STA 360/601: Bayesian and Modern Statistics

Lecture 4: Poisson processes & Non-informative priors

Jeff Miller

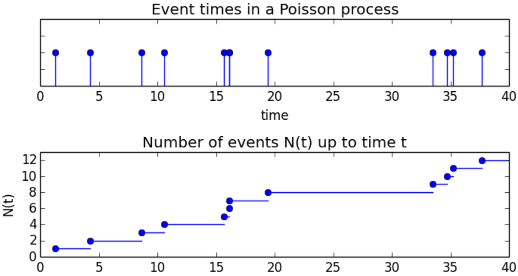
Department of Statistical Science, Duke University

Friday, September 5, 2014

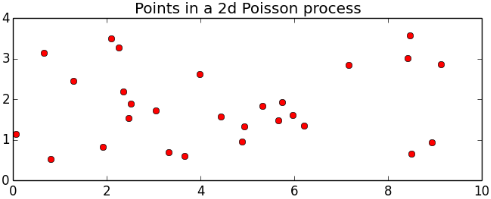
Event time/location data

- Last lecture, we had count data $y_1, \ldots, y_n$, modeled as iid Poisson(θ).
- The Poisson likelihood also arises naturally when the data consist of timing or location of events.
- Examples:
  - y_i = time of the ith traffic accident occurring at an intersection during the study period.
  - y_i = location of a microglia cell in a 3d brain image at snapshot in time.
  - y_i = time and location of a meteorite strike in the United States.
- Often, a Poisson process is a natural model for such data.

1d example



2d example



Intuition for Poisson process

- Remember how Poisson(θ) is the limit of Binomial($n, \theta/n$) as $n \to \infty$?
- Think about that as dividing $[0, 1]$ into n intervals of length $1/n$ and putting a Bernoulli(θ/n) in each independently.
- Intuitively speaking, in terms of the Bernoullis (rather than their sum), the limiting thing you get is a Poisson process.

Poisson process on $[0, 1]$

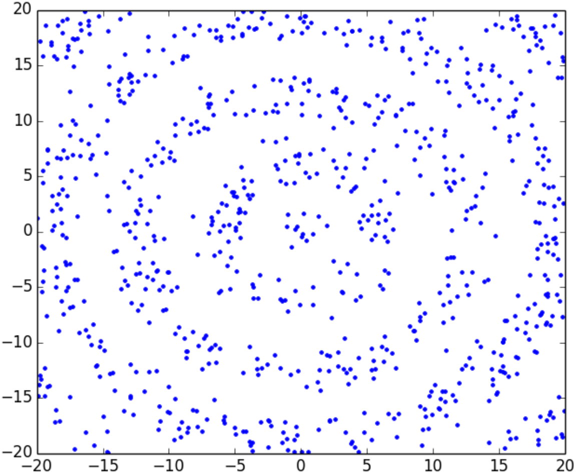
- A random set of points $\{Y_1, \ldots, Y_N\} \subset [0, 1]$ are the points in a Poisson process on $[0, 1]$ with rate $\theta > 0$ if $N \sim \text{Poisson}(\theta)$ and $Y_1, \ldots, Y_N \overset{iid}{\sim} \text{Unif}(0, 1)$ given N.
- Another useful construction (generates the points in order): $X_1, X_2, \ldots \overset{iid}{\sim} \text{Exp}(\theta)$ and $Y_i = \sum_{j=1}^{i} X_j$ for $i = 1, 2, \ldots$ until Y_i is no longer in $[0, 1]$.
- Denote by $N(s, t)$ the $\#$ of points Y_i occurring in interval $(s, t]$. For any $0 \leq t_1 < t_2 < \cdots < t_k \leq 1$, $N(t_i, t_{i+1}) \overset{ind}{\sim} \text{Poisson}((t_{i+1} - t_i)\theta)$ for $i = 1, \ldots, k - 1$.
- There are many equivalent formulations.

Thought experiment

- Let's think bigger. . . Why stop at $[0, 1]$? Why not divide up $[0, \infty)$ into tiny intervals and do the same thing?
- Better yet: What if we took all of $\mathbb{R}^d$, divided it up into tiny boxes, and put an independent Bernoulli(p) in each, where p is θ times the volume of the box?
- Now we're getting somewhere. . . But why limit ourselves to making every box have the same p?
- Let's take a function $r(y) \geq 0$ and define each p as the integral of $r(y)$ over that box.
- This is the intuition behind the multidimensional inhomogeneous Poisson process.

Example

A sample from an inhomogeneous Poisson process on $\mathbb{R}^2$ with rate function $r(y) = \frac{1}{2}(\cos(\|y\|) + 1)$.



Poisson process (multidimensional, inhomogeneous)

- Assume $\int_A r(y)dy < \infty$ for $A \subset \mathbb{R}^d$ bounded.
- A Poisson process on $\mathbb{R}^d$ with rate function $r(y) \geq 0$ is a random countable set of points such that:
  (a) for any $A \subset \mathbb{R}^d$ the number of points $N(A)$ in A is Poisson$(\int_A r(y)dy)$, and
  (b) $N(A_1), \ldots, N(A_k)$ are independent whenever $A_1, \ldots, A_k \subset \mathbb{R}^d$ are disjoint.
- Cool fact: When $c_r = \int_{\mathbb{R}^d} r(y)dy$ is finite, you can sample from a Poisson process by drawing $N \sim$ Poisson(c_r) and then drawing the points $Y_1, \ldots, Y_N$ iid from the pdf $r(y)/c_r$.
- This last property allows us to write down the likelihood:

$$L(y_{1:n}; r) = \text{Pois}(n; c_r) \prod_{i=1}^{n} r(y_i)/c_r.$$

Poisson process (homogeneous)

- Let's specialize to the homogeneous case, where $r(y)$ equals a constant θ on a set $\mathcal{Y}$ of finite volume $v = \int_{\mathcal{Y}} dy$, and is 0 elsewhere.
- Then $c_r = \theta v$ and for $y_i \in \mathcal{Y}$, the likelihood simplifies to

$$L(y_{1:n}; \theta) = \mathrm{Pois}(n; \theta v)/v^n.$$

- The likelihood doesn't care about the locations of the points in $\mathcal{Y}$ — only the number of points matters!
- So, once we choose a prior $\pi(\theta)$ on θ, the posterior is just

$$\pi(\theta | y_{1:n}) \propto \mathrm{Pois}(n; \theta v) \pi(\theta).$$

Applications of Poisson processes

- The "limit of Bernoulli processes" perspective gives intuition into when a Poisson process model might be reasonable.
- Some more examples:
  - times of neuron spikes,
  - locations of mutations in a genome,
  - times of speciation events in phylogenetic history,
  - emission times of radioactively decaying particles,
  - locations of organisms in a habitat at a given time.

Illustration

- Patient A had $n = 8$ seizures on the following days over the past $v = 365$ days, with today as time 0:

  $$y = y_{1:n} = (-349, -297, -289, -251, -249, -202, -81, -69).$$

- A homogeneous Poisson process is a reasonable model (but probably too simplistic in reality).
- Based on a history of many previous patients, you have a prior $\pi(\theta)$ on the rate.
- A certain treatment is known to be fully effective at preventing seizures ($E = 1$) or have no effect ($E = 0$), independently of θ and y, with $\Pr(E = 1) = q = 0.25$.
- Patient A is given the treatment today (time 0).
- 60 days pass with no seizures. What is the probability that the treatment was effective?

Model

$\theta \sim \text{Ga}(a, b)$ with $a = 0.1$, $b = 0.3$.

$Y = Y_{1:N} \sim \text{PP}(\theta)$ on $[-365, 0]$ given θ.

$E \sim \text{Bernoulli}(q)$ independent of θ and Y.

Given θ, Y, E, model future seizure times $Z_1 \leq Z_2 \leq \cdots$ as a PP on $(0, \infty)$ with rate 0 if effective ($E = 1$) and rate θ if not effective ($E = 0$).

Quantity of interest

Probability of ineffective treatment, given no seizures up to t

$$= \Pr(E = 0 \mid Z_1 > t, y) = \frac{\Pr(Z_1 > t \mid E = 0, y)\Pr(E = 0 \mid y)}{\Pr(Z_1 > t \mid y)}.$$

We have $\Pr(E = 0 \mid y) = \Pr(E = 0) = 1 - q$ and

$$\begin{aligned}
\Pr(Z_1 > t \mid E = 0, y) &= \int \Pr(Z_1 > t \mid \theta, E = 0, y)\pi(\theta \mid E = 0, y)d\theta \\
&= \int \Pr(\mathrm{Exp}(\theta) > t)\pi(\theta|y)d\theta \\
&= \int e^{-\theta t}\pi(\theta|y)d\theta,
\end{aligned}$$

where the second step uses the fact that the time to the first event is $\mathrm{Exp}(\theta)$ distributed (given θ and $E = 0$).

Quantity of interest (continued)

The posterior on θ given y is (a lot like your last homework)

$$\pi(\theta|y_{1:n}) \propto \mathrm{Pois}(n; \theta v)\mathrm{Ga}(\theta; a, b) \propto \mathrm{Ga}(\theta; a+n, b+v),$$

so

$$\begin{aligned}\Pr(Z_1 > t \mid E = 0, y) &= \int_0^\infty e^{-\theta t} \frac{(b+v)^{a+n}}{\Gamma(a+n)} \theta^{a+n-1} e^{-(b+v)\theta} d\theta \\ &= \frac{(b+v)^{a+n}}{(b+v+t)^{a+n}},\end{aligned}$$

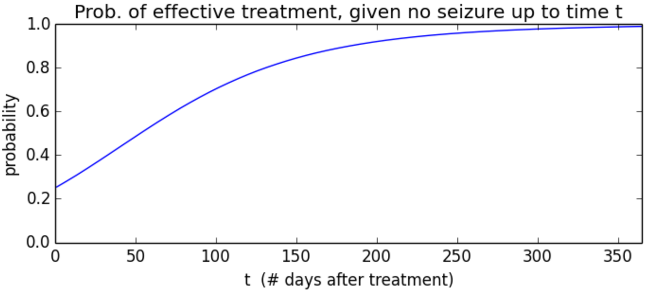
and thus, plugging this into the first equation of the previous slide,

$$\Pr(E = 0 \mid Z_1 > t, y) = \frac{\left(\frac{b+v}{b+v+t}\right)^{a+n}(1-q)}{\Pr(Z_1 > t \mid y)}.$$

Quantity of interest (part trois)

Since $\Pr(E = 1 \mid Z_1 > t, y) = q/\Pr(Z_1 > t \mid y)$, and these two have to sum to 1, we get

$$\Pr(E = 1 \mid Z_1 > t, y) = \frac{q}{q + \left(\frac{b+v}{b+v+t}\right)^{a+n}(1-q)}.$$



A couple points

- Just applying the rules of probability, we can use the posterior to answer pretty much any reasonable question, e.g.:
  - Probability of seizure in next week, given no seizure up to now?
  - If the treatment turns out to be ineffective, mean time to first seizure?
  - One-sided prediction interval for time of first seizure?
  - Patient wants to travel for 2 weeks where there are limited medical facilities; s/he could consider a loss function and make a decision.
- Due to homogeneity, our analysis only used n, not the values $y_1, \ldots, y_n$.

Objective Bayesian inference

- If there is universally-accepted prior information, almost no one would argue with using it.
- But what if you really have no idea at all?
- Or, more likely, what if it is critical that your results not depend on any personal biases? e.g.,
  - clinical trials for a new drug,
  - testing of a medical device,
  - evidence to be presented in a court of law.
- The original motivation of *objective Bayes* was to find priors that contain little, or ideally, no information.
- That has evolved into a more attainable goal of finding "default" priors that provide reliable and interpretable results, to be used as conventions when more specific prior information can't or shouldn't be used.

Non-informative priors

- Such priors $\pi(\theta)$ are called *non-informative*, and they are usually *improper*, in the sense that they do not integrate to a finite value, i.e., $\int \pi(\theta)d\theta = \infty$.
- For example, suppose we choose a gamma prior $\theta \sim \text{Ga}(a, b)$.
- Then, since the prior variance $\text{V}(\theta) = a/b^2$ is finite, in some sense the prior contains some information.
- This is apparent in the *shrinkage* that occurs, with the posterior mean being a convex combination of the sample mean and prior mean.
- If we take $a = b = \varepsilon$ for ε small, the prior mean stays at $a/b = 1$ and the variance becomes large. As $\varepsilon \to 0$, the shape of the prior becomes $\propto 1/\theta$.
- This is one example of an improper prior.

Improper priors

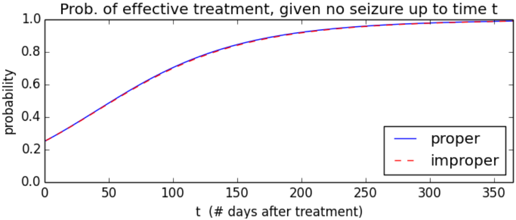
- Improper priors are in some sense not priors at all in that they aren't probability densities.
- There is no prior mean or defined prior variance, we can't sample from an improper prior, and the prior predictive $p(y) = \int p(y|\theta)\pi(\theta)d\theta$ is undefined.
- However, motivated by a desire to avoid having information in our prior, we can plug an improper prior into Bayes' rule.
- In many cases, the resulting "posterior" (defined in a formal sense via Bayes' rule) is a proper density.
- Bayesian inferences can be conducted <u>only if</u> the resulting posterior is proper.

Illustration

Suppose we use $\pi(\theta) = 1/\theta$ in our seizure example. Formally,

$$\begin{aligned}\pi(\theta|y) &\propto \text{likelihood} \times \text{prior} = \text{Pois}(n; \theta v)\pi(\theta) \\ &= e^{-\theta v}\frac{(\theta v)^n}{n!}\frac{1}{\theta} \propto \theta^{n-1}e^{-v\theta} \\ &\propto \text{Ga}(\theta; n, v).\end{aligned}$$

So it's like setting $a = b = 0$ in our posterior from before. How does it compare to the result with our proper prior?



In this case, they are nearly identical.

Comments on improper priors

- Never use them unless you are sure the resulting posterior is proper.
- If the posterior is improper, inferences are typically meaningless — posterior mean, credible intervals, etc., are undefined.
- Even if the posterior is proper, serious issues can arise: contradictory probabilities, prior can dominate for large n, inadmissible estimators, marginalization paradoxes.

Comments on improper priors (continued)

- Bayesian inferences under improper priors are sometimes more similar to frequentist inferences.
- In many (most?) situations a weakly informative prior will outperform a non-informative one.
- In small sample sizes and data sparse situations, weakly informative priors stabilize inferences through mild shrinkage towards the prior mean.
- The age old bias-variance tradeoff — the prior introduces a bit of bias to greatly reduce variance.

Homework exercise

- The Jeffreys prior is a classical non-informative prior, defined (for a univariate parameter) as $\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$ where

$$\mathcal{I}(\theta) = \int \Big(\frac{\partial}{\partial\theta} \log p(y|\theta)\Big)^2 p(y|\theta) dy$$

  is the *Fisher information*.

- Show that for any likelihood $p(y|\theta)$, if $\pi(\theta)$ is the Jeffreys prior, and we have an alternate parametrization, say $q(y|\phi) = p(y|\theta)$ where $\theta = h(\phi)$ and h is a smooth 1-to-1 function, then the Jeffreys prior $\bar{\pi}(\phi)$ for ϕ satisfies

$$\bar{\pi}(\phi) \propto \pi(h(\phi))|h'(\phi)|.$$

  (Hint: Let $\ell(\theta) = \log p(y|\theta)$ and apply the chain rule to compute $\frac{\partial}{\partial\phi}\ell(h(\phi))$.)

- Explain why this property is appealing.

References

Poisson processes

- Grimmett & Stirzaker, *Probability and Random Processes*, Oxford University Press, 2006. (Secs 6.8, 6.13)
- Rick Durrett, *Probability: Theory and Examples*, Duxbury Press, 1996. (pp. 145–148)

Non-informative priors

- Kass & Wasserman, *The selection of prior distributions by formal rules*, JASA, Vol. 91, No. 435, 1996.
- James O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 1985. (Sec 3.3)