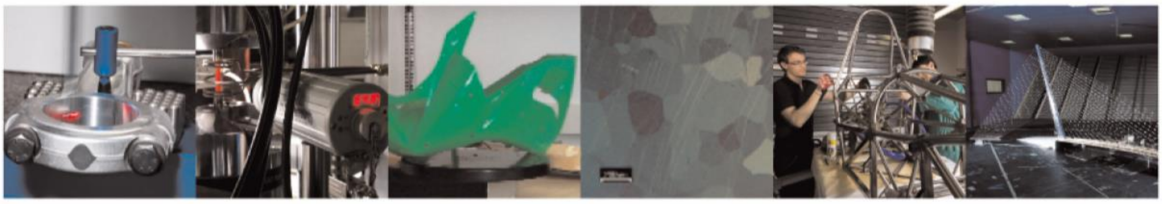

Bayesian statistical process control for Phase I count type data

Tsiamyrtzis, P.; Hawkins, D. M.

Bayesian Statistical Process Control for Phase I Count Type Data

Panagiotis Tsiamyrtzis
Department of Statistics
Athens University of Economics
and Business
76 Patission Str, 10434
Athens, Greece
pt@aueb.gr

Douglas M. Hawkins
School of Statistics
University of Minnesota
313 Ford Hall
224 Church Street S.E.
Minneapolis, MN 55455
dhawkins@umn.edu

Abstract

Count data, most often modeled by a Poisson distribution, are common in statistical process control. They are traditionally monitored by frequentist c or u charts, by cumulative sum (CUSUM) and by exponentially weighted moving average (EWMA) charts. These charts all assume the in-control true mean is known, a common fiction that is addressed by gathering a large Phase I sample and using it to estimate the mean. "Self-starting" proposals that ameliorate the need for a large Phase I sample have also appeared. All these methods are frequentist, they allow only retrospective inference during Phase I and they have no coherent way to incorporate less-than-perfect prior information about the in-control mean. In this paper, we introduce a Bayesian procedure which can incorporate prior information, allow online inference and which should be particularly attractive for short-run settings where large Phase I calibration exercises are impossible or unreasonable.

Key Words: c/u chart, Mixture of Gamma, Online Inference, Poisson, Short Runs.

1 Introduction

Standard univariate Statistical Process Control (SPC) methods split into the two broad categories of variable and attribute type SPC, based on the form of the recorded data (continuous and discrete respectively). Within the attribute (discrete) data two classes arise: binary and count data. The former represents the result of a classification procedure, where the outcome for each individual observation will be either conforming (acceptable) or nonconforming (unacceptable) according to some predetermined standards. In count data, we observe a process that produces non-negative integer values (counts) for each unit inspected. In an industrial application, these counts usually refer to the number of nonconformities per unit. There are numerous applications in several fields that resemble this industrial setup; like in epidemiology/public health (counts of H1N1 symptoms per school district), marketing (arrivals of customers in a shop per time unit), criminology (number of murders in a city) etc.

The usual approach is to assume that we have an underlying Poisson process, where the Poisson distribution ($P(\theta)$) can be used as the counting distribution, with the parameter θ, expressing the average counts per unit. For such a process to be Poisson it has to obey certain rules (see for example Feller, 1968) and there exist various diagnostics (see for example McCullagh and Nelder, 1989) for verifying that the Poisson model is appropriate for a given set of data.

From a SPC point of view, our interest in these types of processes is not only to have some estimate of the underlying parameter θ, but to be able to detect, in an online fashion, when the parameter shifts upwards (denoting process degradation) or downwards (denoting process improvement). We wish to detect shifts as soon as they occur, but at the same time keep the false alarm rate at a low level. Furthermore, in some applications we might be interested in making predictions, i.e. drawing inferences for future observable(s) of the process.

The most popular frequentist SPC tool for count type data is the Shewhart's u control chart (see for example Montgomery, 2012), where the quality characteristic is defined as the average number of nonconformities per inspection unit. In the cases where the inspection unit is constant over time the equivalent c control chart can be used. As is well known the Shewhart type charts are capable of detecting relatively large isolated shifts in the underlying parameter. For persistent shifts in the parameter (of moderate/small magnitude) the use of Poisson-CUSUM (Hawkins and Olwell, 1998) or Poisson-EWMA (Borror et al. 1998) is suggested.

In the frequentist based methods, the parameter θ is assumed to be a prespecified constant. As this constant is generally unknown, the usual practice is to conduct a Phase I study, collecting a large initial set of data, which will be used to estimate the model parameters. These data are assumed to be statistically independent and identically distributed (i.i.d.) observations of the assumed $P(\theta)$ model, i.e. we assume that there is no "special cause" variation due to extraneous factors that can contaminate these estimates. Once this calibration period is over we can start Phase II, where as new observations arrive, actual testing for continued stability of the data stream is done.

There are several serious deficiencies in the existing standard frequentist approach. One is that it has no good provision for changes occurring during Phase I. Another is that it requires large learning data sets and so it does not accommodate short-run settings. A third is that learning stops at the end of Phase I and no use is made of the information potential of Phase II data except for the actual test of stability. Finally, the underlying i.i.d. assumption itself is often unrealistic.

Apart from the fact that the existing methodology has all these restrictive requirements, it can not incorporate in the modelling any prior information regarding the parameter θ, which is often available from a similar process, historic data or from expert opinion.

Various attempts have been made to overcome several of the existing limitations. Woodall (1997) gave a nice bibliographic review of several such approaches related to attribute data. For example, regarding the violation of the Poisson assumption, Kaminsky et al. (1992) showed that this will cause a significant increase of the false alarm rate. Sheaffer and Leavenworth (1976) described real scenarios where the Poisson distribution assumption is inappropriate and proposed the use of Negative Binomial. Ryan and Schwertman (1997) and Schwertman and Ryan (1997) reported that the normal type approximation in determining control limits of attribute data was rather poor on the tails of the distribution and proposed "optimal" control limits for certain cases. Shore (2000) proposed control limit calculations by fitting an appropriate quantile function that preserves all first three moments, in order to improve performance for skewed distributions. In Quesenberry (1991) the Q chart was proposed for short runs of count data. Hawkins and Olwell (1998) described a "self-starting" CUSUM that largely eliminates the need for a separate Phase I study. Alwan and Roberts (1995) examined the effect of violating certain assumptions and misplacing the control limits.

From a Bayesian SPC point of view Hoadley's (1981) Quality Measurement Plan (QMP) was one of the first works that attempted to overcome the constraints of the classical SPC in count data, treating the Poisson parameter as a random variable. QMP adopted an (empirical) Bayes approach and provided an estimation-oriented control chart where at each stage of the process an interval estimate of the process parameter is provided. In Bayarri and García-Donato (2005), a Bayesian and an empirical Bayesian approach is presented: for each stage of the process the modelling of the underlying parameter is considered to be i.i.d. from the prior distribution. Then the predictive distribution is used to infer discrepancies of the underlying distribution parameter from the “in control” situation. In a similar i.i.d. setup but with Jeffrey’s prior, Raubenheimer and Van der Merwe (2014) provided a version of the Bayesian c chart, evaluating its performance against frequentist c chart, using Menzefricke’s (2002) metrics. Finally, if emphasis is placed in detecting the time that the parameter shifts, then the Bayesian change point methods of Shiryaev (1963) and Roberts (1966) could be employed (see for example Kenett and Pollak, 1996 for the Shiryaev-Roberts type control chart of a non-homogeneous Poisson process).

In this work we will treat the parameter of interest as a random variable within a purely Bayesian approach, where observations will be obtained sequentially and inference for the parameter of interest can start as soon as a single observable becomes available, thus eliminating the need for a special Phase I exercise. Inter alia, this provides a solution for the “short-run” setting in which the entire process generates fewer data points than are normally used just for calibrating a frequentist quality control scheme. Further, the parameter of interest will be modelled with a change point model that is subject to random shifts in size and direction. This will relax the i.i.d. assumption (generalizing the previous work of Bayarri et. all, 2005 and Raubenheimer et. all, 2014) while at the same time will provide grounds for capturing various out of control scenarios (outliers, step changes, linear drifts etc.) and accommodating scenarios of departures from the Poisson distribution (like over-dispersion). Bayesian decision theory will provide inference for both the underlying parameter and the potential shift occurrence. In cases where the inference does not raise any concerns about the parameter being “in control”, this posterior will be used as prior for the upcoming stage of the process, leading to a sequentially updated scheme. Last but not least, within the

Bayesian framework we can obtain the formal predictive distribution allowing predictive inference for future observable(s).

In the following section the proposed Bayesian change point model will be presented, while in section 3 inference issues will be explored for this new model. Section 4 will provide the sensitivity analysis to verify the robustness of the new approach. Next, in section 5 we compare the new model against standard self-starting frequentist based methods. In section 6 a real data application will illustrate the proposed method and finally section 7 will conclude this work. The technical details along with extended simulation results on sensitivity and robustness are provided in Appendices.

2 Bayesian change point modeling

We observe sequentially the data $X_1, X_2, \ldots,$ which refer to the number of occurrences (counts) X_n over m_n inspected units, with the m_n, being known and possibly varying over time $n = 1, 2, \ldots$ The usual practice is to model these data assuming a Poisson process, i.e.:

$$X_n|\theta_n \sim Poisson(m_n\theta_n). \tag{1}$$

Our main interest is in drawing inference for the unknown parameter θ_n, in an online fashion and with no need of phase I calibration. At the beginning of the process (i.e. before any observation becomes available), call it time 0, we have an initial prior distribution:

$$\theta_0 \sim Gamma(\alpha_0, \beta_0) \Rightarrow \pi(\theta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\theta_0^{\alpha_0-1}exp\{-\beta_0\theta_0\}. \tag{2}$$

This distribution quantifies (via the choice of α_0 and β_0), all the prior knowledge regarding process performance, before the data arrive. Historic data on this or of a similar process, or expert's opinion could be used to provide appropriate estimates for these hyperparameters. For example if for the parameter θ_0 we believe (a-priori) that it has mean $= \mu_0$ and variance $= \sigma_0^2$, then by moment matching we can get:

$$\alpha_0 = \frac{\mu_0^2}{\sigma_0^2}\ , \qquad \beta_0 = \frac{\mu_0}{\sigma_0^2}\ . \tag{3}$$

In (the rare) cases where no such prior knowledge exists, a non-informative approach could be adopted. For example one could use an improper prior distribution of the form:

$$\pi_0^f(\theta_0) \propto 1 \equiv Gamma(1,0) \qquad \text{or} \qquad \pi_0^J(\theta_0) \propto \frac{1}{\sqrt{\theta_0}} \equiv Gamma\left(\frac{1}{2},0\right) \tag{4}$$

with the former being the flat prior that can be presented as limiting $Gamma(1,v)$ as $v \to 0$ and the latter being the Jeffrey's prior (also a limiting Gamma case) on the positive real numbers. Both distributions are improper (they do not integrate to 1) but for the Poisson likelihood provide proper posteriors (Gamma).

Once the initial prior is set, we relax the usual i.i.d. assumption (required in most standard methods), adopting a change point model for the evolution of the parameter of interest. So, as time evolves, for $n = 1, 2, \ldots$ we assume the following change point model:

$$\theta_n|\theta_{n-1} \sim \left\{ \begin{array}{ll} \theta_{n-1} & \text{with prob. } p_0\,(= 1 - p_1 - p_2) \\ \lambda_1\theta_{n-1} & \text{with prob. } p_1 \\ \lambda_2\theta_{n-1} & \text{with prob. } p_2 \end{array} \right\} \tag{5}$$

where $p_i \in (0,1)$, with $\sum_{i=0}^{2} p_i = 1$, $0 < \lambda_1 < 1 < \lambda_2$.

Thus, between successive times of the process, one of the following three scenarios can occur to the parameter of interest: (i) no shift, (ii) downward shift, or (iii) upward shift. In Poisson type data, the counts usually (but not always) refer to the number of nonconformities in the process. As such in the SPC framework, scenario (i) will represent process stability, while scenarios (ii)/(iii) will refer to process improvement/degradation (the interpretation of scenarios (ii) and (iii) is reversed, in cases where the counts refer to better rather than worse quality).

The values of $(p_0, p_1, p_2)' = \boldsymbol{p}$ determine the relative frequency of the three scenarios, while the values of λ_1, λ_2 define the magnitude of the decrease (in downward shifts) and increase (in upward shifts) of the parameter respectively. The vector $\boldsymbol{\phi}' = (\boldsymbol{p}, \lambda_1, \lambda_2)$ constitutes the nuisance parameters in the model, assumed to be (a-priori) independent of each other.

Specific values for λ_1 and λ_2 could be used, in processes where we know what is the anticipated parameter shift (just like the CUSUM, which is designed for specific shifts). However, we will allow both λ_1 and λ_2 to be random, permitting random sized downward/upward shifts. For λ_1, $(0 < \lambda_1 < 1)$ the natural choice is to use a Beta distribution, i.e. $\lambda_1 \sim Beta(\gamma, \delta)$. Specifically, we will adopt a non-informative (flat) prior setup by selecting $\gamma = \delta = 1$, i.e.:

$$\lambda_1 \sim Beta(1,1) \equiv Uniform(0,1) \Rightarrow \pi(\lambda_1) = 1. \tag{6}$$

The parameter λ_2 $(\lambda_2 > 1)$ will be modeled via the Inverse Beta distribution:

$$\lambda_2 \sim IBeta(\zeta, \eta) \Rightarrow \pi(\lambda_2) = \frac{1}{Be(\zeta,\eta)} \left(\frac{1}{\lambda_2}\right)^{\zeta+1} \left(1 - \frac{1}{\lambda_2}\right)^{\eta-1} \tag{7}$$

where $\zeta > 1, \eta > 0$ are known hyperparameters. The $IBeta$ is the inverse Beta distribution, i.e. if $Y \sim Beta(\zeta, \eta)$ then the random variable $Z = 1/Y$ is distributed as $IBeta(\zeta, \eta)$ (the extra requirement of $\zeta > 1$ is necessary for the first moment existence of the $IBeta$).

The parameter λ_2 is related to the magnitude of the positive jump. Prior knowledge regarding its distribution can be used to elicit the hyperparameter values ζ and η. For example if the prior mean (μ_2) and variance (σ_2^2) of the magnitude of the positive shift are known, then matching the moments will give:

$$\zeta = 2 + \frac{\mu_2(\mu_2 - 1)}{\sigma_2^2} \qquad \text{and} \qquad \eta = \left(1 + \frac{\mu_2(\mu_2 - 1)}{\sigma_2^2}\right)(\mu_2 - 1). \tag{8}$$

Alternatively, if for the positive shifts we know μ_2 (anticipated magnitude) and that the maximum allowable magnitude between any successive times of the process is r, then by setting $r = 6\sigma_2$ (a range that covers at least 90% of this prior) we can obtain an estimate of σ_2 and then by moment matching we will get estimates of ζ, η. Finally for cases where we have a vague idea of what is the range of possible positive shifts we can use a rather large value for r, leading to a vaguely informative prior.

The remaining nuisance parameter, $\boldsymbol{p}$, just as with λ_1, λ_2, could be either estimated (for example, we could use historic data to obtain estimates of how often the process is stable or it drifts downwards/upwards), or alternatively within the Bayesian arena an appropriate prior could be introduced. For the latter approach, a convenient and quite general choice would be to model:

$$\boldsymbol{p} \sim Dirichlet(\boldsymbol{u}) \Rightarrow \pi(\boldsymbol{p}) = \frac{\Gamma(u_0 + u_1 + u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)} p_0^{u_0 - 1} p_1^{u_1 - 1} p_2^{u_2 - 1} \tag{9}$$

where $p_i > 0, i = 0, 1, 2$ and $\sum_{i=0}^{2} p_i = 1$, while $\boldsymbol{u} = (u_0, u_1, u_2)$ refer to the hyperparameters, with $u_i > 0, i = 0, 1, 2$. Just as we did with λ_1, we will assume a non-informative set-up where $u_0 = u_1 = u_2 = 1$ leading to the uniform (flat) distribution over its support set (simplex).

In general, prior knowledge about the process, can lead to more informative settings of the prior distributions in $\boldsymbol{\phi}$. In the proposed scheme we used an objective (flat) prior setting for λ_1 and $\boldsymbol{p}$, but if relevant information regarding these nuisance parameters exists, one can simply modify the values of γ, δ and $u_i, i = 0, 1, 2$ to match this information, moving to a more informative scheme (similarly to what was done for the λ_2 parameter). Sensitivity, regarding the hyperparameter choices will be examined in detail in the related section.

We are interested in drawing inference regarding the unknown parameter θ_n, once the data point X_n becomes available. This can include point/interval estimation (for monitoring purposes), hypothesis testing of whether the parameter exceeds some upper/lower threshold and (upward/downward) shift detection. All the above will simply attempt to summarize aspects of the obtained posterior distribution $\theta_n | X_n$, which can be thought of as the complete inference regarding the unknown parameter. We will elaborate on these issues further in the inference section.

We will adopt a Bayesian sequentially updated procedure, where the data which will arrive sequentially will be fed into the Bayes formula to derive the posterior distribution of the parameter of interest at the current time. When the parameter drifts to “out of control” regions, then we stop the process and take some corrective action. On the other hand, for as long as the unknown parameter provides posterior evidence that the process is under control, we will use the posterior distribution of the current time as a prior for the upcoming observation. The accumulated data observed up to time $n : x_1, x_2, \dots x_n$ will be denoted by $\mathbf{X}_n = (x_1, x_2, \dots, x_n)$.

Since at each stage of the process the parameter of interest can either shift upwards, downwards or not shift, in the Bayesian formulation of the process, the posterior of $\theta_n|\mathbf{X}_n$ will be a mixture with 3^n components. Specifically, we have the following:

Theorem 1 *At time n the posterior distribution of $\theta_n|\mathbf{X}_n$ is a mixture of 3^n Gamma distributions:*

$$p(\theta_n|\mathbf{X}_n) \sim \sum_{j=0}^{3^n-1} w^{(n)}_{\,j} Gamma\left(\alpha^{(n)}_{\,j},\;\; \beta^{(n)}_{\,j}\right)$$

with weights and parameters obeying the recursive equations, for $i = 0, 1, \ldots, 3^{n-1} - 1$:

$$\alpha^{(n)}_{\,3i} = \alpha^{(n)}_{\,3i+1} = \alpha^{(n)}_{\,3i+2} = \alpha^{(n-1)}_{\,i} + x_n$$

$$\begin{aligned}
\beta^{(n)}_{\,3i} &= \beta^{(n-1)}_{\,i} + m_n & w^{(n)}_{\,3i} &= (1 - P_1 - P_2) w^{(n-1)}_{\,i} M_i(x_n)/NC \\
\beta^{(n)}_{\,3i+1} &= \beta^{(n-1)}_{\,i}/\Lambda_1 + m_n & w^{(n)}_{\,3i+1} &= P_1 w^{(n-1)}_{\,i} M^-_i(x_n)/NC \\
\beta^{(n)}_{\,3i+2} &= \beta^{(n-1)}_{\,i}/\Lambda_2 + m_n & w^{(n)}_{\,3i+2} &= P_2 w^{(n-1)}_{\,i} M^+_i(x_n)/NC
\end{aligned}$$

where:

$$\Lambda_1 = \frac{\gamma}{\gamma+\delta}, \qquad \Lambda_2 = \frac{\zeta - 1 + \eta}{\zeta - 1}, \qquad P_1 = \frac{u_1}{u_0 + u_1 + u_2}, \qquad P_2 = \frac{u_2}{u_0 + u_1 + u_2},$$

$$\begin{aligned}
M_i(x_n) &= \frac{\Gamma\left(\alpha^{(n-1)}_{\,i} + x_n\right)}{\Gamma\left(\alpha^{(n-1)}_{\,i}\right)} \frac{1}{x_n!} \left[\frac{m_n}{\beta^{(n-1)}_{\,i} + m_n}\right]^{x_n} \left[\frac{\beta^{(n-1)}_{\,i}}{\beta^{(n-1)}_{\,i} + m_n}\right]^{\alpha^{(n-1)}_{\,i}} \\
M^-_i(x_n) &= \frac{\Gamma\left(\alpha^{(n-1)}_{\,i} + x_n\right)}{\Gamma\left(\alpha^{(n-1)}_{\,i}\right)} \frac{1}{x_n!} \left[\frac{\Lambda_1 m_n}{\beta^{(n-1)}_{\,i} + \Lambda_1 m_n}\right]^{x_n} \left[\frac{\beta^{(n-1)}_{\,i}}{\beta^{(n-1)}_{\,i} + \Lambda_1 m_n}\right]^{\alpha^{(n-1)}_{\,i}} \\
M^+_i(x_n) &= \frac{\Gamma\left(\alpha^{(n-1)}_{\,i} + x_n\right)}{\Gamma\left(\alpha^{(n-1)}_{\,i}\right)} \frac{1}{x_n!} \left[\frac{\Lambda_2 m_n}{\beta^{(n-1)}_{\,i} + \Lambda_2 m_n}\right]^{x_n} \left[\frac{\beta^{(n-1)}_{\,i}}{\beta^{(n-1)}_{\,i} + \Lambda_2 m_n}\right]^{\alpha^{(n-1)}_{\,i}}
\end{aligned}$$

$$NC = \sum_{i=0}^{3^{n-1}-1} \Big[(1-P_1-P_2)\, w_i^{(n-1)} M_i(x_n) + P_1\, w_i^{(n-1)} M_i^-(x_n) + P_2\, w_i^{(n-1)} M_i^+(x_n)\Big]\,.$$

Theorem 1 (proved in Appendix A) was given in its most general form for the cases that we have informative prior settings for all of the nuisance parameters. For the objective priors of λ_1 and $\boldsymbol{p}$ proposed here, we have $\gamma = \delta = 1$ and $u_0 = u_1 = u_2 = 1$ leading to $\Lambda_1 = 1/2$ and $P_1 = P_2 = 1/3$ respectively. Finally, for the cases where point estimates of the nuisance parameters exist in advance, these estimates can be plugged into the respective (uppercase notated) nuisance parameters (i.e. Λ_1, Λ_2, P_1 or P_2) replacing the values obtained from the (non-informative) prior settings assumed here.

2.1 Posterior approximation issues

The number of components in the posterior mixture increases exponentially fast and soon becomes difficult to handle. This is the price we pay for allowing bidirectional shifts at every stage of the process. At time n, the exact posterior distribution has 3^n components and one can obtain the posterior probability of each of the possible 3^n model evolution scenarios. But do we really need to keep track of all these components? In practice, the vast majority of these components will have negligible weights, which as more data become available will become even smaller (since they are multiplied with probabilities). Furthermore, several gamma posterior components will have tiny differences on the parameters. Thus, one could approximate the exact 3^n mixture with a much smaller mixture of K components, without losing significant information for the parameter of interest. We will adopt here the approximation algorithm proposed by Tsiamyrtzis and Hawkins (2010) which is a hybrid of the algorithm proposed by West (1993). If we will call ℓ the number of components in the posterior mixture, then for the stages for which $\ell \leq K$ we use the exact posterior distribution. Once we get for the first time to a stage where $\ell > K$ we initiate the approximation algorithm which consists of the following four steps:

(I) Order the ℓ posterior components in ascending order based on their weights.

(II) Identify the component $i \in \{2, ..., \ell\}$ which is the "nearest neighbor" to component 1.

(III) Pool components 1 and i, to a single new Gamma distribution with updated parameters and set $\ell = \ell - 1$.

(IV) Go to (I) and repeat, until $\ell = K$

Once the approximation algorithm is activated, it will be applied to obtain the posterior distribution for all subsequent data points. Specifically, for each new data point, the mixture will grow from K to $3K$ components and the approximation algorithm will trim it down to K. Similarly to Tsiamyrtzis and Hawkins (2010) the proximity in (II) among two Gamma's: $f_i \sim Gamma(\alpha_i, \beta_i)$, $i = 1, 2$ will be measured by Jeffreys (1948) divergence:

$$J\left(f_1, f_2\right) = (\alpha_1 - \alpha_2)\left[\Psi(\alpha_1) - \Psi(\alpha_2) + log\left(\frac{\beta_2}{\beta_1}\right)\right] + (\beta_1 - \beta_2)\left(\frac{\alpha_2}{\beta_2} - \frac{\alpha_1}{\beta_1}\right) \tag{10}$$

where,$\Psi(\alpha_i) = \Gamma'(\alpha_i)/\Gamma(\alpha_i)$, $i = 1, 2$ is the digamma function. In (III), when we decide to pool the components $G(\alpha_1, \beta_1)$ and $G(\alpha_2, \beta_2)$ with weights w_1 and w_2 respectively, to a single $Gamma(A, B)$, then the new Gamma density will have weight $w_1 + w_2$ and its parameters will be obtained via the method of moments:

$$A = \frac{\left[\left(\frac{w_1}{w_1+w_2}\right)\frac{\alpha_1}{\beta_1} + \left(\frac{w_2}{w_1+w_2}\right)\frac{\alpha_2}{\beta_2}\right]^2}{\left(\frac{w_1}{w_1+w_2}\right)\frac{\alpha_1}{\beta_1^2} + \left(\frac{w_2}{w_1+w_2}\right)\frac{\alpha_2}{\beta_2^2} + \frac{w_1 w_2}{(w_1+w_2)^2}\left(\frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2}\right)^2} \tag{11}$$

$$B = \frac{\left(\frac{w_1}{w_1+w_2}\right)\frac{\alpha_1}{\beta_1} + \left(\frac{w_2}{w_1+w_2}\right)\frac{\alpha_2}{\beta_2}}{\left(\frac{w_1}{w_1+w_2}\right)\frac{\alpha_1}{\beta_1^2} + \left(\frac{w_2}{w_1+w_2}\right)\frac{\alpha_2}{\beta_2^2} + \frac{w_1 w_2}{(w_1+w_2)^2}\left(\frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2}\right)^2}\ . \tag{12}$$

Another issue arising from the approximation algorithm is related to the choice of K. From a practical perspective and especially when we have relatively small number of data points, the recommendation is to to keep a few hundred components. A small simulation study will help us to determine the effect of the choice of K in posterior estimates. Specifically, we generated 1000 runs of length 12 according to the proposed bidirectional change point model, with $m_n = 1, \forall n = 1, \ldots, 12$, using the following hyperparameter values: $(\alpha_0, \beta_0, \Lambda_1, \Lambda_2, P_1, P_2) = (4, 1, 0.5, 1.5, 1/3, 1/3)$. Then at each iteration $i = 1, 2, \ldots, 1000$ we ran the exact model (with 3^{12} components) and we obtained the posterior mean estimate at each stage $n = 1, 2, \ldots, 12$ of the data, $\hat{\theta}^{[i]}_{n,E}$ (n refers to the stage of the process, E stands for the exact distribution while [i] refers to the iteration number). We repeated the posterior mean calculations for the approximate distribution with K components, leading to the estimation of the $\hat{\theta}^{[i]}_{n,K}$. In Table 1 we provide the Mean Absolute Error: $MAE_K = \sum_{i=1}^{1000} \left|\hat{\theta}^{[i]}_{n,E} - \hat{\theta}^{[i]}_{n,K}\right| \Big/ 1000$ for the values of $K = 100, 500$ and 1000, along with the variability of the exact posterior mean estimate at each of the 12 stages of the process (for these K, the approximation algorithm is activated at the fifth stage or later, so we skip the first 4 stages, where the error is $MAE = 0$).

Table 1 about here

This table suggests that even a hundred components should suffice for most practical purposes, when the number of data points is small.

3 Inference

The unknown parameter θ_n within the frequentist based SPC is assumed to be a fixed constant, which is estimated via a phase I exercise (where we typically assume iid data coming from the "in control" distribution). The goal of the traditional frequentist methods is to detect as soon as possible when the unknown parameter drifts from this estimated "in-control" value. Within the Bayesian approach though the unknown parameter is treated as a random variable. This offers great flexibility in terms of inference.

The posterior distribution obtained at each stage n of the process is considered to be the complete inference of the unknown parameter θ_n. Thus we could start by simply plotting the posterior distribution over time and qualitatively inspect the unknown parameter evolution (Apley, 2012 proposed such plots for continuous data). Apart from visual inspection, we can also provide quantitative summaries of these posterior distributions. If our interest is in monitoring and we would like to have a point estimate of the posterior distribution, then we could use the posterior mean, which under the squared error loss function is the Bayes rule, i.e.:

$$\hat{\theta}_n = E\left[\theta_n|\mathbf{X}_n\right] = \sum_{i=0}^{L-1} w_i^{(n)} \frac{\alpha_i^{(n)}}{\beta_i^{(n)}} \tag{13}$$

where $L = min\left\{3^n, K\right\}$, depending on whether the exact or the approximate scheme is used at stage n.

The fact that the unknown parameter, θ_n, is a random variable, allows us to move from the frequentist's "in control value" concept to the Bayesian "in control region", so instead of talking about the "target value" we are allowed to talk about the (more realistic) "target region". This is actually not a new concept. For instance in the 6σ philosophy, talking about the 3.4 defective parts per million opportunities (DPMO) in a normal process we refer to 4.5σ from the edges of $\pm 1.5\sigma$ out to 6σ. Thus we allow the process mean to drift 1.5σ of its target value (Tennant, 2001 and Lucas, 2002). So when we perform hypothesis testing we can move from a point to an interval null hypothesis, i.e.

$$\left\{ \begin{array}{ll} H_{0,n}: & L \leq \theta_n \leq U \\ H_{1,n}: & \theta_n < L, \text{ or } \theta_n > U \end{array} \right\} \tag{14}$$

where L (U) denote a lower (upper) threshold value related to process improvement (degradation) for the cases where the data x_n refer to the count of defects.

These (predetermined) constants are related to the unknown parameter θ_n and should not be confused with the traditional specification limits (referring to the observations x_n) and which, of necessity, will be smaller (larger) than L (U). The actual value of L (U) can come from expert opinion, or from economic criteria trading off the costs of stopping and adjusting the process with those of continuing a possibly deteriorated production. For this sequence of hypotheses, one can compute the posterior probability of the null hypothesis at each time n, $P(H_{0,n}|\mathbf{X}_n)$, and decide to reject/accept the null, based on whether this probability is too small/large. A time series plot of these probabilities might also give some insight on the evolution of the process performance. Furthermore, Bayes factors (Jeffreys, 1948) could be employed or alternatively a Bayes test could be derived when a loss function is introduced.

When the exact distribution is used, the weights $w_i^{(n)}$ play an important role, since they can be seen as posterior evidence of each of the 3^n possible model evolution scenarios from time 1 to n that the particular gamma density represents. Once the approximation starts though, we lose this ability to track the exact scenario. In this case, we will have K components in the mixture, which will increase to $3K$ when a new data point arrives and then the approximation algorithm will reduce them down to K. Before the approximation is applied though, it is easy to identify which of these $3K$ components refer to scenarios of: (i) no shift, (ii) downward or (iii) upward shift and by summing their weights, one can obtain the marginal probability of each of the three scenarios at the current state, independently of the earlier history.

Finally, one can derive the predictive distribution $X_{n+1}|\mathbf{X}_n$, of the next (unseen) observation x_{n+1}, based on the available data $\mathbf{X}_n = (x_1, x_2, \ldots, x_n)$ for forecasting or model assessment purposes.

4 Sensitivity Analysis

In this section we focus our attention on the robustness of the proposed methodology to mispecifications of the hyper-parameters (parameters of the prior distributions).

Adopting the objective prior setting for $\lambda_1 \sim U(0,1) \equiv Beta(1,1)$ and $\boldsymbol{p} \sim Dirichlet(1,1,1)$ we still need to specify the parameter values for the initial prior $\theta_0 \sim G(\alpha_0, \beta_0)$ and the prior regarding the positive jumps $\lambda_2 \sim IBeta(\zeta, \eta)$. So one needs to tune in the hyperparameters in these two priors for the method to run. In this section we will examine the robustness of the posterior mean estimation to various misspecifications of the hyperparameter values using a small simulation study.

We generate 1,000 series of length 10 according to the proposed model with the hyperparameter values given in the third column of Table 2. Then we run the exact (i.e. without approximating) model where the hyperparameters of both priors are correctly specified and we obtain the posterior mean estimate of each stage n of the process, call it $\hat{\theta}_{n,c}^{[i]}$ (n refers to the stage of the process, c stands for correct hyperparameter specification in the model while [i] refers to the iteration number). The boxplots of errors $\hat{\theta}_{n,c}^{[i]} - \theta_n^{[i]}$ (where $\theta_n^{[i]}$ refers to the true simulated value at stage n of iteration i) for all stages of the process can be seen in Figure 2 (scenario S1).

Table 2 about here

Next, we rerun the exact model three times, where we misspecify the prior hyperparameters of θ_0 only, λ_2 only, and both θ_0 and λ_2, called scenarios S2, S3 and S4 respectively. The hyperparameter misspecifications are presented in column four of Table 2. The misspecifications attempt to give a form in the prior which will look sufficiently different from the correct one. In Figure 1 we provide the plots of the correct and misspecified priors of the parameters θ_0 and λ_2.

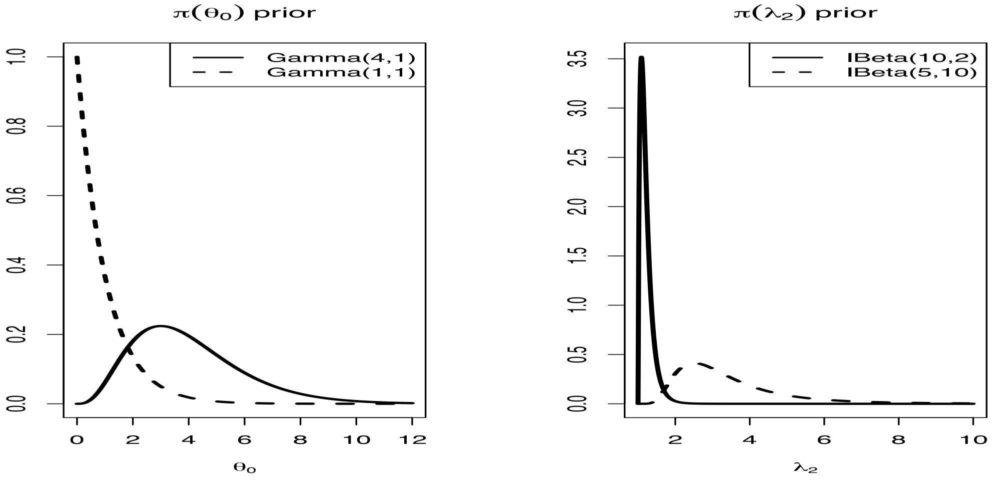

Figure 1: The pdfs of the correct (solid lines) and misspecified (dash lines) prior settings for the parameters θ_0 and λ_2

Next, for each of these three scenarios of misspecifications we run the exact model and we obtain the respective posterior mean estimates, at each stage of the process. The boxplots of the error terms (posterior mean − true theta) for each stage and each of the four scenarios (S1-S4) of correct/miss-specifications can be seen in Figure 2.

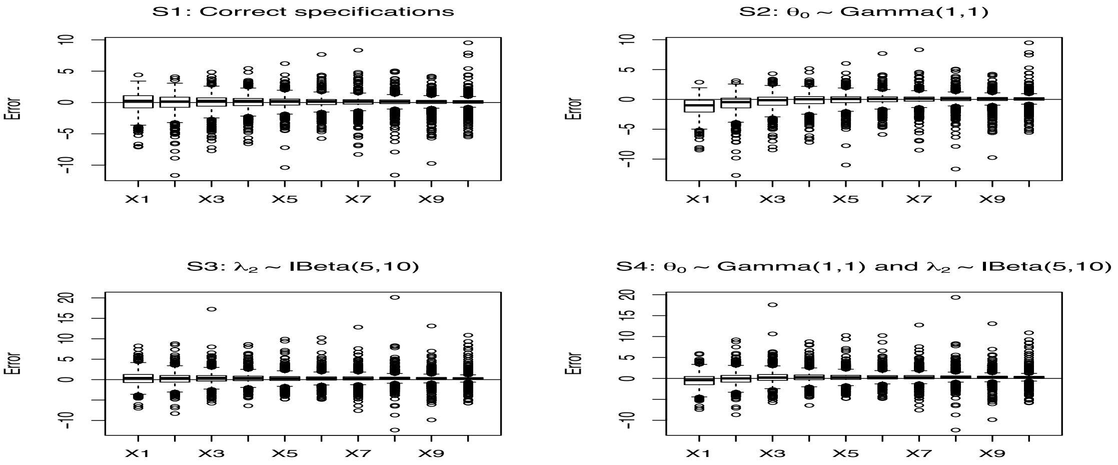

Figure 2: The error boxplots of the posterior mean estimate minus the true parameter value, when all hyper-parameters are correctly specified or one/two priors (denoted in the title of each subplot) is misspecified.

The mean and standard deviation of the absolute error, for each stage $n = 1, 2, \ldots, 10$ and for each of the misspecification scenarios (S2-S4) and correct specifications (S1) can be seen in Table 3.

Table 3 about here

We observe that in general the results appear to be quite robust when the hyperparameters are misspecified. The only case where the prior choice seemed to affect (partially) the results was the choice of the initial prior θ_0. In this simulation scenario the misspecified θ_0 was (on average) underestimating the correct θ_0, something which is evident in the respective error plots. However, due to the dynamic update mechanism, the effect of this misspecification quickly washes out, affecting only the very few initial data points.

More simulation results regarding sensitivity and robustness are presented in Appendices. Specifically in Appendix B we examine the robust performance to various parameter evolution scenarios (step change, linear drift, etc) while in Appendix C we examine the sensitivity of the proposed methodology when the Poisson assumption is violated (over-dispersed data).

5 Competing methods

When we analyze short run type of data in absence of a Phase I calibration stage a frequentist based proposal is to use a self starting approach, where the calibration and monitoring is performed simultaneously. The idea behind such techniques, is to transform the data and obtain a pivotal statistic that will be free of the underline unknown parameter. We will attempt to compare the performance of the proposed Bayesian scheme against such self starting methods. Let's assume that we have a short run of length 30 for Poisson count data:

$$X_i|\theta_i \sim P(\theta_i), \qquad i = 1, 2, \ldots, 30 \tag{15}$$

for which under the "in control" state (θ_{IC}) let's assume that $\theta_i = 4$. Our goal will be to detect a step change of size $\{0.5\sigma, 1\sigma, \text{ or } 2\sigma\}$ (leading to θ_{OOC}) that can occur at location $\{5 \text{ or } 15\}$ in the data stream. Since our interest is in detecting a step change we will make use of self starting cusum methods that fulfill certain optimality criteria. Specifically, we will run two self starting Poisson cusum schemes, based on the scores proposed by Quesenberry (1991, 1995) and Hawkins and Olwell (1997) that we will denote as Qcusum and HOcusum respectively. The scores on which we run the Poisson self starting cusums are calculated as follows: Initially we obtain the cumulative probabilities:

$$A_n = Pr\left[Bin\left(\sum_{i=1}^{n} X_i, \frac{1}{n}\right) \leq X_n\right], \qquad \text{for} \qquad n = 2, 3, \ldots, 30 \tag{16}$$

where $Bin(\mathcal{A}, \mathcal{B})$ denotes the binomial distribution with $\mathcal{A}$ trials and success probability $\mathcal{B}$. Then:

Qcusum: obtain $Q_n = \Phi^{-1}(A_n)$, where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative standard normal distribution, for $A_n < 1$. The value of $A_n = 1$ is suggested to be winsorized (since for this value the Q_n score is undefined). Then we use the Q_n scores to obtain the Qcusum with:

$$K_Q^+ = \left(\frac{\theta_{OOC} - \theta_{IC}}{2}\right)\frac{1}{\sigma} + 0.2 \tag{17}$$

with the correction constant 0.2 proposed in Hawkins and Olwell (1997) to encounter the fact that the true mean of Q_n is around 0.2 as opposed to 0.

HOcusum: we make use of an "educated guess", call it m, of the underline parameter, and we transform the data X_n to Y_n such that the Y_n will minimize:

$$\left|\sum_{j=0}^{Y_n} \frac{e^{-m} m^j}{j!} - A_n\right| \tag{18}$$

for $A_n < 1$. In case of $A_n = 1$, it is suggested to put $Y_n = X_n$. Then we use the Y_n scores to obtain the HOcusum with:

$$K_{HO}^+ = \frac{\theta_{OOC} - \theta_{IC}}{ln(\theta_{OOC}) - ln(\theta_{IC})}\ . \tag{19}$$

To the best of our knowledge no ARL related information exists for the above charts (or in general for self starting methods). Furthermore, since our concern is in short run type data, ARL discussion is probably not relevant here. For the decision interval type cusums, we need to determine the related decision constants h^Q and h^{HO}, so that we have predetermined false alarm rate over a horizon of N data points. Specifically, for the short run scenario ($N = 30$) under study, we will simulate the process under the "in control" state and then derive these decision thresholds so that both achieve the same (predetermined) false alarm rate. Precisely, we generate 10,000 sequences of "in control" data $X_i \sim P(4)$, for $i = 1, 2, \ldots, 30$ and obtain the scores Q_n and Y_n. Then we run the respective cusums and obtain the threshold values h^Q and h^{HO} by fixing the false alarm rate at 5% (see table 4).

To compare against the self starting cusums we need to formulate a decision making scheme for the proposed Bayesian approach. Since our interest is to detect shifts from $\theta_{IC} = 4$ to $\theta_{OOC} = \{5, 6 \text{ or } 8\}$ we can consider this as a hypothesis testing problem with:

$$\left\{ \begin{array}{ll} H_{0,n}: & \theta_n < \theta_{OOC} \\ H_{1,n}: & \theta_n \geq \theta_{OOC} \end{array} \right\}. \tag{20}$$

Then using the proposal given in the inference section we can obtain the posterior coverage probability of $H_{1,n}$ and decide to reject the null hypothesis when this probability is above a stopping threshold, that we will derive via simulations. For the nuisance parameters we make use of flat priors for λ_1 and $\boldsymbol{p}$ (i.e. $\gamma = \delta = 1$ and $p_0 = p_1 = p_2 = 1/3$ respectively) while for θ_0 we chose the vaguely informative $\alpha_0 = 4$, $\beta_0 = 1$, and for λ_2 we chose $\zeta = 11$ and $\eta = 5$ (so that $E[\lambda_2] = 1.5$, a symmetric choice with respect to $E[\lambda_1] = 0.5$). Finally, we keep only $K = 100$ components in the approximation scheme.

Similarly to what we did earlier, we use the 10,000 sequences of "in control" data to obtain the $P\left(H_{1,n}|\mathbf{X}_n\right)$ and then derive the decision threshold h^B so that we have a 5% false alarm rate (sse table 4).

Table 4 about here

Once the decision constants have been determined so that all three methods have the same false alarm rate (5%) when the process is "in control", we will examine their detection power when a step change of size $\{0.5\sigma, 1\sigma, \text{ or } 2\sigma\} \equiv \{1, 2, \text{ or } 4\}$ occurs at location $\{5 \text{ or } 15\}$.

Table 5 about here

The results (see Table 5) indicate that the Bayesian Poisson change point (BPCP) scheme outperforms both Qcusum and HOcusum in all but one scenario, while the Q and HO cusums appear to have similar performance. The correct detection rate for the new approach is significantly higher from the competing cusum methods when the step change occurs early in the process (at time 5) and it is still higher when we are at location 15. Only for the step change of size 0.5σ at location 15 the new approach shows slightly smaller detection power from the respective cusums. Apart from the detection rate though we are interested in the time it takes to signal for each method. For this reason in the cases that we had shift detection, we obtain the mean and standard deviation of the delay in identifying the shift. Just as with the detection rates, the new approach has smaller mean delay in all scenarios but the one where the step change of size 1σ occurred at location 5 (for which case we should note that the correct detection percentage of BPCP is more than double from the respective percentages of the cusums).

6 Real Data Application

The developed methodology will be illustrated in the monitoring of crime statistics, specifically on monthly counts of murders, recorded in the city of Houston, TX. The data filed from the Houston Police Department and are available at http://www.houstontx.gov/police/cs/stats2.htm. Let's assume that as statisticians we are called on Jan 2014 to provide online monitoring and control of this crime index from that time onwards. The data we will use for this illustration consist of 16 months (Jan 2014 - Apr 2015, see Table 6), and we assume that they become available to us in a sequential manner.

Table 6 about here

So starting from Jan 2014, the data arrive sequentially and our goal is in drawing inference (control/monitor the process) each time a new data point becomes available. We will make use of the proposed Bayesian methodology with the objective prior choices for $\lambda_1 \sim U(0,1)$ and $\boldsymbol{p} \sim Dirichlet(1,1,1)$, while for the remaining parameters, the historic data of Jan 2010 -Dec 2013 (see Table 7) will be used to estimate/elicitate them.

Table 7 about here

In terms of controlling the data stream, beginning on Jan 2014, via hypothesis testing, our interest will be to raise an alarm whenever the mean crime statistic exceeds an upper threshold value U. So the hypothesis testing (14) will take in this application the form of one sided hypothesis:

$$\left\{ \begin{array}{ll} H_{0,n}: & \theta_n \leq U \\ H_{1,n}: & \theta_n > U \end{array} \right\}. \tag{21}$$

The upper threshold value U could be estimated in various ways. In this illustration we set this value to be the 85th percentile of the 2010-2013 historic data. This turns out to be $U = 22.95$, i.e. we would like to raise an alarm in an online fashion when the true mean (θ_n) exceeds this value U.

Next, we will elicitate the initial prior distribution for the unknown parameter, θ_0. For this reason we will make use of the most recent annual data, i.e. of the year 2013. Specifically, the $n = 12$ data points of 2013 are assumed to have $Y_j \sim Poisson(\theta)$ likelihood, which we combine with a flat prior $\pi_0^f(\theta) \propto 1 \equiv Gamma(1, v)$ as $v \to 0$ (see (4)) leading to the posterior distribution $\theta|(Y_1, \ldots, Y_{12}) \sim Gamma\left(\sum Y_j + 1, n\right) = Gamma(210, 12)$, which will play the role of the initial prior θ_0 before the Jan 2014 data point arrives, i.e. $\theta_0 \sim Gamma(210, 12)$.

Finally, we need to elicitate the parameter λ_2 which reflects on the expected parameter inflation when an upward shift occurs. In this example the initial expected value of the unknown parameter is $E[\theta_0] = \alpha_0/\beta_0 = 210/12 = 17.5$, while the upper threshold that we are interested in not to exceed is $U = 22.95$. We will define a point estimate (via the expected value) of λ_2 as the ratio of $U/E[\theta_0] = 22.95/17.5 = 1.311$ which will indicate the factor that can lead (in a single step) the unknown parameter from the initial prior mean value to the OOC region. In general the ratio of $U/E[\theta_0]$ is expected to be bigger than one (necessary condition for λ_2) or otherwise we will have a process which has prior mean in the OOC region indicating that some corrective action is required before even we start running the process.

Summarizing the parameter setup in this example via Theorem (1) we have:

$$\theta_0 \sim Gamma(210, 12), \qquad P_1 = P_2 = \frac{1}{3} \qquad \Lambda_1 = \frac{1}{2}, \qquad \Lambda_2 = 1.311\ .$$

We used $K = 1,000$ components in the approximate posterior distribution, giving the exact posterior for the first 6 data points and the approximate from there on. Our main concern is to perform sequential hypothesis testing at each month and raise an alarm if the unknown parameter (mean murder count) $\theta_n > U$. From the new proposed methodology, at each point of the process n, we can obtain the posterior distribution of $\theta_n|\mathbf{X}_n$ which will be a mixture of $min\left\{3^n, K\right\}$ Gamma distributions. This posterior distribution from a Bayesian perspective is the complete inference regarding the unknown parameter θ_n. In Figure 3 we provide the posterior distributions for each month of the 2014-15, as the data become available sequentially.

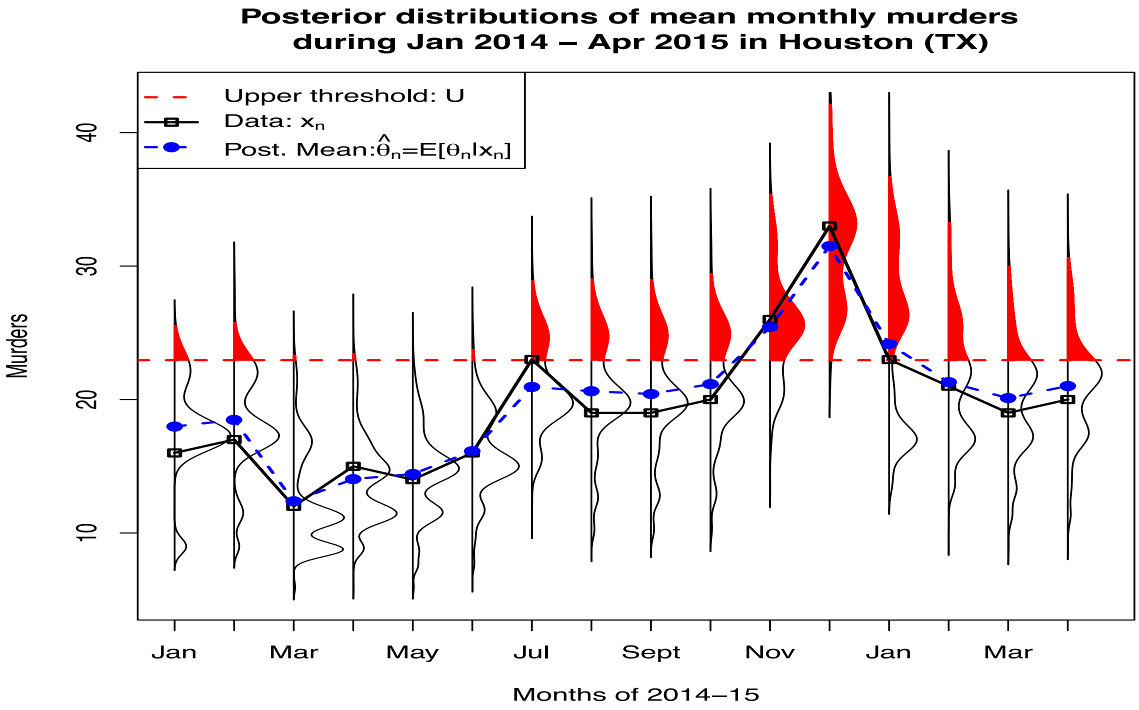

Figure 3: Posterior distribution plots of the mean murder counts (θ_n) at each month n of 2014-15 with the posterior coverage of the alternative hypothesis ($\theta_n > U$) being highlighted. The line plots of the posterior mean (point estimate) $\hat{\theta}_n = E[\theta_n|\mathbf{X}_n]$ (solid points) and the data x_n (square points) are also added.

Next, we obtained the posterior mean, which is Bayes rule under squared error loss for this parameter, and plot it in Figure 3 (its values appear in column three of Table 8). Furthermore, we calculated the posterior coverage probabilities of the alternative hypothesis: $Pr(\theta_n > U|\mathbf{X}_n)$, which we highlight in Figure 3 and report the exact values in column four of Table 8.

Table 8 about here

In addition, thanks to the weights in the posterior mixture, the new model provides the marginal probabilities of whether the unknown parameter θ_n shifted downward, upward or not at each time n, independently of the past. These probabilities are provided in the last three columns of Table 8 and the downward, upward shift probabilities are plotted in Figure 4.

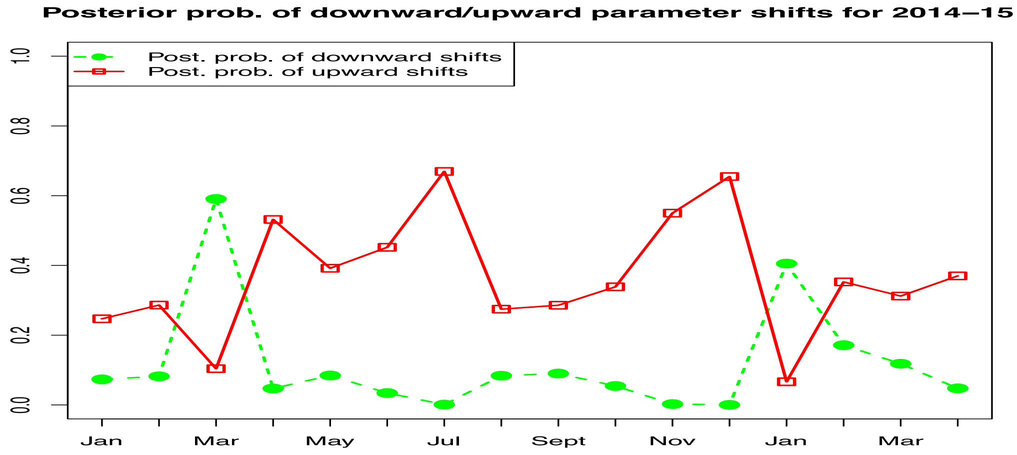

Figure 4: The marginal posterior probabilities of having a downward (dashed line) or upward (solid line) shift at each month of the 2014-15 murder counts at the city of Houston, TX.

The combination of Figures 3 and 4 provides a much more informative setup to control/monitor the ongoing process, compared to what the traditional say c control chart would do and most importantly it can do that in real time, allowing inference for Phase I data and/or short production runs.

From a decision making point of view regarding the hypothesis testing one could proceed in various ways with the posterior probabilities $Pr(\theta_n > U|\mathbf{X}_n)$. One alternative is to consider the $E[\theta_0]$ as the IC value and the U as the OOC parameter value and run simulations to determine the value of the decision threshold h^B so that the false alarm rate (FAR) is at some predetermined level for a finite horizon of N observations (similarly to what was done in section 5). For example for FAR=0.05 and running 10,000 simulations of IC data sequences of length 16, we obtained $h^B = 0.842$, indicating that in Dec 2014 the criminal statistic under study exceeds the predetermined threshold U. In a decision theory based approach, if the costs of type I and II errors are known to be c_I and c_{II} respectively, then the Bayes test under the generalized $0-1$ loss function would reject H_0 when $Pr(\theta_n > U|\mathbf{X}_n) > c_{II}/(c_I + c_{II})$. In the field of public security it is natural to expect $c_I < c_{II}$, so that $c_{II}/(c_I + c_{II}) > 0.5$.

From Figure 4 we observe that in the last two months of 2014 there exist successive high probabilities of upward parameter shifts leading eventually sufficient posterior mass (Figure 3) to be above the threshold $h^B = 0.842$. Furthermore, in Jul 2014 the highest probability of upward shift is recorded indicating the largest increase of the unknown parameter in succesive stages, without trigerring an alarm though.

7 Conclusions

Traditional frequentist models for count data emphasize restrictive assumptions – of independent observations from a Poisson distribution whose parameter is known exactly. The assumption of known parameter then necessitates large Phase I studies, and the frequentist approach is unable to incorporate prior information that is less specific than that of the Phase I study.

The Bayesian model proposed here allows for evolution in the process parameter, and distinguishes shifts that are too small to matter from those that do matter. Also as a Bayesian approach, it gives a conceptually sound way of incorporating partial prior information, removing completely the need for any Phase I study, and allowing monitoring to start with the first process reading.

We believe these features make the approach particularly compelling for short-run problems in settings where there are processes that are somewhat like processes that were used previously. Examples are where different machines and/or operators are used, and in scale-up settings where pilot information is available.

Appendix A: Proof of Theorem 1

Initially, we will need to derive the distribution of $\theta_n|\theta_{n-1}$ marginalizing out the nuisance parameters $\boldsymbol{\phi}' = (\boldsymbol{p}, \lambda_1, \lambda_2)$. The proof will be done using the general prior setup of:

$$\lambda_1 \sim Beta(\gamma, \delta), \qquad \lambda_2 \sim IBeta(\zeta, \eta) \qquad \boldsymbol{p} \sim Dirichlet(u_0, u_1, u_2)$$

for the cases where prior info regarding these nuisance parameters exists and eventually we will replace $\gamma = \delta = 1$ and $u_0 = u_1 = u_2 = 1$ to obtain the proposed objective form of the theorem. First, we will prove (using induction) that:

$$\theta_n|\theta_{n-1} \sim \left\{ \begin{array}{ll} \theta_{n-1} & \text{with prob. } u_0/(u_0+u_1+u_2) \\ \left(\frac{\gamma}{\gamma+\delta}\right)\theta_{n-1} & \text{with prob. } u_1/(u_0+u_1+u_2) \\ \left(\frac{\zeta-1+\eta}{\zeta-1}\right)\theta_{n-1} & \text{with prob. } u_2/(u_0+u_1+u_2) \end{array} \right\} .$$

For $n = 1$ it is easy to show that it holds. We assume that it is true for $n-1$ and we will show that it holds for n. Specifically, we have:

$$\begin{aligned} \theta_n|\theta_{n-1}, \lambda_1, \lambda_2, \mathbf{p} &\sim p_0\theta_{n-1} + p_1\lambda_1\theta_{n-1} + p_2\lambda_2\theta_{n-1} \\ \pi(\lambda_1) &\sim Beta(\gamma, \delta) \\ \pi(\lambda_2) &\sim IBeta(\zeta, \eta) \\ \pi(\boldsymbol{p}) &\sim Dirichlet(\boldsymbol{u}) . \end{aligned}$$

We will obtain the distribution of $\theta_n|\theta_{n-1}$ by integrating out all the nuisance parameters, i.e.:

$$\begin{aligned}
\pi(\theta_n|\theta_{n-1}) &= \int\int\int \pi(\theta_n, \lambda_1, \lambda_2, \boldsymbol{p}|\theta_{n-1}) d\boldsymbol{p}\; d\lambda_2\; d\lambda_1 \\
&= \int \left[\int \left[\int \pi(\theta_n|\theta_{n-1}, \lambda_1, \lambda_2, \boldsymbol{p})\pi(\boldsymbol{p}) d\boldsymbol{p}\right] \pi(\lambda_2) d\lambda_2\right] \pi(\lambda_1) d\lambda_1\ .
\end{aligned}$$

If we will call the inner integral I_1 then we have:

$$\begin{aligned}
I_1 &= \int \pi(\theta_n|\theta_{n-1}, \lambda_1, \lambda_2, \boldsymbol{p})\pi(\boldsymbol{p}) d\boldsymbol{p} \\
&= \int \left[p_0\theta_{n-1} + p_1\lambda_1\theta_{n-1} + p_2\lambda_2\theta_{n-1}\right] \frac{\Gamma(u_0+u_1+u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)} p_0^{u_0-1}p_1^{u_1-1}p_2^{u_2-1} d\boldsymbol{p} \\
&= \frac{\Gamma(u_0+u_1+u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)}\ \theta_{n-1} \int p_0^{(u_0+1)-1}p_1^{u_1-1}p_2^{u_2-1} d\boldsymbol{p} \\
&+ \frac{\Gamma(u_0+u_1+u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)}\ \lambda_1\theta_{n-1} \int p_0^{u_0-1}p_1^{(u_1+1)-1}p_2^{u_2-1} d\boldsymbol{p} \\
&+ \frac{\Gamma(u_0+u_1+u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)}\ \lambda_2\theta_{n-1} \int p_0^{u_0-1}p_1^{u_1-1}p_2^{(u_2+1)-1} d\boldsymbol{p} \\
&= \frac{\Gamma(u_0+u_1+u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)}\ \frac{\Gamma(u_0+1)\Gamma(u_1)\Gamma(u_2)}{\Gamma(u_0+u_1+u_2+1)}\ \theta_{n-1} \\
&+ \frac{\Gamma(u_0+u_1+u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)}\ \frac{\Gamma(u_0)\Gamma(u_1+1)\Gamma(u_2)}{\Gamma(u_0+u_1+u_2+1)}\ \lambda_1\theta_{n-1} \\
&+ \frac{\Gamma(u_0+u_1+u_2)}{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2)}\ \frac{\Gamma(u_0)\Gamma(u_1)\Gamma(u_2+1)}{\Gamma(u_0+u_1+u_2+1)}\ \lambda_2\theta_{n-1} \\
&= \left(\frac{u_0}{u_0+u_1+u_2}\right)\theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2}\right)\lambda_1\theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2}\right)\lambda_2\theta_{n-1} \\
&= \pi(\theta_n|\theta_{n-1}, \lambda_1, \lambda_2)\ .
\end{aligned}$$

Then, if we will call I_2 the middle integral we will have:

$$\begin{aligned}
I_2 &= \int \pi(\theta_n|\theta_{n-1}, \lambda_1, \lambda_2)\pi(\lambda_2)d\lambda_2 \\
&= \int \left[\left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2} \right) \lambda_1\theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2} \right) \lambda_2\theta_{n-1} \right] \times \\
&\quad \times \frac{1}{Be(\zeta,\eta)} \left(\frac{1}{\lambda_2} \right)^{\zeta+1} \left(1 - \frac{1}{\lambda_2} \right)^{\eta-1} d\lambda_2 \\
&= \left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2} \right) \lambda_1\theta_{n-1} + \\
&+ \left(\frac{u_2}{u_0+u_1+u_2} \right) \theta_{n-1} \frac{1}{Be(\zeta,\eta)} \int \left(\frac{1}{\lambda_2} \right)^{(\zeta-1)+1} \left(1 - \frac{1}{\lambda_2} \right)^{\eta-1} d\lambda_2 \\
&= \left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2} \right) \lambda_1\theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2} \right) \frac{Be(\zeta-1,\eta)}{Be(\zeta,\eta)} \theta_{n-1} \\
&= \left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2} \right) \lambda_1\theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2} \right) \left(\frac{\zeta-1+\eta}{\zeta-1} \right) \theta_{n-1} \\
&= \pi(\theta_n|\theta_{n-1}, \lambda_1) \; .
\end{aligned}$$

Then for the outer integral we will have: $\pi(\theta_n|\theta_{n-1}) =$

$$\begin{aligned}
&= \int \pi(\theta_n|\theta_{n-1}, \lambda_1)\pi(\lambda_1)d\lambda_1 \\
&= \int \left[\left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2} \right) \lambda_1\theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2} \right) \left(\frac{\zeta-1+\eta}{\zeta-1} \right) \theta_{n-1} \right] \times \\
&\quad \times \frac{1}{Be(\gamma,\delta)} \lambda_1^{\gamma-1} (1-\lambda_1)^{\delta-1} d\lambda_1 \\
&= \left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2} \right) \left(\frac{\zeta-1+\eta}{\zeta-1} \right) \theta_{n-1} + \\
&+ \left(\frac{u_1}{u_0+u_1+u_2} \right) \theta_{n-1} \frac{1}{Be(\gamma,\delta)} \int \lambda_1^{(\gamma+1)-1} (1-\lambda_1)^{\delta-1} d\lambda_1 \\
&= \left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2} \right) \frac{Be(\gamma+1,\delta)}{Be(\gamma,\delta)} \theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2} \right) \left(\frac{\zeta-1+\eta}{\zeta-1} \right) \theta_{n-1} \\
&= \left(\frac{u_0}{u_0+u_1+u_2} \right) \theta_{n-1} + \left(\frac{u_1}{u_0+u_1+u_2} \right) \left(\frac{\gamma}{\gamma+\delta} \right) \theta_{n-1} + \left(\frac{u_2}{u_0+u_1+u_2} \right) \left(\frac{\zeta-1+\eta}{\zeta-1} \right) \theta_{n-1} \; .
\end{aligned}$$

To ease the notation we will call:

$$\Lambda_1 = \frac{\gamma}{\gamma + \delta}, \qquad \Lambda_2 = \frac{\zeta - 1 + \eta}{\zeta - 1}, \qquad P_1 = \frac{u_1}{u_0 + u_1 + u_2}, \qquad P_2 = \frac{u_2}{u_0 + u_1 + u_2}$$

and then we have:

$$\theta_n | \theta_{n-1} \sim \left\{ \begin{array}{ll} \theta_{n-1} & \text{with prob. } P_0 = 1 - P_1 - P_2 \\ \Lambda_1 \theta_{n-1} & \text{with prob. } P_1 \\ \Lambda_2 \theta_{n-1} & \text{with prob. } P_2 \end{array} \right\} . \qquad (I)$$

Next we will prove (via induction) the form of the posterior distribution of $\theta_n | \mathbf{X}_n$ given in the theorem. For $n = 1$ it is easy to show that it holds. Assume that the theorem holds for $n - 1$, i.e. the posterior distribution of $\theta_{n-1} | \mathbf{X}_{n-1}$ is a mixture of 3^{n-1} Gamma distributions with the weights and parameters obeying the respective relationships, i.e.:

$$p(\theta_{n-1} | \mathbf{X}_{n-1}) \sim \sum_{i=0}^{3^{n-1}-1} w_i^{(n-1)} Gamma\left(\alpha_i^{(n-1)}, \ \beta_i^{(n-1)}\right) .$$

Then we will prove that it holds for n. From (I) We have that:

$$\begin{aligned} \theta_n | \theta_{n-1} \ &\sim \ (1 - P_1 - P_2) \ \ Post(\theta_{n-1} | \mathbf{X}_{n-1}) + P_1 \ \ Post(\Lambda_1 \theta_{n-1} | \mathbf{X}_{n-1}) + P_2 \ \ Post(\Lambda_2 \theta_{n-1} | \mathbf{X}_{n-1}) \\ \theta_{n-1} | \mathbf{X}_{n-1} \ &\sim \ \sum_{i=0}^{3^{n-1}-1} w_i^{(n-1)} Gamma\left(\alpha_i^{(n-1)}, \ \beta_i^{(n-1)}\right) . \end{aligned}$$

Thus for the distribution of $\theta_n | \mathbf{X}_{n-1}$ which will be the updated (based on the proposed model) prior distribution of stage n of the process we have:

$$\begin{aligned}\pi(\theta_n|\mathbf{X}_{n-1}) &\sim \sum_{i=0}^{3^{n-1}-1}\Big[(1-P_1-P_2)\; w_i^{(n-1)} Gamma\left(\alpha_i^{(n-1)}, \beta_i^{(n-1)}\right) \\ &\qquad + P_1\; w_i^{(n-1)} Gamma\left(\alpha_i^{(n-1)},\; \frac{\beta_i^{(n-1)}}{\Lambda_1}\right) + P_2\; w_i^{(n-1)} Gamma\left(\alpha_i^{(n-1)},\; \frac{\beta_i^{(n-1)}}{\Lambda_2}\right)\Big] \\ &\sim \sum_{i=0}^{3^{n-1}-1}\Big[(1-P_1-P_2)\; w_i^{(n-1)}\pi_i(\theta_n) + P_1\; w_i^{(n-1)}\pi_i^-(\theta_n) + P_2\; w_i^{(n-1)}\pi_i^+(\theta_n)\Big]\end{aligned}$$

where $\pi_i(\theta_n) \equiv Gamma\left(\alpha_i^{(n-1)},\;\; \beta_i^{(n-1)}\right)$, $\pi_i^-(\theta_n) \equiv Gamma\left(\alpha_i^{(n-1)},\;\; \beta_i^{(n-1)}/\Lambda_1\right)$ and $\pi_i^+(\theta_n) \equiv Gamma\left(\alpha_i^{(n-1)},\;\; \beta_i^{(n-1)}/\Lambda_2\right)$. Then at time n the count $X_n = x_n$ will be observed over m_n inspected units and will form the likelihood:

$$f(X_n|\theta_n) \sim Poisson(m_n\theta_n)\ .$$

The posterior distribution will be given by:

$$\begin{aligned}p(\theta_n|\mathbf{X}_n) &= \frac{f(x_n|\theta_n)\pi(\theta_n)}{\int f(x_n|\theta_n)\pi(\theta_n)d\theta_n} \\ &\propto \sum_{i=0}^{3^{n-1}-1}\Big[(1-P_1-P_2)\; w_i^{(n-1)} f(x_n|\theta_n)\pi_i(\theta_n) \\ &\qquad + P_1\; w_i^{(n-1)} f(x_n|\theta_n)\pi_i^-(\theta_n) + P_2\; w_i^{(n-1)} f(x_n|\theta_n)\pi_i^+(\theta_n)\Big]\ .\end{aligned}$$

We will call:

$$\begin{aligned} M_i(x_n) &= \int f(x_n|\theta_n)\pi_i(\theta_n)d\theta_n = \frac{\Gamma\left(\alpha_i^{(n-1)}+x_n\right)}{\Gamma\left(\alpha_i^{(n-1)}\right)x_n!}\left[\frac{m_n}{\beta_i^{(n-1)}+m_n}\right]^{x_n}\left[\frac{\beta_i^{(n-1)}}{\beta_i^{(n-1)}+m_n}\right]^{\alpha_i^{(n-1)}} \\ M_i^-(x_n) &= \int f(x_n|\theta_n)\pi_i^-(\theta_n)d\theta_n = \frac{\Gamma\left(\alpha_i^{(n-1)}+x_n\right)}{\Gamma\left(\alpha_i^{(n-1)}\right)x_n!}\left[\frac{\Lambda_1 m_n}{\beta_i^{(n-1)}+\Lambda_1 m_n}\right]^{x_n}\left[\frac{\beta_i^{(n-1)}}{\beta_i^{(n-1)}+\Lambda_1 m_n}\right]^{\alpha_i^{(n-1)}} \\ M_i^+(x_n) &= \int f(x_n|\theta_n)\pi_i^-(\theta_n)d\theta_n = \frac{\Gamma\left(\alpha_i^{(n-1)}+x_n\right)}{\Gamma\left(\alpha_i^{(n-1)}\right)x_n!}\left[\frac{\Lambda_2 m_n}{\beta_i^{(n-1)}+\Lambda_2 m_n}\right]^{x_n}\left[\frac{\beta_i^{(n-1)}}{\beta_i^{(n-1)}+\Lambda_2 m_n}\right]^{\alpha_i^{(n-1)}} . \end{aligned}$$

Then we have:

$$\begin{aligned} p(\theta_n|\mathbf{X}_n) \propto \sum_{i=0}^{3^{n-1}-1} \Big[&(1-P_1-P_2)\, w_i^{(n-1)} M_i(x_n) p_i(\theta_n|\mathbf{X}_n) \\ &+ P_1\, w_i^{(n-1)} M_i^-(x_n) p_i^-(\theta_n|\mathbf{X}_n) + P_2\, w_i^{(n-1)} M_i^+(x_n) p_i^+(\theta_n|\mathbf{X}_n)\Big] . \end{aligned}$$

If we will call NC to be the normalizing constant of the posterior distribution i.e.

$$NC = \sum_{i=0}^{3^{n-1}-1}\Big[(1-P_1-P_2)\, w_i^{(n-1)} M_i(x_n) + P_1\, w_i^{(n-1)} M_i^-(x_n) + P_2\, w_i^{(n-1)} M_i^+(x_n)\Big]$$

then we get:

$$\begin{aligned} p(\theta_n|\mathbf{X}_n) = \sum_{i=0}^{3^{n-1}-1}\Bigg[&\left(\frac{(1-P_1-P_2)\, w_i^{(n-1)} M_i(x_n)}{NC}\right) p_i(\theta_n|\mathbf{X}_n) \\ &+ \left(\frac{P_1\, w_i^{(n-1)} M_i^-(x_n)}{NC}\right) p_i^-(\theta_n|\mathbf{X}_n) + \left(\frac{P_2\, w_i^{(n-1)} M_i^+(x_n)}{NC}\right) p_i^+(\theta_n|\mathbf{X}_n)\Bigg] = \mathbf{(I)}. \end{aligned}$$

Given that the Gamma is conjugate prior for the Poisson, from standard Bayes theory it is easy to show that:

$$\begin{aligned}
p_i(\theta_n|\mathbf{X}_n) &\sim \; Gamma\left(\alpha^{(n-1)}_{i} + x_n,\;\; \beta^{(n-1)}_{i} + m_n\right) \equiv Gamma\left(\alpha^{(n)}_{3i},\;\; \beta^{(n)}_{3i}\right) \\
p_i^-(\theta_n|\mathbf{X}_n) &\sim \; Gamma\left(\alpha^{(n-1)}_{i} + x_n,\;\; \frac{\beta^{(n-1)}_{i}}{\Lambda_1} + m_n\right) \equiv Gamma\left(\alpha^{(n)}_{3i+1},\;\; \beta^{(n)}_{3i+1}\right) \\
p_i^+(\theta_n|\mathbf{X}_n) &\sim \; Gamma\left(\alpha^{(n-1)}_{i} + x_n,\;\; \frac{\beta^{(n-1)}_{i}}{\Lambda_2} + m_n\right) \equiv Gamma\left(\alpha^{(n)}_{3i+2},\;\; \beta^{(n)}_{3i+2}\right) \;.
\end{aligned}$$

Thus the posterior distribution **(I)** will become:

$$p(\theta_n|\mathbf{X}_n) = \sum_{i=0}^{3^{n-1}-1} \left[w^{(n)}_{3i}\; G\left(\alpha^{(n)}_{3i},\; \beta^{(n)}_{3i}\right) + w^{(n)}_{3i+1}\; G\left(\alpha^{(n)}_{3i+1},\; \beta^{(n)}_{3i+1}\right) + w^{(n)}_{3i+2}\; G\left(\alpha^{(n)}_{3i+2},\; \beta^{(n)}_{3i+2}\right)\right]$$

where:

$$w^{(n)}_{3i} = \frac{(1-P_1-P_2)w^{(n-1)}_{i} M_i(x_n)}{NC}, \quad w^{(n)}_{3i+1} = \frac{P_1 w^{(n-1)}_{i} M_i^-(x_n)}{NC} \quad \text{and} \quad w^{(n)}_{3i+2} = \frac{P_2 w^{(n-1)}_{i} M_i^+(x_n)}{NC} \;.$$

Appendix B: Robustness of the Parameter Evolution Model

The goal of the proposed random size and occurrence change point model for the underlying parameter (θ_n), was to be general enough to cover various realistic alternatives usually encountered in SPC. In this Appendix we will assess the performance/robustness of the proposed model to certain types of model evolution misspecifications. Precisely, we will examine four (quite) different modeling alternatives regarding θ with a process of length 30:

M1: Step change model: The parameter θ starts at $\theta_1 = 4$ and it is piecewise constant with three step changes of size $+4, -2$ and $+3$ occurring at locations 7, 15 and 22 respectively.

M2: Ramp change model: The parameter θ is piecewise constant with values 4, 10 and 6, during the time segments [1,6], [12,17] and [25,30] respectively. For the remaining two segments: [6,12] and [17,25] it changes linearly with slopes $+1$ and -0.5 respectively.

M3: Sinusoidal model: The parameter changes according to the sinusoidal model, for $n = 1, 2, \ldots, 30$

$$\theta_n = 4 + 2sin\left(\frac{2\pi}{29}(n-1)\right) .$$

M4: AR(1) model: The parameter varies according to the autoregressive of order 1 model:

$$\theta_n = c + \phi\theta_{n-1} + \epsilon_n$$

with $c = 1.2$, $\phi = 0.7$ (so that $E[\theta_n] = c/(1-\phi) = 4$) and $\epsilon_n \sim N(0, 1)$.

M1-M3 present a deterministic sequence of θ_n while M4 is a random sequence – the plots of M1-M4 can be seen in Figure 5. For each scenario we use the true vector $(\theta_1, \ldots, \theta_{30})$ to generate 1,000 data sequences $\left(x_1^{[i]}, \ldots, x_{30}^{[i]}\right)$, where $x_j^{[i]} \sim Poisson\,(\theta_j)$, with $j = 1, \ldots, 30$ and $[i]$ denoting the iteration number ($i = 1, 2, \ldots, 1,000$). Before running the proposed model for each data stream, we need to specify the priors for the nuisance parameters which are set to:

$$\theta_0 \sim Gamma(4,1), \qquad \boldsymbol{p} \sim Dirichlet(1,1,1), \qquad \lambda_1 \sim Beta(1,1) \ \text{ and } \ \lambda_2 \sim IB(10,5)\ .$$

Apart from the objective prior choice of $\boldsymbol{p}$ and λ_1 we have a somewhat informative prior for θ_0 (since all models starts at $\theta_1 = 4$) a vaguely informative choice for λ_2 (due to absence of available prior information).

Next, for each scenario, we apply the proposed model to each sequence of data, approximating the exact posterior distribution with a mixture of 100 components (i.e. we have the exact posterior for the first four data points and the approximate from there on). The posterior distribution is summarized by the posterior mean estimate (which is Bayes rule under squared error loss). The boxplots of the posterior means over all 1,000 iterations, along with the true underlying θ values, for each of the models M1-M4, are provided in Figure 5.

As we observe, the proposed model appears to be quite robust to various misspecifications of the underlying model parameter θ_n. Specifically, based on the posterior mean summary boxplots, we observe that on average the proposed model is capable of detecting the (various) changes of the underlying parameter quite satisfactorily.

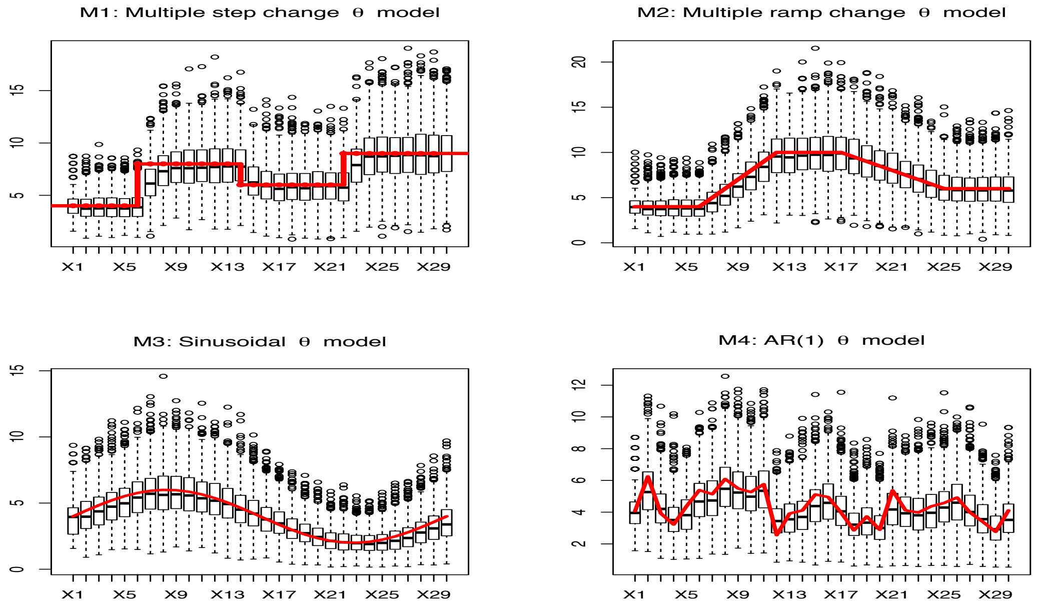

Figure 5: The boxplots of the posterior mean estimate when the model evolution regarding the parameter θ_n is misspecified. The true evolution of θ_n is denoted with the boldface line and corresponds to models M1-M4.

Appendix C: Sensitivity to non-Poisson data

The performance of various frequentist approaches for count data are well known to suffer seriously in cases where the data appear to diverge from the typical Poisson assumption. For example, in Fang (2003) and Hawkins and Olwell (1998), the effect of non-Poison data to the performance of the u control chart, and Poisson CUSUM respectively is explored. On the other hand, quite often in practice the data appear not to conform to the Poisson assumption (for example when failures come in clusters, or when we have a mixture of various types of nonconformities, Jackson, 1972). Most often, count type data appear overdispersed, affecting seriously the performance of standard frequentist methods. The usual practice in such cases is to adopt a distribution with somewhat heavier tails than the Poisson, with the Negative

Binomial being the most popular choice. If we use the pdf form of the Negative Binomial that counts the number of failures before the r-th success (so that we have the same support with the Poisson distribution) then for $Y \sim \text{NB}\big(r, \frac{o}{o+1}\big)$ we have:

$$f_Y(y) = \binom{r+y-1}{y} \left(\frac{o}{o+1}\right)^r \left(\frac{1}{o+1}\right)^y \quad \text{where } r > 0,\ \ o > 0 \ \text{ and } \ y = 0, 1, 2, \ldots$$

with:

$$E[Y] = \frac{r}{o} = \mu \qquad \text{and} \qquad V[Y] = \mu \left(1 + \frac{1}{o}\right)$$

where for any positive o we get $V[Y] > E[Y]$. The parameter o relates to the overdispersion, where the smaller the value the bigger the overdispersion.

Next, we will examine the robustness of our approach in cases where the likelihood is overdispersed compared to Poisson. Specifically, just as we did in Appendix B, we will use models M1-M4 to generate the true vector $(\theta_1, \ldots, \theta_{30})$ for each of the four scenarios. Then we will generate 1,000 sequences of data using three alternative likelihood models, corresponding to three different overdispersion factors o, with 25%, 50% and 100% variance increase (compared to the mean) respectively.

O1: $\left(y_1^{[i]}, \ldots, y_{30}^{[i]}\right)$, where $y_j^{[i]} \sim \text{NB}(4\theta_n, 4/5)$, with $E[Y] = \theta_n$ and $V[Y] = 1.25\theta_n$

O2: $\left(z_1^{[i]}, \ldots, z_{30}^{[i]}\right)$, where $z_j^{[i]} \sim \text{NB}(2\theta_n, 2/3)$, with $E[Z] = \theta_n$ and $V[Z] = 1.5\theta_n$

O3: $\left(w_1^{[i]}, \ldots, w_{30}^{[i]}\right)$, where $w_j^{[i]} \sim \text{NB}(\theta_n, 1/2)$, with $E[W] = \theta_n$ and $V[W] = 2\theta_n$

with $j = 1, \ldots, 30$ and $i = 1 \ldots, 1000$ denoting the data point and iteration respectively. We will adopt the identical prior setup to the one presented in Appendix B and we will approximate the exact posterior distribution using only 100 components. In Figure 6 we provide the boxplots of the posterior means over all 1,000 iterations, along with the true underlying θ values, for each of the models M1-M4, when the data are overdispersed according to scenario O3 (where the variance is twice the mean). For the other two scenarios the performance is quite similar. In Figure 7 the Mean Absolute Error (MAE) for the Poisson and the overdispersion scenarios O1-O3 has been plotted for all scenarios M1-M4. As we observe, the MAE appears somewhat inflated as we move to more dispersed data but it is still comparable to the Poisson's MAE. Thus in general the proposed method appears rather robust to various degrees of overdispersion.

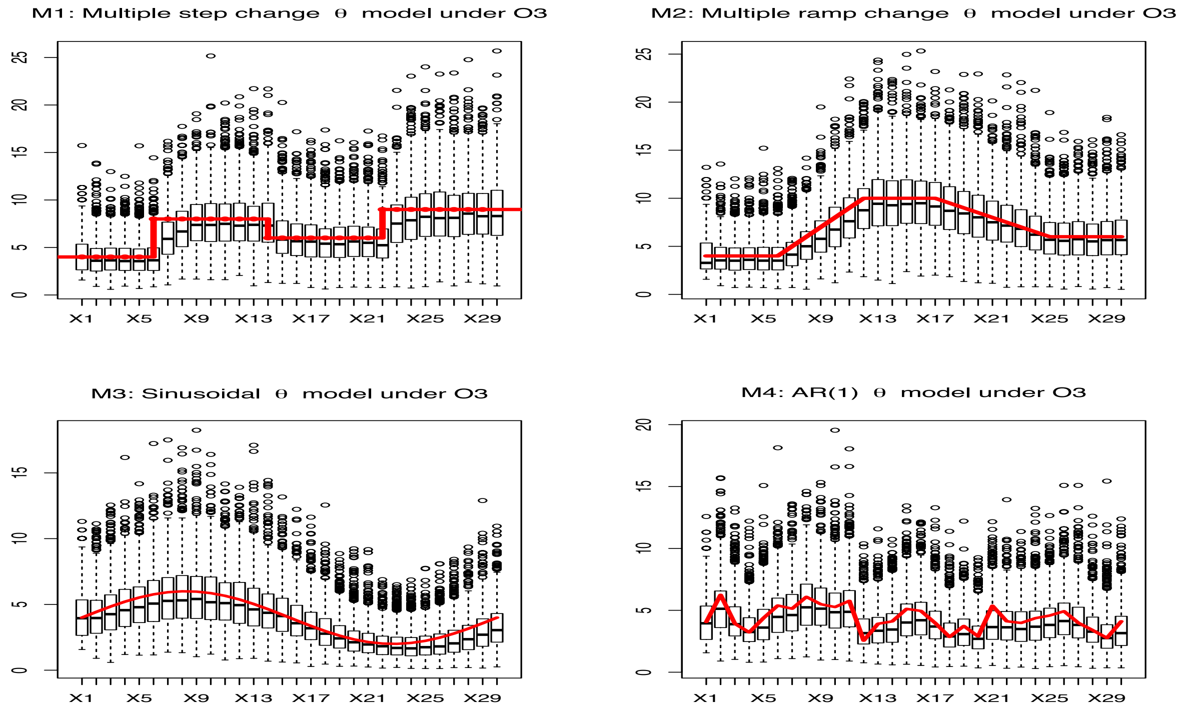

Figure 6: The boxplots of the posterior mean estimate when the model evolution regarding the parameter θ_n is misspecified and the data are overispersed (scenario O3). The true evolution of θ_n is denoted with the boldface line and corresponds to models M1-M4.

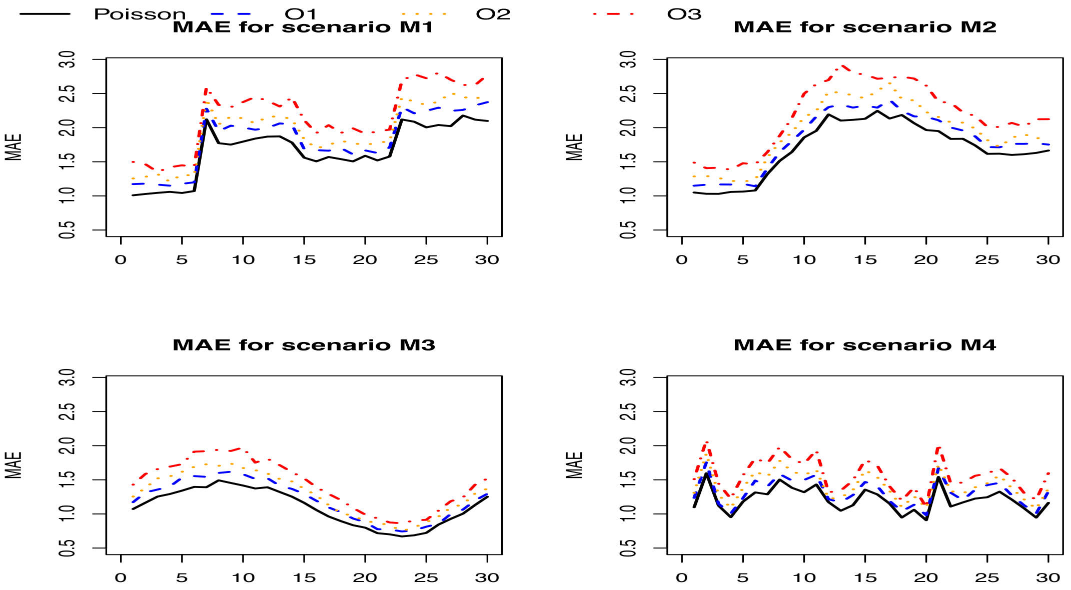

Figure 7: The MAE of the Poisson (solid line) and overdispersed (dashed, dotted and dash-dotted for overdispersed scenarios O1-O3 respectively) data for models M1-M4.

8 Acknowledgments

This research was partially funded by the Research Center of the Athens University of Economics and Business.

References

[1] Feller W. An Introduction to Probability Theory and its Applications, Vol. 1. New York: John Wiley & Sons; 1968.

[2] McCullagh P, Nelder JA. Generalized Linear Models. Chapman & Hall/CRC; 1989.

[3] Montgomery DC. Introduction to Statistical Quality Control, Seventh edition. New York: John Wiley & Sons; 2012

[4] Hawkins DM, Olwell DH. Cumulative Sum Charts and Charting for Quality Improvement. New York: Springer; 1998.

[5] Borror CM, Champ CW, Rigdon SE. Poisson EWMA Control Charts. Journal of Quality Technology 1998;30(4):352-361.

[6] Woodall WH. Control Charts Based on Attribute Data: Bibliography and Review. Journal of Quality Technology 1997;29(2):172-183.

[7] Kaminsky F, Benneyan J, Davis R, Burke R. Statistical Control Charts Based on a Geometric Distribution. Journal of Quality Technology 1992;24(2):63-69.

[8] Sheaffer R, and Leavenworth R. The Negative Binomial Model for Counts in Units of Varying Size. Journal of Quality Technology, 1976;8(3):158-163.

[9] Ryan TP, Schwertman NC. Optimal Limits for Attributes Control Charts. Journal of Quality Technology 1997;29(1):86-98.

[10] Schwertman NC, Ryan TP. Implementing Optimal Attribute Control Charts. Journal of Quality Technology 1997;29(1):99-104.

[11] Shore H. General Control Charts for Attributes. IIE Transactions 2000;32(12):1149-1160.

[12] Quesenberry CP. SPC Q Charts for a Poisson Parameter: Short or Long Runs. Journal of Quality Technology 1991;23(4):296-303.

[13] Alwan LC, Roberts HV. The Problem of Misplaced Control Limits. Journal of Royal Statistical Society, Series C (Applied Statistics) 1995;44:269-278.

[14] Hoadley B. Quality Measurement Plan (QMP). Bell System Technical Journal 1981;60:215-274.

[15] Bayarri MJ, García-Donato G. A Bayesian sequential Look at u-Control Charts. Technometrics 2005;47(2):142-151.

[16] Raubenheimer L, Van der Merwe AJ. Bayesian Control Chart for Nonconformities. Quality Reliability Engneering International 2014, DOI: 10.1002/qre.1668.

[17] Menzefricke U. On Properties of Poisson Q Charts for Attributes. Communications in Statistics - Theory and Methods 2002;31(8):1423-1440.

[18] Shiryaev AN. On Optimum Methods in Quickest Detection Problems. Theory of Probability and its Applications 1963;8:22-46.

[19] Roberts SW. A Comparison of Some Control Chart Procedures. Technometrics 1966;8(3):411-430.

[20] Kenett RS, Pollak M. Data-Analytic Aspects of the Shiryayev-Roberts Control Chart: Surveillance of a Non-Homogeneous Poisson Process. Journal of Applied Statistics 1996;23(1):125-137.

[21] Tsiamyrtzis P, Hawkins DM. Bayesian Startup Phase Mean Monitoring of an Autocorrelated Process That Is Subject to Random Sized Jumps. Technometrics 2010;52(4):438-452.

[22] West M. Approximating posterior distributions by Mixtures. Journal of Royal Statistical Society, Series B 1993;55:409-422.

[23] Apley DW. Posterior Distribution Charts: A Bayesian Approach for Graphically Exploring a Process Mean. Technometrics 2012;54(3):296-310.

[24] Tennant G. Six Sigma: SPC and TQM in Manufacturing and Services. Gower Publishing; 2001.

[25] Lucas JM. The essential Six Sigma. Quality Progress 2002;35(1):27-31.

[26] Jeffreys H. Theory of Probability, Second Edition. Oxford: University Press; 1948.

[27] Quesenberry CP. On Properties of Poisson Q Charts for Attributes. Journal of Quality Technology 1995;27(4):293-303.

[28] Fang Y. c-Charts, X-Charts, and the Katz Family of Distributions. Journal of Quality Technology 2003;35(1):104-114.

[29] Jackson JE. All Count Distributions are not Alike. Journal of Quality Technology 1972;4(2):86-92.

List of Tables

	5	6	7	8	9	10	11	12
MAE_{100}	0.00221	0.01325	0.02570	0.04878	0.06200	0.08639	0.10783	0.12337
MAE_{500}	0	0.00004	0.00065	0.00184	0.00375	0.00893	0.01672	0.02419
MAE_{1000}	0	0	0.00006	0.00031	0.00087	0.00277	0.00593	0.01107
$sd\left(\hat{\theta}_{n,E}\right)$	4.54709	5.31826	5.60457	5.54657	6.31890	6.63497	6.37572	7.06447

Table 1: The Mean Absolute Error at each stage (columns) of the process, for three choices of K : 100, 500 and 1000, along with the standard deviation of the exact posterior mean estimate, $\hat{\theta}_{n,E}$.

Parameter	Prior	Data Simulation settings	Model Misspecification settings
θ_0	$Gamma\,(\alpha_0, \beta_0)$	$Gamma(4, 1)$	$Gamma(1, 1)$
λ_2	$IBeta\,(\zeta, \eta)$	$IBeta(10, 2)$	$IBeta(5, 10)$

Table 2: Sensitivity regarding the choice of hyperparameters: the parameters along with the adopted priors (in the first two columns), the hyperparameters used to simulate the data (column three) and the misspecifications (column four) that were used to run the model.

	1	2	3	4	5	6	7	8	9	10
S1	1.181	1.067	0.965	0.881	0.799	0.729	0.676	0.630	0.595	0.541
	(0.930)	(1.012)	(0.908)	(0.862)	(0.868)	(0.816)	(0.869)	(0.867)	(0.846)	(0.851)
S2	1.462	1.144	0.968	0.867	0.776	0.714	0.663	0.621	0.588	0.535
	(1.295)	(1.202)	(1.015)	(0.920)	(0.907)	(0.834)	(0.878)	(0.876)	(0.853)	(0.853)
S3	1.313	1.164	1.126	1.017	0.926	0.859	0.832	0.793	0.736	0.666
	(1.138)	(1.096)	(1.169)	(1.070)	(1.020)	(1.004)	(1.064)	(1.233)	(1.039)	(1.062)
S4	1.298	1.116	1.099	0.992	0.909	0.847	0.823	0.785	0.730	0.661
	(1.124)	(1.122)	(1.213)	(1.077)	(1.031)	(1.005)	(1.069)	(1.226)	(1.038)	(1.064)

Table 3: The mean and the standard deviation in parenthesis of the absolute error at each stage (columns) of the process for the correct specifications (S1) and each of the three misspecification (S2-S4) scenarios of the hyperparameters.

Size of step change	h^Q	h^{HO}	h^B
0.5σ	7.632	16.299	0.959
1σ	4.715	11.337	0.877
2σ	2.491	7.229	0.569

Table 4: The h thresholds for each of the three methods and each of the three step change scenarios obtained via simulations, so that all methods have the same (5%) false alarm rate under the “in control” scenario.

							Mean (SD)
Method	Loc	Size	h limit	CD	FA	MA	Delay in CD
Qcusum	5	0.5 σ	7.632	15.79%	0.00%	84.21%	14.09 (6.13)
HOcusum	5	0.5 σ	16.299	15.78%	0.00%	84.22%	14.05 (6.17)
BPCP	5	0.5 σ	0.959	28.34%	0.60%	71.06%	12.61 (7.00)
Qcusum	5	1 σ	4.715	31.32%	0.13%	68.55%	9.23 (5.94)
HOcusum	5	1 σ	11.337	32.12%	0.00%	67.88%	9.35 (5.91)
BPCP	5	1 σ	0.877	74.62%	0.58%	24.80%	10.60 (6.70)
Qcusum	5	2 σ	2.491	56.81%	0.55%	42.64%	5.18 (4.69)
HOcusum	5	2 σ	7.229	62.19%	0.21%	37.60%	5.41 (4.65)
BPCP	5	2 σ	0.569	99.40%	0.59%	0.01%	4.17 (2.86)
Qcusum	15	0.5 σ	7.632	20.87%	0.99%	78.14%	8.85 (3.94)
HOcusum	15	0.5 σ	16.299	21.85%	0.87%	77.28%	8.87 (3.97)
BPCP	15	0.5 σ	0.959	18.07%	2.19%	79.74%	7.87 (4.13)
Qcusum	15	1 σ	4.715	52.56%	2.04%	45.40%	7.52 (3.63)
HOcusum	15	1 σ	11.337	53.44%	1.69%	44.87%	7.46 (3.62)
BPCP	15	1 σ	0.877	54.93%	2.20%	42.87%	7.32 (4.05)
Qcusum	15	2 σ	2.491	86.41%	2.47%	11.12%	4.46 (2.98)
HOcusum	15	2 σ	7.229	90.00%	2.11%	7.89%	4.50 (2.94)
BPCP	15	2 σ	0.569	97.09%	2.20%	0.71%	4.05 (2.61)

Table 5: The performance of the three competing methods under various step change scenarios. The columns refer to: location of the step change (col. 2), size of step change (col. 3), the h threshold obtained by Table 4 (col. 4), the percent of Correct Detection (col. 5), False Alarm (col. 6) and Missed Alarm (col. 7) rates. In the last column the mean and standard deviation of the location delay in signaling an alarm for the correctly detected cases is reported.

Time (Month)	1	2	3	4	5	6	7	8	9	1 0	11	12	13	14	15	16
Murder Count	16	17	12	15	14	16	23	19	19	20	26	33	23	21	19	20

Table 6: The available monthly murder counts from Jan 2014 till Apr 2015 in the city of Houston, TX (http://www.houstontx.gov/police/cs/stats2.htm) as reported by the Houston Police Department

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
2010	19	27	18	16	29	26	18	23	20	29	20	15
2011	13	17	12	7	22	15	15	22	18	19	19	13
2012	14	13	27	11	18	18	20	24	16	14	12	18
2013	16	17	16	13	21	18	19	20	17	9	19	24

Table 7: The available historic monthly murder counts in the city of Houston, TX (http://www.houstontx.gov/police/cs/stats2.htm) as reported by the Houston Police Department

Month	x_n	$E[\theta_n \mid \mathbf{X}_n]$	$Pr(\theta_n > U \mid \mathbf{X}_n)$	Pr(no shift)	Pr(down. shift)	Pr(up. shift)
Jan 2014	16	17.978	0.078	0.680	0.073	0.247
Feb 2014	17	18.475	0.111	0.632	0.082	0.286
Mar 2014	12	12.377	0.009	0.305	0.591	0.104
Apr 2014	15	14.042	0.010	0.421	0.047	0.532
May 2014	14	14.418	0.005	0.523	0.085	0.392
Jun 2014	16	16.138	0.017	0.514	0.034	0.452
Jul 2014	23	20.947	0.274	0.329	0.001	0.670
Aug 2014	19	20.624	0.281	0.642	0.084	0.275
Sept 2014	19	20.420	0.279	0.624	0.090	0.286
Oct 2014	20	21.157	0.337	0.607	0.054	0.339
Nov 2014	26	25.419	0.750	0.447	0.003	0.550
Dec 2014	33	31.503	0.987	0.345	0.000	0.655
Jan 2015	23	24.164	0.578	0.528	0.405	0.066
Feb 2015	21	21.304	0.344	0.476	0.171	0.353
Mar 2015	19	20.104	0.226	0.570	0.118	0.312
Apr 2015	20	21.013	0.271	0.582	0.048	0.370

Table 8: The first column reports the date (n), x_n is the murder count during month n in the city of Houston, TX. Column three is the posterior mean at stage n of the process, $\hat{\theta}_n = E[\theta_n|\mathbf{X}_n]$. Next we provide the posterior coverage of the alternative hypothesis, $Pr(\theta_n > U|\mathbf{X}_n)$. In the last three columns the Pr(no shift), Pr(down. shift) and Pr(up. shift) are the marginal probabilities of no, downward and upward shifts respectively of the unknown parameters θ_n, at each month n, independently of the past history.