DERIVATION OF THE EXPECTATION-MAXIMIZATION ALGORITHM FOR THE ETAS MODEL

KRISTIAN LUM

1. Introduction

In this short document, we rederive the algorithm given in [1] that is used for predicting the locations of crimes. Because we suspect there is a typo in the article, we provide support for our interpretation of the intended model with quotes from the paper in footnotes.

2. Model

We observe data $X = \{t_i^{(n)}, n = 1, ..., k, i = 1, ..., N_n\}$, where $t_i^{(n)}$ is the time of the ith crime that occurs in location n. We define N_n to be the total number of crimes that occur in bin n during the period of observation.

In this model, crimes arise either via a baseline process or as the "offspring", "children", or "aftershocks" of previous crimes. We define latent indicator variables, $Z_{ij}^{(n)}$, which denote the provenance of the jth crime, as either arising from the baseline process or as a child of the ith crime in bin n for $0 < i < j$. Under this model, child crimes only occur in the same bin in which their parent crime occurred.[1] Specifically,

$$Z_{0j}^{(n)} = \begin{cases} 1 \text{ if the } j\text{th crime arose from the baseline process} \\ 0 \text{ otherwise} \end{cases}$$

$$Z_{ij}^{(n)} = \begin{cases} 1 \text{ if the } i\text{th crime is the parent of the } j\text{th crime} \\ 0 \text{ otherwise} \end{cases}$$

The baseline process is defined to be a Poisson process with rate μ_n, which is constant across time within each bin. [2] Under this model, the number of crimes

[1] Bottom right of page 1402: "The conditional intensity, or probabilistic rate $\lambda_n(t)$ of events in box n at time t was determined by $\lambda_n(t) = \mu_n + \sum_{t_n^i < t} \theta\omega e^{-\omega(t-t_n^i)}$." The rate in bin n is only a function of other crimes that occurred in bin n (t_n^i) and the baseline rate, so children must only be born in the same bin as their parent, otherwise the rate would include events from other bins as well.

[2] Top of page 1402: "In particular, the background rate μ is a nonparametric histogram estimate of a stationary Poisson process (Marsan and Lengline 2008)". Bottom of page 1402: "First-generation events occur according to a Poisson process with constant rate μ".

arising from the baseline process in location n over a time period of duration T is distributed as Poisson with rate $\mu_n T$. Using the latent variables, Z, we define

$$\sum_{j=1}^{N_n} Z_{0j}^{(n)} \sim \text{Poisson}(\mu_n T),$$

and the contribution to the full Z-augmented log-likelihood of the events arising from the baseline process is given by

$$l_{\text{baseline}}(X) = \sum_{n=1}^{k} \left(-\mu_n T + \sum_{j=1}^{N_n} Z_{0j}^{(n)} \log \mu_n T - \log((\sum_{j=1}^{N_n} Z_{0j}^{(n)})!) \right)$$

A crime that occurs at time t in bin n, regardless of whether the crime arose from the baseline or as a child event, gives rise to c_t child events in bin n, where $c_t \sim \text{Pois}(\theta)$. [3] The time until each of the child events are random draws from an exponential distribution with rate parameter ω, so

$$p(t_j^{(n)} \mid t_i^{(n)}, Z_{ij}^{(n)} = 1) \sim \text{Exp}(\omega)$$

and

$$\sum_{j:t_i^{(n)} < t_j^{(n)}} Z_{ij}^{(n)} \sim \text{Poisson}(\theta),$$

where, by definition, for fixed i and n, $\sum_{j:t_i^{(n)} < t_j^{(n)}} Z_{ij}^{(n)}$ is the number of child events of crime i from location n. Then, the contribution to the log-likelihood of the child events is given by

$$\begin{aligned}
l_{\text{children}}(X) &= l_{\text{exponential}}(X) + l_{\text{poisson}}(X) \\
l_{\text{exponential}}(X) &= \sum_{n=1}^{k} \sum_{i=1}^{N_n} \sum_{j:t_i^{(n)} < t_j^{(n)}} Z_{ij}^{(n)} \left(\log \omega - \omega(t_i^{(n)} - t_j^{(n)}) \right) \\
l_{\text{poisson}}(X) &= \sum_{n} \sum_{i=1}^{N_n} -\theta + (\sum_{j:t_i^{(n)} < t_j^{(n)}} Z_{ij}^{(n)}) \log \theta - \log((\sum_{j:t_i^{(n)} < t_j^{(n)}} Z_{ij}^{(n)})!)
\end{aligned}$$

Combining all of these pieces, then the full augmented log-likelihood is given by

$$l(X; \theta, \omega, \mu) = l_{\text{baseline}}(X) + l_{\text{children}}(X)$$

Under this model, the expected rate at time t at location n given all other previously observed crimes in location n is given by

[3] Bottom left on page 1402: "Events (from all generations) each give birth to N direct offspring events, where N is a Poisson random variable with parameter θ."

$$\lambda_n(t) = \mu_n + \sum_{i:t_i^{(n)}<t} \theta\omega e^{-\omega(t-t_i^{(n)})}.^4$$

We derive this as the expected number of crimes arising from the baseline process μ_n plus the number of crimes arising as children of previous events in bin n. The expected number of children arising from crime i is given by θ, and on expectation $e^{-\omega(t-t_i^{(n)})}$ of them will occur at time t, giving $\theta e^{-\omega(t-t_i^{(n)})}$ as the expected number of child events at time t arising from crime i.

3. Estimation using the E-M Algorithm

Estimation takes place using the Expecation-Maximization algorithm, an iterative procedure. At iteration s, define the current set of parameter values to be $\{\theta^{[s]}, \omega^{[s]}, \mu_n^{[s]}\}$. Then, for the E-step, we calculate the expectation of each of the latent variables, $Z_{ij}^{(n)}$.

3.1. **E-step.**

$$p_{ij}^{(n)} = E[Z_{ij}^{(n)} \mid X, \theta^{[s]}, \omega^{[s]}, \mu_n^{[s]}]$$

where

$$\begin{aligned} p_{ij}^{(n)} &= \frac{\theta^{[s]}\omega^{[s]}e^{-\omega^{[s]}(t_i^{(n)}-t_j^{(n)})}}{\lambda_n^{[s]}(t_j^{(n)})} \\ p_{0j}^{(n)} &= \frac{\mu_n^{[s]}}{\lambda_n^{[s]}(t_j^{(n)})} \\ \lambda_n^{[s]}(t) &= \mu_n^{[s]} + \textstyle\sum_{i:t_i^{(n)}<t} \theta^{[s]}\omega^{[s]}e^{-\omega^{[s]}(t-t_i^{(n)})5} \end{aligned}$$

for $i = 0, ..., N_n$, $j > i$.

3.2. **M-step.** Then, in order to derive the EM algorithm associated with the stated log-likelihood, we must calculate

$$\begin{aligned} Q(\theta, \omega, \mu) &= E_{Z|X,\theta^{[s]},\omega^{[s]},\mu_n^{[s]}} l(X, Z; \theta, \omega, \mu) \\ &= E_{Z|X,\theta^{[s]},\omega^{[s]},\mu_n^{[s]}} l_{\text{baseline}}(X) + \\ &\quad E_{Z|X,\theta^{[s]},\omega^{[s]},\mu_n^{[s]}} + l_{\text{exp}}(X) + \\ &\quad E_{Z|X,\theta^{[s]},\omega^{[s]},\mu_n^{[s]}} l_{\text{poisson}}(X) \end{aligned}$$

This is easily achieved by plugging in $p_{ij}^{(n)}$ of equations 1 and 1 for many of the corresponding value of $Z_{ij}^{(n)}$. For those components of the sum that are non-linear in $Z_{ij}^{(n)}$, do not plug in $p_{ij}^{(n)}$, as this would be mathematically inappropriate. In

[4] We note that this is the same intensity function given on the bottom right of page 1401, though we use slightly different notation. Specifically, we denote t_n^i as $t_i^{(n)}$.

those cases where the term is not a function of μ_n, ω, or θ, we shorten notation by denoting these by $C(Z)$, a function that is constant in Z. Plugging in the $p_{ij}^{(n)}$s where appropriate gives

$$\begin{aligned}
E_{Z|X,\theta^{[s]},\omega^{[s]},\mu_n^{[s]}} l_{\text{baseline}}(X) &= \sum_{n=1}^{k} -\mu_n T + (\sum_{j=1}^{N_n} p_{0j}^{(n)}) \log \mu_n T - C_1(Z) \\
E_{Z|X,\theta^{[s]},\omega^{[s]},\mu_n^{[s]}} l_{\exp}(X) &= \sum_{n=1}^{k} \sum_{i=1}^{N_n} \sum_{j:t_i^{(n)}<t_j^{(n)}} p_{ij}^{(n)} \left(\log \omega - \omega(t_i^{(n)} - t_j^{(n)})\right) \\
E_{Z|X,\theta^{[s]},\omega^{[s]},\mu_n^{[s]}} l_{\text{pois}}(X) &= \sum_{n=1}^{k} \sum_{i=1}^{N_n} -\theta + (\sum_{j:t_i^{(n)}<t_j^{(n)}} p_{ij}^{(n)}) \log \theta - C_2(Z)
\end{aligned}$$

In order to update our parameters, $\theta^{(s+1)}$, $\omega^{(s+1)}$, $\mu_n^{(s+1)}$, we must optimize $Q(\theta, \omega, \mu)$ with respect to μ, θ, ω. This is easily done by taking partial derivatives, setting the resulting equation equal to zero, and solving for the parameter.

Differentiating Q with respect to μ_n,

$$\frac{\partial Q}{\partial \mu_n} = -T + \frac{\sum_{j=1}^{N_n} p_{0j}^{(n)}}{\mu_n} \tag{1}$$

This gives an updated value of each μ_n as

$$\mu_n^{(s+1)} = \frac{\sum_{j=1}^{N_n} p_{0j}^{(n)}}{T}$$

This "makes sense" as it is the expected number of crimes arising from the baseline process from bin n per unit time. [6]

Differentiating Q with respect to ω,

$$\frac{\partial Q}{\partial \omega} = \sum_{n} \sum_{i=1}^{N_n} \sum_{j:t_i^{(n)}<t_j^{(n)}} \frac{p_{ij}^{(n)}}{\omega} - p_{ij}^{(n)}(t_i^{(n)} - t_j^{(n)})$$

Once again, setting equal to zero and solving gives an updated value of ω as

$$\omega^{(s+1)} = \frac{\sum_n \sum_{i=1}^{N_n} \sum_{j:t_i^{(n)}<t_j^{(n)}} p_{ij}^{(n)}}{\sum_n \sum_{i=1}^{N_n} \sum_{j:t_i^{(n)}<t_j^{(n)}} p_{ij}^{(n)}(t_i^{(n)} - t_j^{(n)})}$$

[6] This differs from the equation given on page 1402, though it is clear that the original equation cannot possibly be correct. In the E-step given in Mohler et. al, a separate parameter, μ_n is given for each bin. However, in the M-step, only one global parameter, μ is updated, thus under the original algorithm, μ_n would have to be set *a priori*, which is unlikely.

Once again, this value intuitively makes sense in the context of the described model, as it is the inverse of the inverse of the expected time between parent and child events.[7]

Finally, differentiating Q with respect to θ,

$$\frac{\partial Q}{\partial \theta} = \sum_n \sum_{i=1}^{N_n} -1 + \frac{\sum_{j:t_i^{(n)}<t_j^{(n)}} p_{ij}^{(n)}}{\theta}$$

Setting equal to zero and solving gives an updated value of θ as

$$\theta^{(s+1)} = \frac{\sum_n \sum_{i=1}^{N_n} \sum_{j:t_i^{(n)}<t_j^{(n)}} p_{ij}^{(n)}}{\sum_n \sum_{i=1}^{N_n} 1}$$

This equation gives the expected number of child events divided by the total number of potential parent events, which also makes sense as an estimate of the number of child events per parent event.[8]

References

[1] G. O. Mohler, M. B. Short, Sean Malinowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015.

[7] This equation also differs from the analogous equation given in Mohler et. al, though they appear very similar. Our derivation differs in that we include the sum over i, whereas the original does not. This may just be a difference in notation, as it is possible that the original paper meant for the notation $\sum_{i<j}$ to denote a summation over all i and j such that $i < j$. If that is the case, then what we derived matches exactly with the original algorithm.

[8] This equation differs from the update for θ given in Mohler et. al in that we again have a third summand over i. Once again, this may simply be a notational difference. If we are to interpret $\sum_{i<j}$ as the summation over all possible i and j such that $i < j$, then these would match.