis(2)
```