Adaptive Event Detection with Time-Varying Poisson Processes

Math 263

Spring 2016

San José State University

Joey Fitch

Hongzhe Liu

Austin Powell

May 17, 2016

Contents

1 Introduction

Everyone is familiar with daily routines. Many people wake up at the same time every morning, or funnel through the entrance to their workplace every day at 8am, or sit in traffic for an hour on their evening commute home. These phenomena all operate on a predictable, periodic schedule. Occasionally, though, an external event may occur that disturbs the regular routine. A late-night party may cause students to enter a dorm building at a higher-than-normal rate, or a local sports game might cause unusual traffic on their commute home. The goal of our model is to identify and describe such deviations from standard behavior. We attempt to detect the presence of external events, to measure the duration of such events, and to quantify the changes in activity due to the presence of the event.

We consider a model $N(t)$ composed as the sum of two non-homogenous Poisson processes:

$$N(t) = N_0(t) + N_E(t)$$

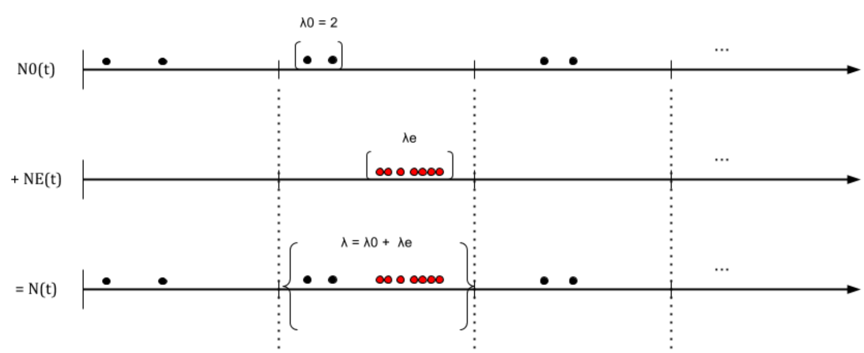
The first process N_0 counts the baseline activity, where we simply model the standard expected behavior as a function of time. The second process N_E counts any additional observations on top of the baseline behavior, caused by the presence of some external event.

The event process N_E normally takes on a value of zero and has no effect, until an event occurs and begins adding observations at a nonzero rate. We can consequently model the distribution of $N_E(t)$ as

$$N_E(t) \sim z(t) \cdot \text{Poisson}(\gamma(t))$$

where z(t) is an indicator function that equals 1 only when there is an event occurring at time t (and equals zero otherwise), and $\gamma(t)$ is the additional rate of observations due to the presence of the event (notice that we allow for non-homogeneity).

You may think of this as "stacking" the usual behavior plus any extra event behavior:



2 Basic Threshold Model

In the simplest approach, we can use N_0 to calculate the probability of observed data under non-event circumstances. If a set of observed data is very unlikely under N_0 - i.e. some data deviates strongly from the norm - then we conclude that an event is probably occurring at that time.

By assuming a non-homogenous Poisson Process, we can say that the number of observed events per hour follows a Poisson distribution with rate $\lambda(t)$. We approximate $\lambda(t)$ according to the maximum likelihood estimate, which is the mean number of observations per time unit:

$$\lambda(t) = \lambda_0 \cdot \delta_{day} \cdot \eta_{day,time}$$

where λ_0 is the universal mean, δ_{day} is the specific day-of-week effect, and $\eta_{day,time}$ is the time-of-day effect specific to each day of the week.

If we break apart the day into k discrete time intervals, then we have $24 \cdot k$ different approximations for $\lambda(t)$, applied according to each specific time-of-day and day-of-week combination. It is important to note that we are considering time in discrete intervals, breaking it up into k segments per day.

For example, for this paper's data sets, they break the "building entry data" into half-hour intervals, giving a total of 48 time increments per day. This means that a week contains $48 \cdot 7 = 296$ different estimates for $\lambda(t)$. In the traffic data, they bin the data into 5-minute intervals, giving 288 time increments per day and $288 \cdot 7 = 2016$ different estimates for lambda (a unique estimate for each time-of-day and day-of-week).

Assuming a Poisson distribution, we can calculate the probability of observing the given data under normal conditions (a non-event period):

$$\mathrm{P}(N = n \,|\, \text{non-event conditions}) \;=\; \mathrm{P}(N = n \,|\, N \sim \text{Poisson}(\,\lambda(t)\,)) \;=\; e^{-\lambda} \cdot \lambda^n / n!$$

where n would be the observed value in the data.

If an observation is large enough such that we obtain a very low probability of observing this data under normal behavior, then it is likely that some external event is affecting the rate of observations.

Within this model, it makes sense to assign a threshold ϵ where, for an observed n, if

$$P(N = n) \;<\; \epsilon$$

according the Poisson pmf, then we conclude an event is present. Note that $\epsilon \in (0, 1)$.

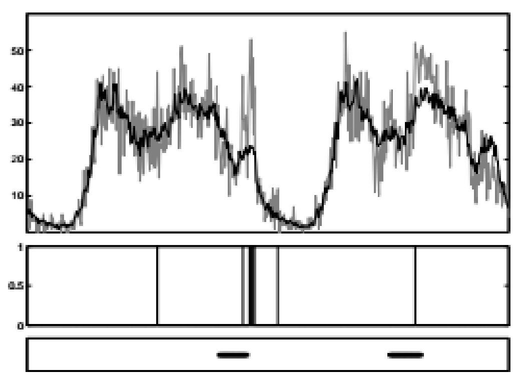
If we conclude that an event is present, we can estimate the added rate of observations due to the event at time t by calculating $N(t) - \lambda(t)$, and we can estimate the duration of an event by summing up how many hours we observe a significant deviation from the norm (where significant is defined according to the choice of threshold, ϵ).

Notice that making ϵ smaller will tighten the threshold, only flagging very unlikely observations, while raising ϵ will loosen the threshold and flag a wider range of observations.

2.1 Application of the Basic Model

2.1.1 Traffic Data

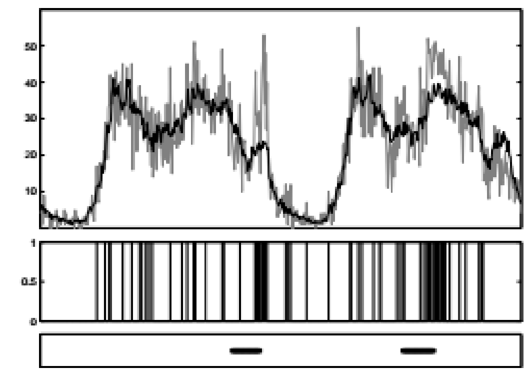
Consider the following analysis of freeway data:



In this figure, we display 3 graphics. The top graphic shows a subset of observed data (light gray) plotted against the time-variant mean (black line). The middle graphic flags the data that exceeds our probability threshold. The third graphic reveals the two events that we know caused traffic (two baseball games).

These results are obviously imperfect. It detects a false positive early on, and when the second event occurs, it detects 3 separate events rather than one continuous event. It almost wholly fails to capture the real duration of the third event as well.

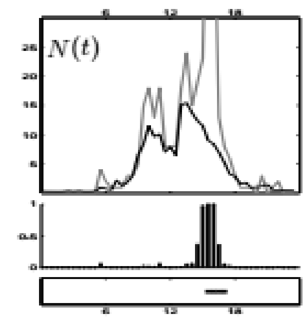
We can relax the threshold to address the second and third concerns:



By relaxing the threshold, we see large continuous event detections right around the two known events. However, we also end up with a tremendous amount of false positives, compounding the first concern we mentioned before.

2.1.2 Building Entry Data

The following graphic details the number of observed people entering a building at time t:



Again, the top graphic shows the observed data (light gray) plotted against the mean (black line).

The second graphic shows the flagged data, where observations exceed the probability threshold of baseline behavior.

The third graphic shows the known ground truth where, in reality, only one event occurred that externally affected the influx of people into this building.

Here, the authors arbitrarily chose a threshold that accurately depicted the behavior of this data subset. However, applying a threshold universally can be problematic, as a threshold that works well for this subset may cause false positives or missed events in other subsets of the data.

2.2 Problems with the Basic Model

The basic approach works very nicely under ideal circumstances, but it faces many problems in real-world conditions.

One major issue is that the approximation of λ in N_0 depends on the mean calculation being a good estimate of normal non-event behavior. However, we cannot decouple our observations into event and non-event data when calculating this mean. The mean is supposed to represent only baseline non-event behavior, but it inherently includes event-data in its calculation, so our basis for "normal" behavior is distorted by the presence of event data. If there are too many events, or if the events have too large of an effect, then our average will be heavily skewed and the model will fail to sufficiently differentiate between usual and unusual behavior.

Another issue is that not all events are drastic. Our model is sensitive to large deviations from the norm, but it fails to account for events which have a smaller but sustained effect. To detect less dramatic events, the threshold must be loosened to capture more data, which increases the capture of false positives. For example, a threshold of .05 would assign the event label to all observations with a $< 5\%$ chance of occurring naturally; however, by definition, 5% of normal non-event data is expected to take on these values anyway. By the nature of probability, some observations will be large simply by randomness, not due to any external factors. This model inherently presents a tradeoff between capturing all events accurately and the flagging of false positives.

Lastly, an issue occurs in the process by which we estimate the duration of an event. Noisy data will present spikes in the observations, such that even during a single event we may repeatedly dip above/below an assigned threshold. As stated, our method would identify this as multiple events occurring in succession, even though it was actually only a single event with substantial variance. To prevent this breakup of event detection, we would have to loosen the threshold, but this would again cause an increase in the flagging of false positives.

3 Main Model

The end-goal of applying a much more complicated process in classifying events or non-events is to find posterior, or "more informed" distributions for z_1 (probability of an event) and z_0 (probability of a non-event) along with a non-homogenous Poisson Process $\lambda(t)$. This is done by applying several different maximization methods which help to better model rare and persistent events, and distinguish these from just "noise". By way of example, noise in the paper's building data set might be a random spike in activity such as larger number of people entering a building at almost the same time. While complicated, this model has the flexibility to provide an event modeler the opportunity to better describe and classify what is eventful much further than the base/threshold model. Finally, in addition to estimating probabilities for an event taking place, it is possible to inadvertently provide a metric for how popular an event is through $N_E(t)$, that is the index of counts due to an event.

3.1 A Markov-modulated Poisson Process (MMPP)

The steps for obtaining the posterior variables of interest are shown roughly as follows:

1. Decide on how informative (or non-informative) the prior information should be.

$$\lambda_0 \sim \Gamma(\lambda; a^L, b^L)$$
$$\tfrac{1}{7}[\delta_1, ...\delta_7] \sim Dir(\alpha_1^d, ..., \alpha_7^d)$$
$$\tfrac{1}{D}[\eta_{j,1}, ..., \eta_{j,D}] \sim Dir(\alpha_1^h, ..., \alpha_D^h)$$

2. In order to sample the hidden variables $z(t)$, the probability of there being an event at time t, our paper uses a variant of the forward-backward algorithm (this is a maximization technique for probabilistic functions of Markov Chains) in order to maximize those probabilities:

   Computing Forward Probabilities Compute the probability of observing $N(t)$ counts given the probability distribution $z(t)$. For $t \in \{1, ..., T\}$

$$p(N(t)|z(t)) = \begin{cases} P(N(t); \lambda(t)) & z(t) = 0 \\ \sum_i P(N(t) - i; \lambda(t)) \text{NBin}(i) & z(t) = 1 \end{cases}$$

   Computing Backward Probabilities Next, draw samples Z(t) from the probability of being in each state $z(t)$, that is the probability of being in a "event" period or not at time t. This is used for determining in a sense the most probable state at any time t $t \in \{T, ..., 1\}$ (t starting at T and incrementally going backwards)

$$\{Z(t) \sim p(z(t)|z(t+1) = Z(t+1) \quad \{N(t'), t' \leq t\}$$

   Computing Event and Non-event counts ($N_0(t)$ and $N_E(t)$ respectively) Given the distribution from the backward section, get $N_0(t)$ and $N_E(t)$:

   $z(t) = 0$ by letting non-event counts ($N_0(t)$) equal all counts $N(t)$

   $z(t) = 1$ $N_0(t)$ gets drawn from $P(N(t)-i; \lambda(t))\text{NBin}(i; a^E, b^E/(1+b^E))$ (Note: $\text{NBin}(i; a^E, b^E/(1+b^E))$. $N_E(t)$ is easily computed by $N_E(t) = N(t) - N_E(t)$.

3. **Complete Data Likelihood** Using the information from finding the hidden variables, it is then possible to possible to maximize the parameters for $\lambda(t)$. In other words, find the parameters that maximize the quantity:

$$\prod_t e^{-\lambda(t)}\lambda(t)^{N_0(t)} \prod_t p(Z(t)|Z(t-1)) \prod_{Z(t)=1} \text{NBin}(N_E(t))$$

   Where the first product maximizes the counts $\lambda(t)$:

$$e^{-T\lambda_0}\lambda_0^{N_0(t)} \prod_i \delta_i^{\sum_{d(t)=i} N_0(t)} \prod_{j,i} \eta_{j,i}^{(...)}$$

4. **Posterior Parameters**

$$\lambda_0 \sim \Gamma(\lambda; a^L + S, b^L + T)$$
$$\tfrac{1}{7}[\delta_1, ...\delta_7] \sim Dir(\alpha_1^d + S_1, ..., \alpha_7^d + S_7)$$
$$\tfrac{1}{D}[\eta_{j,1}, ..., \eta_{j,D}] \sim Dir(\alpha_1^h + S_{j,1}, ..., \alpha_D^h + S_{j,D})$$

5. **Final Model** After iterating between sampling ($z(t), N_0(t), N_E(t)$, $z(t)$ being either z_0 and z_1) (the MCMC part) it is possible to compute Z_{ij}, and to obtain the posterior distributions. These are the main equations of interest in the MMPP model:

$$z_0 \sim \beta(z; a_0^Z + Z_{01}, b_0^Z + Z_{00})$$
$$z_1 \sim \beta(z; a_1^Z + Z_{10}, b_1^Z + Z_{11})$$

   Along with the updated $\lambda(t)$ from the posterior parameters above. Where the Z_{ij} are time-variant: $Z_{ij} = \sum_{t:z(t)=i,z(t+1)=j} 1$ for $i = 0, 1$ $j = 0, 1$.

3.2 Model Results

Since the paper's researchers had access to the information about when an event actually did occur, they were able to perform some benchmarking. They tested their model on a reserved set of data that had not been used in the creating the model so they could more accurately assess how their model would perform "in the real world".

The results are displayed in Tables 1 and 2. In general, the MMPP model was very accurate, only loosing accuracy when there were less events they wanted to detect (more sparse data). Although the paper is somewhat vague on why they chose these values in particular, it helps to gather more information on what they describe as "its inability to solve a 'chicken and egg' problem" for the baseline model (also called threshold). This is essentially the idea that detecting events requires knowledge of what "normal" behavior is which you don't really have unless you know what not-normal behavior is. The threshold model did reasonably well, however it should be pointed out that the threshold model would probably not be as accurate without the information from MMPP since the percentages shown reflect adjusting the epsilon or cutoff value to best match the results from the MMPP.

3.3 Model Assessment

3.3.1 Testing Heterogeneity

Obviously, the application of the model discussed is for time processes where there is an additional element of cycles of counts. By way of example, the building's data set will likely share similar building entry counts on all Sundays that are different from Mondays. A similar argument could be made for certain times of the day. It then becomes important for the author's of the paper to test for heterogeneity to decide on how to combine data in order to obtain better predictions.

Specifically, they perform 6 tests for heterogeneity. Three for a day effect (all days the same, weekends and weekdays the same, or all day effects separate). Three for a time-of-day effect (any time during the day is the same, times during weekends/weekdays the same, or all time effects separate). The tests were performed on the marginal likelihoods which are the likelihood of the data under the model. Theses marginal likelihoods were from samples drawn from the MCMC process.

3.3.2 Performance

The threshold model performed surprisingly well considering how much less complexity it had. It should be noted however that the percentage accuracies reported for the threshold were made more accurate by updating: "...to find the same number of events as the MMPP model by adjusting its threshold ϵ." It is hard to say how well it did without the MMPP to aid it. The MMPP model performs very well however. It seems that a very high accuracy can be achieved with the MMPP by in a sense "over-predicting" how many events are going to take place.

Predicted Building Events Totals (by adjusted MMPP model parameters)	**MMPP Model**	**Threshold Model**
104	100.0%	86.2%
70	96.9%	75.9%
48	79.3%	65.5%

Table 1: Accuracies of predictions for the building data: the percentage of the 29 known events correctly predicted by each model, for different numbers of total events predicted.

Predicted Traffic Events totals (by adjusted MMPP model parameters)	**MMPP Model**	**Threshold Model**
203	100.0%	86.2%
186	100.0%	75.9%
134	100.0%	65.5%
98	98.5%	60.5%

Table 2: Accuracies of predictions for the freeway traffic data: the percentage of the 76 known events correctly predicted by each model, for different numbers of total events predicted.

3.3.3 Potential Improvements

The paper is using a sort of Bayesian approach to their model. Figuring out how to incorporate more prior knowledge of their data so that they have a more informative prior might be a good way to improve on the model. This is especially the case since deciding what constituted an event to begin with might be an exercise in of itself. One other item that might need to be looked into further, was how to deal with data for more sparse events since it seemed both models struggled a little with the building data set.

Also, the paper mentions using multiple correlated time-series arising from different sources (sensors, doors, freeway entrances, etc.). Incorporating this would definitely add to the complexity of the model but would provide, as the authors say, "richer information [about] behavioral patterns."

4 Conclusion

In some ways the purpose of this paper was to model a very limited type of data. The data consists only of time-stamps for when a count occurred. From that information, the goal was to determine when something out of the ordinary occurred but to avoid counting random noise. The baseline model, although somewhat effective, was a limited approach and essentially an extension of how we (or anyone) intuitively would say that an event had occurred by looking at a plot of data counts. In order to more accurately predict events and avoid false-positives, the authors developed the MMPP which proved to be highly effective in established events in two different sets of data.