Continuous-time Estimation of a Change-point in a Poisson Process

By R. Webster West and R. Todd Ogden
Department of Statistics
University of South Carolina
Columbia, SC 29208, U.S.A.

Summary

The problem of estimation of change-points in a sequence of Poisson random variables is approached by allowing the change-point τ to range over the continuous time interval $(0, T)$. A maximum-likelihood point estimator is derived, along with a Bayesian-based interval estimator. Simulation studies confirm that the true coverage probability is close to the nominal one for several locations of the change-point within the sequence and various rate changes. Finally, the procedure is applied to the British coal-mining disaster data and the results are shown to perform extremely well in comparison with previous estimates based on more complete information.

Some key words: British coal mining disaster data, change-point, continuous-time estimation, Poisson process.

1 Introduction

The task of detecting a change-point in the number of daily defects in an industrial process or in the number of annual cases of a particular genetic disease may be considered in the context of a Poisson process. Henderson and Matthews (1993) consider this problem in standard change-point style by restricting the change-point to fall at the endpoint of a time interval. While this is very natural, this restriction only allows for the estimation of the interval in which the change occurred. In order to establish reasons for the

industrial or epidemiological shift a more precise estimate of the change-point is desirable. For example, a refined estimate of the exact time rather than the day in which the number of defects increases in an industrial process allows one to pinpoint and eliminate problems in a more straightforward manner. Methods for obtaining point estimates and confidence intervals for a change-point which may take on any of a continuum of values are developed in this paper.

Suppose data over T unit time periods, $X_1, \ldots, X_T$, are observed where the observation X_i represents the number of events that occurred in the ith time period. A natural model for X_i is the Poisson distribution. The question of interest is whether there has been an abrupt change in the rate parameter defining the Poisson distribution over the T periods. Let τ, $0 < \tau < T$, represent such a change-point, θ_0 represent the Poisson rate parameter before the change, and θ_1 represent the rate parameter after the change.

Denote by $[x]$ the greatest integer function of x. If a change occurs at time point τ, then the change occurs in the $([\tau]+1)$st interval. The observation for this period can be thought of as a sum of two independent Poisson random variables which are not observed directly: $X_{[\tau]+1} = X_{i0} + X_{i1}$, where $X_{i0} \sim$ Poisson $(p(\tau)\,\theta_0)$, $X_{i1} \sim$ Poisson $((1-p(\tau))\,\theta_1)$, and $p(\tau) = \tau - [\tau]$. The sum of independent Poisson random variables is also Poisson so that

$$X_i \sim \begin{cases} \text{Poisson}(\theta_0) & i = 1, \ldots, [\tau] \\ \text{Poisson}(p(\tau)\,\theta_0 + (1-p(\tau))\,\theta_1) & i = [\tau]+1 \\ \text{Poisson}(\theta_1) & i = [\tau]+2, \ldots, T \end{cases}$$

where all of the observations are mutually independent.

2 Point Estimation

Using the model described in Section 1, it is possible to derive maximum likelihood estimators for the three parameters, τ, θ_0, and θ_1. The log likelihood is given by

$$\log L(\theta_0, \theta_1, \tau | X_1, \ldots, X_n) = \tag{1}$$

$$-\tau\theta_0 + \left(\sum_{i=1}^{[\tau]} X_i\right)\log(\theta_0) - (T-(\tau+1))\,\theta_1 + \left(\sum_{i=[\tau]+2}^{T} X_i\right)\log(\theta_1)$$

$$-p(\tau)\,\theta_0 - (1-p(\tau))\,\theta_1 + X_{[\tau]+1}\log(p(\tau)\,\theta_0 + (1-p(\tau))\,\theta_1) - \log\sum_{i=1}^{T} X_i!.$$

If τ is known, then (1) can be differentiated with respect to the two parameters θ_0 and θ_1 and the maximum likelihood estimators of θ_0 and θ_1 can be shown to satisfy

$$\hat{\theta}_0 = \frac{1}{\tau}\left\{\sum_{i=1}^{[\tau]} X_i + \frac{p(\tau)\,\hat{\theta}_0}{p(\tau)\,\hat{\theta}_0 + (1-p(\tau))\,\hat{\theta}_1} X_{[\tau]+1}\right\}$$

and

$$\hat{\theta}_1 = \frac{1}{T-\tau}\left\{\sum_{i=[\tau]+2}^{T} X_i + \frac{(1-p(\tau))\,\hat{\theta}_1}{p(\tau)\,\hat{\theta}_0 + (1-p(\tau))\,\hat{\theta}_1} X_{[\tau]+1}\right\}.$$

Estimates of θ_0 and θ_1 can be found by solving the resulting quadratic expression. The above expressions are similar in some sense to what one would typically expect in a Poisson rate estimate in that they are the estimated number of events in an interval divided by the length of the interval.

Of course in most usual situations, the change-point, τ, is not known but rather it is the most important quantity to be estimated. Even though τ is defined on a continuous space, the likelihood is not differentiable with respect to τ because of discontinuities in the function $p(\tau)$, so standard calculus techniques used to find maximum likelihood estimates are not applicable in this situation. One approach to finding the maximum likelihood estimators of all three parameters is to let τ range over a set of points on the interval $(0, T)$, computing $\hat{\theta}_0(\tau)$ and $\hat{\theta}_1(\tau)$ and the likelihood for each value of τ.

An example of this is in Figure 1, in which $\log L(X|\hat{\theta}_0(\tau), \hat{\theta}_1(\tau), \tau)$ is plotted as a function of τ. The data are from Henderson and Matthews (1993), and represent the number of haemolytic uraemic syndrome (HUS) cases in Newcastle from 1970–1989. This plot demonstrates the discontinuities of the first derivative which appear at the set of integers. Based on this type of analysis, the maximum likelihood estimates would be $\hat{\tau} = 14.944$ with $\hat{\theta}_0 = 2.429$ and $\hat{\theta}_1 = 12.600$, which closely corresponds with the estimate of

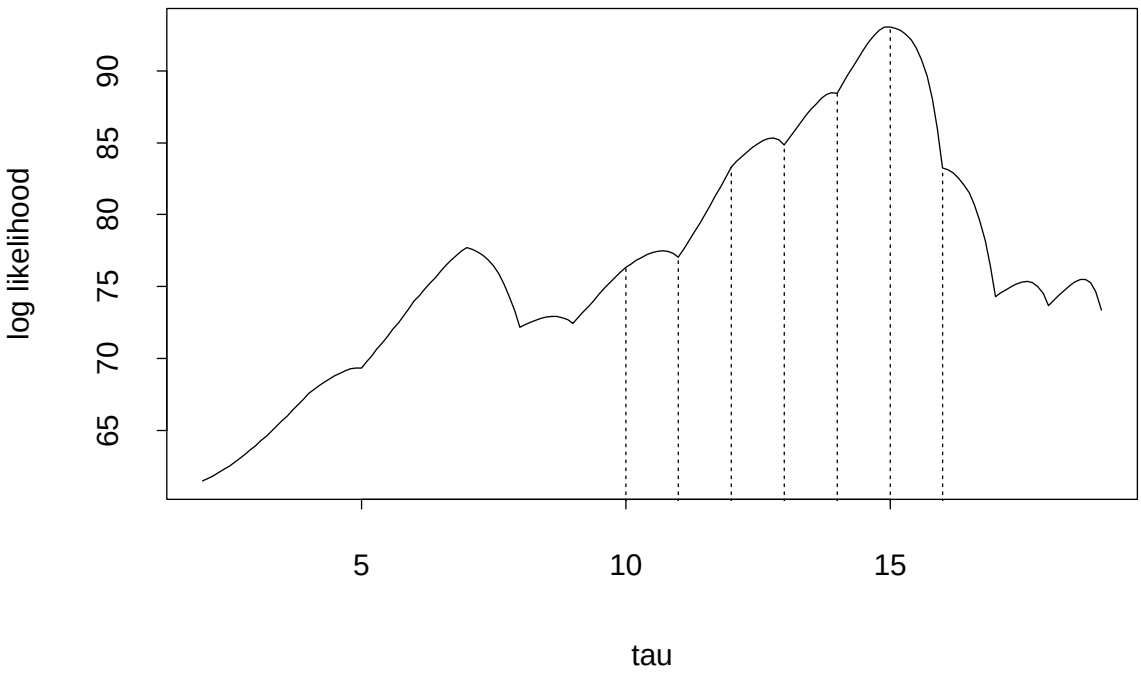

Figure 1: Profile log-likelihood over τ for HUS data.

15 given by Henderson and Matthews (1993) who considered only the set of integers as possible change-points.

The shape of the plot in Figure 1 provides insight into an alternative approach to finding maximum likelihood estimates in that the likelihood appears piecewise smooth over each of the time intervals. In fact this is the case, so a grid search as described before is not the most efficient way to obtain parameter estimates: the problem can be considered one interval at a time. Considering the problem in this way yields some interesting results.

If it is known that $\tau \in [i^*, i^*+1)$, then the maximum likelihood estimators of θ_0, θ_1 and τ are given by

$$\begin{aligned} \hat{\theta}_0 &= \frac{S_{i^*}}{i^*} \\ \hat{\theta}_1 &= \frac{S_T - S_{i^*+1}}{T - (i^*+1)} \\ \hat{\tau} &= i^* + \frac{X_{i+1} - \hat{\theta}_1}{\hat{\theta}_0 - \hat{\theta}_1}, \end{aligned}$$

where $S_i = X_1 + \cdots + X_i$. Of course, the expressions above do not guarantee that the estimate of τ will in fact lie in the specified interval, so in that case, the endpoints of the interval must be examined.

This result is interesting for several reasons. First, the estimate of the Poisson parameter before (after) the change-point depends only on observations that occur before (after) the interval which contains τ. Second, the data point representing the interval which contains τ, X_{i^*+1}, is only used for estimation of τ.

By approaching the problem in this manner, a complicated grid search is reduced to a search involving only one calculation for each of the T intervals.

3 Interval estimation

Since the likelihood function over τ is not smooth, usual normal theory may not be applied to the maximum likelihood estimator for asymptotic interval estimates. Krishnaiah and Miao (1988) review Bayesian approaches to the problem of estimating τ. Let $\pi(\tau)$ denote the prior for the change-point. If $f(X|\tau)$ represents the likelihood function of the data, then the posterior distribution of τ is proportional to $\pi(\tau) f(X|\tau)$. If the prior distribution of

τ is U $(0, T)$, then the Bayes estimator of τ is just the maximum likelihood estimator of Section 2.

Bayesian confidence intervals are formed by computing the highest posterior density region. This involves a computationally intensive algorithm which will not be described here. Simulation studies were conducted to explore the effects of change-point location within the sequence and the difference in rate parameters on the coverage probability for the interval estimation procedure.

Table 1 displays some Monte Carlo results for the coverage probability and mean width of the Bayesian interval as three general change-point locations are considered, corresponding roughly to the first quarter of the data $\tau = T/4 + 0.25$, the middle of the data $\tau = T/2 + 0.5$ and the last quarter of the data $\tau = 3T/4 - 0.25$. The adding of fractions is to ensure that the change-point will not fall at the endpoint of any interval. For this study, the Poisson parameters used were $\theta_0 = 10$ and $\theta_1 = 20$. The sample sizes for the simulation ranged from $T = 10$ to $T = 100$, along multiples of 10. In each case, 1000 Poisson samples were simulated and 95% confidence intervals were computed. In these simulated examples, as in the real-data example of the next section, the uniform prior is applied to the change-point. As seen in Table 1, the coverage probabilities are in general close to the nominal 95%, ranging from a low of 0.932 to a high of 0.961, indicating that the method seems to work well for a variety of sample sizes. It is expected that the mean width of the confidence intervals should decrease as T increases. Though not immediately evident from Table 1, this is seen to be taking place in this situation if it is taken into account that the number of observations increases by 10 at each step. To compare widths of confidence intervals, a more appropriate measure might be the proportion of the data set covered by the interval, or width/T.

In Table 2 are simulated results using a single sample size $T = 50$ and constant $\theta_0 = 10$, allowing θ_1 to increase. The change-point for the simulation reported in Table 2 was 25.5 and 1000 replications were performed for each choice of θ_1. All the simulated coverage probabilities in Table 2 except one are reasonably close to the nominal 0.95 and, as to be expected, the average interval widths are seen to decrease dramatically as θ_1 increases which means the change-point is easier to identify for larger differences between θ_0 and θ_1.

Table 1: Coverage and average interval width for simulation study.

	$T/4 + 0.25$		$T/2 + 0.5$		$3T/4 - 0.25$	
T	empirical coverage	mean width	empirical coverage	mean width	empirical coverage	mean width
10	0.957	3.619	0.958	3.324	0.955	3.650
20	0.961	3.504	0.957	3.084	0.953	3.328
30	0.958	3.152	0.958	2.762	0.951	3.010
40	0.961	2.807	0.940	2.693	0.947	2.759
50	0.945	2.740	0.952	2.660	0.957	2.621
60	0.940	2.639	0.942	2.592	0.945	2.588
70	0.952	2.534	0.938	2.577	0.950	2.538
80	0.947	2.499	0.948	2.526	0.950	2.535
90	0.948	2.493	0.936	2.249	0.960	2.423
100	0.949	2.494	0.932	2.530	0.948	2.487

Table 2: Coverage, mean and median width for $T = 50$, $\theta_0 = 10$ as θ_1 changes.

θ_1	coverage	mean width
12.0	0.984	34.236
16.0	0.948	6.742
20.0	0.952	2.660
24.0	0.944	1.679
28.0	0.931	1.288
32.0	0.914	1.043
36.0	0.939	0.877
40.0	0.952	0.761

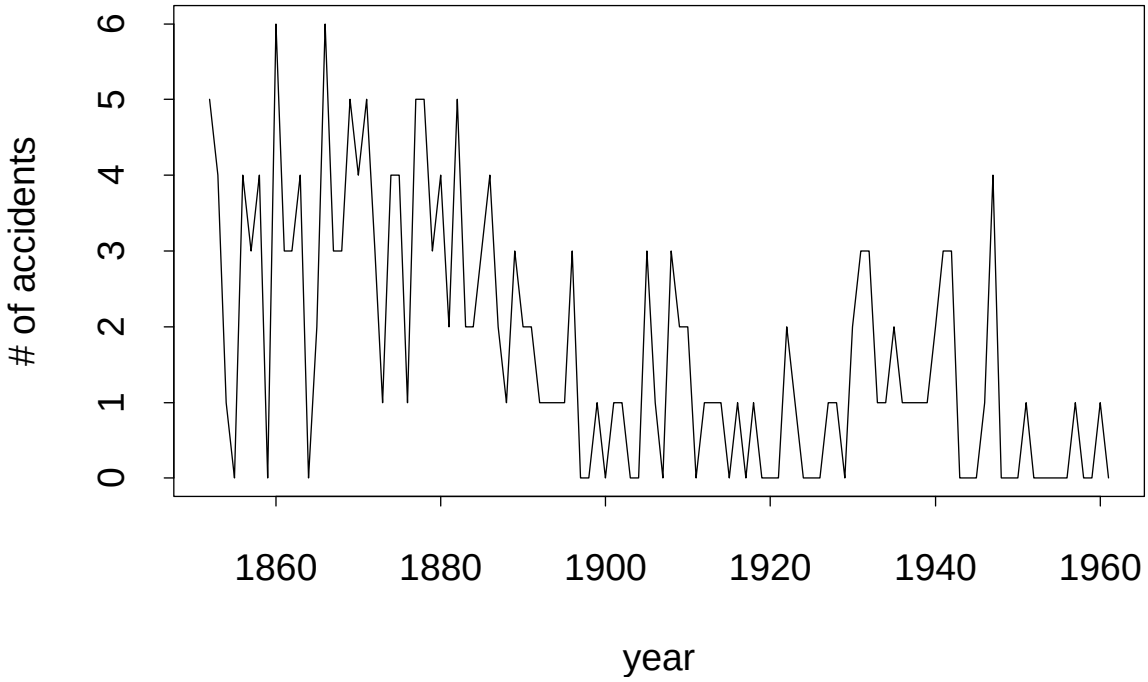

Figure 2: Plot of British coal mining accident data.

4 Application to coal-mining disaster data

To allow a comparison of the proposed method with other existing methods, a classical data set is studied. The data set, originally reported in Maguire, Pearson, and Wynn (1952), consists of intervals (in days) between coal-mining disasters in which more than 10 men were killed. A corrected and expanded version of this data is tabulated in Jarrett (1979). The yearly number of such disasters between 1852 and 1961 are displayed in Figure 2.

Both data sets are given in terms of time between successive disasters, so prior approaches have involved using an exponential distribution to model the data. Previous estimates and confidence limits for the change-point are presented in Table 3. It should be noted that the estimate of Worsley (1986) was based on the original data of Maguire et al. (1952).

The data here are binned so as to allow application of the procedures discussed in this paper. Such a procedure is not the best way to analyze the data — indeed, it amounts to "throwing away" a certain amount of informa-

Table 3: Previous point and interval estimates of the change-point.

Method	$\hat{\tau}$	Left limit	Right limit
Loader (1992)	Mar 10, 1890	Length 8.08 years	
Akman and Raftery (1992)	Mar 10, 1890	Oct 6, 1886	Dec 17, 1898
Raftery and Akman (1986)	Aug 27, 1890	May 15, 1887	Aug 3, 1895
Worsley (1986)	1890	1884	1895

Table 4: Estimates and confidence intervals for τ with bin sizes from 1 to 10 years.

years/bin	$\hat{\tau}$	Left limit	Right limit
10	Apr 24, 1889	May 26, 1885	Aug 7, 1894
9	Jan 28, 1891	Nov 6, 1884	Aug 25, 1895
8	Dec 14, 1888	Jul 9, 1885	May 26, 1894
7	May 17, 1889	Dec 2, 1883	Aug 25, 1893
6	Dec 26, 1889	Sep 13, 1884	Jun 26, 1894
5	May 12, 1889	Apr 20, 1886	Dec 13, 1895
4	Nov 27, 1889	Aug 7, 1885	Dec 2, 1893
3	Jun 2, 1889	Mar 3, 1886	Feb 28, 1896
2	Dec 25, 1890	Jun 24, 1886	Jun 2, 1894
1	Jul 1, 1891	Jan 26, 1886	Jun 24, 1895

tion. It is performed only to allow comparison between the proposed methods based on a sequence of Poisson random variables and existing methods based on a sequence of exponential random variables. The proposed methods are applied to binned data, representing the number of accidents in each k-year interval, for k ranging from one to ten. Point estimates and corresponding 95% confidence limits are listed in Table 4. The estimates for the number of events per year before and after the change are presented in Table 5.

Table 5: Estimates of yearly rates before and after $\hat{\tau}$ for the various bin sizes.

years/bin	$\hat{\theta}_0$/year	$\hat{\theta}_1$/year
10	3.23	0.90
9	3.19	0.87
8	3.25	0.95
7	3.23	0.94
6	3.19	0.91
5	3.23	0.90
4	3.19	0.91
3	3.19	0.93
2	3.13	0.90
1	3.10	0.90

5 Discussion

It is striking to notice from Table 4 the amount of consistency among estimates from annual data (110 data points) to those based on 10-year data (11 data points). In Table 5, a remarkable consistency among estimates of the rate parameters for the various bin sizes is also displayed. This is somewhat surprising, considering the amount of information that is being "thrown away" for the larger bin sizes.

Perhaps the most striking feature of the estimates given in Table 4 is their close correspondence to the previous estimates in Table 3. Due to the binning of data, one might expect much lower precision in the interval estimates. However, all but the nine-year binned data have 95% confidence intervals of less than 10 years width, and all of the confidence intervals compare favorably with existing estimates. In many situations, such detailed data as the time between events are not available, but rather the only observations are the number of events in a given time interval. The proposed methods apply to this more general setting and compare favorably to existing procedures designed for use with more detailed information.

References

Akman, V. E., & Raftery, A. E. (1992). Asymptotic inference for a change-point Poisson process. *The Annals of Statistics*, **14**, 1583–1590.

Henderson, R., & Matthews, J. N. S. (1993). An investigation of change-points in the annual number of cases of haemolytic uraemic syndrome. *Applied Statistics*, **42**, 461–471.

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika*, **66**, 191–193.

Krishnaiah, P. R., & Miao, B. Q. (1988). Review about estimation of change-points. In P. R. Krishnaiah, & C. R. Rao (Eds.), *Handbook of Statistics, Vol. 7*, pp. 375–402. Elsevier, Amsterdam.

Loader, C. R. (1992). A log-linear model for a Poisson process change point. *The Annals of Statistics*, **20**, 1391–1411.

Maguire, B. A., Pearson, E. S., & Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika*, **38**, 168–180.

Raftery, A. E., & Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change point. *Biometrika*, **73**, 85–89.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, **73**, 91–104.