Mixed Poisson Processes with Panel Flow

Anatoly Zhigljavsky and Vippal Savani *

Abstract— **The problem of parameter estimation and statistical inference when fitting an $M/G/\infty$ queuing process to data is considered in the situation where the times of arrival and departure are unknown; instead recurrent events, which occur according to a mixed Poisson process, are observed between the times of arrival and departure.**

Keywords: Mixed Poisson process, panel flow, $M/G/\infty$ system

1 Introduction

This paper investigates the modeling of arrival and departure times of individuals in the case where precise arrival and departure times are unknown; instead we observe recurrent events between the times of arrival and departure. In the case where arrival and departure times are known, it is straightforward to fit a standard queueing model of the form $M/G/\infty$ where the arrival process is Markovian, the time spent between arrival and departure follows a general distribution G and there are no queues.

The situation described above occurs in many fields. In the field of market research we observe purchases made by individuals during an observation period. The time when an individual begins or ceases to buy (i.e. time of arrival and departure), however, is unknown. Such an example occurs for purchases of products such as baby food, where individuals "flow" in and out of the panel.

1.1 Panel flow

Consider a panel of individuals consisting of subjects that have recurrent events. Panel flow occurs when individuals move into and out of the panel. We assume that panel flow occurs according to an $M/G/\infty$ process so that: individuals enter the panel according to a Poisson process with intensity ν; individuals stay in the panel for a random period of time σ (independent and identically distributed for each individual) and that there are an infinite number of servers so that an individual will immediately enter into the panel.

When the problem of panel flow exists then recurrent events may no longer be analyzed under the assumption of stationarity. In this case, to model recurrent events, we require knowledge of N_t, the number of individuals in the panel at time t. If we know the intensity of the Poisson arrival process (ν) and the distribution of the time spent in the panel (G) then we can use the properties of $M/G/\infty$ processes to model N_t.

Let us assume that N_t, modeled by an $M/G/\infty$ process, is at a steady stationary state. For all t the random variables N_t have Poisson distribution with mean $\nu E\sigma$ and the covariances $cov(N_t, N_{t+s}) = \nu E[\max(0, \sigma - s)]$ for all $s > 0$ (see e.g. [4]).

The assumption that individuals enter the panel according to a Poisson process with intensity ν is a very natural assumption in practice. If the arrival process is non-stationary, it can be easily deseasonalised (that is, made stationary). The intensity ν of the arrival Poisson is also simple to estimate.

1.2 Recurrent events

In estimating the parameters of the $M/G/\infty$ process we require the arrival and departure times for each individual. In our problem, the arrival and departure times are unknown. We consider the use of recurrent events to estimate these times.

We consider mixed Poisson processes as a model for the recurrent events. We define a mixed Poisson process as a pure Poisson process with random intensity λ.

In the case of panel data analysis, the mixed Poisson process has a natural interpretation in that each individual has events according to a Poisson process; however, the intensity of each process

*School of Mathematics, Cardiff University, CF24 4AG Cardiff, UK, Email: Zhigljavskyaa@cf.ac.uk, SavaniV@cf.ac.uk

varies randomly across processes according to some distribution.

The fitting of mixed Poisson processes to recurrent data has been studied in fields such as: market research where events are purchases of products (see e.g. [2, 5]); health care where events may be the occurrence of cancer tumours or arrivals at hospitals (see e.g. [1, 6]) and insurance where events may be the occurrence of accidents or insurance claims (see e.g. [3]).

1.3 Observation period

In practice, we collect panel data over a fixed time period called the observation period. When analyzing flow through, we must assess whether or not an individual is in the panel at the start and at the end of the observation period.

For individuals whose recurrent events occur completely within the observation period, we may use the full realization of recurrent events to estimate the time of their arrival and the duration of their service time. For all other individuals we have i) a censored observation for the length of duration in the panel and ii) an estimate for either the time of arrival or the time of departure (but not both).

For a given individual, let t_* be the time when an individual enters the panel and let t^* ($t^* \geq t_*$) be the time when this individual leaves the panel. During the period of time $[t_*, t^*]$, the individual has recurrent events according to a mixed Poisson process. The individual is observed during $[T_*, T^*]$.

Layout

We consider the following problems related to panel flow (modeled by a stationary $M/G/\infty$ system) where individuals have recurrent events in accordance to a mixed Poisson process:

- estimation of individual intensities and the underlying distribution of these intensities;
- estimation of the time period an individual spends in a panel;
- estimation of the distribution G (the distribution for a random individual to stay in the panel);
- estimation of the distribution for the number of recurrent events in a given time interval;
- estimation and forecast of the characteristics of recurrence.

2 Estimation of intensities

Let $[T_*, T^*]$ be a fixed time interval during which we observe recurrent events. For a given individual, let t_* be the time when the individual enters the panel and t^* ($t^* \geq t_*$) be the time when this individual leaves the panel. We assume that any individual on the panel has a positive probability of having recurrent events during the time interval $[t_*, t^*]$. These events occur according to a Poisson process with constant intensity λ, where λ is a random variable. That is, the rate of events I_t for a particular individual is

$$I_t = \begin{cases} \lambda & \text{if } t \in [t_*, t^*] \\ 0 & \text{otherwise.} \end{cases}$$

We assume that if the same individual enters the panel more than once then the individual is assigned a new interval of random duration σ and a new random intensity λ; both σ and λ are independent of the values used previously.

Consider an individual who has n recurrent events at times $t_1, \ldots, t_n$ during the time interval $[T_*, T^*]$. Here $n \geq 0$ and

$$T_* \leq t_1 \leq \ldots \leq t_n \leq T^*. \tag{1}$$

The sequence of recurrent events (1) observed during the interval $[T_*, T^*]$ is a subset of the realization of the Poisson process during the interval $[t_*, t^*]$.

2.1 Estimators for the intensity

In the estimation of λ for a given individual, as we have no data prior to T_* and beyond T^*, we can only use recurrent events observed during the time interval $[T_*, T^*]$. If we know that $[T_*, T^*] \subset [t_*, t^*]$ then the number of events n has a Poisson distribution with intensity λT where $T = T^* - T_*$. Therefore, to estimate λ we can use the maximum likelihood estimator (MLE) $\hat{\lambda} = n/T$. For this estimator, we have $E\hat{\lambda} = \lambda$ and $var(\hat{\lambda}) = \lambda/T$.

If, however, $[T_*, T^*] \not\subseteq [t_*, t^*]$ then the estimator $\hat{\lambda}$ is biased: indeed,

$$E\hat{\lambda} = \frac{S}{T}\lambda \;\text{ and }\; var(\hat{\lambda}) = \frac{S}{T}\lambda\,. \tag{2}$$

where $S = S^* - S_*$, $S_* = \max\{t_*, T_*\}$, $S^* = \min\{t^*, T^*\}$. Note that, for any individual, we have

$S \leq T$. If S is very different from T, the quality of the estimator $\hat{\lambda}$ is poor. In many cases, we can improve the quality of this estimator by using only the observations that belong to the interval $[t_1, t_n]$.

As an alternative estimator to $\hat{\lambda}$ we consider the estimator

$$\tilde{\lambda} = \begin{cases} n/T & \text{for } n = 0, 1, 2, 3 \\ (n-2)/(t_n - t_1) & \text{for } n > 3 \end{cases} \quad (3)$$

This estimator is based upon the fact that (for $n > 3$) individuals are, with certainty, in the panel during the period (t_1, t_n) and hence events during the time interval (t_1, t_n) occur according to a Poisson process.

2.2 Efficiency and bias of estimators: no panel flow

To understand the efficiency of the $\tilde{\lambda}$, we need to compare the estimators $\hat{\lambda}$ and $\tilde{\lambda}$ in the case $[T_*, T^*] \subset [t_*, t^*]$ (that is, when $S = T$). For $n = 0, 1, 2, 3$ we have $E\tilde{\lambda} = E\hat{\lambda} = \lambda$ and $var(\tilde{\lambda}) = var(\hat{\lambda}) = \lambda/T$. Thus we only need to compare estimators for the case $n > 3$.

Lemma 1. If $[T_*, T^*] \subset [t_*, t^*]$, then the estimator (3) is an unbiased estimator of λ and

$$var(\tilde{\lambda}) = \frac{\lambda}{T}\left(1 + 2\exp(-\lambda T)\sum_{n=3}^{\infty}\frac{(\lambda T)^n}{(n-2)n!}\right) \quad (4)$$

Proof. If $[T_*, T^*] \subset [t_*, t^*]$, then the number of events n in the interval $[T_*, T^*]$ has a Poisson distribution Poisson(λT).

Note that $t_n - t_1$ is a random variable, thus to analyze the statistical properties of $\tilde{\lambda}$ we require the distribution of $1/(t_n - t_1)$. The random variables $t_1/T, \ldots, t_n/T$ (where $T = T^* - T_*$) have the same distribution as n order statistics corresponding to an independent sample of size n from the uniform distribution on $[0, 1]$. Consequently, $(t_n - t_1)/T$ has the same distribution as the second largest order statistics from this sample. Assume that $n > 3$, then the density of $\xi_n = (t_n - t_1)/T$ is

$$p(x) = n(n-1)x^{n-2}(1-x), \;\; 0 < x < 1.$$

This yields that the density of the random variable $1/\xi_n = T/(t_n - t_1)$ is

$$\phi_n(t) = n(n-1)t^{-n}(1 - 1/t), \;\; 1 < t < \infty.$$

Proof of mean: $E\tilde{\lambda}$.

For all $n > 3$, we have

$$E(1/\xi_n) = n(n-1)\int_1^{\infty} t^{1-n}(1 - 1/t)dt = \frac{n}{n-2}.$$

and hence

$$E(\tilde{\lambda}|n) = [(n-2)/T]\, E(1/\xi_n) = n/T.$$

Thus for all $n = 0, 1, 2, \ldots$ we have $E(\tilde{\lambda}|\, n) = n/T$ which implies

$$E(\tilde{\lambda}) = E_n(E(\tilde{\lambda}|n)) = E_n(n/T) = E_n(n)/T = \lambda.$$

Proof of variance: $var(\tilde{\lambda})$.

For the computation of the variance we require $E(1/\xi_n^2)$ and $E(n^2)$. We have $E(n^2) = \lambda T(\lambda T + 1)$ and

$$E(1/\xi_n^2) = \int t^2\phi_n(t)dt = n(n-1)\int_1^{\infty} t^{2-n}(1 - 1/t)dt$$
$$= \frac{n(n-1)}{(n-2)(n-3)}.$$

Let $\mu = \lambda T$ and $p_n(\mu) = \exp(-\mu)\mu^n/n!$ $(n = 0, 1, 2, \ldots)$ then

$$E(\tilde{\lambda}^2) = E_n(E(\tilde{\lambda}^2|n))$$
$$= \sum_{n=0}^{3} p_n(\mu)\frac{n^2}{T^2} + \sum_{n=4}^{\infty} p_n(\mu)\frac{(n-2)^2}{T^2}E(1/\xi_n^2)$$

and

$$E(\tilde{\lambda}^2) = \sum_{n=0}^{3} p_n(\mu)\frac{n^2}{T^2} + \sum_{n=4}^{\infty} p_n(\mu)\frac{n(n-1)(n-2)}{T^2(n-3)}$$

$$= \sum_{n=0}^{\infty} p_n(\mu)\frac{n^2}{T^2} + \sum_{n=4}^{\infty} p_n(\mu)\left(\frac{n(n-1)(n-2)}{T^2(n-3)} - \frac{n^2}{T^2}\right)$$

$$= \frac{1}{T^2}\left(\mu(\mu+1) + 2\exp(-\mu)\sum_{n=4}^{\infty}\frac{n\,\mu^n}{(n-3)n!}\right)$$

which implies (4). ■

The efficiency of the estimator $\tilde{\lambda}$ can be defined as

$$eff(\tilde{\lambda}) = var(\hat{\lambda})/var(\tilde{\lambda})$$

Then the formulae $var(\hat{\lambda}) = \lambda/T$ and (4) imply

$$eff(\tilde{\lambda}) = 1/\left(1 + 2\exp(-\lambda T)\sum_{n=3}^{\infty}\frac{(\lambda T)^n}{(n-2)n!}\right).$$

For large λ we have

$$eff(\tilde{\lambda}) = 1 - \frac{3}{\lambda T} + O\left(\frac{1}{\lambda^2}\right) \text{ as } \lambda \to \infty.$$

2.3 Bias of estimators: panel flow

In the general case, when the inclusion $[T_*, T^*] \subset [t_*, t^*]$ does not necessarily hold, $n \sim Poisson(\lambda S)$ and the estimator $\hat{\lambda}$ may be heavily biased, see (2).

As $\tilde{\lambda} = \hat{\lambda}$ for $n = 0, 1, 2, 3$, the estimator $\tilde{\lambda}$ is biased for $n = 0, 1, 2, 3$. However, since $\tilde{\lambda}$ is an unbiased estimator for λ for $n > 4$, the estimator $\tilde{\lambda}$ has a smaller bias than that of $\hat{\lambda}$. For the mean of $\tilde{\lambda}$, we have

$$E(\tilde{\lambda}) = \sum_{n=0}^{3} p_n(\lambda S)\frac{n}{T} + \sum_{n=4}^{\infty} p_n(\lambda S)\frac{n}{S} =$$
$$\sum_{n=0}^{\infty} p_n(\lambda S)\frac{n}{S} + \sum_{n=0}^{3} n p_n(\lambda S)\left(\frac{1}{T} - \frac{1}{S}\right) =$$
$$\lambda - \left(\frac{1}{S} - \frac{1}{T}\right)\left(\lambda S + (\lambda S)^2 + \frac{1}{2}(\lambda S)^3\right) e^{-\lambda S}.$$

3 Time length an individual spends on the panel

3.1 Estimation of the time length

Consider an individual with the sequence of events (1). Let us estimate t_* and t^*, the times of arrival of the individual to the panel and his departure from the panel, using only this data. Under the assumption that the sequence of events (1) is a realization of a Poisson process on $[t_*, t^*]$, the distribution of all differences $t_{i+1} - t_i$ $(i = 0, \ldots, n;\ t_0 = t_*,\ t_{n+1} = t^*)$ is identical. The empirical mean of the differences $t_{i+1} - t_i$ $(i = 1, \ldots, n-1)$ is $(t_n - t_1)/(n-1)$ and hence MLE estimators of t_* and t^* are

$$\tilde{t}_* = t_1 - (t_n - t_1)/(n-1),\ \ \tilde{t}^* = t_n + (t_n - t_1)/(n-1).$$

If $\tilde{t}_* < T_*$ we conclude that the individual has been on the panel before the observation period started; in this case, we cannot use the estimator $\tilde{t}_*$ as we suspect that there were unregistered events for this individual prior to T_*. Similarly, if $\tilde{t}^* > T^*$, the individual will most probably have more events at times $t_{n+j} > T^*$, $j = 1, 2, \ldots$.

3.2 Distribution of the time length

Let $\sigma = t^* - t_* \sim G$ be the period of time a particular individual stays in the panel and let $[T_*, T^*]$ be the observation period. We assume that there is a non-empty intersection of the intervals $[t_*, t^*]$ and $[T_*, T^*]$ so that the individual is in the panel for at least a part of the whole period $[T_*, T^*]$. Let $S(\sigma)$ denote the period defined by the intersection of the time intervals $[T_*, T^*]$ and $[t_*, t^*]$.

Let $[t_*, t^*]$ be fixed and consider the random placement of the observation window $[T_*, T^*]$ of fixed length T at a random starting point T_*. Let $\xi = \sigma/T$ and define $s(t) = S(\sigma)/T$ where $t = (t^* - T_*)/T$ with $t \in [0, 1+\xi]$. Then, depending on whether $\xi <= 1$ or $\xi >= 1$, we have two possible laws for the function $s(t)$:

$$\text{if } \xi \le 1: \quad s(t) = \begin{cases} t & \text{for } 0 \le t \le \xi \\ \xi & \text{for } \xi \le t \le 1 \\ 1+\xi-t & \text{for } 1 \le t \le 1+\xi \end{cases}$$

$$\text{if } \xi \ge 1: \quad s(t) = \begin{cases} t & \text{for } 0 \le t \le 1 \\ 1 & \text{for } 1 \le t \le \xi \\ 1+\xi-t & \text{for } \xi \le t \le 1+\xi \end{cases}$$

with all values of t being equiprobable.

This implies that for fixed ξ, the random variable $s(t)$ has the following distribution:

$$\text{if } \xi \le 1:\ s(t) \text{ is } \begin{cases} = \xi & \text{w. p. } \frac{1-\xi}{1+\xi} \\ \text{uniform on } [0, \xi] & \text{w. p. } \frac{2\xi}{1+\xi} \end{cases}$$

$$\text{if } \xi \ge 1:\ s(t) \text{ is } \begin{cases} = 1 & \text{w. p. } \frac{\xi-1}{1+\xi} \\ \text{uniform on } [0, 1] & \text{w. p. } \frac{2}{1+\xi} \end{cases}$$

To obtain the distribution of the length $s(t)$ we need to integrate this distribution with respect to the distribution of ξ; the c.d.f. of ξ is $G(x/T)$ where $G(\cdot)$ is the c.d.f. of σ.

References

[1] R. J. Cook and W. Wei. Conditional analysis of mixed Poisson processes with baseline counts: implications for trial design and analysis. *Biostatistics*, 4:479–494, 2003.

[2] A. S. C. Ehrenberg. *Repeat-Buying: Facts, Theory and Applications.* Oxford University Press., 1988.

[3] J. Grandell. *Mixed Poisson processes.* Chapman & Hall, London, 1997.

[4] M. Parulekar and A. M. Makowski. Tail probabilities for $M/G/\infty$ input processes. I. Preliminary asymptotics. *Queueing Systems Theory Appl.*, 27(3-4):271–296 (1998), 1997.

[5] V. Savani and A. A. Zhigljavsky. Modeling recurrent events in panel data using mixed Poisson models. *Present volume.*

[6] A. Woehl. Modeling health care events using mixed Poisson models. *Present volume.*