UNIVERSITY OF CALIFORNIA
RIVERSIDE

Statistical Process Control Methods for Network Monitoring Using Generalized Linear Mixed Model

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Yingzhuo Fu

March 2013

Dissertation Committee:
Dr. Daniel R. Jeske, Chairperson
Dr. James Flegal
Dr. Gloria Gonzalez-Rivera

The Dissertation of Yingzhuo Fu is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgement

I would like to express the deepest appreciation to my advisor, Professor Daniel R. Jeske, a gracious mentor guiding me throughout the research work. Without his excellent advice, enthusiasm and persistent help, this dissertation would not have been possible. I gratefully acknowledge all members of my committee who have given their time to read this manuscript and provided valuable advice. I give thanks to Mazda Marvasti from VMware Inc., who provided the motivating problem and valuable data sources. I warmly appreciate all my friends in the Department of Statistics for their kindly help, concern and companion through my Ph.D. study. Finally, I want to thank my husband who has been a constant source of love, support and strength all these years and helped me bear the pressure throughout this endeavor.

To my beloved husband and parents

ABSTRACT OF THE DISSERTATION

Statistical Process Control Tools for Network Monitoring
Using Generalized Linear Mixed Model

by

Yingzhuo Fu

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, March 2013
Professor Daniel R. Jeske, Chairperson

Network surveillance algorithms are becoming increasingly important as the ability to monitor a wide variety of data is rapidly expanding. Traffic metrics are usually count data that display a non-stationary pattern in their mean structure. We propose to model traffic counts using a generalized linear mixed model to capture these features. We then develop three tracking statistics proposed for anomaly detection. Two of the statistics are derived variants of a Bartlett-type sequential probability ratio approach, which itself is not computationally tractable. The first of these variants is based on an approximation to the integrated likelihood while the second is based on the concept of h-likelihood. We also consider a tracking statistic that is an exponentially weighted moving average. We investigate the properties of the three tracking statistics from the point of view of false alarm rate and detection power, and compare the proposed tracking statistics with current literature. Our comparisons show that the two Bartlett-type probability ratio variants are

preferred choices as SPC tools for network surveillance. Computational aspects of the three procedures are also discussed

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The health of a network is crucial for providing the services it delivers. Intuitively, one could anticipate that monitoring various network traffic metrics would provide opportunities to detect problems and signal alarms as appropriate. Barford *et al.* (2002) mentioned that a common technique for handling network surveillance is periodically plotting data and using locally authored rules to determine if those data appear to be consistent with expectations. One of the problems with this approach is that the rules are often ad hoc and heavily rely on the expert knowledge of network operators. Another problem is the manual nature of the procedure which greatly limits the scope of monitoring that could be done. Statistical process control (SPC) methods could potentially develop automatic monitoring algorithms with some optimal properties such as minimal detection delay. However, special features of network traffic make the development of SPC tools very complicated. Brutlag (2000) points out that any statistical model for network traffic should at least capture a likely non-stationary mean structure and take into account that most streams of network traffic are discrete correlated data.

With these complex characteristics, using simple SPC tools on network traffic metrics would not be appropriate.

Motivated by our own real network traffic data traces, we derived a GLR based change point detection method in conjunction with the use of a generalized linear mixed model (GLMM) for network counts. In the current era of data acquisition capabilities, it is reasonable to expect access to in-control historical data, and we incorporate use of that type of data into our proposed solutions. In chapter 2, related work on network monitoring and classic change point methods are discussed. In chapter 3, we introduce our GLMM for describing network traffic. In chapter 4, tracking statistics and their implementation details are proposed followed by an illustrative example in a normal linear mixed model context and comparison of ILR, JLR and EWMA with Lambert and Liu's method in chapter 5. We investigate the performance of each method with respect to false alarm rate (FAR) and power in chapter 6. We conclude in chapter 7 with a discussion of implementation issues, including how a practitioner might determine if the amount of historical data on hand is sufficient to proceed with the use of our tracking statistics, and also how much computational complexity is required to implement the proposed tracking statistics. Chapter 8 provides a summary and discussion of our proposed algorithm with recommendation among the three tracking statistics. Suggestions for future work are described in chapter 9.

Chapter 2

Literature Review

2.1 Related Network Surveillance Work

Problems that utilize SPC tools are generally referred to as change point detection problems. In general, attempts to solve change point detection problems can be classified into two groups, classical or Bayesian. Among the classical approaches, the goal is to control the FAR when the process is in-control and maximize the speed of detection when there is an anomalous event. The change point is considered fixed but unknown. Many commonly implemented change-point detection methods such as the Shewhart Chart, cumulative sum (CUSUM) algorithms, exponentially weighted moving averages (EWMA) and generalized likelihood ratio (GLR) algorithms fall into the classical approaches category. These methods are studied extensively by many authors. Details of those methods can be found in references such as Montgomery (1996) and Basseville and Nikiforov (1993). Bayesian approaches consider that the change point is a random variable and aim to minimize the expected value of a loss function. For example, Shiryaev (1963) and Roberts (1966) independently presumed each post-change

observation has a positive constant cost and sought to minimize $E(N-v \mid N \geq v)$ where v is the change point and N is the stopping time.

We now discuss some of the SPC methods that have been proposed specifically for network anomaly detection. Feather *et al.* (1993) use historical data to establish in-control thresholds for the data stream. The thresholds are obtained by a scheme based on mean and variance estimates for each time point. The degree of anomalous behavior is then determined by how far a new observation is from the estimated mean in terms of the variance. Observations are scored on the basis of five anomaly values $(2, 1, 0, -1, -2)$. For instance, if it is between 3 and 6 standard deviation away it is scored an anomaly value of 1, and so on. Similarly, other metrics are scored and then a fault feature vector is encoded that represents the current behavior of the system. This vector is input into a pattern matching system to determine if it resembles a pattern that is a-priori known to be associated with a specific fault. Because daily patterns are assumed to repeat themselves, the templates of means and standard deviations are updated every day to adapt to network churns. The adequacy of this approach will rely heavily on how consistently a given fault will reproduce the same pattern, and also on the depth of the library of fault patterns. In addition, the thresholds are estimated using variance estimates that do not address potential correlations in the data.

Thottan and Ji (1998) presume the data stream can be divided into batches of data that follow piece-wise normal-theory auto-regressive models. Let $\{Y_t\}_{t=1}^{n}$ denote the sequence being monitored. A sequence of N_b (they use $N_b = 10$) observations makes up

a batch. Suppose $\{Y_t\}_{t=n-9}^{n}$ and $\{Y_t\}_{t=n-19}^{n-10}$ are observations from two adjacent batches. An AR(1) model is used to describe the observations within each batch. Then residual errors $\{\varepsilon_t\}_{t=n-9}^{n}$ and $\{\varepsilon_t\}_{t=n-19}^{n-10}$ are obtained from the two AR(1) models and are considered to follow $N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$ respectively. A hypothesis test of the form

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_1 : \sigma_1^2 \neq \sigma_2^2$$

was performed using a generalized likelihood ratio test statistics

$$\lambda = 2(N_b - 1)(\log \hat{\sigma}_1^2 - \log \hat{\sigma}_2^2)$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the maximum likelihood estimates. Using a threshold H, a change between the two batches is observed if $\lambda > H$. The threshold H is chosen experimentally using two data sets of network traffic to achieve a maximum fault detection rate with a low false alarm rate for those two data sets. They claim this threshold works well other data sets they examined. The adequacy of this approach will rely heavily on the plausibility that observations within a batch can be modeled as a stationary normal-theory process, and also on the delicate balance between having a batch size large enough to fit the hypothesized model and the incurred detection delay by waiting for that amount of data to accrue.

Lambert and Liu *et al* (2006) propose an approach which they suggest simultaneously addresses the non-stationarity, discreteness and autocorrelation of network stream. By an example from real data, they suggested network counts could reasonably be assumed to follow a negative binomial distribution and that autocorrelation in the data can be ignored

if each observation is allowed to have its own mean. The authors used iterated Hanning smoothing methods on eight weeks of minute counts to provide empirical evedence the counts behave like independent random variables conditional on their means. By iteration number 360 of Hanning smoothing, the first five estimated autocorrelation coefficients of the counts get close to zero. Even though they show that autocorrelations become negligible using iterative Hanning smoothing method, they propose a different method, interpolation of grid values, to estimate the mean values. Detailed procedure is summarized below.

Let $y_1,\ldots,y_n$ denote the observations in time order. First, the author argued that $y_1,\ldots,y_n$ follow independent Poisson distributions conditional on their means.

$$y_1,\ldots,y_n \mid \mu_1,\ldots\mu_n \sim \prod_{t=1}^{n} Poisson(\mu_t)$$

where $\mu_1,\ldots\mu_n$ are the corresponding means and variances of Poisson distributions. Second, they argue that replacing the means $\{\mu_t\}_{t=1}^{n}$ with estimates and ignoring the uncertainty in the estimates would give unrealistically short tails and hence too many false alarms. Therefore they incorporate uncertainty about the means by treating the means $\mu_1,\ldots\mu_n$ as random variables. For simplicity, they use a gamma distribution for the conditional Poisson means. Together, a conditional Poisson distribution for the count y_t and a Gamma(α_t,β_t) distribution for the conditional mean μ_t of the count imply that the marginal distribution of y_t is negative binomial. Third, the mean and variance of negative binomial distribution are estimated by interpolating the grid values. Here a grid is a similar idea to a time slot that we will introduce later. Basically, it is assumed that

observations have a similar mean and variance within each grid. However, they do not clearly mentioned how to get the grid values. Based on the meaning of grid values, we assumed the grid means and variances are calculated using all available historical data for each grid.

To simplify the terminology, a cycle of 24 hours (grids) is used. Within each hour, there are 60 (m^*) minute counts, and each of them has a different mean and variance. There are 24 (h^*) pairs of grid values for the mean and variance. Historical data and interpolation are then used to obtain the mean and variance of each observation at each time point. By design, the interpolation is unbiased in the sense that arithmetic average of the M interpolated means for an hour equals the stored grid value U_h for the hour, and the average of the M interpolated variances equals the grid value V_h . The interpolation is carried out like this. Take three consecutive hours, $(-1,0]$, $(0,1]$, $(1,2]$ and define the quadratic interpolation coefficients A, B, C by

$$\frac{1}{m^*}\int_{-1}^{0}\left(At^2+Bt+c\right)dt=\frac{1}{m^*}\left(\frac{A}{3}-\frac{B}{2}+C\right)=U_{-1}$$

$$\frac{1}{m^*}\int_{0}^{1}\left(At^2+Bt+c\right)dt=\frac{1}{m^*}\left(\frac{A}{3}-\frac{B}{2}+C\right)=U_{0}$$

$$\frac{1}{m^*}\int_{1}^{2}\left(At^2+Bt+c\right)dt=\frac{1}{m^*}\left(\frac{A}{3}-\frac{B}{2}+C\right)=U_{1}.$$

Solve for A, B, C gives,

$$A=\frac{m^*\left(U_{-1}-2U_0+U_1\right)}{2}$$

$$B = m^*\left(U_0 - U_{-1}\right)$$

$$C = \frac{m^*\left(2U_{-1} + 5U_0 - U_1\right)}{6}\ .$$

Let $q = (m-1)/m^*$ and $r = m/m^*$ Then the interpolated mean $\hat{\mu}_{h,m}$ at time t corresponding to minute m of hour h is,

$$\hat{\mu}_{h,m} = \int_q^r (At^2 + Bt + c)dt$$

$$= \frac{A}{3m^*}(r^2 + rq + q^2) + \frac{B}{2m^*}(r+q) + \frac{C}{m^*}.$$

Using stored variance grid values V_{-1}, V_0, V_1 in place of U_{-1}, U_0, U_1 gives an estimated variance $\hat{\sigma}^2_{h,m}$.

After establishing the reference distribution for each observation, the tracking statistic and threshold is constructed as follow. Let us suppose that large counts suggest abnormal network behavior. Let $F_t(.)$ be the marginal cumulative distribution function (c.d.f.) of y_t. For each count y_t, calculate $F_t(y_t)$, and then define a normal score $Z_t = \Phi^{-1}(F_t(y_t))$. An EWMA tracking statistic based on the normal scores is proposed as follows

$$S_t = (1-w)S_{t-1} + wZ_t$$

for a weight of w. An alarm is raised if $S_t > L\sigma_w$, where $\sigma_w = w/(2-w)$. The paper use $w = 0.25,\ L = 3.68$ which gives an average run length of 1 false alarm per $10,000$ counts, approximately one false alarm per week for a metric measured each minute. It is very easy to develop a threshold for this tracking statistic because no matter how the reference distributions change, if the motivating assumptions hold, the inputs of the EWMA

tracking statistic are always conditionally independently distributed standardized normal random variables.

The main problem with Lambert and Liu's method is centered on the estimated reference distributions. Even the reference distributions are updated at the end of each cycle to keep up with the network churn, the updating mechanism in the approach inherently lags behind the real-time monitoring period and our simulation results in Chapter 4 shows this can severely inflate the conditional FAR.

Jeske *et al.* (2009) propose a non-parametric CUSUM approach. The authors acknowledge that network data are correlated and exhibit non-stationarity in the mean structure. They choose a certain length of time interval as a time slot, such that the counts within each time interval can be considered identically distributed. It is assumed autocorrelations can be removed within a timeslot through the use of a suitable application dependent transformation. The timeslot structure is assumed to repeat itself as cycles in the network stream, and hence the distribution for each interval could be estimated by collecting enough historical cycles of counts. A CUSUM tracking statistic based on the empirical probability integral transformation is used for change point detection. In this way, the procedure becomes asymptotically distribution free. The tracking statistic is built as follows. First, K cycles within a sliding window of historical data need to be collected $H = \{Y_{ijk}, i = 1,\ldots m, j = 1,\ldots n_i, k = 1,\ldots, K\}$ where n_i denote the number of observation within time slot i. The author's premise is that under normal operating situation, after the transformation to remove autocorrelation, the data within timeslots independently follow heterogeneous distribution function $\{F_j\}_{j=1}^{m}$. The

transformed CUSUM (TC) tracking statistics is defined as (if only consider large observations are abnormal)

$$T_n = \max\{0, T_{n-1} + \hat{F}_{\tau_n}(Y_n) - \alpha$$

where α is a suitably chose reference value with $\alpha \in (0,1)$. Consider an arbitrary observation Y_n which maps to timeslot τ_n. Here $\hat{F}_{\tau_n}(Y_n)$ is the empirical c.d.f of Y_n and is estimated in the following way. Let $\{X_{\tau_n(k)}\}_{k=1}^{n_{\tau_n}}$ denote the ordered historical data associated with timeslot τ_n. Define $X_{\tau_n(0)} = 0$ and $X_{\tau_n(n_{\tau_n}+1)} = \infty$, it follows that

$$P_0 = \left(\hat{F}_{\tau_i}(Y_i) = t / n_{\tau_i} \mid H\right) = F_{\tau_i}\left(X_{\tau_i(t+1)}\right) - F_{\tau_i}\left(X_{\tau_i(t)}\right), \quad t = 0,1,\ldots n_{\tau_i}.$$

If the depth of the historical data as measured by n_{τ_n} is sufficiently large, the distribution can be approximated by

$$P_0 = \left(\hat{F}_{\tau_i}(Y_i) = t / n_{\tau_i} \mid H\right) \approx 1/\left(n_{\tau_i} + 1\right), \ \ t = 0,\ldots n_{\tau_i}.$$

Then the random variable $\hat{F}_{\tau_n}(Y_n)$ is an approximately discrete uniform variable and the transformed CUSUM tracking statistics is asymptotically distribution free.

Potential limitations in this approach are that the transformation approach for dealing with autocorrelations leaves edge effects in the correlation structure and the assumption of homogeneity within pre-defined timeslots may not always be tenable.

2.2 Classical CUSUM Algorithm for Statistical Process Control

Classical statistical process control methods use sequence of independent observations $\{Y_i\}_{i=1}^{\infty}$ to detect a departure from in-control situation. When the process is in-control, observations are considered independent identically distributed (i.i.d.) random variables with known in-control density function f_0. Subsequent to the change, observations are considered i.i.d. from an out-of control distribution with density function f_1. Algorithms are designed to raise a signal at the earliest possible time after enough evidence in the observations has accumulated to confidently declare that a change from f_0 to f_1 has occurred. The algorithm by Page (1954) uses the tracking statistic

$$S_n = \sum_{j=1}^{n} \log \frac{f_1(y_j)}{f_0(y_j)} - \min_{1 \le k \le n} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)} \right\}.$$

An alarm is raised the first time the tracking statistic S_n exceeds a threshold, H, which is chosen to achieve a specified average run length (ARL) under the null. Formally, the stopping time t_a is expressed as $t_a = \inf\{ n \ge 1 : t_n \ge H \}$. This algorithm is known to be optimal [see Lorden (1971)] in the sense that among all algorithms with a specified average run length under f_0 , say ARL_0, it achieves the asymptotic lower limit of the expected stopping time. More simply said, it minimizes the average run length under f_1, say ARL_1. An alternative, computationally useful, recursive form of the tracking statistic is $S_n = \max\left(0 , S_{n-1} + \log[f_1(y_n) / f_0(y_n)] \right)$. Detailed derivation of this recursive form is

provided in section 2.2.2. After the algorithm alarms, the most recent time at which the tracking statistic was zero is an estimate of the time the change occurred.

Most commonly, Page's CUSUM algorithm is presented in the context where the in-control distribution is $N(\mu_0, \sigma^2)$ and the out-of-control distribution is $N(\mu_0 + d\sigma, \sigma^2)$, with $d > 0$. Here, the targeted shift is expressed as a multiple of the process standard deviation and the constant d is referred to as the shift, expressed as units of the standard deviation. The tracking statistic simplifies to $S_n = \max[0, S_{n-1} + (X_n - \mu_0) - D]$, where $D = d\sigma / 2$, and the increment to the CUSUM can be seen as the deviation of null-centered observations from one-half of the targeted shift. A similar representation can be derived for the case where $d < 0$.

In order to use Page's CUSUM algorithm, the pair of functions (f_0, f_1) must be completely specified. Once given, the threshold H can be obtained through a search procedure where null sample paths for T_n are simulated and used to estimate ARL_0 for alternative choices of H.

2.2.1 Classic CUSUM in a Recursion Form

We show the Page CUSUM

$$S_n = \sum_{j=1}^{n} \log \frac{f_1(y_j)}{f_0(y_j)} - \min_{1 \le k \le n} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)} \right\}$$

has a convenient recursion form

$$S_n = \max \left\{ 0, S_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)} \right\}$$

with $S_0 = 0$. We prove this in two different ways.

Proof 1: Let S_n^* denote the proposed recursion from of S_n. We must show that $S_n = S_n^*$ for all $n \geq 1$. First consider the case $n = 1$,

$$S_1 = \log \frac{f_1(y_1)}{f_0(y_1)} - \min\left\{0, \log \frac{f_1(y_1)}{f_0(y_1)}\right\}.$$

If $\log \frac{f_1(y_1)}{f_0(y_1)} < 0$, then $S_1 = \log \frac{f_1(y_1)}{f_0(y_1)} - \log \frac{f_1(y_1)}{f_0(y_1)} = 0$, $S_1^* = 0$, so we have $S_1 = S_1^*$.

If $\log \frac{f_1(y_1)}{f_0(y_1)} \geq 0$, then $S_1 = \log \frac{f_1(y_1)}{f_0(y_1)}$, and again $S_1 = S_1^*$.

Therefore, in the case $n = 1$, it's clear that $S_1 = S_1^*$. Suppose $S_1 = S_1^*$ for $n = m-1, m \geq 2$, it will be shown as below that $S_n = S_n^*$ still holds for $n = m$.

$$S_m^* = \max\left\{0, S_m^* + \log \frac{f_1(y_m)}{f_0(y_m)}\right\} = 0$$

$$S_m = \sum_{j=1}^{m} \log \frac{f_1(y_j)}{f_0(y_j)} - \min_{1 \leq k \leq m}\left\{0, \sum_{j=1}^{m} \log \frac{f_1(y_j)}{f_0(y_j)}\right\}$$

$$= \sum_{j=1}^{m-1} \log \frac{f_1(y_j)}{f_0(y_j)} - \min_{1 \leq k \leq m-1}\left\{0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)}, \sum_{j=1}^{m} \log \frac{f_1(y_m)}{f_0(y_m)}\right\} + \log \frac{f_1(y_m)}{f_0(y_m)}.$$

If $S_{m-1}^* + \log \frac{f_1(y_m)}{f_0(y_m)} < 0$, then

$$S_m^* = \max\left\{0, S_{m-1}^* + \log \frac{f_1(y_m)}{f_0(y_m)}\right\} = 0.$$

Since we assume that, $S_n = S_n^*$ holds for $n = m-1$,then $S_{m-1}^* + \log\frac{f_1(y_m)}{f_0(y_m)} < 0$ is equivalent to

$$S_{m-1} + \log\frac{f_1(y_m)}{f_0(y_m)} < 0$$

i.e. $$\sum_{j=1}^{m-1}\log\frac{f_1(y_j)}{f_0(y_j)} - \min_{1\le k\le m-1}\left\{0,\ \sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)}\right\} + \log\frac{f_1(y_m)}{f_0(y_m)} < 0$$

i.e. $$\sum_{j=1}^{m}\log\frac{f_1(y_j)}{f_0(y_j)} < \min_{1\le k\le m-1}\left\{0,\ \sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)}\right\}.$$

Under this condition,

$$\min_{1\le k\le m-1}\left\{0,\ \sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)},\ \sum_{j=1}^{m}\log\frac{f_1(y_m)}{f_0(y_m)}\right\} = \sum_{j=1}^{m}\log\frac{f_1(y_m)}{f_0(y_m)}.$$

Then we have

$$S_m = \sum_{j=1}^{m-1}\log\frac{f_1(y_j)}{f_0(y_j)} - \sum_{j=1}^{m}\log\frac{f_1(y_m)}{f_0(y_m)} + \log\frac{f_1(y_m)}{f_0(y_m)} = 0.$$

So when $S_{m-1} + \log\frac{f_1(y_m)}{f_0(y_m)} < 0$,

$$S_m = S_m^* = 0.$$

If $S_m^* + \log\frac{f_1(y_m)}{f_0(y_m)} \ge 0$, then

$$S_m^* = \max\left\{0, S_{m-1}^* + \log\frac{f_1(y_m)}{f_0(y_m)}\right\}$$

$$= S_m^* + \log\frac{f_1(y_m)}{f_0(y_m)}.$$

Similarly to the previous condition, we know $S_{m-1}^{*}+\log\frac{f_1(y_m)}{f_0(y_m)}\geq 0$ is equivalent to

$$\sum_{j=1}^{m}\log\frac{f_1(y_j)}{f_0(y_j)}>\min_{1\leq k\leq m-1}\left\{0,\ \sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)}\right\}.$$

Then ,

$$\min_{1\leq k\leq m-1}\left\{0,\ \sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)},\ \sum_{j=1}^{m}\log\frac{f_1(y_m)}{f_0(y_m)}\right\}=\min_{1\leq k\leq m-1}\left\{0,\ \sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)}\right\}.$$

So we have

$$\begin{aligned}S_m&=\sum_{j=1}^{m-1}\log\frac{f_1(y_j)}{f_0(y_j)}-\min_{1\leq k\leq m-1}\left\{0,\ \sum_{j=1}^{m-1}\log\frac{f_1(y_j)}{f_0(y_j)}\right\}+\log\frac{f_1(y_m)}{f_0(y_m)}\\&=S_{m-1}+\log\frac{f_1(y_m)}{f_0(y_m)}\\&=S_{m-1}^{*}+\log\frac{f_1(y_m)}{f_0(y_m)}\\&=S_m^{*}.\end{aligned}$$

Therefore, when $S_n=S_n^{*}$ holds for $n=m-1,m\geq 2$, we have shown that $S_n=S_n^{*}$ still holds for $n=m$. With $S_n=S_n^{*}$ holds for $n=1$, we have $S_n=S_n^{*}$ holds for all n by induction.

Proof 2: Another way of proving the results is now presented.

$$\begin{aligned}S_n&=\sum_{j=1}^{n-1}\log\frac{f_1(y_j)}{f_0(y_j)}+\frac{f_1(y_n)}{f_0(y_n)}-\min_{1\leq k\leq n}\left\{0,\sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)}\right\}\\&=\sum_{j=1}^{n-1}\log\frac{f_1(y_j)}{f_0(y_j)}+\frac{f_1(y_n)}{f_0(y_n)}-\min_{1\leq k\leq n-1}\left\{0,\sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)},\sum_{j=1}^{n}\log\frac{f_1(y_j)}{f_0(y_j)}\right\}\end{aligned}$$

$$= \begin{cases} = \sum_{j=1}^{n-1} \log \frac{f_1(y_j)}{f_0(y_j)} + \frac{f_1(y_n)}{f_0(y_n)} - \min_{1 \le k \le n-1} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)} \right\}; & \text{if } \min_{1 \le k \le n-1} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)} \right\} < \sum_{j=1}^{n} \log \frac{f_1(y_j)}{f_0(y_j)} \\ 0 \quad ; \quad \text{otherwise} \end{cases}.$$

But the side condition is saying

$$\sum_{j=1}^{n} \log \frac{f_1(y_j)}{f_0(y_j)} - \min_{1 \le k \le n-1} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)} \right\} > 0$$

which is equivalent to

$$S_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)} > 0$$

then we have

$$S_n = \begin{cases} S_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)}; & \text{if } S_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)} > 0 \\ 0 \quad ; \quad \text{otherwise} \end{cases}$$

$$= \max \left(0, S_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)} \right).$$

2.2.2 Classic CUSUM from a Likelihood Ratio Perspective

The previously described setting could be viewed as the following hypothesis test setting

$$\begin{aligned} H_0 &: Y_1, Y_2, ..., Y_n \sim f_0 \\ H_1 &: Y_1, Y_2, ..., Y_{k-1} \sim f_0 \\ & \quad Y_k, Y_{k+1}, ..., Y_n \sim f_1 \end{aligned}$$

Under the null hypothesis, there is no change in the underlying distribution for all the observations. In the alternative hypothesis, the underlying distribution f_0 changes to f_1 at a particular unknown time point k, $k = 0, 1, ..., n+1$. If a likelihood ratio test is used, H_0

will be favored when the likelihood ratio $\frac{L_0}{L_{0\cap 1}}$ is too small, i.e. when $\log\frac{L_{0\cap 1}}{L_0}$ is too big, where

$$L_0 = \prod_{j=1}^{n} f_0(y_j)$$

$$L_{0\cap 1} = \max_{1\le k\le n+1}\left\{\prod_{j=1}^{k-1} f_0(y_j)\prod_{j=k}^{n} f_1(y_j)\right\}, \quad s.t.\ \prod_{j=n+1}^{n} f_1(y_j)=1 \text{ and } \prod_{j=1}^{0} f_0(y_j)=1.$$

In the following we will show that $\log\frac{L_{0\cap 1}}{L_0}$ is equivalent to the CUSUM defined by Page (1954) ,

$$S_n = \sum_{j=1}^{n}\log\frac{f_1(y_j)}{f_0(y_j)} - \min_{1\le k\le n}\left\{0, \sum_{j=1}^{k}\log\frac{f_1(y_j)}{f_0(y_j)}\right\}.$$

Proof:

$$\log\frac{L_{0\cap 1}}{L_0} = \log\frac{\max_{1\le k\le n+1}\left\{\prod_{j=1}^{k-1} f_0(y_j)\prod_{j=k}^{n} f_1(y_j)\right\}}{\prod_{j=1}^{n} f_0(y_j)}$$

$$= \log\frac{\max_{1\le k\le n+1}\left\{\prod_{j=1}^{k-1} f_0(y_j)\prod_{j=1}^{n} f_1(y_j) / \prod_{j=1}^{k-1} f_1(y_j)\right\}}{\prod_{j=1}^{n} f_0(y_j)}$$

$$= \log\frac{\prod_{j=1}^{n} f_1(y_j)\max_{1\le k\le n+1}\left\{\prod_{j=1}^{k-1} f_0(y_j) / \prod_{j=1}^{k-1} f_1(y_j)\right\}}{\prod_{j=1}^{n} f_0(y_j)}$$

$$= \sum_{j=1}^{n}\log\frac{f_1(y_j)}{f_0(y_j)} + \max_{1\le k\le n+1}\left\{\sum_{j=1}^{k-1}\log\frac{f_0(y_j)}{f_1(y_j)}\right\}$$

$$= \sum_{j=1}^{n} \log \frac{f_1(y_j)}{f_0(y_j)} + \max_{2 \le k \le n+1} \left\{ 0, \sum_{j=1}^{k-1} \log \frac{f_0(y_j)}{f_1(y_j)} \right\}$$

$$= \sum_{j=1}^{n} \log \frac{f_1(y_j)}{f_0(y_j)} + \max_{1 \le k \le n} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)} \right\}$$

$$= \sum_{j=1}^{n} \log \frac{f_1(y_j)}{f_0(y_j)} - \min_{1 \le k \le n} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_1(y_j)}{f_0(y_j)} \right\}$$

$$= S_n .$$

2.2.3 Classical CUSUM Extension

We note here in a more generally situation, even if $Y_1, Y_2, \ldots Y_n$ follows different distributions respectively $F_1, F_2 \ldots F_n$ with density functions $f_{Y_1}^0, f_{Y_2}^0, \ldots f_{Y_n}^0$, as long as the inputs $Y_1, Y_2, \ldots Y_n$ are independent, in order to test the following hypotheses,

$$H_0 : Y_1, Y_2, ..., Y_n \sim f_{Y_1}^0, f_{Y_2}^0, \ldots f_{Y_n}^0$$
$$H_1 : Y_1, Y_2, ..., Y_{k-1} \sim f_{Y_1}^0, f_{Y_2}^0, \ldots f_{Y_{k-1}}^0$$
$$Y_k, Y_{k+1}, ..., Y_n \sim f_{Y_k}^1, f_{Y_{k+1}}^1, \ldots f_{Y_n}^1$$

the CUSUM based on LR perspective

$$L_n = \sum_{j=1}^{n} \log \frac{f_{Y_j}^1(y_j)}{f_{Y_j}^0(y_j)} - \min_{1 \le k \le n} \left\{ 0, \sum_{j=1}^{k} \log \frac{f_{Y_j}^1(y_j)}{f_{Y_j}^0(y_j)} \right\}$$

will have a recursion form as below,

$$S_n = \max \left\{ 0, S_{n-1} + \log \frac{f_{Y_j}^1(y_j)}{f_{Y_j}^0(y_j)} \right\}.$$

For example, a sequence of counts can be considered independently following a negative binomial distribution with its own mean. A possible out-of-control situation is that the

means shift up from some time point. As long as the counts could still be assumed to be independent, this extended CUSUM could be used for a change-point-detection in this sequence.

2.2.4 Classical CUSUM Interpreted as a Repeated SPRT

Besides Page's original derivation of the classic CUSUM, there is another thread of theoretical work to provide insight on the CUSUM algorithm, namely, the sequential probability ratio test (SPRT) by Wald (1945).

Summary of Wald's SPRT

Wald's SPRT uses data sequentially to test a simple null hypothesis H_0 versus a simple alternative hypothesis H_1. The data are independently observed in sequential order with known density function f_0 under the null hypothesis and a known density function f_1 under the alternative hypothesis. A probability ratio is calculated sequentially at each time a new observation is available

$$\Lambda_n = \prod_{i=1}^{n} \frac{f_1(y_i)}{f_0(y_i)}.$$

When Λ_n become equal or less than a predetermined cutoff point A, the null hypothesis H_0 will be accepted. When Λ_n become equal or greater than a predetermined cutoff point B, the alternative hypothesis H_1 will be accepted. Otherwise, data collection continues until a decision between H_0 and H_1 is made. Usually, it is easier to work with $\log \Lambda_n$,

$$\log \Lambda_n = \sum_{i=1}^{n} \log \frac{f_1(y_i)}{f_0(y_i)}.$$

Let $Z_i = \log \frac{f_1(y_i)}{f_0(y_i)}$. Then $\log \Lambda_n = \sum_{i=1}^{n} Z_i$, and we could consider that the test is based on a cumulative sum of Z_i. If the null hypothesis H_0 is true, the expected value of Z_i is negative. As observations accumulated, $\log \Lambda_n$ tends to drift downward and eventually will cross the lower boundary A. In the opposite situation, if the alternative hypothesis H_1 is true, $\log \Lambda_n$ tends to drift upwardsand eventually cross the upper boundary B. The upper and lower boundary values are chosen according to the desired Type I error α and Type II error β. Approximately values are,

$$A = \log \frac{\beta}{1-\alpha}, \quad B = \log \frac{1-\beta}{\alpha}.$$

Wald and Wolfowitz (1948) proved that, the SPRT minimize the expected number of observations required before a decision is a made when a change from f_0 to f_1 has occurred. Wald's SPRT is the optimal test for simple versus simple hypotheses test.

CUSUM and Wald's SPRT

As shown in section 2.2.2, Page's CUSUM has the recursion form

$$S_n = \max\left\{0, S_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)}\right\}.$$

Page recognized that this CUSUM could be viewed as a repeated Wald SPRT with a slight difference from the original Wald SPRT. The difference is Wald SPRT will stop when a decision is made. But for the purpose of change-point detection in a monitoring process, we are not interested in ultimately accepting H_0, i.e. deciding the process is in

control. We want to stop and take a look at the process when there is evidence of out of control. Thus we want to repeat Wald's test when the lower bound is encountered.

The choice of lower and upper bound for each Wald test is also different from the original Wald test. In change point detection, the average run length during an in-control period is of more interest than Type I and Type II error for one hypothesis test. Intuitively Page suggested the optimal lower bound for the Wald test statistic should be zero when it is used for change point monitoring. This result was formally proved by Shiryaev(1961), Lorden(1971), Moustakides(1986), Ritov (1990). Once the SPRT statistic becomes negative, it is reset to zero and a new Wald SPRT begins. This means that previous observations will be ignored and only new observations going forward are used to calculate the new SPRT statistic. With this definition, a tracking statistic for repeated use of Wald SPRT can be written in the following form of g_n

$$g_n = \max\{0, \quad \log \Lambda_n\} = \max\left\{0, \quad g_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)}\right\}.$$

It can be seen this is the same as the recursion form of the CUSUM

$$S_n = \max\left\{0, S_{n-1} + \log \frac{f_1(y_n)}{f_0(y_n)}\right\}.$$

The upper boundary of the repeated SPRT is computed to control the average run length under H_0

2.3 GLR Algorithm for Change-Point Detection

Another intensively studied area for change point detection is the generalized likelihood ratio (GLR) approach. Using the setting for change-point detection as before, finite sequence of observations $\{Y_t\}_{t=1}^{n}$ are either already in hand, or will be sequentially collected. Unlike CUSUM, GLR detection algorithms generally do not specify an out-of-control distribution. Instead, they seek to identify any type of change that occurs. Hinkley (1970), Hawkins (1977), Worsely (1979), Siegmund (1986), James *et al.* (1988) discussed this problem at length in the context of normal models.

The GLR algorithm is defined within a parametric setting. Let $f(y;\theta,\xi)$ denote a class of distributions for the observations, indexed by a (possibly vector-valued) informative parameter θ which is the parameter of interest and a (possibly vector-valued) nuisance parameter ξ. The null hypothesis H_0 is that $\theta=\theta_0$ for the entire sequence, while the alternative hypothesis H_1 is that $\theta=\theta_0$ for $1\le t\le k$ and then $\theta=\theta_1$ for $k+1\le t\le n$. Here, $k\in\{1,2,\ldots,n-1\}$ is unknown, and the most typical case also has θ_1 unknown.

Three cases are usually treated, depending on whether the parameters of interest θ_0, θ_1 and nuisance parameter ξ are considered known. In case 1, when θ_0 is known and θ_1 is unknown, the generalized likelihood ratio test (GLRT) statistic of H_0 vs. H_1 of can be obtain as follow,

$$g_n = \max_{1\le k\le n} \log \frac{\sup_{\theta_1} \prod_{i=1}^{k} f_0(y_i;\theta_0,\xi) \prod_{i=k+1}^{n} f_1(y_i;\theta_1,\xi)}{\prod_{i=1}^{n} f_1(y_i;\theta_0,\xi)}$$

$$= \max_{1\le k\le n} \sup_{\theta_1} \sum\nolimits_{i=k}^{n} \log\left[\frac{f_1(y_i;\theta_1)}{f_0(y_i;\theta_0)}\right].$$

The critical point for g_N can be more simply obtained through the use of Monte-Carlo methods, simulating values of g_N under H_0.

In case 2 when both the θ_0 and θ_1 are unknown, the LRT statistic becomes

$$g_n^* = \max_{1\le k\le n} \sup_{\theta_0,\theta_1} \sum\nolimits_{i=k}^{n} \log\left[\frac{f_1(y_i;\theta_1,\xi)}{f_0(y_i;\theta_0,\xi)}\right].$$

In case 3, with the presence of nuisance parameters, GLR approach is based upon the maximization of the likelihood ratio with respect to the all unknown parameters. Suppose the hypotheses testing problem is composite for the parameter of interest θ :

$$H_o : \{\theta = \Theta_0, \xi \in \Xi_0\} \text{ versus } H_1 : \{\theta = \Theta_1, \xi \in \Xi_1\} .$$

To solve this hypotheses testing problem, the GLR algorithm is

$$g_n^{**} = \max_{1\le k\le n} \log\left[\frac{\sup_{\theta_0\in\Theta_0,\xi\in\Xi_0} \prod_{i=k}^{n} f_1(y_i;\theta_0,\xi)}{\sup_{\theta_1\in\Theta_1,\xi\in\Xi_1} \prod_{i=k}^{n} f_0(y_i;\theta_1,\xi)}\right].$$

The critical point for g_n^* and g_n^{**} cannot be simulated using Monte-Carlo methods due to unknown parameters under null hypothesis. Under certain distribution assumption of the monitored observations, distribution of tracking statistics in case 2 and case 3 can be determined analytically when. Otherwise, asymptotic distributions need to be obtained to

set threshold for GLR tracking statistics. We will show an example for case 3 under the normal distribution context. Let $\{Y_t\}_{t=1}^{n}$ be a sequence of independent normal random variables with mean $\mu_1,...,\mu_n$ respectively and a common unknown variance σ^2. Here the mean value is the parameter of interest and variance is treated as a nuisance parameter. The hypotheses being tested for a change point detection problem is set up as follows,

$$H_0 : \mu_1 = \mu_2 = ... = \mu_n = \mu \text{ versus } H_1 : \mu_1 = \mu_2 = ... = \mu_k \neq \mu_{k+1} = ... = \mu_n$$

where k is the unknown change point. The means before and after change are both unknown. The likelihood function under H_0 is

$$L_0(\mu, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n}(y_i - \mu)^2 / (2\sigma^2)\right)$$

and the MLE of μ and σ^2 is

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

Under the alternative hypothesis, the likelihood function is

$$L_1(\mu_1, \mu_n, \sigma_1^2) = (2\pi)^{-n/2}(\sigma_1^2)^{-n/2} \exp\left\{-\left[\sum_{i=1}^{k}(y_i - \mu_1)^2 + \sum_{i=k+1}^{n}(y_i - \mu_n)^2\right] / (2\sigma_1^2)\right\}$$

and the MLE of μ_1, μ_n and σ_1^2 are

$$\hat{\mu}_1 = \bar{y}_k = \frac{1}{k}\sum_{i=1}^{k} y_i$$

$$\hat{\mu}_n = \overline{y}_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n} y_i$$

$$\hat{\sigma}_1^2 = \frac{1}{n}\left(\sum_{i=1}^{k}(y_i - \overline{y}_k)^2 + \sum_{i=k+1}^{n}(y_i - \overline{y}_{n-k})^2\right).$$

It is then easy to show the likelihood ratio test is based on

$$g_n^{**} = \max_{1\le k\le n-1}\left(\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sum_{i=1}^{k}(y_i - \overline{y}_k)^2 + \sum_{i=k+1}^{n}(y_i - \overline{y}_{n-k})^2}\right)^{1/2}.$$

Let $V^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2$ denote the total sum of squares between all the observations and $V_k^2 = \sum_{i=1}^{k}(y_i - \overline{y}_k)^2 + \sum_{i=k+1}^{n}(y_i - \overline{y}_{n-k})^2$ be the within-group sum of squares of observations split by change point k , simple algebra will show

$$V_k^2 = \sum_{i=1}^{k}(y_i - \overline{y}_k)^2 + \sum_{i=k+1}^{n}(y_i - \overline{y}_{n-k})^2$$

$$= \sum_{i=1}^{k}(y_i - \overline{y} + \overline{y} - \overline{y}_k)^2 + \sum_{i=k+1}^{n}(y_i - \overline{y} + \overline{y} - \overline{y}_{n-k})^2$$

$$= \sum_{i=1}^{k}\left(y_i - \overline{y} - \frac{1}{k}\sum_{i=1}^{k}(y_i - \overline{y})\right)^2 + \sum_{i=k+1}^{n}\left(y_i - \overline{y} - \frac{1}{n-k}\sum_{i=k+1}^{n}(y_i - \overline{y})\right)^2$$

$$= V^2 - \frac{1}{k}\left[\sum_{i=1}^{k}(y_i - \overline{y})\right]^2 - \frac{1}{n-k}\left[\sum_{i=k+1}^{n}(y_i - \overline{y})\right]^2.$$

Since it is easy to show that $0 = \sum_{i=1}^{n}(y_i - \overline{y}) = \sum_{i=1}^{k}(y_i - \overline{y}) + \sum_{i=k+1}^{n}(y_i - \overline{y})$

We can write

$$V_k^2 = V^2 - \frac{n}{k(n-k)}\left[\sum_{i=1}^{k}(y_i - \bar{y})\right]^2$$

$$= V^2 - \frac{1}{kn(n-k)}\left[nk\bar{y}_k - k(k\bar{y}_k + (n-k)\bar{y}_{n-k}\right]^2$$

$$= V^2 - \frac{k(n-k)}{n}\left[\bar{y}_k - \bar{y}_{n-k}\right]^2$$

$$\triangleq V^2 - Q_k^2 \ .$$

Then the generalized likelihood ratio test statistics can be written as

$$g_n^{**} = \max_{1\le k\le n-1}\left(1 + \frac{|Q_k|}{V_k}\right) .$$

Worsley (1979) treat

$$g_n^{W**} = \max_{1\le k\le n-1}(n-2)^{1/2}\frac{|Q_k|}{V_k}$$

as the generalized likelihood ratio test statistic for H_0 versus H_1. Note that within normal context, $Q_k \sim N(0,\sigma^2)$ under H_0 and V_k^2/σ^2 follows a χ^2 distribution with $n-2$ degrees of freedom. With Q_k independent of V_k , $(n-2)^{1/2}Q_k/V_k$ has a t distribution with $n-2$ degrees of freedom. Let H^w denote the threshold for the test statistic g_n^{W**}. Using a Bonferroni inequality,

$$P\left(\max_{1\le k\le n-1}(n-2)^{1/2}\frac{|Q_k|}{V_k}>H^w\right)=P\left(\bigcup_{k=1}^{n-1}(n-2)^{1/2}\frac{|Q_k|}{V_k}>H^w\right)\le\sum_{k=1}^{n-1}P\left((n-2)^{1/2}\frac{|Q_k|}{V_k}>H^w\right)=\alpha$$

a conservative threshold H^w for level α test can be obtained through the upper $\frac{\alpha}{2(n-1)}$ quartile of the t distribution with $n-2$ degrees of freedom. Worsley (1979) illustrated the exact values and Bonferroni approximations are very close for small n and α when $n<10$ and $\alpha\le 0.1$. In other cases, the threshold needs to be simulated under the null.

As we've shown here, even in a simple example with normal observations, using GLR tracking statistics and obtaining the threshold are not that straight forward. Chen and Gupta (2000) summarize analytical derivations for different normal distribution contexts. They also present derivations for multivariate normal settings, regression settings, gamma distribution settings, and Poisson distribution settings.

2.4 H-likelihood

Random effects are often used in models that describe temporal or special correlated data. For instance, correlated network data, clustered epidemiology data, longitudinal data in economics and survival analysis frequently use random-effect models. To save intensive computation effort for inference of parameters in this kind of models, Lee and Nelder (1996) proposed to use an h-likelihood in the form

$$L_h=L(\underset{\sim}{y},\underset{\sim}{v}\mid\underset{\sim}{\beta},\underset{\sim}{\phi},\underset{\sim}{\theta})=\log f(\underset{\sim}{y}\mid\underset{\sim}{v};\underset{\sim}{\beta},\underset{\sim}{\phi})+\log f(\underset{\sim}{v};\underset{\sim}{\theta})$$

where $f(\underset{\sim}{y}|\underset{\sim}{v};\underset{\sim}{\beta},\underset{\sim}{\phi})$ denote the conditional density of $\underset{\sim}{y}$ given random effects $\underset{\sim}{v}$, $f(\underset{\sim}{v};\underset{\sim}{\theta})$ denotes the density function of $\underset{\sim}{v}$, $\underset{\sim}{\beta}$ is the vector of fixed effects, $\underset{\sim}{\phi}$ is the dispersion parameter and $\underset{\sim}{\theta}$ denotes the parameters for random effect $\underset{\sim}{v}$.

One of the attractive features of h-likelihood, compared to marginal likelihood, is to obtain parameter estimates for both the fixed effects and random effect without a computation demanding procedure. Even though inference on fixed effect can be made using marginal likelihood that integrates out the random effects, it is computationally too intensive.

Chapter 3

GLMM for Non-stationary Correlated Counts

3.1 Real Data of Motivation

Figure 1 shows two weeks of network data that are 5-minute counts of the number of live users on a particular network server. The data is provided by the VMware Company and it shows a clear weekly pattern.

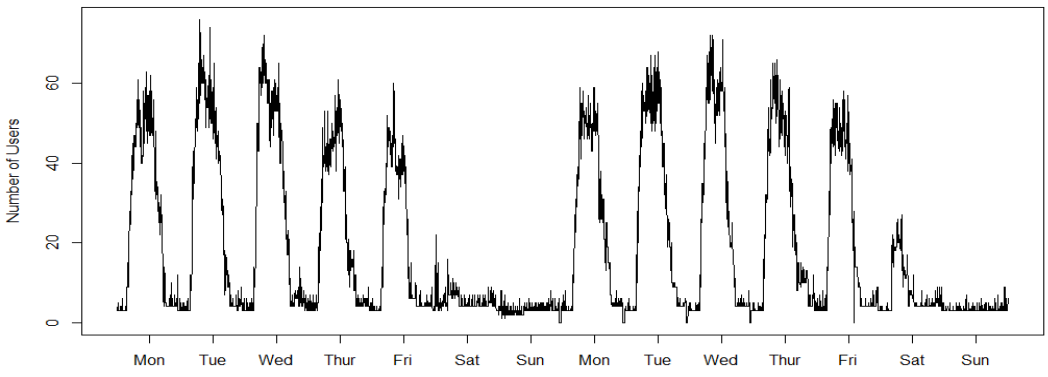


Figure 1. Two weeks of 5-minute Counts for Number of Users with Strong Daily Pattern

From Monday to Friday, the counts have a similar pattern with slight difference in magnitude, while Saturdays and Sundays have different patterns and dramatically

reduced magnitudes. Other types of networks will likely have similar patterns. Motivated by these observations, we propose to use a flexible GLMM model to describe the data stream.

3.2 Notation and Generalized Linear Mixed Model

To model network counts, we use a GLMM that captures a daily/weekly pattern through a non-stationary mean structure and autocorrelations through the random effects that appear in the associated link function. We emphasize that although we suggest using a negative binomial GLMM, the tracking statistics we subsequently propose are applicable to a larger class of GLMMs that could apply to a variety of other types of applications.

We suppose the data stream can be organized as cycles that exhibit a timeslot structure. Let m denote the number of timeslots and let $\{n_i\}_{i=1}^{m}$ denote the number of observations within the timeslots. For example, if the data shows a weekly pattern as in Figure 1, then cycles correspond to weeks and the 168 hours during the week could be viewed as timeslots. If data are available every minute, then each n_i would be a maximum of 60, depending on what (if any) observations might be missing. Alternatively, adjacent hours might be grouped together to form a smaller number of timeslots and larger n_i values.

Let Y_{ij} be the j-th observation from the i-th timeslot. Our negative binomial GLMM model is constructed as follow. Let β_{ij} denote a fixed effect for time slot i and

observation j and let S_i denote a common random effect for all of the observations in the i-th time slot. Conditional on all the random effects $\underset{\sim}{S} = (S_i, \cdots S_m)'$, the counts are independently distributed as negative binomial with mean μ_{ij} and dispersion parameter κ. For simplicity, we have assumed the same dispersion parameter for each timeslot, though that is not a critical assumption. That is, the conditional probability function of Y_{ij} is given by $f_{Y_{ij}}(y_{ij} \mid \mu_{ij}, \kappa) = \frac{\Gamma(y_{ij}+\kappa)}{\Gamma(y_{ij}+1)\Gamma(\kappa)}\left(\frac{\mu_{ij}}{\mu_{ij}+\kappa}\right)^{y_{ij}}\left(\frac{\kappa}{\mu_{ij}+\kappa}\right)^{\kappa}$, where a link function $g(\cdot)$ is selected such that $g(\mu_{ij}) = \beta_{ij} + S_i$. There is a lot of flexibility in choosing a model for β_{ij}. If the counts do not vary too much within each timeslot we could take $\beta_{ij} = \beta_i$. If we assume a linear trend within each time slot, then we could take $\beta_{ij} = \beta_{i0} + \beta_{i1} j$. A smooth function for fixed effects in the entire cycle can also be considered. But the smoothness will be broken due to the presence of random effects in the proposed GLMM. On the other hand, the abruptness between discontinuous fixed effects on the edge of each timeslot is eased by the inclusion of random effects in the GLMM. Therefore, we are not using a smooth function for the fixed effects in later size and power studies.

We assume $\underset{\sim}{S}$ follows a multivariate normal distribution with mean $\underline{0}$ and covariance matrix G. Let $\underset{\sim}{\theta}$ denote parameters needed to specify the covariance matrix G. The G matrix influences the autocorrelation between observations from the same timeslot and from different timeslots within the same cycle. This connection will be made more

explicit in section 3.4. The choice of G will vary from application to application, but our proposed tracking statistics are sufficiently general to accommodate the many options for G that are available. Indeed, the flexibility in G is a compelling feature of using a GLMM to describe network data.

3.3 Model Sanity

To verify the applicability of the proposed GLMM, we considered eight Mondays of traffic from the network server discussed in section 3.1. We considered each hour as a time slot and tried fitting a negative binomial GLMM with a linear trend in each time slot using a log-link mean function. We chose a heterogeneous AR(1) covariance structure for the G matrix with parameter $\underset{\sim}{\theta} = (\underset{\sim}{\sigma}^2, \rho)$. Parameter estimates for the model were obtained using the method of pseudo likelihood described in Wolfinger (1993). Because the fitted κ was large, we used the reduced Poisson GLMM as the fitted model. The estimated intercepts and slopes for the linear trend fixed effects are shown in Table 1, the heterogeneous variances are also reported in Table 1 and and correlation parameter estimate was $\hat{\rho} = 0.84$. The fitted model reveals sharp differences in the intercepts, as expected, and five of the slopes are statistically significant. The overall incremental value of adding a linear trend is debatable. More important is the benefit achieved by fitting a heterogeneous variance structure.

In Figure 2, the light gray lines represent observed counts from the eight Mondays. The dark black lines are obtained from the fitted model by generating 1000 cycles of data and then extracting the lower and upper 10^{th} percentiles of the counts at each time point.

We can see that the percentile lines nicely capture the eight traces of observed data with 13% of the observed data points falling outside the percentile limits.

Table 1. Intercepts, Slopes and Variances for Each Hour from Fitted Poisson GLMM

		Slope					Slope		
Hour	Intercept	Estimate	p-value	$\hat{\sigma}_i^2$	Hour	Intercept	Estimate	p-value	$\hat{\sigma}_i^2$
1	2.040	-0.001	0.907	0.000	13	4.100	-0.013	0.006	0.001
2	2.090	-0.014	0.125	0.008	14	3.930	-0.021	<.0001	0.010
3	2.000	-0.003	0.462	0.001	15	3.620	-0.009	0.195	0.034
4	1.830	0.064	<.0001	0.001	16	3.500	-0.016	0.001	0.056
5	2.920	0.052	<.0001	0.030	17	3.180	-0.019	0.001	0.141
6	3.570	0.019	<.0001	0.012	18	2.780	-0.026	<.0001	0.256
7	3.860	0.008	0.111	0.005	19	2.330	-0.016	0.037	0.168
8	4.010	0.002	0.389	0.008	20	2.260	-0.008	0.373	0.112
9	4.050	-0.008	0.027	0.007	21	2.210	-0.006	0.303	0.000
10	3.950	0.005	0.085	0.006	22	2.200	-0.010	0.241	0.000
11	4.050	0.002	0.701	0.006	23	2.350	-0.036	<.0001	0.142
12	4.070	0.000	0.934	0.008	24	2.090	-0.009	0.396	0.016

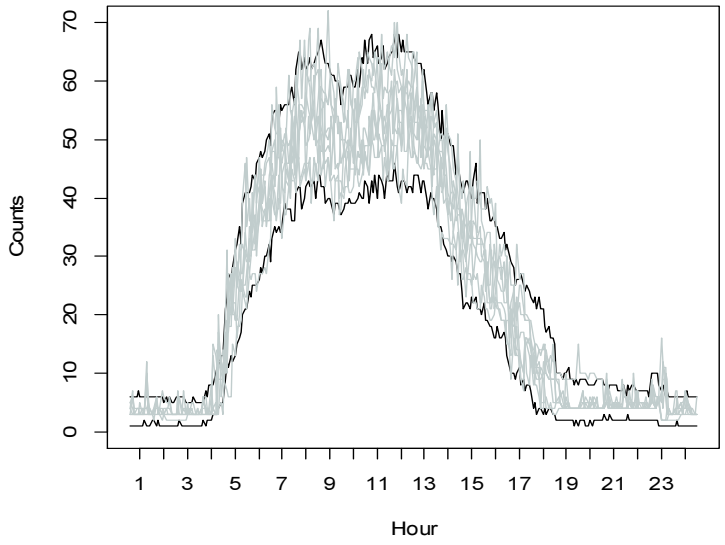


Figure 2. Eight Mondays of Observed Data with Upper and Lower 10th Percentiles of Data Generated from Fitted GLMM

3.4 Correlation Structure

Consider the data in a single cycle. Let g_{uv} denote the element on the uth row and vth column of the $m \times m$ matrix G.

The covariance of counts within the same timeslot is

$$\operatorname{cov}(Y_{ij}, Y_{ij'}) = E\left[\operatorname{cov}(Y_{ij}, Y_{ij'} \mid s_i)\right] + \operatorname{cov}\left[E(Y_{ij} \mid s_i), E(Y_{ij'} \mid s_i)\right].$$

Give random effects, counts are independent, so

$$\operatorname{cov}(Y_{ij}, Y_{ij'}) = 0 + \operatorname{cov}\left[E(Y_{ij} \mid s_i), E(Y_{ij'} \mid s_i)\right]$$

$$= \operatorname{cov}\left[\exp(\beta_{ij} + s_i), \exp(\beta_{ij'} + s_i)\right]$$

$$= \exp(\beta_{ij} + \beta_{ij'}) \operatorname{cov}\left[\exp(s_i), \exp(s_i)\right].$$

Since s_i follows normal distribution mean zero variance g_{ii}, $\exp(s_i)$ follows a log-normal distribution with mean $\exp(\frac{1}{2} g_{ii})$ and variance $\exp(g_{ii})\left[\exp(g_{ii}) - 1\right]$. So we have

$$\operatorname{cov}(Y_{ij}, Y_{ij'}) = \exp(\beta_{ij} + \beta_{ij'} + g_{ii})\left[\exp(g_{ii}) - 1\right].$$

Variance of count Y_{ij} could be obtained by

$$\operatorname{var}(Y_{ij}) = E\left[\operatorname{var}(Y_{ij} \mid s_i)\right] + \operatorname{var}\left[E(Y_{ij} \mid s_i)\right]$$

$$= E\left[\exp\left(\beta_{ij} + s_i\right) + \frac{1}{\kappa} \exp 2\left(\beta_{ij} + s_i\right)\right] + \operatorname{var}\left[\exp\left(\beta_{ij} + s_i\right)\right]$$

$$\operatorname{var}(Y_{ij}) = \exp\left(\beta_{ij}\right) E\left[\exp(s_i)\right] + \frac{1}{\kappa} \exp\left(2\beta_{ij}\right) E\left[\exp(2s_i)\right] + \exp\left(2\beta_{ij}\right) \operatorname{var}\left[\exp\left(s_i\right)\right]$$

$$= \exp\left(\beta_{ij} + g_{ii} / 2\right) + \frac{1}{\kappa} \exp\left(2\beta_{ij} + 2g_{ii}\right) + \exp\left(2\beta_{ij} + g_{ii}\right)\left[\exp\left(g_{ii}\right) - 1\right]$$

$$= \exp\left(2\beta_{ij} + g_{ii}\right)\left(\exp\left(-\beta_{ij} - g_{ii}/2\right) + \frac{1}{\kappa}\exp\left(g_{ii}\right) + \exp\left(g_{ii}\right) - 1 \right).$$

And finally the correlation of counts from the same time slot is,

$$corr(Y_{ij}, Y_{ij'}) = \frac{\operatorname{cov}\left(Y_{ij}, Y_{ij'}\right)}{\sqrt{VAR(Y_{ij})VAR(Y_{ij'})}}$$

$$= \frac{\exp(g_{ii}) - 1}{\sqrt{\left(\exp\left(-\beta_{ij} - \frac{1}{2}g_{ii}\right) + (1 + \frac{1}{\kappa})\exp\left(g_{ii}\right) - 1 \right)\left(\exp\left(-\beta_{ij'} - \frac{1}{2}g_{ii}\right) + (1 + \frac{1}{\kappa})\exp\left(g_{ii}\right) - 1 \right)}}.$$

With an AR(1) covariance structure of parameter $\left(\sigma^2, \rho\right)$ for the random effects, the correlation between of counts within the same timeslot $(Y_{ij}, Y_{ij'})$ could be written as

$$corr\left(Y_{ij}, Y_{ij'}\right) = \frac{\exp(\sigma^2) - 1}{\sqrt{\left(\exp\left(-\beta_{ij} - \frac{1}{2}\sigma^2\right) + \frac{1+\kappa}{\kappa}\exp\left(\sigma^2\right) - 1 \right)\left(\exp\left(-\beta_{ij'} - \frac{1}{2}\sigma^2\right) + \frac{1+\kappa}{\kappa}\exp\left(\sigma^2\right) - 1 \right)}}.$$

For the case when counts are from different timeslot,

$$\operatorname{cov}(Y_{ij}, Y_{i'j'}) = \exp(\beta_{ij} + \beta_{i'j'})\operatorname{cov}\left[\exp(s_i), \exp(s_{i'})\right].$$

Using multivariate log-normal property,

$$\operatorname{cov}\left[\exp(s_i), \exp(s_{i'})\right] = \exp\left(\frac{1}{2}\left(g_{ii} + g_{i'i'}\right)\right)\left(\exp(g_{ii'}) - 1\right)$$

so the covariance of counts from different timeslot is

$$\operatorname{cov}(Y_{ij}, Y_{i'j'}) = \exp\left((\beta_{ij} + \beta_{i'j'}) + \frac{1}{2}\left(g_{ii} + g_{i'i'}\right)\right)\left(\exp(g_{ii'}) - 1\right).$$

The correlation of counts from different timeslots is

$$corr(Y_{ij}, Y_{i'j'}) = \frac{\exp(g_{ii'}) - 1}{\sqrt{\left(\exp\left(-\beta_{ij} - \frac{1}{2}g_{ii'}\right) + (1+\frac{1}{\kappa})\exp(g_{ii'}) - 1\right)\left(\exp\left(-\beta_{i'j'} - \frac{1}{2}g_{ii'}\right) + (1+\frac{1}{\kappa})\exp(g_{ii'}) - 1\right)}}.$$

With an AR(1) covariance structure of parameter $\left(\sigma^2, \rho\right)$ for the random effects, the correlation between of counts within differenct timeslots $(Y_{ij}, Y_{i'j'})$ could be written as

$$corr\left(Y_{ij}, Y_{i'j'}\right) = \frac{\exp(\sigma^2 \rho^{|i-i'|}) - 1}{\sqrt{\left(\exp\left(-\beta_{ij} - \frac{1}{2}\sigma^2\right) + \frac{1+\kappa}{\kappa}\exp\left(\sigma^2\right) - 1\right)\left(\exp\left(-\beta_{i'j'} - \frac{1}{2}\sigma^2\right) + \frac{1+\kappa}{\kappa}\exp\left(\sigma^2\right) - 1\right)}}.$$

If we examine this formula closely, we can see that aside from the fact that counts further away from each other have diminishing correlation, it is also evident that larger fixed effects will lead to a higher correlation. Although the special case $\rho = 0$ may not have broad applicability in network monitoring contexts, even here we can see the model allows correlation amongst observations within the same timeslot. Finally, when the dispersion parameter κ goes to infinity, the negative binomial distribution will reduce to Poisson and the correlations will reach their maximum values.

Chapter 4

Proposed Tracking Statistics

Three tracking statistics are proposed 1) Integrated Likelihood Ratio (ILR), 2) Joint Likelihood Ratio (JLR), and 3) Exponential Weighted Moving Average (EWMA). We assume K cycles of historical data are available to characterize in-control characteristics of the data stream, and that the historical data are updated with a sliding window mechanism to account for network churn. Montes de Oca *et al.* (2010) discusses the implementation of a particular updating scheme. The in-control parameters of interest could be assumed known with sufficient amount of historical data. In chapter 5 we give guidelines on the amount of historical data that is necessary to largely mitigate the effect of estimation errors associated with the model parameters. Consequently, we use the estimated parameters from the historical data as the true in-control values when calculating and calibrating the in-control characteristics of the tracking statistics.

Two of our proposed tracking statistics, ILR and JLR represent repeated Bartlett-type sequential probability ratio tests (SPRT). Taking into account the complexity of network traffic that was discussed earlier, the Bartlett-type SPRT itself would be too computationally demanding to use in practice. In particular, the high dimensional integration is required to get the integrated likelihood, and then the subsequent

maximizations of that function would be quite formidable. Our third tracking statistic is similar to Lambert and Liu (2006) in the sense that it is a EWMA based on normal scores. A major difference is that during the monitoring period we use predicted random effects to adjust for real-time variations and thereby hope to achieve a more robust FAR. The rest of this chapter provides derivation details, comparisons and discussion for our three tracking statistics.

4.1 Integrated Likelihood Ratio (ILR) Tracking Statistic

Denote the current set of observations during the monitoring cycle by $\{Y_t\}_{t=1}^{n}$. For each t, let (i_t, j_t) respectively correspond to the timeslot and observation within timeslot indices for Y_t. In the context of the GLMM notations introduced in section 3.2, the conditional distribution (given the random effects) of Y_t is negative binomial with mean $\mu_{i_t j_t}$ and dispersion parameter κ. Suppose K cycles of historical data are available $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_K$. Our null hypothesis is that the $\{Y_t\}_{t=1}^{n}$ observations are in-control and we represent this by the situation that the fixed effects $\underset{\sim}{\beta}$ in the GLMM have not changed relative to the historical data. Let $\underset{\sim}{\beta}_0$ denote the pre-specified in-control value of $\underset{\sim}{\beta}$, possibly obtained through analysis of the historical data. The alternative hypothesis we consider is that fixed effects change from $\underset{\sim}{\beta}_0$ to $c\underset{\sim}{\beta}_0$. Formally, our hypothesis is $H_0 : \underset{\sim}{\beta} = \underset{\sim}{\beta}_0$ versus $H_1 : \underset{\sim}{\beta} = c\underset{\sim}{\beta}_0$, where c is a specified constant which represent the minimum degree of change, we call it the inflation factor, that is desired to be detected.

Let $L_k(\underset{\sim}{x}_k;\underset{\sim}{\beta}_0,\kappa,\underset{\sim}{\theta})$ denote the integrated likelihood of the historical data in cycle k. For the monitoring cycle, let $L(y_1,\ldots y_n;\underset{\sim}{\beta}_0,\kappa,\underset{\sim}{\theta})$ denote the in-control integrated likelihood. The corresponding out of control likelihood during the monitoring cycle is $L(y_1,\ldots y_n;c\underset{\sim}{\beta}_0,\kappa,\underset{\sim}{\theta})$. With the presence of unknown nuisance parameters, a Bartlett-type sequential probability ratio test (SPRT) statistic can be set up as follows

$$T_n^{BLR}=\frac{\max\limits_{\underset{\sim}{\theta},\kappa}\left\{L(y_1,\ldots y_n;c\underset{\sim}{\beta}_0,\ \kappa,\underset{\sim}{\theta})\prod_{k=1}^{K}L_k(\underset{\sim}{x}_k;\underset{\sim}{\beta}_0,\kappa,\underset{\sim}{\theta})\right\}}{\max\limits_{\underset{\sim}{\theta},\kappa}\left\{L(y_1,\ldots y_n;\underset{\sim}{\beta}_0,\kappa,\underset{\sim}{\theta})\prod_{k=1}^{K}L_k(\underset{\sim}{x}_k;\underset{\sim}{\beta}_0,\kappa,\underset{\sim}{\theta})\right\}}.$$

The integrated likelihoods for both the historical cycle and the monitoring cycle involve high-dimension integration over the joint distribution of the random effects, which are not generally going to be distributed independently. The integrated likelihoods in both the numerator and denominator of T_n^{BLR} also need to be maximized over nuisance parameters after the integrated likelihoods are obtained. In addition, it is difficult to compute the threshold under the null due to unknown nuisance parameters. The challenging computational aspects make the BLR test statistic generally intractable. However, it does inspire practical variations that we now develop.

First, since the historical data is likely to be substantial, we suggest approximating the maximizations associated with the BLR by using the in-control historical data to obtain estimated parameters. Substituting these values as true values into the numerator and denominator of T_n^{BLR} yields the following approximate BLR testing statistic

$$T_n^{ABLR} = \frac{L(y_1,\ldots y_n; c\underset{\sim}{\beta}_0, \kappa, \underset{\sim}{\theta})}{L(y_1,\ldots y_n; \underset{\sim}{\beta}_0, \kappa, \underset{\sim}{\theta})} = \frac{\int \prod_{t=1}^{n} f_{Y_t}\left(y_t \mid c\underset{\sim}{\beta}_0, \underset{\sim}{s}, \kappa\right) f_{\underset{\sim}{S}}(\underset{\sim}{s} \mid \underset{\sim}{\theta})\, d\underset{\sim}{s}}{\int \prod_{t=1}^{n} f_{Y_t}(y_t \mid \underset{\sim}{\beta}_0, \underset{\sim}{s}, \kappa) f_{\underset{\sim}{S}}(\underset{\sim}{s} \mid \underset{\sim}{\theta})\, d\underset{\sim}{s}},$$

which is the integrated likelihood ratio under H_1 and H_0 the for the monitoring period. Note that with this approximation, we are assuming the nuisance parameters behave relatively the same under the null and alternative hypotheses. Negative values of $\log T_n^{ABLR}$ favor H_0 and would suggest no inspection of the network is needed, whereas when this quantity becomes large and positive it suggests an out-of-control situation where network inspection is needed. We suggest a sequential tracking statistic by the resetting the ABLR testing statistic (to zero) when it becomes negative and starting again with a test of H_0 vs. H_1. The motivation for thinking about using a repeated Bartlett-type SPRT as a change-point problem stems from the interpretation of Page's CUSUM as a repeated Wald SPRT[see, for example, Basseville and Nikiforov (1993)]. Specifically, our ILR tracking statistic is defined as follows

$$T_n^{ILR} = \max\left\{0,\ \log \frac{\int \prod_{t=r_n}^{n} f_{Y_t}(y_t \mid c\underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{\underset{\sim}{S}^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta})\, d\underset{\sim}{s}^*}{\int \prod_{t=r_n}^{n} f_{Y_t}(y_t \mid \underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{\underset{\sim}{S}^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta})\, d\underset{\sim}{s}^*}\right\}.$$

Here, r_n is the index of the first observation of the monitoring cycle after the most recent reset and $\underset{\sim}{\beta}_0^*$ and $\underset{\sim}{s}^*$ represent the fixed effects and random effects that correspond to observations $\{Y_t\}_{t=r_n}^{n}$. Defined this way, the ILR tracking statistic achieves two attractive features. First, by resetting it prevents a large negative run during a sustained in-control

period that would otherwise delay an alarm when an out-of-control situation emerges. Second, since the number of observations in the monitoring cycle that are used at any given time is $n-r_n+1$, the dimension of the required integration to evaluate the test statistic is significantly reduced and will often only require a few of the correlated random effects.

Pushing practicality a bit more, the integrals in the numerator and denominator of T_n^{ILR} could be approximated using a Laplace approximation. First, let $h_0(\underset{\sim}{s}^*)$ denote the joint likelihood function in the integral in the denominator of T_n^{ILR}. That is,

$$h_0(\underset{\sim}{s}^*) = \prod_{t=r_n}^{n} f_{Y_t}(y_t \mid \underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{\underset{\sim}{S}^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta}).$$

Let $l_0(\underset{\sim}{s}^*) = \log h_0(\underset{\sim}{s}^*)$ and $A_0^{-1}(s^*) = \dfrac{\partial^2 l_0(\underset{\sim}{s}^*)}{\partial \underset{\sim}{s}^{*2}}$. Use a Taylor expansion of $l_0(\underset{\sim}{s}^*)$ around $\underset{\sim}{s}^* = \underset{\sim}{s}_0^*$ we have

$$l_0(\underset{\sim}{s}^*) \approx l_0(\underset{\sim}{s}_0^*) - \frac{1}{2}(\underset{\sim}{s}^* - \underset{\sim}{s}_0^*)' A_0^{-1}(\underset{\sim}{s}_0^*)(\underset{\sim}{s}^* - \underset{\sim}{s}_0^*)$$

where $\underset{\sim}{s}_0^*$ is the value of $\underset{\sim}{s}^*$ that makes gradient of $l_0(\underset{\sim}{s}^*)$ equal zero and maximize the joint likelihood. Then

$$\int h_0(\underset{\sim}{s}^*)\, d\underset{\sim}{s}^* \approx \int \exp\left(l_0(\underset{\sim}{s}_0^*) - \frac{1}{2}(\underset{\sim}{s}^* - \underset{\sim}{s}_0^*)' A_0^{-1}(\underset{\sim}{s}_0^*)(\underset{\sim}{s}^* - \underset{\sim}{s}_0^*) \right) d\underset{\sim}{s}^*$$

$$= \max_{\underset{\sim}{s}^*} \left(\prod_{i=r_n}^{n} f_{Y_i}(y_i \mid \underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{S^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta}) \right) (2\pi)^{n/2} \left| A_0(\underset{\sim}{s}_0^*) \right|^{1/2}.$$

Similarly, we could start with

$$h_1(\underset{\sim}{s}^*) = \prod_{t=r_n}^{n} f_{Y_t}(y_t \mid c\underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{\underset{\sim}{S}^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta}) \ ,$$

let $A_1^{-1} = \dfrac{\partial^2 l_1(\underset{\sim}{s}^*)}{\partial \underset{\sim}{s}^{*2}}$ and $\underset{\sim}{s}_1^*$ be the value of $\underset{\sim}{s}^*$ that could makes gradient of $l_1(\underset{\sim}{s}^*) = \log h_1(\underset{\sim}{s}^*)$ vanish. Then

$$\int h_1(\underset{\sim}{s}^*)\, d\underset{\sim}{s}^* \approx \int \exp\left(l_1(\underset{\sim}{s}_1^*) - \frac{1}{2}(\underset{\sim}{s}^* - \underset{\sim}{s}_1^*)' A_1^{-1}(\underset{\sim}{s}_1^*)(\underset{\sim}{s}^* - \underset{\sim}{s}_1^*) \right) d\underset{\sim}{s}^*$$

$$= \max_{\underset{\sim}{s}^*} \left(\prod_{t=r_n}^{n} f_{Y_t}(y_t \mid \underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{\underset{\sim}{S}^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta}) \right) (2\pi)^{n/2} \left| A_1(\underset{\sim}{s}_1^*) \right|^{1/2} .$$

Putting the two Laplace approximations back into the tracking statistic, we can and approximated T_n^{ILR} as follow

$$T_n^{AILR} = \max\left\{ 0,\ \log \frac{\max_{\underset{\sim}{s}^*} \left(\prod_{t=r_n}^{n} f_{Y_t}(y_t \mid c\underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{S^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta}) \right)}{\max_{\underset{\sim}{s}^*} \left(\prod_{t=r_n}^{n} f_{Y_t}(y_t \mid \underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{S^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta}) \right)} + \frac{1}{2}\left(\log\left| A_0(\underset{\sim}{s}_0^*) \right| - \log\left| A_1(\underset{\sim}{s}_1^*) \right| \right) \right\} .$$

In our later simulation study, we use AILR tracking statistic mitigate the potentially high-dimension integration that would otherwise be required.

The threshold for the AILR tracking statistic can be obtained by using the historical data to estimate the in-control distribution of T_n^{AILR}. More precisely, the procedure to get threshold is laid out in Algorithm 1.

Algorithm 1: Determining threshold H for ILR

1) Use historical data to fit an appropriate GLMM
2) Simulate a cycle of data $n = 1, \ldots, N$ from the fitted GLMM where N is cycle length.
3) Run T_n^{AILR} along the cycle and obtain the maximum value of $\left\{T_n^{AILR}\right\}_{n=1}^{N}$
4) Repeat steps (2) – (3) B_1 (we use $B_1 = 1000$) times and set the threshold H for this set of historical data to be the $(1-\alpha)$ percentile of the B_1 maximum values.

4.2 JLR Tracking Statistic

The JLR test of H_0 v.s. H_1 uses the h-likelihood [see section 2.4] rather than the integrated likelihood. The idea is to replace the random effects in the monitoring period by predictions rather than by integrating them out. In particular, we propose the following tracking statistic,

$$T_n^{JLR} = \max\left\{0,\ \log\frac{\max_{\underset{\sim}{s}^*}\left(\prod_{t=r_n}^{n} f_{Y_t}(y_t \mid c\underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{S^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta})\right)}{\max_{\underset{\sim}{s}^*}\left(\prod_{t=r_n}^{n} f_{Y_t}(y_t \mid \underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{S^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta})\right)}\right\}.$$

The numerator is the joint density of observed counts and random effects maximized over the random effects under the alternative hypothesis. The denominator is the same joint density maximized over the random effects under H_0. Threshold estimation and parameter updating procedures are exactly the same as was described for T_n^{ILR} test statistics. A key difference between T_n^{ILR} and T_n^{JLR} is that the latter utilizes predictions

for the actual random effects realized during the monitoring period whereas ILR averages out the random effects with respect to the distribution of all possible values.

However, comparing T_n^{AILR} with T_n^{JLR}, the difference is just the increment $\left(\log\left|A_1(\underset{\sim}{s}_0^*)\right| - \log\left|A_0(\underset{\sim}{s}_1^*)\right|\right)/2$. We learn from a simulation study reported in section 6.1 that this difference between AILR and JLR is very small relative to the magnitude of both tracking statistics. Figure 3 is an illustrative graph of AILR and JLR for the same set of simulated observations from a negative binomial GLMM with fixed effects generated from the following smooth function

$$\exp(\beta_n) = 1.87 + 0.0066\left[3804e^{\sin(-2.27\pi n/1440)} - 119e^{\sin(12\pi n/1440)}\right],$$

G chosen to have an $\mathrm{AR}(1)$ structure with $(\sigma^2, \rho) = (0.1, 0.7)$, and dispersion parameter $\kappa = 100$. It can be seen that the two tracking statistics are very close to each other. Therefore, we expect ILR and JLR might perform similarly with respect of false alarm rate and detection power.

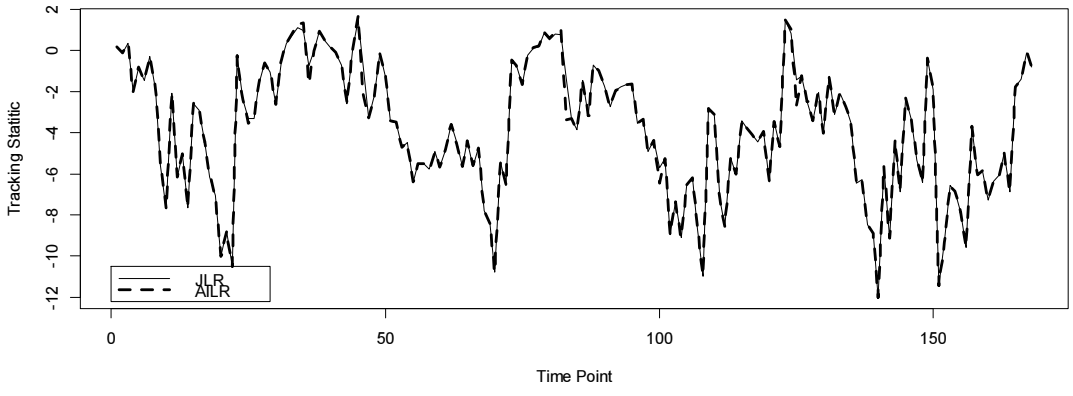


Figure 3. JLR and AILR tracking statistics for one simulated cycle to illustrate that their difference is quite small.

4.3 EWMA of Normal Scores

Our EWMA tracking statistic is related to the previously discussed work of Lambert and Liu (2006) [see section 2.1]. A key difference, however, is that we will use predictions of the random effects from the monitoring period to improve the accuracy of the conditional means and variances of the negative binomial distributions for each of the observation. Every time a new observation y_n is gathered during the monitoring cycle, we maximize the joint conditional in-control likelihood $\prod_{t=1}^{n} f_{Y_t}(y_t \mid \underset{\sim}{\beta}_0^*, \underset{\sim}{s}^*, \kappa) f_{\underset{\sim}{S}^*}(\underset{\sim}{s}^* \mid \underset{\sim}{\theta})$ over $\underset{\sim}{s}^*$. This step is the same as maximizing the joint likelihood under the null hypothesis in the JLR tracking statistic. Later in this section, we will propose a strategy to optionally reduce the dimension of this optimization. Conditional on the random effect s_{i_n}, the observation y_n is conditionally independent negative binomial random variables with mean $\exp(\beta_{i_n,j_n} + s_{i_n})$ and dispersion parameter κ. Let F_n denote the conditional distribution of Y_n which can be calculated as

$$F_{Y_n}(y_n) = \sum_{b=0}^{y_n} P(Y_n = b \mid \beta_{i_n,j_n}, s_{i_n}, \kappa) .$$

We estimate this distribution with $\hat{F}_{Y_n}(y_n)$ by replacing β_{i_n,j_n} and κ with estimates from the historical data and s_{i_n} with its predictor, obtained as described above. If large observations imply network degradation, the EWMA tracking statistic is defined as

$$T_n^{EWMA} = (1-w)T_{n-1} + wZ_n \ ,$$

where $Z_n = \Phi^{-1}(\hat{F}_n(y_n))$, $w \in (0,1]$ and Φ^{-1} is the inverse of the normal c.d.f.. If small counts imply network degradation, then Z_n is redefined as $Z_n = \Phi^{-1}(1-\hat{F}_n(y_n))$. Note that with either definition of Z_n, the tracking statistic will always tend to become large during an out-of-control situation.

During an in-control monitoring cycle the input variables Z_n can be regarded as approximately conditionally independent identically distributed $N(0,1)$ random variables and a threshold can be obtained to achieve a given conditional FAR as is done in the context of classical EWMA implementations [see, for example, Robinson and Ho (1978) , Vardeman and Jobe (1998)]. However, the parameters β_{i_n,j_n} and κ that are involved in the estimated conditional distribution $\hat{F}_n$ will need to be updated as the historical data gets updated.

Returning to the issue of finding predictions for the random effects during the monitoring cycle, we suggest the following strategy which shares similar motivation as in Xie *et al.* (2013) to reduce the optimization dimension. Taking into account that observations are not only correlated within each time slot but also between timeslots, all previous observations $y_1,...,y_{n-1}$ contain information about the random effect corresponding to y_n. Because the correlation of the random effects decays, not all random effect $s_1,\ldots,s_{i_n}$ are equally influential. We take advantage of that by only utilizing those closest to the time slot i_n. We suggest that a sliding window scheme of q

timeslots $i_n - q + 1, \ldots, i_n$ could be used when predicting s_{i_n}. Depending on the strength of the correlation between random effects, which can be estimated by analysis of historical data, q can be relatively small if the correlation vanishes quickly.

Since Y_n is discrete, a continuity correction is needed when approximating the in-control distribution of Z_n by $N(0,1)$. Failure to use a continuity correction will result in conditional FARs that are too high. We suggest a continuity correction through adding a random uniform $U(0,1)$ variable U_n to the observed counts Y_n and calculate the c.d.f of this modified observation, $W_n = Y_n + U_n$, given by $G_{W_n}(w_n) = \int_0^1 F_{Y_n}(w_n - u)\,du$. The integral could be easily approximated through a standard numerical integration method. Using the continuity correction implies using T_n^{EWMA} after replacing $Z_n = \Phi^{-1}(\hat{F}_n(y_n))$ with $Z_n = \Phi^{-1}(\hat{G}_n(w_n))$.

Chapter 5

Discussion about ILR, JLR and EWMA

5.1 An Example of using ILR for Normal Distributed Observations

An example is provided here to illustrate the implementation of ILR and the relationship between ILR and classical SPC approaches when the observations can be assumed as normally distributed.

Consider a sequence of data $\{Y_t\}_{t=1}^{n}$ which has the timeslot structure described in section 3.2 such that the pair (i_t, j_t) represents the timeslot and time point corresponding to observation Y_t , $i_t = 1,2,...,i_n$, $j_t = 1,2,...,n_i$. Assume that Y_t is conditionally independent random Normal$\left(\mu_{i_t,j_t}, \sigma^2_{e,i_t}\right)$ given random effect s_{i_t} and μ_{i_t,j_t} is obtained through the following link function $\mu_{i_t,j_t} = \beta_{i_t,j_t} + s_{i_t}$, where β_{i_t,j_t} the fixed effect and s_{i_t} is a random effect from $N(0,\sigma^2_{s,i_t})$. The hypotheses test for a change-point detection problem is set up as follows

$$H_0 : Y_1,...,Y_n \sim f^0_{Y_1}(y_1;\beta^0_{i_1,j_1},\sigma^2_{e,j_1},\sigma^2_{s,j_1}),\ldots,f^0_{Y_n}(y_n;\beta^0_{i_n,j_n},\sigma^2_{e,j_n},\sigma^2_{s,j_n})$$

$$H_1 : Y_1,...,Y_k \sim f^0_{Y_1}(y_1;\beta^0_{i_1,j_1},\sigma^2_{e,j_1},\sigma^2_{s,j_1}),\ldots f^0_{Y_k}(y_k;\beta^0_{i_k,j_k},\sigma^2_{e,j_k},\sigma^2_{s,j_k})$$

$$Y_{k+1},...,Y_n \sim f^1_{Y_{k+1}}(y_{k+1};\beta^1_{i_{k+1},j_{k+1}},\sigma^2_{e,j_{k+1}},\sigma^2_{s,j_{k+1}}),\ldots,f^1_{Y_n}(y_n;\beta^1_{i_n,j_n},\sigma^2_{e,j_n},\sigma^2_{s,j_n}),$$

which means the fixed effects in the link function changed to a different set of value after time point k. For this example, assume infinite historical data is available, so all parameters are known. It is also assumed that fixed effects in the out-of-control situation are known.

From a classical SPC perspective, the vector of the observations $\underset{\sim}{Y} = (Y_1,...,Y_n)'$ can be considered having a marginal distribution of multivariate normal with mean $\underset{\sim}{\beta} = (\beta_{i_1,j_1},...,\beta_{i_n,j_n})'$ and variance covariance matrix $\Sigma = diag\left[\left(\sigma^2_{e,i} I_{n_i} + \sigma^2_{s,i} J_{n_i}\right)\right]$, with $i = 1,2,....,i_n$, where I_{n_i} is a n_i dimensional diagonal matrix and J_{n_i} is an n_i by n_i matrix with each element equals 1. With Cholesky decomposition, $\Sigma = \Gamma\Gamma'$, a vector of transformed observations $\underset{\sim}{Z} = \Gamma^{-1}\underset{\sim}{Y}$ can be obtained which follows multivariate normal $(\Gamma^{-1}\underset{\sim}{\beta},\ I_n)$. One attractive property of Cholesky decomposition, we call it the invariant property here, is that when a new observation Y_{n+1} is available, the element $z_1,...,z_n$ in the new transformed observation $(z_1,...,z_{n+1})$ is the same as when only n observations is available. This property makes sure a sequential change-point algorithm can be applied to this transformed sequence of observations $\{Z_t\}_{t=1}^n$. Notice that observations in $\{Z_t\}_{t=1}^n$ can be considered as identically distributed normal random variables, extended CUSUM in section 2.2.3 could be applied directly with a tracking statistic

$$S_n = \max\left\{0, S_{n-1} + \log\frac{f^1_{Z_n}(z_n)}{f^0_{Z_n}(z_n)}\right\}.$$

Now if we want to directly apply the ILR tracking statistic to this change-point detection, we have the integrated likelihood for $\{Y_t\}_{t=1}^{n}$ in the most general form as follow

$$L(\underset{\sim}{y}) = \prod_{i_t=1}^{i_n}\left\{\int_{-\infty}^{\infty}(2\pi\sigma^2_{e,i_t})^{-n_{i_t}/2}\exp\left[-\frac{1}{2\sigma^2_{e,i_t}}\sum_{j_t=1}^{n_{i_t}}(y_{i_t,j_t}-\beta_{i_t,j_t}-s_{i_t})^2\right](2\pi\sigma^2_{s,i_t})^{-1/2}\exp\left\{-s_{i_t}{}^2/(2\sigma^2_{s,i_t})\right\}ds_{i_t}\right\}$$

$$= (2\pi)^{-n/2}\Sigma^{-1/2}\exp\left\{-\frac{1}{2}(\underset{\sim}{y}-\underset{\sim}{\beta})'\Sigma^{-1}(\underset{\sim}{y}-\underset{\sim}{\beta})\right\}$$

$$\propto (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(\underset{\sim}{z}-\Gamma^{-1}\underset{\sim}{\beta})'I_n(\underset{\sim}{z}-\Gamma^{-1}\underset{\sim}{\beta})\right\}$$

$$= L(\underset{\sim}{z}) = \prod_{t=1}^{n} f_Z(z)$$

Due to the fact that ILR tracking statistic will be reset to zero when it becomes negative, we use $\underset{\sim}{y}^*$ and $\underset{\sim}{z}^*$ to denote the observations after the most recent resetting and r_n denote the time point after the most recent resetting. Finally, the ILR tracking statistic can be written as

$$T_n^{ILR} = \max\left\{0,\ \log\frac{\prod_{t=r_n}^{n} f^0_{Z_t}(z_t)}{\prod_{t=r_n}^{n} f^1_{Z_t}(z_t)}\right\}$$

where $f^0_{Z_t}$ and $f^1_{Z_t}$ denote the distribution of Z_t under the null and the alternative respectively. This is of the same form of a repeated Wald SPRT tracking statistics. Based on the discussion in section 2.2.4, classic CUSUM can be also interpreted as repeated

Wald SPRT with lower bound zero, we can have a equivalent from of ILR tracking statistic as

$$T_n^{ILR} = \max\left\{0, T_{n-1}^{ILR} + \log\frac{f_{Z_n}^1(z_n)}{f_{Z_n}^0(z_n)}\right\},$$

which is exactly the same as the tracking statistics obtain from a classical SPC perspective. Therefore, we show in this example, that within normal distribution context, the ILR tracking statistic is exactly the same as the tracking statistic that would be obtained from a classical SPC approach.

5.2 Compare AILR, JLR and EWMA with Lambert and Liu's method

In this section we compare the proposed AILR, JLR and EWMA algorithms with the method described in Lambert and Liu (2006) (LL). Similar to their setting, a day with 24 hours (timeslots) and 60 one-minute counts in each hour is considered as a cycle. The counts Y_t, $t = 1,2,\ldots 1440$ in a cycle are independently generated from a negative binomial GLMM with conditional means μ_t of the form $\log(\mu_t) = \beta_t + s_{i_t}$. Here, the fixed effect β_t is calculated from a smooth function similar to that used in Lambert and Liu (2006). In our case, we fit the smooth function using the eight traces of VMware Monday data displayed in Figure 2. The fitted smooth function is

$$\exp(\beta_t) = 1.87 + 0.0066\left[3804e^{\sin(-2.27\pi t/1440)} - 119e^{\sin(12\pi t/1440)}\right],$$

with the means obtained from this function is displayed in Figure 5 as follows.

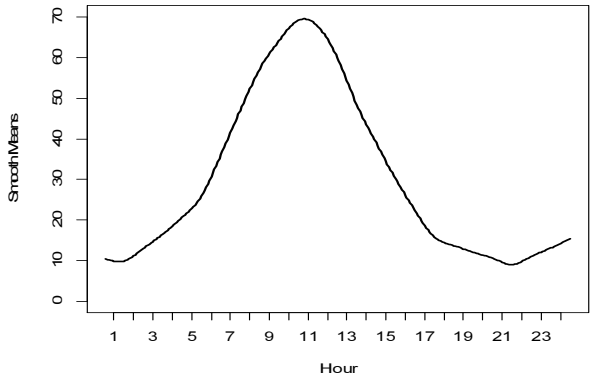


Figure 4. The Fitted Smoothed Function

The random effects within a cycle are generated from a multivariate normal distribution with mean $\underset{\sim}{0}$ and an AR(1) covariance matrix (σ^2, ρ). For our comparison we assume infinite historical data are available. Thus, true parameter values are available to calculate the tracking statistics which is sufficient to make our point without complications associated with how much historical is needed to get accurate parameters. Thresholds are chosen for each algorithm to achieve a target false alarm rate of 10%.

Table 2. FAR Comparison of AILR,JLR, EWMA and LL

κ σ^2 ρ	50 E-10 0.4	50 0.001 0.4	50 0.005 0.4
AILR	0.099	0.085	0.102
JLR	0.104	0.103	0.096
EWMA	0.098	0.096	0.097
LL	0.101	0.194	0.625

Table 2 shows FAR estimates based on simulating 1000 in-control sample paths for each of the four methods. By construction, the AILR, JLR and EWMA algorithms achieve the

target FAR value. The interesting finding in Table 2 is that the LL method has inflated FAR values when σ^2 deviates away from zero. In fact, the LL method is only designed for the case $\sigma^2 = 0$. The introduction of random effects into the model, which adds flexibility when wanting to account for normal network churn and evolution, causes variation in the conditional means that cannot be accounted for in the LL method. Because of this finding, we will only focus on the size and power properties of AILR, JLR and EWMA tracking statistics in the following chapter.

Chapter 6

Simulation Analysis of Size and Power

6.1 Size Analysis

Because the threshold computed for AILR and JLR methods depend on the estimated parameters associated with the in-control GLMM, the FAR observed during a monitoring cycle will be conditional on that set of historical data. Likewise, because the transformation associated with the EWMA method depends on the fitted model, it too will also have an FAR that is conditional on the historical data. Clearly, the more cycles of historical data that available to estimate the in-control parameters, the more precise a conditional FAR will be relative to its target value.

To compare the tracking statistics with respect to how quickly they calibrate to varying levels of historical data, we examined the distributions of their conditional FARs for different depths of historical data. In order to do this, it was necessary to define a feasibility criterion for what a satisfactory distribution of conditional FARs would look like. The criterion we used was that for a target FAR of 10% we should have at least 90% of the conditional FARs between 7% and 13% and at least 95% of the conditional FARs

between 6% and 14%. A similar criterion can be defined for any other target FAR level. Two simulation studies were carried out as follows. First, we chose a week as a cycle and assumed that counts are gathered every five minutes. The counts were modeled with a negative binomial GLMM using a AR(1) covariance structure for the random effects. We considered eight scenarios for the parameters (κ, σ^2, ρ) by choosing $\kappa \in \{50, 100\}$, $\sigma^2 \in \{0.01, 0.001\}$ and $\rho \in \{0.4, 0.7\}$. For the fixed effects, we used the smooth function in section 5.2 for each of the five weekdays, and then decreased that function by 50% for each of the two weekend days.

The second simulation study assumed the same weekly cycle and 168-hour timeslot structure as described above, but the five-minute counts were modeled with a Poisson GLMM, again using a AR(1) covariance structure for the random effects. Four scenarios for the covariance parameters were considered, as per the combinations of $\sigma^2 \in \{0.01, 0.001\}$ and $\rho \in \{0.4, 0.7\}$. Fixed effects for the five weekdays were taken to be identical and modeled with a linear trend according to the intercepts and slopes shown in Table 1. The fixed effect for each time point obtained from Table 1 is then lowered by a constant to generate a sequence of counts with lower magnitude. Fixed effects for the weekend days were again reduced 50% relative to the weekdays.

The main difference between the negative binomial setting and the Poisson setting is that expected value of the counts ranged between 9 and 70 in the former as compared to between 2 and 22 in the latter. In this way, the two studies reflect metrics of a different nature. The reason that σ^2 in the Poisson study was chosen to be smaller than in the

negative binomial setting was to keep the conditional means of the Poisson counts from overlapping with the conditional means of the negative binomial counts. A piecewise linear GLMM is fitted for each scenario in both the Negative Binomial and Poisson settings. Algorithm 2 summarizes the steps used for executing the simulation studies. As mentioned earlier, the threshold is fixed for EWMA, and is a function of *w*. We chose $w = 0.25$, for which $H = 1.45$, thus eliminating the need for step 2.

Algorithm 2: Evaluating conditional FAR for AILR, JLR and EWMA tracking statistics

---

1) Choose a scenario for the parameters of the in-control distribution
   Generate K cycles (we use $K = 20$ and 30) of historical data
2) Determine the threshold H as described in Algorithm 1 (we use target FAR = 10%)
3) Simulate B_2 (we use $B_2 = 1000$) cycles of monitoring data using true parameters
4) Run T_n^{ILR}, T_n^{JLR}, T_n^{EWMA} tracking statistics along each simulated cycle
5) Record the conditional FAR as the proportion of the B_2 cycles that alarmed
6) Repeat step (1) – (5) M times (we use $M = 25$)
7) Summarize the distribution of the M conditional FARs and get the mean of them as the unconditional FAR

In the negative binomial and Poisson settings, GLMMs with linear trend in each timeslot is fitted. Based on the conditional FAR results and power results obtained later, this model works well for the both settings. Figure 5 shows that for the negative binomial setting with parameters $(\kappa, \sigma^2, \rho) = (50, 0.001, 0.4)$, the conditional FAR from EWMA tracking statistic achieved the feasibility criterion with 20 weeks of historical data, while

the AILR and JLR need 30 weeks of historical data to achieved the feasibility criteria. Figure 6 shows that with 30 weeks of historical data, all three methods consistently achieve the feasibility criterion.

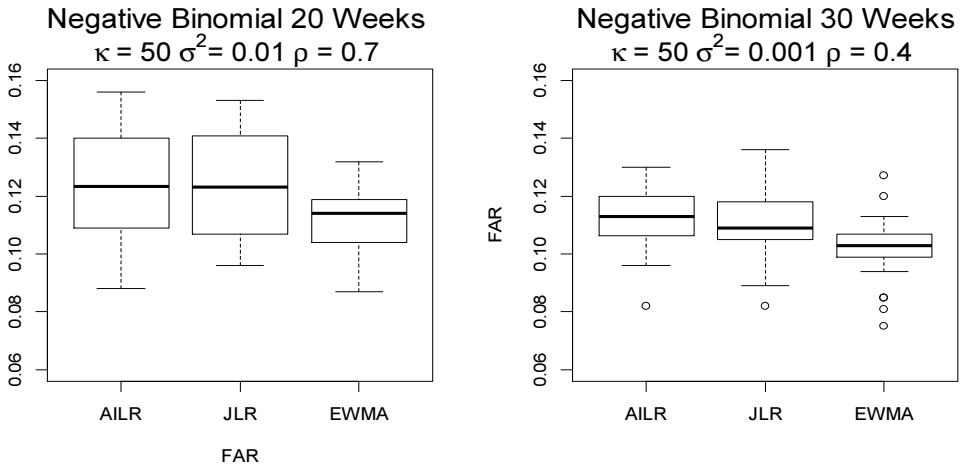
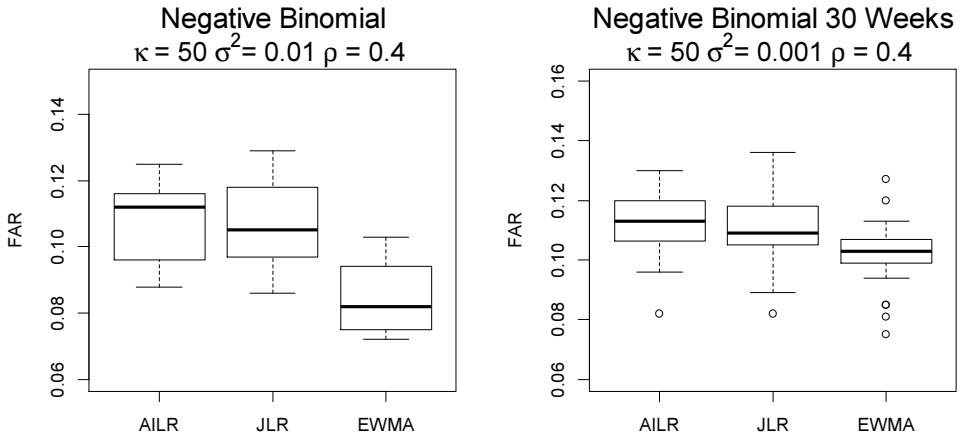
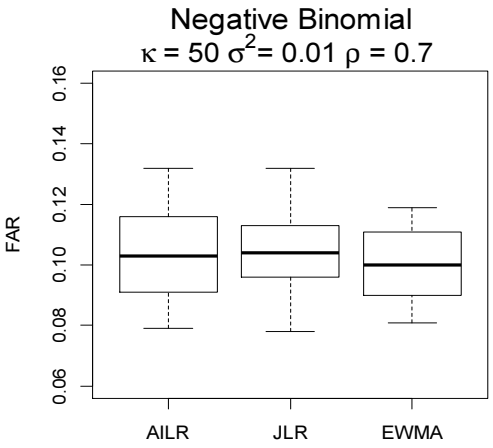
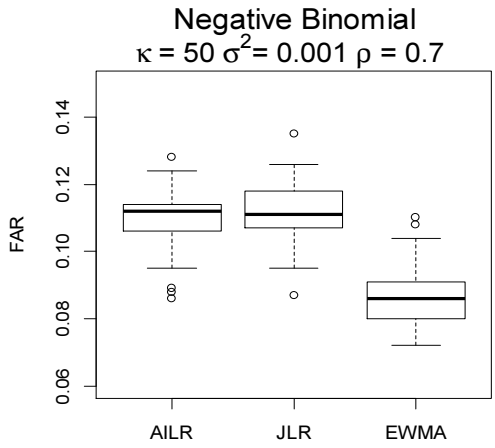
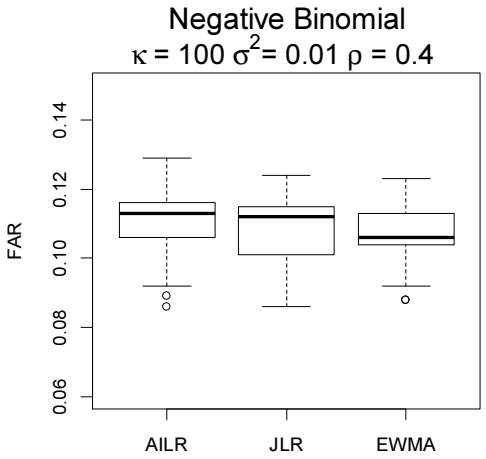


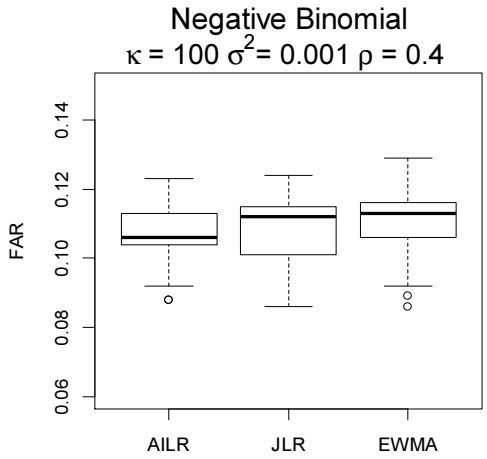
Figure 5. Conditional FAR for EWMA, AILR and JLR with 20 weeks and 30 weeks of historical data in one of the negative binomial Setting

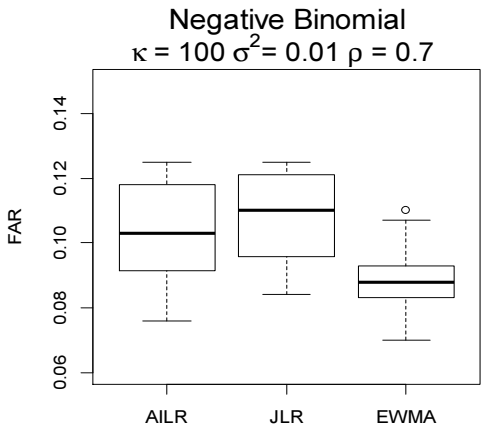


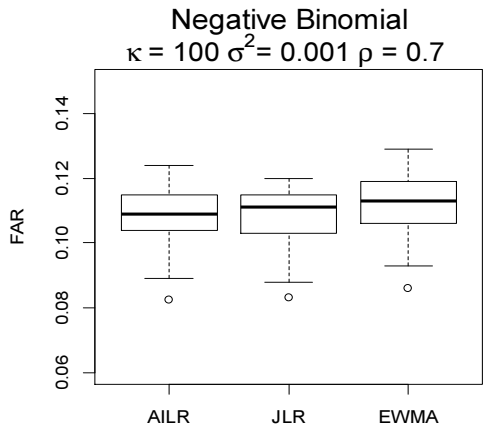

Negative Binomial
κ = 50 σ2= 0.01 ρ = 0.7
FAR
0.06
0.08
0.10
0.12
0.14
0.16
AILR
JLR
EWMA


Negative Binomial
κ = 50 σ2= 0.001 ρ = 0.7
FAR
0.06
0.08
0.10
0.12
0.14
AILR
JLR
EWMA


Negative Binomial
κ = 100 σ2= 0.01 ρ = 0.4
FAR
0.06
0.08
0.10
0.12
0.14
AILR
JLR
EWMA


Negative Binomial
κ = 100 σ2= 0.001 ρ = 0.4
FAR
0.06
0.08
0.10
0.12
0.14
AILR
JLR
EWMA


Negative Binomial
κ = 100 σ2= 0.01 ρ = 0.7
FAR
0.06
0.08
0.10
0.12
0.14
AILR
JLR
EWMA


Negative Binomial
κ = 100 σ2= 0.001 ρ = 0.7
FAR
0.06
0.08
0.10
0.12
0.14
AILR
JLR
EWMA

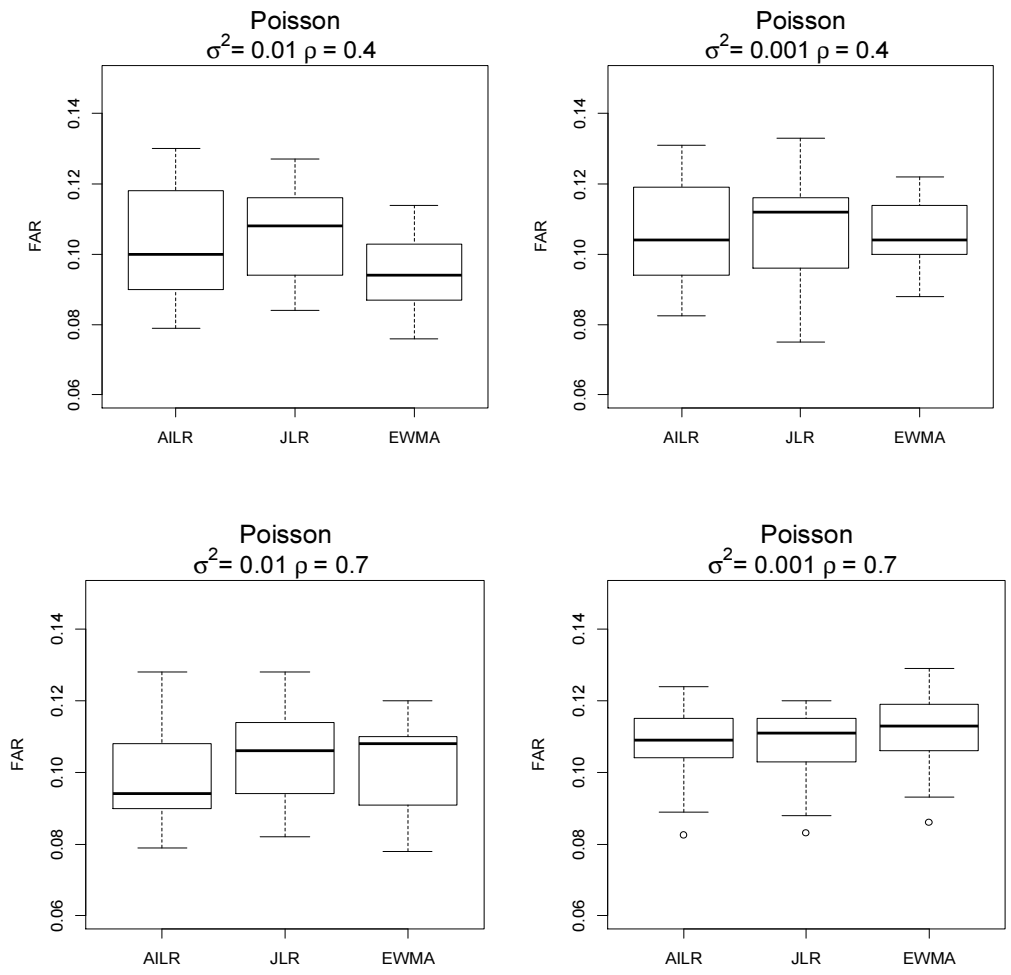


Figure 6. Conditional FAR for EWMA, AILR and JLR 30 weeks of historical data

6.2 Power Study

Besides FAR, ability of a tracking statistic to alarm as soon as possible when there is an anomalous event is another critical property. We investigate this property for AILR, JLR and EWMA with the following simulation study. We consider the same twelve scenarios described in the size study. To give a fair comparison of power for the three tracking statistics, consider that infinite historical data are available so thresholds are not

affected by accuracy of parameter estimation from GLMM. The AILR and JLR methods are constructed to detect a fault that corresponds to an inflation parameter of $c = 1.1$ for the negative binomial setting and $c = 1.2$ for the Poisson setting. Two sets of fault are injected in the data stream. One set of faults starts on Monday morning 7 a.m., which is a timeslot where counts begin to increase toward their daily maximum. The second set of faults is injected at 12 p.m. on Friday which is a time period that could be considered as a peak period. Injected faults have two different durations, 60 observations and 120 observations, and injected faults vary with respect to the magnitude of their inflation factors of the fixed effects. We consider $\{1.05,\ 1.07,\ 1.085,\ 1.1,\ 1.13,\ 1.16\}$ for the inflation parameter of the fixed effects as anomaly events for the negative binomial setting and $c = \{1.08,\ 1.1,\ 1.12,\ 1.2,\ 1.22,\ 1.24\}$ for the Poisson setting, which approximately correspond to increases of the in-control timeslot mean values by $\{20\%,\ 30\%,\ 40\%,\ 50\%,\ 60\%,\ 80\%\}$, respectively.

For one of the eight negative binomial scenarios, Table 3 shows detection rates where each conditional detection rate is based on 1000 sample paths. As one would expect, the detection rate is larger when the inflation factor or duration of the fault is larger. We also see that the detection rate is larger for faults introduced earlier in the week (Monday) compared to later in the week (Friday). AILR and JLR are quite similar in their detection rates, and both are substantially better than EWMA. Also shown in Table 3 is the average detection time for all of the faults that are detected. Here again we see there is not much difference between AILR and JLR. Recall that in section 4.2 we anticipated this similarity between AILR and JLR.

Table 3. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 100,\ \sigma^2 = 0.01,\ \rho = 0.7$

Negative Binomial			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 6am	60 obs	1.05	30.5	0.24	30.3	0.25	18.8	0.03
		1.07	26.8	0.56	27.3	0.59	17.5	0.11
		1.085	22.7	0.8	24.8	0.71	12.7	0.19
		1.1	17.1	0.94	18.3	0.92	13.4	0.39
		1.13	9.2	1.00	10.2	1.00	10.4	0.74
		1.6	5.1	1.00	5.8	1.00	6.7	0.96
	120 obs	1.05	48.4	0.4	41.4	0.37	35.4	0.04
		1.07	40.8	0.77	38.9	0.75	27.5	0.11
		1.085	30.1	0.94	29.7	0.93	21.4	0.24
		1.1	20.9	0.99	21.3	0.99	18.7	0.39
		1.13	9.1	1.00	11.3	1.00	14.5	0.79
		1.16	5.1	1.00	5.5	1.00	7.9	0.98
Fri 12 pm	60 obs	1.05	20.8	0.2	19.2	0.18	10.3	0.04
		1.07	18.0	0.45	16.4	0.48	8.7	0.16
		1.085	19.1	0.71	13.1	0.73	7.3	0.32
		1.1	8.7	0.91	9.4	0.9	6.5	0.54
		1.13	3.3	0.99	4.1	1.00	4.1	0.91
		1.16	1.6	0.99	1.8	1.00	2.7	1.00
	120 obs	1.05	32.4	0.27	34.4	0.23	18.8	0.05
		1.07	23.7	0.6	29.1	0.57	10.2	0.18
		1.085	19.1	0.86	21.6	0.85	9.8	0.35
		1.1	14.3	0.94	10.4	0.95	8.4	0.56
		1.13	4.1	0.99	3.4	1.00	4.4	0.92
		1.16	1.9	1.00	1.7	1.00	2.8	1.00

Simulation results for the other scenarios are reported in Appendix C. Under scenarios with $\sigma^2 = 0.01$, power results consistently show the same finding, namely that the performances of AILR and JLR are nearly identical and far superior to EWMA. In other scenarios, the detection rates of the three tracking statistics are comparable, but AILR and JLR have faster detection time than EWMA.

Chapter 7

Implementation Aspects

Computation time is another essential property of online change-point detection algorithm that especially affects the applicability of the algorithm and detection power. In this section, we will report detailed breakdown of computation time for the proposed algorithm using AILR, JLR and EWMA tracking statistics. In our change-point detection algorithm, there are three main steps. Step 1 is to fit the GLMM and obtain estimated parameters for fixed and random effects of the in-control situation. With the estimated parameters on hand, step 2 is to simulate the sample path of AILR or JLR tracking statistics and extract the threshold H for AILR and JLR for a target FAR. Step 3 is to apply the threshold to the observation from live network stream and raise an alarm when the tracking statistic goes across the threshold. The first two steps are offline computation and step 3 is on-line process.

Under the simulation scenario described in section 6.1, average computation time of those 3 steps are shown in Table 4. Step 1 is performed by SAS 9.2 on a computer with

8GB installed memory and Intel® Core™ CPU @ 2.50GHz. Step 2 and 3 are R processes that run on a single core of a 4-core, 16GB memory with Intel Xeon CPU at a processing speed of 2.4GHz. Total number of cycles of historical data is chosen to be 30, which is the depth that yields desirable conditional FAR for AILR and JLR tracking statistics.

The computation time of step 1 is influenced by the depth of historical data and also number of time slots in each cycle. Less cycles of historical data or less time slots in each cycle will lead to less computation time to fit the GLMM. In the following table, 6 hours are the average computation time for 30 cycles of historical data with 168 timeslots in each cycle for the negative binomial scenarios, but it only needs 15 minutes on average for Poisson scenarios. This difference is due to the fact that the likelihood function of negative binomial scenario has one more dispersion parameter κ to be estimated and the dispersion parameter involves in multiple terms of the likelihood in a non-linear way which make the likelihood maximization more complicated than the Poisson scenarios

Comparing the two distributions settings in our simulation analysis, there is also a big time difference for step 2. It only takes about half an hour on average to get the threshold for AILR and JLR in the negative binomial settings. But in the Poisson scenarios, it takes 33 hours on average to obtain the threshold for JLR and 53 hours for AILR. This is due to the fact that Poisson scenarios have smaller counts so the tracking statistics do not reset as often as in negative binomial scenarios.

In step 3, both settings require negligible time, a small fraction of a second, to process each new observation even though negative binomial takes much less time than the

Poisson settings. Due to the fact that AILR need to do additional determinant calculation than JLR, it takes a little bit longer time in step 3.

Note the EWMA tracking statistic does not require updating the threshold at the end of each cycle. The threshold does not require parameter estimates so it only need to be obtained offline once using standardized normal random variables, which saves a portion of offline preparation time than AILR and JLR.

Table 4. Breakdown of Computation Time Using AILR, JLR and EWMA

Negative Binomial	AILR	JLR	EWMA
Step 1: SAS Estimation	6 hours	6 hours	6 hours
Step 2: Find H (1000 iterations)	35 min	32 min	--
Step 3: Process a new Observation	1.04E-03 sec	9.52E-04 sec	9.95E-05 sec
Poisson	AILR	JLR	EWMA
Step 1: SAS Estimation	2 hours	2 hours	2 hours
Step 2: Find H (1000 iterations)	2.7 hours	2.5 hours	--
Step 3: Process a new Observation	4.82E-03 sec	4.64E-03 sec	4.61E-04 sec

When the offline preparation time is very long, for example in our Poisson simulation setting, and there is a certain time period in each cycle that are unlikely to experience network problems or with a low cost of not detecting anomaly events, we suggest starting the offline calculation for the next monitoring cycle at the beginning of this period instead of waiting to the end of a cycle, and then use the data from corresponding time period in the last cycle to complete a new cycle for GLMM fitting and threshold simulation. Another option to reduce the offline preparation time is by using multiple processers in step 2. The iterations can be performed separately on different processors which would significantly cut down the computation time in a scenario like our Poisson simulation setting.

Chapter 8

Summary and Future work

We proposed a change-point detection method with three choices of tracking statistics, AILR, JLR and EWM to monitor non-stationary correlated discrete data in network traffic. Compared with Lambert and Liu's algorithm, our methods demonstrate satisfactory false alarm rates and good tolerance for negligible change in the non-stationary mean structure in network traffic. Based on simulation studies of size and power, we find that AILR and JLR tracking statistics demonstrate better performance than EWMA with competitively strong detection ability. Therefore, if historical data are enough to obtain a satisfactory conditional FAR, we would not recommend using EWMA tracking statistics. With the approximation of integrals in AILR tracking statistics, AILR and JLR only differ by a tiny number but it requires additional determinants computation. Moreover, according to our power analysis, the JLR has comparable fault detection ability of AILR. Based on these findings, we recommend JLR tracking statistic among all the three proposed ones.

Obviously, the assumptions of conditional distributions for network counts are not necessarily limited to negative binomial and Poisson distribution. AILR, JLR and

EWMA algorithms could be easily extended to other suitable distributions for real application. The assumption we have for constructing the GLMM is that data are correlated within the same cycle but independent from cycle to cycle. Further investigation could be placed on different correlation structures and also correlated cycle of data. One of the challenges for correlated cycles of data is that estimation of fixed and random effects become more complicated. An example of Poisson data with random effects correlated between cycles is demonstrated in Appendix A. Covariance structure of random effects which is required in the parameter estimation is calculated. It also hints the computation intensity of handling correlated cycles of data.

Simulation studies of different depth of historical data indicate that EWMA tracking statistics requires less historical data than AILR and JLR tracking statistics to obtain a desired false alarm rate. As shown in section 6.1, with 20 weeks of historical data, an average of 11% false alarm rate could be achieved using EWMA. However, AILR and JLR tracking statistics require 30 weeks of historical data to yield desired conditional FARs. This is partially due to the fact that EWMA use the parameter estimates from GLMM in a different way. Only in the monitoring week, conditional distributions of counts rely on the GLMM parameter estimates from historical data. In contrast, thresholds for AILR and JLR need to be obtained with parameter estimates from the GLMM, which makes the performance of these two tracking statistics heavily depend on depth of historical data.

Typical practitioners will not have any idea of how many cycles of historical would be needed to provide adequate conditional FAR. Suppose W cycles of in-control

historical data are available to the practitioner, we suggest the following procedure to determine the adequacy of the depth of historical data.

Algorithm 3: Sanity Check for Depth of Historical Data

---

1. Use all W cycles of historical data and fit the GLMM
2. Bootstrap M (we use $M = 25$) sets of W cycles of data from the fitted GLMM
3. Fit another GLMM to each of the M sets of data, find a threshold and calculate conditional FAR respectively for each of the M sets of data.
4. If those M conditional FARs satisfy the feasibility criteria introduced in section 6.1, then call W cycles is sufficient. Otherwise, gather more historical data and repeat this procedure.

Based on some findings of the size studies, we learned that if smooth fixed effects are used to generated data, which better represent the true nature, piecewise linear GLMM fit the data very well and yield satisfactory conditional FARs. A proposed method is as follows. In Algorithm 3 step 1, we suggest that a practitioner estimate the smooth fixed effects by the interpolation as in Lambert and Liu's method, and then fit a GLMM with piecewise linear fixed effects to obtain other nuisance parameters. Specifically, let $X_{ijw}, i = 1,2,..,m \;\; j = 1,2,...,n_j$, $w = 1,2,...,W$ denote the observation in timeslot i , time point j and historical cycle W . In step 1, suppose that the practitioner decides to use a Poisson GLMM with a log-link function for the means and AR(1) structure for the random effects. First, the overall mean $U_i, i = 1,2,...,m$ for each timeslot can be estimated by taking the average of observations in timeslot i from all W cycles. Using the interpolation method introduced in section 2.1, smooth mean β_{ij}^* for each time point can

be obtained. However, with a log-link function, the smooth fixed effect for each time point should be $\log(\beta_{ij}^{*})$. Second, assume the fixed effects are known and fit a piecewise linear Poisson GLMM and obtain the estimates for covariance parameter (σ^2, ρ) of the random effects. In step 2, generate M set of W cycles of data from the Poisson GLMM with smooth fixed effects and covariance parameter estimates obtained in step1. In step 3, fit a GLMM with non-smooth fixed effects such as piecewise linear and obtain conditional FARs. If the conditional FARs satisfy the feasibility criteria, the practitioner could claim the amount of historical data is sufficient.

Bibliography

Barford *et al*, (2002), "A Signal Analysis of Network Traffic Anomalies," in *Proceedings of ACM SIGC- OMM, Internet Measurement workshop*.

Basseville, M. and Nikiforov, L.V. (1993), *Detection of Abrupt Changes-Theory and Application*, Prentice-Hall, Inc.

Brutlag, J.D. (2000), "Aberrant Behavior Detection in Time Series for Network Monitoring," in Proceedings of the 14th Systems Administration Conference, New Orleans, Louisiana, 139-146.

Chen J. and Gupta, A.K. (2000), *Parametrical Statistical Change Point Analysis*, Birkhauser, Boston.

Feather, F.W., Siewiorek, D. and Maxion, R. (1993), "Fault Detection in Ethernet Networks Using Ano-maly Signature Matching," in Proceedings of the SIGCOMM93, San Franciscok CA, 279-288.

Montgomery, D.C. (1996), *Introduction to Statistical Quality Control*, 3rd edition, John Wiley & Sons, New York.

Hawkins, D.M. (1977), "Testing a Sequence of Observations for a Shift in Location," *Journal of the American Statistician Association*, 72, 180-186.

Hawkins, D.M., Qiu, P., and Kang C.W. (2003), "The Changepoint Model for Statistical Process Control", *Journal of Quality Technology*, 35, 355-366.

Hinkley, D.V. (1970), "Inference About the Change Point in a Sequence of Random Variables," Biometrika, 57, 1-17.

JAMES, B. et al. (1988), "Conditional Boundary Crossing Probabilities with Applications to Change-Point Problems," Annals Probability, 16, 825-839.

Jeske, D.R., *et al.* (2009), " CUSUM Techniques for Timeslot Sequences with Applications to Network Surveillance," Computational Statistics and Data Analysis, 53, 4332-4344.

Lambert, D. and Liu, C. (2006), "Adaptive Thresholds: Monitoring Streams of Network Counts," *Journal of the American Statistician Association*, 101, 78-88.

Lee, Y. and Nelder, J.A. (1996), "Hierarchical Generalized Linear Models (with discussion)," *Journal of Royal Statistical Society*, 58, 619--678.

Lucas, J.M. and Saccucci, M.S. (1990), "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements" (with discussion), *Technometrics*, 32, 1-29.

Lorden, G. (1971). "Procedures for reacting to a change in distribution," *Annals Mathematical Statistics*, 42, 1897-1908.

Montes de Oca, V. et al. (2010), "*A CUSUM Change-Point Detection Algorithm for Non-Stationary Sequences with Application to Network Surveillance*," *Journal of Software Systems*, 83, 1288-1298.

Moustakides, G. (1986), "Optimal procedures for detecting changes in distributions," *Annals Statistics*, 14, 1379-1387.

Page, E. S. (1954a), "Continuous Inspection Schemes," *Biometrika*, 41, 100-115.

Page, E. S. (1954b), "An Improvement to Wald's Approximation for Some Properties of Sequential Tests," *Journal of Royal Statistical Society*, B-16, 136-139.

Shiryaev, A. N. (1961), "The Problem of the Most Rapid Detection of a Disturbance in a Stationary Process," *Soviet Math. Dokl*., 2, 795-799.

Shiryaev, A.N. (1963), "On Optimum Methods in Quickest Detection Problems", *Theory of Probability and Its Applications*, 8, 22-46.

Ritov, Y. (1990). "Decision Theoretic Optimality of the CUSUM Procedure," *Annals of Statistics*, 18, 1464-1469.

Robinson, P.B. and Ho, T.Y. (1978), "Average Run Lengths of Geometric Moving Average Charts by Numerical Methods," *Technometrics*, 20, 85-93.

Roberts,S.W. (1966), "A Comparison of some control chart procedures", *Technometrics*, 8, 411-430.

Siegmund, D. (1985), *Sequential Analysis - Tests and Confidence Intervals*, Series in Statistics, Springer, New York.

Thottan, M. and Ji, C. (1998), "Adaptive Thresholding for Proactive Network Problem Detection", *Proceedings of the IEEE Third International Workshop on Systems Management*, 108-116.

Xie,Y., et al. (2013), "Change-Point Detection for High-Dimensional Time Series with Missing Data", *IEEE Journal of Selected Topics in Signal Processing*, 7, 12-26.

Vardeman, S. B. and Jobe, J. M. (1998), "*Statistical Quality Assurance Methods for Engineers*," New York: Wiley.

Wald, A. (1945), "Sequential Tests of Statistical Hypotheses," *Annals of Mathematical Statistics*, 16, 117-186.

Wald, A. and Wolfowitz, J. (1948), "Optimum Character of the Sequential Probability Ratio Test," *Annals of Mathematical Statistics*, 19, 326-339.

Wolfinger, R. (1993), ”Generalized Linear Mixed Models: A Pseudo-likelihood Approach,” *Journal of Statistical Computation and Simulation*, 48, 233-243.

Worsley. K. J. (1979), “On the Likelihood Ratio Test for a Shift in Location of Normal Populations”, *Journal of the American Statistician Association*, 74, 365-367

Appendix

A. GLMM and Covariance structure for Random Effects in Correlated-Cycle Context

Consider the situation that network traffic are not only correlated within each week but also correlated between weeks. A generalized linear mixed model can also be adopted for this kind of structure. To incorporate the correlation between weeks, the random effects will also be assumed to be correlated between weeks. In particular, suppose the network counts are modeled with a Poisson generalized linear mixed model, which means for the counts Y_{tk} ,which matches the i th observation in the j th timeslot of the k th week, $i = 1,2,\ldots I$ $j = 1,2,\ldots J$, $k = 1,2,\ldots K$ given the mean μ_{ijk} , the counts are independently following Poisson distribution with mean μ_{ijk} . The link function of the Poisson GLMM is

$$\log\left(\mu_{ijk}\right) = \beta + s_{j(k)}$$

where we assume the simplest case that all timeslot have the same fixed effect β , $s_{j(k)}$ is the random effect for timeslot i of week k . For different cycles, the vectors of random effect $\left(s_{1(k)}, s_{2(k)}, \ldots, s_{J(k)}\right)'$ form a multivariate time series

$$\begin{pmatrix} s_{1(k)} \\ s_{2(k)} \\ \vdots \\ s_{J(k)} \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 & 0 & 0 \\ 0 & \alpha_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \alpha_J \end{pmatrix} \begin{pmatrix} s_{1(k)} \\ s_{2(k)} \\ \vdots \\ s_{J(k)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1(k)} \\ \varepsilon_{2(k)} \\ \vdots \\ \varepsilon_{J(k)} \end{pmatrix}$$

Suppose the error term have an AR(1) type of covariance structure, then the random effects within each week and between weeks are all correlated, which would make the counts have correlation within week and also between week. The covariance structure of the error term could also take on other forms in application, here we would like to have a structure that make the correlation decreases as the time distance increase within a week and at the same time, the correlation of random effects from the same timeslot of different week have higher correlation than the that of random effects from different timeslot of different week.

Next, we can have a more clear idea of how the correlation structure looks like. Let $\underset{\sim}{\varepsilon}_k$ denote the error vector with variance-covariance matrix Ω, $\underset{\sim}{s}_k$ denote the random vector of cycle k with variance-covariance matrix Σ,

$$\Omega = E\left(\underset{\sim}{\varepsilon}_k \underset{\sim}{\varepsilon}_k'\right) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{J-1} \\ \rho & 1 & \rho & \cdots & \rho^{J-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{J-2} & \rho^{J-3} & \rho^{J-4} & \cdots & \rho \\ \rho^{J-1} & \rho^{J-2} & \rho^{J-3} & \cdots & 1 \end{pmatrix}$$

$$\Sigma = E\left(\underset{\sim}{s}_k \underset{\sim}{s}_k'\right) = \begin{pmatrix} \operatorname{cov}\left(s_{1k}, s_{1k}\right) & \operatorname{cov}\left(s_{1k}, s_{2k}\right) & \cdots & \operatorname{cov}\left(s_{1k}, s_{Jk}\right) \\ & \operatorname{cov}\left(s_{2k}, s_{2k}\right) & \cdots & \operatorname{cov}\left(s_{1k}, s_{Jk}\right) \\ \vdots & \vdots & \ddots & \vdots \\ & & & \operatorname{cov}\left(s_{Jk}, s_{Jk}\right) \end{pmatrix}$$

and A denote the coefficient matrix in the multivariate time series ,

$$\mathrm{A}=\begin{pmatrix}\alpha_1 & 0 & 0 & 0\\ 0 & \alpha_2 & 0 & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \alpha_J\end{pmatrix}$$

Take variance on both sides of the time series we have

$$\Sigma=\mathrm{A}\Sigma\mathrm{A}'+\Omega$$

To obtain a analytical form of Σ, a proposition could be used here:

Proposition: Let $A,B,$and C be matrices whose dimensions are such that the product of ABC exist. Then

$$vec(ABC)=(C'\otimes A)\bullet vec(B)$$

Where $\otimes$is Kronecker Product, vec is the vector operator that if A is an $(m\times n)$matrix, then $vec(A)$ is an $(mn\times 1)$column vector obtained by stacking the columns of A ,one below the other, with the column ordered from left to right, i.e.

$$A=\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\\ a_{31} & a_{32}\end{pmatrix}\qquad vec(A)=\begin{pmatrix}a_{11}\\ a_{21}\\ a_{31}\\ a_{12}\\ a_{22}\\ a_{32}\end{pmatrix}$$

Apply this proposition to $\Sigma=\mathrm{A}\Sigma\mathrm{A}'+\Omega$ we have

$$vec(\Sigma)=(\mathrm{A}\otimes\mathrm{A})\,vec(\Sigma)+vec(\Omega)$$

So that

$$vec(\Sigma)=(I-\mathrm{A}\otimes\mathrm{A})^{-1}\,vec(\Omega)$$

With

$$I - \mathrm{A} \otimes \mathrm{A} = \begin{pmatrix} 1-\alpha_1\alpha_1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1-\alpha_1\alpha_J & & & & & & & \\ & & & 1-\alpha_1\alpha & & & & & & \\ & & & & \ddots & & & & & \\ & & & & & 1-\alpha_1\alpha_J & & & & \\ & & & & & & \ddots & & & \\ & & & & & & & 1-\alpha_J\alpha_1 & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & 1-\alpha_J\alpha_J \end{pmatrix}$$

and

$$vec(\Omega) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{J-1} & \rho & 1 & \cdots & \rho^{J-2} & \ldots & \rho^{J-1} & \rho^{J-2} & \cdots & 1 \end{pmatrix}'$$

We could get the formula of the (i, j) element of Σ as follows,

$$\mathrm{cov}\left(s_{ik}, s_{jk}\right) = \frac{\rho^{|i-j|}}{1-\alpha_i\alpha_j}\sigma^2$$

In the parameter estimation stage for this GLMM model, suppose the total number of historical data available is K. Then the covariance matrix for all random effect is

$$D = E\left\{\left(s_{11}, s_{12}, \ldots s_{1J}, \ldots s_{1K}, s_{2K}, \ldots s_{JK}\right)' \left(s_{11}, s_{12}, \ldots s_{1J}, \ldots s_{1K}, s_{2K}, \ldots s_{JK}\right)\right\}$$

It could be written in block of matrices that could be expressed by Σ and A that we defined before as follow

$$D = \begin{pmatrix} E\left(\underset{\sim}{s}_1\underset{\sim}{s}_1'\right) & E\left(\underset{\sim}{s}_1\underset{\sim}{s}_2'\right) & \cdots & E\left(\underset{\sim}{s}_1\underset{\sim}{s}_K'\right) \\ E\left(\underset{\sim}{s}_2\underset{\sim}{s}_1'\right) & E\left(\underset{\sim}{s}_2\underset{\sim}{s}_2'\right) & \cdots & E\left(\underset{\sim}{s}_2\underset{\sim}{s}_K'\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(\underset{\sim}{s}_K\underset{\sim}{s}_1'\right) & E\left(\underset{\sim}{s}_K\underset{\sim}{s}_2'\right) & \cdots & E\left(\underset{\sim}{s}_K\underset{\sim}{s}_K'\right) \end{pmatrix}$$

Since we know that

$$E\left(\underset{\sim}{s}_t \underset{\sim}{s}_t'\right) = \Sigma$$

$$E\left(\underset{\sim}{s}_t \underset{\sim}{s}_{t-u}'\right) = E\left(\left(\mathrm{A}^u \underset{\sim}{s}_{t-u} + \underset{\sim}{\varepsilon}_t + \underset{\sim}{\varepsilon}_{t-1} + \cdots \underset{\sim}{\varepsilon}_{t-u+1}\right) \underset{\sim}{s}_{t-u}'\right)$$

$$= E\left(\left(\mathrm{A}^u \underset{\sim}{s}_{t-u}\right) \underset{\sim}{s}_{t-u}'\right) = \mathrm{A}^u \Sigma$$

The covariance matrix for all random effects in the historical weeks could be written as

$$D = \begin{pmatrix} \Sigma & \mathrm{A}\Sigma & \cdots & \mathrm{A}^{K-1}\Sigma \\ \mathrm{A}\Sigma & \Sigma & \cdots & \mathrm{A}^{K-2}\Sigma \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{A}^{K-1}\Sigma & \mathrm{A}^{K-2}\Sigma & \cdots & \Sigma \end{pmatrix}$$

B. Pseudo-likelihood

In GLMM , consider a vector of observations $\underset{\sim}{y}$ of length n with mean $\underset{\sim}{\mu}$

$$\underset{\sim}{y} = \underset{\sim}{\mu} + \underset{\sim}{\varepsilon}$$

and a differentiable monotonic link function $g()$ such that

$$g(\underset{\sim}{\mu}) = X\underset{\sim}{\alpha} + Z\underset{\sim}{\beta}$$

Here $\underset{\sim}{\alpha}$ is a vector of unknown fixed effects in the link function whereas $\underset{\sim}{\beta}$ is a vector of unknown random effects with mean $\underset{\sim}{0}$ and unknown covariance matrix D. Given $\underset{\sim}{\mu}$, the error term $\underset{\sim}{\varepsilon}$ have mean $\underset{\sim}{0}$ and variance

$$\text{var}(\underset{\sim}{\varepsilon}) = R_{\underset{\sim}{\mu}}^{1/2} R R_{\underset{\sim}{\mu}}^{1/2}$$

For example, if we assume the observations follow a Poisson distribution, then

$$\text{var}(\underset{\sim}{\varepsilon}) = diag(\underset{\sim}{\mu})$$

and we have

$$R_{\underset{\sim}{\mu}} = diag(\underset{\sim}{\mu}),\ R = I_{n\times n}.$$

When the random effects are highly correlated, the marginal likelihood function of $\underset{\sim}{y}$ involves high dimension integration of the random effects. The merit of pseudo-likelihood is to estimate the fixed effect $\underset{\sim}{\alpha}$ and the variance-covariance matrix for random effect without the direct computation of high dimension integration. The pseudo-likelihood is constructed by several analytic and probabilistic approximations of the exact likelihood and parameters will be iteratively estimated.

The first approximation is an analytic approximation of the unknown mean vector $\underset{\sim}{\mu}$ using first order Taylor expansion.

Suppose $\hat{\underset{\sim}{\alpha}}$ and $\hat{\underset{\sim}{\beta}}$ are known estimates of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$, then estimate of $\underset{\sim}{\mu}$ through the known link function $g()$ is,

$$\hat{\underset{\sim}{\mu}} = g^{-1}\left(X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}}\right)$$

which is a vector consisting of evaluations of function g^{-1} at each component of the vector $X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}}$. The first order Taylor expansion of $\underset{\sim}{\mu}$ at value $\hat{\underset{\sim}{\mu}}$ is as follow

$$\underset{\sim}{\mu} = g^{-1}(X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}}) \approx \hat{\underset{\sim}{\mu}} + \left[(g^{-1})'(X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}})\right] \bullet (X\underset{\sim}{\alpha} + Z\underset{\sim}{\beta} - X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}})$$

where $(g^{-1})'(X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}})$ is a diagonal matrix with the i-th element as $(g^{-1})'(X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}})$. Here X_i is the i-th row vector in the X matrix, Z_i is the i-th row vector in the Z matrix.

The second approximation to obtain pseudo-likelihood is to apply a probabilistic approximation to the error term with a Normal distribution. Let $\hat{\underset{\sim}{\varepsilon}}$ be the approximation to $\underset{\sim}{\varepsilon}$ which satisfies $\underset{\sim}{y} = \underset{\sim}{\mu} + \underset{\sim}{\varepsilon}$ using the Taylor approximation of $\underset{\sim}{\mu}$, we have

$$\hat{\underset{\sim}{\varepsilon}} = \underset{\sim}{y} - \hat{\underset{\sim}{\mu}} - (g^{-1})'(X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}})(X\underset{\sim}{\alpha} + Z\underset{\sim}{\beta} - X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}})$$

Following Laird and Louis (1982) and Lindstrom and Bates (1990), approximate the conditional distribution of $\hat{\underset{\sim}{\varepsilon}}$ given $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ with a Normal Distribution which have the same first and second moments as $\underset{\sim}{\varepsilon} \mid \underset{\sim}{\alpha}, \underset{\sim}{\beta}$. In particular, we assume that $\underset{\sim}{\varepsilon} \mid \underset{\sim}{\alpha}, \underset{\sim}{\beta}$ is $Normal\left(\underset{\sim}{0}, R_{\underset{\sim}{\mu}}^{1/2} R R_{\underset{\sim}{\mu}}^{1/2}\right)$.

For each component $X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}}$ in $X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}}$, we could write

$$g(\hat{\mu}_i) = X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}}$$

So

$$(g^{-1})'(X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}}) = (g^{-1})'\left(g(\hat{\mu}_i)\right)$$

Also, with simple derivation and transformation, we know that

$$1 = \frac{\partial\hat{\mu}_i}{\hat{\mu}_i} = \frac{\partial g^{-1}(g(\hat{\mu}_i))}{\hat{\mu}_i} = (g^{-1})'(g(\hat{\mu}_i))\bullet g'(\hat{\mu}_i)$$

Then

$$(g^{-1})'(X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}}) = (g^{-1})'(g(\hat{\mu}_i)) = \frac{1}{g'(\hat{\mu}_i)}$$

Put this relationship back into the Taylor approximation of $\hat{\underset{\sim}{\varepsilon}}$, for each component $\hat{\underset{\sim}{\varepsilon}}_i$ we could write

$$\hat{\underset{\sim}{\varepsilon}}_i = \underset{\sim}{y}_i - \hat{\underset{\sim}{\mu}}_i - \frac{1}{g'(X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}})}(X_i\underset{\sim}{\alpha} + Z_i\underset{\sim}{\beta} - X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}})$$

Move the part with $\underset{\sim}{y}_i - \hat{\underset{\sim}{\mu}}_i$ to the one side and multiply both sides by $g'(X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}})$, i.e. $g'(\hat{\mu}_i)$, we have

$$g'(\hat{\mu}_i)(\underset{\sim}{y}_i - \hat{\underset{\sim}{\mu}}_i) = g'(\hat{\mu}_i)\hat{\underset{\sim}{\varepsilon}}_i + (X_i\underset{\sim}{\alpha} + Z_i\underset{\sim}{\beta} - X_i\hat{\underset{\sim}{\alpha}} + Z_i\hat{\underset{\sim}{\beta}})$$

We can write

$$g'(\hat{\underset{\sim}{\mu}})(\underset{\sim}{y} - \hat{\underset{\sim}{\mu}}) = g'(\hat{\mu})\hat{\underset{\sim}{\varepsilon}} + (X\underset{\sim}{\alpha} + Z\underset{\sim}{\beta} - X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}})$$

As we assumed that given $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$, $\hat{\underset{\sim}{\varepsilon}}$ follows $Normal\left(\underset{\sim}{0}, R_{\underset{\sim}{\mu}}^{1/2} R R_{\underset{\sim}{\mu}}^{1/2}\right)$, now the left hand side would also follow Normal distribution

$$g'(\hat{\underset{\sim}{\mu}})\left(\underset{\sim}{y} - \hat{\underset{\sim}{\mu}}\right) | \underset{\sim}{\alpha}, \underset{\sim}{\beta} \sim Normal\left(X\underset{\sim}{\alpha} + Z\underset{\sim}{\beta} - X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}}, \quad g'(\hat{\underset{\sim}{\mu}}) R_{\underset{\sim}{\mu}}^{1/2} R R_{\underset{\sim}{\mu}}^{1/2} g'(\hat{\underset{\sim}{\mu}})\right)$$

where $g'(\hat{\underset{\sim}{\mu}})$ is a diagonal matrix with the i-th element as $g'(\hat{\underset{\sim}{\mu}}_i)$.

The last approximation, which is related to the iteratively estimating parameters fixed and random effects, is to use $\hat{\underset{\sim}{\mu}}$ for $\underset{\sim}{\mu}$ in the variance matrix $R_{\underset{\sim}{\mu}}^{1/2} R R_{\underset{\sim}{\mu}}^{1/2}$. Now if substitute $X\hat{\underset{\sim}{\alpha}} + Z\hat{\underset{\sim}{\beta}}$ in the mean of Normal distribution with $g(\hat{\underset{\sim}{\mu}})$ and define

$$\underset{\sim}{y} = g(\hat{\underset{\sim}{\mu}}) + g'(\hat{\underset{\sim}{\mu}})\left(\underset{\sim}{y} - \hat{\underset{\sim}{\mu}}\right).$$

It could be verified that

$$\underset{\sim}{y} | \underset{\sim}{\alpha}, \underset{\sim}{\beta} \sim Normal\left(X\underset{\sim}{\alpha} + Z\underset{\sim}{\beta}, \quad g'(\hat{\underset{\sim}{\mu}}) R_{\underset{\sim}{\mu}}^{1/2} R R_{\underset{\sim}{\mu}}^{1/2} g'(\hat{\underset{\sim}{\mu}})\right)$$

The Gaussian log-likelihood corresponding to the linear mixed model for v is

$$l(\underset{\sim}{\alpha}, D, R, \underset{\sim}{y}) = -\frac{1}{2}\log|V| - \frac{1}{2}\left(\underset{\sim}{y} - X\underset{\sim}{\alpha}\right)' V^{-1}\left(\underset{\sim}{y} - X\underset{\sim}{\alpha}\right) - \frac{n}{2}\log 2\pi$$

where

$$V = g'(\hat{\underset{\sim}{\mu}}) R_{\underset{\sim}{\mu}}^{1/2} R R_{\underset{\sim}{\mu}}^{1/2} g'(\hat{\underset{\sim}{\mu}}) + ZDZ'$$

The log likelihood can be maximized analytically for $\underset{\sim}{\alpha}$, resulting the log likelihood as follow

$$l(D, R; \underset{\sim}{y}) = -\frac{1}{2}\log|V| - \frac{1}{2}\underset{\sim}{\gamma}' V^{-1}\underset{\sim}{\gamma} - \frac{n}{2}\left(1 + \log\left(2\pi / n\right)\right)$$

where

$$\underset{\sim}{\gamma} = \underset{\sim}{y} - X\left(X'V^{-1}X\right)^{-1}X'V^{-1}\underset{\sim}{y}$$

In summary, pseudo-likelihood approach is to approximate the exact likelihood function of $\underset{\sim}{y}$ with an analytic approximation of the unknown mean vector $\underset{\sim}{\mu}$, a probabilistic approximation to the error term with a Normal distribution and numeric approximation of $\underset{\sim}{\mu}$ to avoid high-dimension integral computation for the exact likelihood function of $\underset{\sim}{y}$. Combined with numerical methods, the final goal of estimating the fixed effects and covariance parameter for random effects will be achieved.

C. Power Study Results

Table 5. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 100,\ \sigma^2 = 0.001,\ \rho = 0.7$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.05	22.8	0.92	19.2	0.97	28.1	0.75
		1.07	13.8	1.00	10.2	1.00	15.7	0.99
		1.085	8.8	1.00	7.2	1.00	10.3	1.00
		1.1	6.6	1.00	5.1	1.00	7.1	1.00
		1.13	4.1	1.00	3.3	1.00	4.4	1.00
		1.6	3.0	1.00	2.3	1.00	3.2	1.00
	120 obs	1.05	31.7	1.00	22.0	0.98	35.7	0.89
		1.07	14.1	1.00	10.7	1.00	16.5	1.00
		1.085	9.2	1.00	7.5	1.00	9.9	1.00
		1.1	6.7	1.00	5.1	1.00	7.0	1.00
		1.13	4.2	1.00	3.6	1.00	4.4	1.00
		1.16	3.0	1.00	2.3	1.00	3.2	1.00
Fri 12 am	60 obs	1.05	16.5	0.97	15.8	0.98	24.6	0.81
		1.07	6.4	1.00	6.0	1.00	13.2	1.00
		1.085	3.7	1.00	3.2	1.00	8.8	1.00
		1.1	2.6	1.00	2.5	1.00	4.3	1.00
		1.13	1.7	1.00	1.7	1.00	2.4	1.00
		1.16	1.2	1.00	1.2	1.00	1.6	1.00
	120 obs	1.05	21.4	0.99	20.8	0.99	28.9	0.93
		1.07	6.7	1.00	6.6	1.00	13.1	1.00
		1.085	3.8	1.00	3.9	1.00	8.8	1.00
		1.1	2.6	1.00	2.9	1.00	4.5	1.00
		1.13	1.7	1.00	1.6	1.00	2.1	1.00
		1.16	1.2	1.00	1.2	1.00	1.6	1.00

Table 6. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 100,\ \sigma^2 = 0.01,\ \rho = 0.4$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.05	32.1	0.32	32.0	0.39	27.7	0.05
		1.07	30.2	0.72	28.4	0.77	24.8	0.14
		1.085	25.2	0.90	22.8	0.94	26.5	0.30
		1.1	20.3	0.98	17.6	0.99	23.7	0.48
		1.13	12.2	1.00	9.5	1.00	17.7	0.86
		1.6	8.1	1.00	5.7	1.00	10.4	1.00
	120 obs	1.05	53.0	0.55	51.4	0.61	50.4	0.09
		1.07	41.3	0.95	36.8	0.96	51.0	0.21
		1.085	27.7	0.99	25.0	1.00	39.5	0.43
		1.1	20.7	0.99	17.3	1.00	37.1	0.64
		1.13	12.4	1.00	9.8	1.00	24.1	0.96
		1.16	7.9	1.00	5.8	1.00	10.4	1.00
Fri 12 am	60 obs	1.05	28.4	0.32	25.0	0.33	19.2	0.05
		1.07	23.1	0.71	20.5	0.78	17.7	0.13
		1.085	16.3	0.93	14.9	0.94	15.4	0.27
		1.1	10.3	0.99	9.5	1.00	14.4	0.45
		1.13	4.2	1.00	3.4	1.00	9.9	0.86
		1.16	2.5	1.00	1.8	1.00	5.1	1.00
	120 obs	1.05	44.4	0.46	40.9	0.47	36.7	0.05
		1.07	32.8	0.91	29.0	0.88	29.5	0.18
		1.085	20.4	0.98	17.1	0.98	27.9	0.33
		1.1	11.2	1.00	9.83	1.00	20.6	0.55
		1.13	4.2	1.00	3.4	1.00	12.1	0.90
		1.16	2.5	1.00	1.6	1.00	4.9	1.00

Table 7. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 100,\ \sigma^2 = 0.001,\ \rho = 0.4$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.05	25.2	0.91	26.5	0.90	27.5	0.86
		1.07	12.6	1.00	13.4	1.00	15.3	0.92
		1.085	8.8	1.00	8.9	1.00	10.0	1.00
		1.1	6.5	1.00	6.5	1.00	7.1	1.00
		1.13	4.2	1.00	4.1	1.00	4.5	1.00
		1.6	2.9	1.00	3.0	1.00	3.2	1.00
	120 obs	1.05	29.4	0.98	29.5	1.00	33.8	0.97
		1.07	13.3	1.00	13.3	1.00	15.2	1.00
		1.085	9.0	1.00	8.5	1.00	10.1	1.00
		1.1	6.4	1.00	6.4	1.00	7.3	1.00
		1.13	4.1	1.00	4.2	1.00	4.5	1.00
		1.16	3.0	1.00	3.1	1.00	3.1	1.00
Fri 12 am	60 obs	1.05	17.5	0.93	16.4	0.92	15.4	0.87
		1.07	6.5	1.00	6.6	1.00	6.4	0.98
		1.085	3.6	1.00	3.8	1.00	4.3	1.00
		1.1	2.6	1.00	2.6	1.00	3.2	1.00
		1.13	1.7	1.00	1.6	1.00	2.2	1.00
		1.16	1.2	1.00	1.2	1.00	1.7	1.00
	120 obs	1.05	21.0	0.99	21.1	0.96	17.7	0.88
		1.07	6.4	1.00	6.5	1.00	6.5	1.00
		1.085	3.6	1.00	3.9	1.00	4.2	1.00
		1.1	2.5	1.00	2.6	1.00	3.1	1.00
		1.13	1.7	1.00	1.6	1.00	2.2	1.00
		1.16	1.2	1.00	1.2	1.00	1.7	1.00

Table 8. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 50,\ \sigma^2 = 0.01,\ \rho = 0.7$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.05	32.5	0.36	32.8	0.29	26.1	0.05
		1.07	27.7	0.82	28.5	0.78	22.6	0.14
		1.085	24.7	0.91	23.4	0.94	16.8	0.27
		1.1	19.2	0.95	18.1	0.99	15.3	0.48
		1.13	10.2	1.00	11.2	1.00	11.8	0.79
		1.6	5.9	1.00	7.3	1.00	7.9	0.97
	120 obs	1.05	53.6	0.65	51.4	0.61	38.1	0.08
		1.07	42.0	0.89	36.8	0.86	29.9	0.15
		1.085	31.2	0.43	25.0	1.00	26.2	0.32
		1.1	22.4	0.82	17.3	1.00	23.5	0.50
		1.13	11.2	0.95	9.8	1.00	15.7	0.85
		1.16	6.2	1.00	5.8	1.00	12.8	0.07
Fri 12 am	60 obs	1.05	21.5	0.31	25.0	0.33	11.1	0.22
		1.07	17.2	0.59	20.5	0.78	8.3	0.38
		1.085	12.8	0.82	14.9	0.94	7.1	0.62
		1.1	9.4	0.95	9.5	0.98	5.0	0.93
		1.13	3.8	1.00	3.4	1.00	3.3	1.00
		1.16	2.2	1.00	1.8	1.00	14.3	0.43
	120 obs	1.05	36.9	0.36	40.9	0.47	12.7	0.20
		1.07	28.3	0.68	29.1	0.88	26.2	0.07
		1.085	20.5	0.96	17.1	0.98	14.9	0.23
		1.1	12.1	0.98	9.8	1.00	9.3	0.61
		1.13	4.1	1.00	3.4	1.00	5.3	0.93
		1.16	2.0	1.00	1.6	1.00	3.2	1.00

Table 9. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 50,\ \sigma^2 = 0.001,\ \rho = 0.7$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.05	27.9	0.80	27.9	0.83	28.1	0.75
		1.07	15.7	0.99	15.8	1.0	15.7	0.85
		1.085	10.4	1.00	10.5	1.00	9.3	1.00
		1.1	7.9	1.00	7.5	1.00	7.1	1.00
		1.13	4.9	1.00	4.9	1.00	5.4	1.00
		1.6	3.6	1.00	3.5	1.00	4.2	1.00
	120 obs	1.05	35.4	0.97	34.8	0.96	35.7	0.89
		1.07	16.6	1.00	16.4	1.00	18.5	1.00
		1.085	10.5	1.00	10.8	1.00	11.9	1.00
		1.1	7.8	1.00	7.7	1.00	10.0	1.00
		1.13	5.0	1.00	4.9	1.00	7.0	1.00
		1.16	3.6	1.00	3.5	1.00	4.4	1.00
Fri 12 am	60 obs	1.05	19.4	0.80	19.5	0.89	22.3	0.77
		1.07	8.8	0.99	8.8	1.00	11.4	0.93
		1.085	5.1	1.00	5.0	1.00	8.1	1.00
		1.1	3.5	1.00	3.5	1.00	6.7	1.00
		1.13	2.2	1.00	2.1	1.00	3.4	1.00
		1.16	1.6	1.00	1.6	1.00	2.3	1.00
	120 obs	1.05	25.0	0.89	25.1	0.89	29.7	0.81
		1.07	8.6	0.99	8.7	1.00	13.5	0.98
		1.085	5.0	1.00	5.0	1.00	9.2	1.00
		1.1	3.5	1.00	3.5	1.00	6.3	1.00
		1.13	2.2	1.00	2.2	1.00	3.3	1.00
		1.16	1.6	1.00	1.5	1.00	2.1	1.00

Table 10. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 50,\ \sigma^2 = 0.01,\ \rho = 0.4$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.05	32.8	0.35	33.8	0.38	33.0	0.06
		1.07	28.5	0.72	28.9	0.72	30.5	0.17
		1.085	22.8	0.91	21.9	0.91	26.3	0.31
		1.1	20.0	0.99	18.4	0.99	25.9	0.53
		1.13	12.3	1.00	10.8	1.00	19.1	0.88
		1.6	6.7	1.00	6.5	1.00	11.1	0.99
	120 obs	1.05	52.7	0.62	50.3	0.62	53.6	0.08
		1.07	37.1	0.95	35.2	0.98	48.5	0.20
		1.085	24.5	1.00	28.6	1.00	43.7	0.47
		1.1	18.6	1.00	18.2	1.00	38.3	0.77
		1.13	14.1	1.00	13.5	1.00	24.0	0.91
		1.16	6.3	1.00	6.0	1.00	10.4	1.00
Fri 12 am	60 obs	1.05	30.9	0.36	23.2	0.37	19.5	0.07
		1.07	25.5	0.75	21.9	0.74	16.9	0.18
		1.085	17.2	0.93	16.3	0.94	16.2	0.34
		1.1	10.7	0.99	7.9	1.00	14.7	0.53
		1.13	4.1	1.00	4.1	1.00	9.2	0.89
		1.16	2.5	1.00	2.2	1.00	5.4	1.00
	120 obs	1.05	46.2	0.43	41.7	0.47	42.4	0.11
		1.07	34.2	0.86	29.4	0.89	33.5	0.29
		1.085	23.2	0.97	18.3	0.99	24.2	0.70
		1.1	14.0	1.00	10.4	1.00	21.3	0.97
		1.13	6.2	1.00	4.0	1.00	10.0	1.00
		1.16	3.5	1.00	2.1	1.00	4.7	1.00

Table 11. Performance Comparison of AILR, JLR and EWMA in the Negative Binomial Setting with $\kappa = 50,\ \sigma^2 = 0.001,\ \rho = 0.4$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.05	27.3	0.85	26.7	0.88	29.3	0.75
		1.07	14.3	0.99	11.4	1.00	15.7	0.85
		1.085	10.9	1.00	8.4	1.00	9.3	1.00
		1.1	7.4	1.00	6.0	1.00	7.1	1.00
		1.13	4.9	1.00	4.6	1.00	5.4	1.00
		1.6	3.6	1.00	3.6	1.00	4.2	1.00
	120 obs	1.05	32.0	0.98	32.1	0.98	35.7	0.89
		1.07	14.9	0.99	11.6	1.00	18.5	1.00
		1.085	10.0	0.99	10.1	1.00	11.9	1.00
		1.1	7.6	1.00	5.8	1.00	10.0	1.00
		1.13	4.8	1.00	3.9	1.00	7.0	1.00
		1.16	3.6	1.00	3.5	1.00	4.4	1.00
Fri 12 am	60 obs	1.05	18.0	0.86	14.3	0.89	22.3	0.77
		1.07	8.3	0.99	8.3	0.99	11.4	0.93
		1.085	5.1	0.99	5.0	1.00	8.1	1.00
		1.1	3.6	0.99	3.6	1.00	6.7	1.00
		1.13	2.2	0.99	2.2	1.00	3.4	1.00
		1.16	1.6	1.00	1.6	1.00	2.3	1.00
	120 obs	1.05	21.5	0.91	22.4	0.91	29.7	0.81
		1.07	8.2	0.99	8.5	1.00	13.5	0.98
		1.085	5.0	1.00	4.9	1.00	9.2	1.00
		1.1	3.6	1.00	3.5	1.00	6.3	1.00
		1.13	2.3	1.00	2.2	1.00	3.3	1.00
		1.16	1.8	1.00	1.7	1.00	2.1	1.00

Table 12. Performance Comparison of AILR, JLR and EWMA in the Poisson Setting with $\sigma^2 = 0.01,\ \rho = 0.7$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.08	27.2	0.41	28.2	0.37	18.0	0.05
		1.1	24.0	0.50	25.0	0.52	15.0	0.08
		1.12	20.0	0.79	21.4	0.76	14.2	0.19
		1.2	6.0	1.00	6.4	1.00	7.1	0.92
		1.22	4.7	1.00	5.0	1.00	5.3	1.00
		1.24	3.9	1.00	3.8	1.00	4.3	1.00
	120 obs	1.08	38.4	0.54	42.8	0.51	34.0	0.03
		1.1	35.8	0.72	37.9	0.68	20.4	0.08
		1.12	22.4	0.87	29.5	0.88	27.1	0.20
		1.2	5.6	1.00	6.4	1.00	8.3	0.98
		1.22	4.6	1.00	4.8	1.00	6.4	1.00
		1.24	3.9	1.00	3.88	1.00	5.3	1.00
Fri 12 am	60 obs	1.08	16.7	0.33	16.6	0.29	8.6	0.06
		1.1	13.2	0.49	15.7	0.45	8.8	0.13
		1.12	10.7	0.71	13.0	0.67	7.5	0.33
		1.2	2.4	1.00	2.6	1.00	4.1	1.00
		1.22	2.0	1.00	1.9	1.00	3.5	1.00
		1.24	1.6	1.00	1.5	1.00	3.2	1.00
	120 obs	1.08	22.6	0.37	25.4	0.33	11.9	0.07
		1.1	23.5	0.50	20.8	0.49	9.7	0.15
		1.12	10.4	0.76	17.3	0.73	8.6	0.35
		1.2	2.6	1.00	3.1	1.00	4.1	1.00
		1.22	2.0	1.00	1.9	1.00	3.6	1.00
		1.24	1.6	1.00	1.5	1.00	3.2	1.00

Table 13. Performance Comparison of AILR, JLR and EWMA in the Poisson Setting with $\sigma^2 = 0.001,\ \rho = 0.7$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.08	28.2	0.72	26.6	0.80	25.4	0.81
		1.1	22.2	0.93	20.9	0.95	21.4	0.94
		1.12	11.9	1.00	12.0	1.00	11.3	1.00
		1.2	3.6	1.00	3.7	1.00	5.0	1.00
		1.22	3.1	1.00	3.1	1.00	4.5	1.00
		1.24	2.7	1.00	2.6	1.00	4.0	1.00
	120 obs	1.08	38.4	0.82	34.3	0.92	38.7	0.91
		1.1	24.3	0.98	21.7	1.00	28.9	0.98
		1.12	12.3	1.00	12.6	1.00	12.9	1.00
		1.2	3.6	1.00	3.7	1.00	5.1	1.00
		1.22	3.0	1.00	3.1	1.00	4.4	1.00
		1.24	2.7	1.00	2.6	1.00	4.1	1.00
Fri 12 am	60 obs	1.08	37.0	0.81	29.2	0.79	23.2	0.76
		1.1	14.8	0.92	14.5	0.91	17.8	0.88
		1.12	7.4	0.99	7.9	0.99	8.0	1.00
		1.2	1.6	1.00	1.7	1.00	3.2	1.00
		1.22	1.4	1.00	1.4	1.00	3.0	1.00
		1.24	1.2	1.00	1.2	1.00	2.8	1.00
	120 obs	1.08	20.1	0.69	23.3	0.76	23.0	0.76
		1.1	16.1	0.90	15.9	0.93	18.8	0.94
		1.12	7.1	0.99	7.5	1.00	7.1	1.00
		1.2	1.6	1.00	1.7	1.00	4.2	1.00
		1.22	1.4	1.00	1.4	1.00	4.0	1.00
		1.24	1.2	1.00	1.2	1.00	2.8	1.00

Table 14. Performance Comparison of AILR, JLR and EWMA in the Poisson Setting with $\sigma^2 = 0.01,\ \rho = 0.4$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.08	28.5	0.41	30.0	0.45	25.5	0.04
		1.1	25.0	0.62	27.5	0.61	26.5	0.08
		1.12	19.1	0.83	23.4	0.84	25.7	0.22
		1.2	5.9	1.00	6.7	1.00	10.1	0.99
		1.22	4.5	1.00	5.1	1.00	7.0	1.00
		1.24	3.9	1.00	4.1	1.00	5.1	1.00
	120 obs	1.08	38.8	0.67	40.9	0.63	41.4	0.05
		1.1	35.0	0.87	38.8	0.84	43.2	0.12
		1.12	24.6	1.00	26.7	0.96	34.5	0.25
		1.2	5.8	1.00	6.4	1.00	10.1	1.00
		1.22	4.6	1.00	5.2	1.00	7.0	1.00
		1.24	3.8	1.00	4.1	1.00	5.2	1.00
Fri 12 am	60 obs	1.08	18.2	0.53	21.0	0.46	20.3	0.02
		1.1	22.0	0.73	24.8	0.72	14.9	0.07
		1.12	15.5	0.86	16.5	0.87	14.0	0.24
		1.2	2.4	1.00	3.0	1.00	4.6	1.00
		1.22	1.9	1.00	2.0	1.00	3.3	1.00
		1.24	1.6	1.00	1.6	1.00	2.6	1.00
	120 obs	1.08	31.1	0.57	32.6	0.53	20.5	0.03
		1.1	29.2	0.81	30.8	0.86	22.1	0.06
		1.12	22.0	0.92	21.3	0.96	16.2	0.25
		1.2	2.5	1.00	2.9	1.00	4.7	1.00
		1.22	2.0	1.00	2.0	1.00	3.3	1.00
		1.24	1.6	1.00	1.5	1.00	2.7	1.00

Table 15. Performance Comparison of AILR, JLR and EWMA in the Poisson Setting with $\sigma^2 = 0.001,\ \rho = 0.4$

			AILR		JLR		EWMA	
Start Time	Duration	Increment(c)	Det Time	Det Rate	Det Time	Det Rate	Det Time	Det Rate
Mon 7am	60 obs	1.08	25.4	0.80	25.6	0.84	25.3	0.81
		1.1	19.0	0.98	19.4	0.98	21.8	0.93
		1.12	11.4	1.00	11.8	1.00	13.2	1.00
		1.2	3.7	1.00	3.7	1.00	7.0	1.00
		1.22	3.1	1.00	3.0	1.00	5.1	1.00
		1.24	2.6	1.00	2.6	1.00	4.2	1.00
	120 obs	1.08	33.5	0.95	32.1	0.97	32.8	0.91
		1.1	20.5	0.98	19.9	1.00	24.1	1.00
		1.12	11.6	1.00	11.6	1.00	19.0	1.00
		1.2	3.6	1.00	3.6	1.00	7.1	1.00
		1.22	3.0	1.00	3.0	1.00	5.7	1.00
		1.24	2.6	1.00	2.7	1.00	4.2	1.00
Fri 12 am	60 obs	1.08	19.3	0.75	18.1	0.73	17.2	0.74
		1.1	14.5	0.94	13.4	0.94	14.7	0.93
		1.12	7.7	1.00	7.3	1.00	7.3	0.35
		1.2	1.6	1.00	1.7	1.00	3.2	1.00
		1.22	1.3	1.00	1.3	1.00	2.7	1.00
		1.24	1.2	1.00	1.2	1.00	2.2	1.00
	120 obs	1.08	24.8	0.82	25.5	0.82	22.7	0.84
		1.1	15.6	1.00	15.4	0.96	13.3	1.00
		1.12	7.2	1.00	7.5	1.00	8.3	1.00
		1.2	1.7	1.00	1.6	1.00	3.1	1.00
		1.22	1.4	1.00	1.3	1.00	2.6	1.00
		1.24	1.2	1.00	1.2	1.00	2.2	1.00