Journal of Quality Technology
A Quarterly Journal of Methods, Applications and Related Topics

ISSN: 0022-4065 (Print) 2575-6230 (Online) Journal homepage: https://www.tandfonline.com/loi/ujqt20

On computing multiple change points for the gamma distribution

Xun Xiao, Piao Chen, Zhisheng Ye & Kwok-Leung Tsui

To cite this article: Xun Xiao, Piao Chen, Zhisheng Ye & Kwok-Leung Tsui (2020): On computing multiple change points for the gamma distribution, Journal of Quality Technology, DOI: 10.1080/00224065.2020.1717398

To link to this article: https://doi.org/10.1080/00224065.2020.1717398

JOURNAL OF QUALITY TECHNOLOGY
https://doi.org/10.1080/00224065.2020.1717398

On computing multiple change points for the gamma distribution

Xun Xiao[a], Piao Chen[b], Zhisheng Ye[c], and Kwok-Leung Tsui[d]

[a]School of Fundamental Sciences, Massey University, Palmerston North, New Zealand; [b]Institute of High Performance Computing, Singapore; [c]Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore; [d]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon Tong, Hong Kong

ABSTRACT

This study proposes an efficient approach to detect one or more change points for gamma distribution. We plug a closed-form estimator into the gamma log-likelihood function to obtain a sharp approximation to the maximum of log-likelihood. We further derive a closed form calibration of approximate likelihood which is asymptotically equivalent to the exact log-likelihood. This circumvents iterative optimization procedures to find maximum likelihood estimates which can be a burden in detecting multiple change points. The simulation study shows that the approximation is accurate and the change points can be detected much faster. Two case studies on the time between events arising from industrial accidents are presented and extensively investigated.

KEYWORDS

approximate likelihood; calibration; change point analysis; industrial accidents; likelihood ratio test; time between events

1. Introduction

Offline change point detection is a powerful tool for finding one or more abrupt changes that take place in an ordered data sequence. The detection concerns the inference of epochs where statistical properties of a data stream change. Techniques developed for change point detection have received wide applications in many areas, including economics (Andrews 1993; Bai and Perron 1998) and signal processing (Jackson et al. 2005). In manufacturing engineering, change point detection is closely related to statistical process control. It serves as a complementary tool to the standard control charts in the production process. As argued in Woodall and Montgomery (1999), change point detection focuses on estimating the occurrence time of a change once it is detected, while control charting emphasizes on detecting the change as soon as possible after it occurs. As a result, change point models have been embedded into the design of many advanced control charts, *e.g.*, see Hawkins, Qiu, and Kang (2003), Zou, Zhang, and Wang (2006), and Zou, Qiu, and Hawkins (2009). A review on change point analysis for the control chart post-signal diagnostics is provided in Amiri and Allahyari (2012).

Statistically, the offline change point analysis concerns testing, estimation, and computation of possible change points in the parameters of a statistical model given an ordered sequence of realisations from this model. A hypothesis test is usually designed to check the null of no change in the process. Once the null is rejected, the number, times, and magnitudes of the shifts will be further estimated. A number of nonparametric methods that require minimum assumptions on the data generation process have been developed for this purpose. See Brodsky and Darkhovsky (1993) and Zou et al. (2014), among others. Nevertheless, if the parametric law for the data generation process is known, a parametric model is usually more efficient in detecting the changes. Following the pioneer work of Hinkley (1970) on normal distributions, numerous parametric change point procedures have been developed, including the binomial (Hinkley and Hinkley 1970), exponential (Worsley 1986), and gamma distributions (Hsu 1979). A book-length account on this topic can be found in Chen and Gupta (2000). In this line of research, the likelihood based approaches have become the major tool for detecting change points in a parametric model.

This study is confined to change point analysis for a sequence of independent gamma distributed random variables. The gamma distribution is one of the most important distributions for lifetime data (Meeker and Escobar 1998). It is also a very useful model in signal processing and image processing (Vaseghi 2008). Change point analysis of the gamma distribution has been investigated in a series of studies. Hsu (1979)

CONTACT Zhisheng Ye yez@nus.edu.sg Department of Industrial Systems Engineering and Management, National University of Singapore, E1A 06-25, 1 Engineering Drive 2, West Coast, Singapore 129790, Singapore.
Piao Chen is now affiliated with Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

discussed detecting one change point in the gamma scale parameter, and demonstrated the applications in finance and transportation. The work of Hsu (1979) was extended to the Bayesian setting in Diaz (1982). Ramanayake (2005) concerned testing the change in the gamma shape parameter while fixing the scale parameter. In statistical process control, Zhang et al. (2007) studied a control chart for gamma-distributed time between events and Huang et al. (2013) investigated the run-length distribution of CUSUM charts for gamma distributions. Tan, Miao, and Zhou (2013) further studied the consistency of the change-point estimator when the shape parameter changes in negatively associated gamma distribution. Simultaneous change detection of the gamma shape and scale parameters has not been well-addressed in the literature. Although Csörgö and Horváth (1997) studied the likelihood ratio test for a single change point in the exponential family, their analysis is general and not tailored for the gamma distribution.

All the above studies focus on a single change point. In real applications, there can be multiple change points. Hawkins (2001) mentioned that dynamic programing can be used to detect multiple change points when the gamma shape parameter is fixed and known. A more efficient dynamic programing algorithm was implemented in an R package *changepoint* for detecting multiple change points in the gamma scale parameter in Killick and Eckley (2014). However, their package still requires that the shape parameter is unchanged and prespecified. In other words, only the scale parameter can shift in their set-up. We suspect that this restriction may be blamed to the unavailability of closed-form maximum likelihood estimator (MLE) for gamma distribution. The lack of an algorithm for multiple gamma change points that allows both parameters to shift can potentially limit our capability to handle a modern data stream with multiple change points and versatile parameter shift patterns. A motivating example is the time between mine disasters reported in NIOSH (2013) which shows a decreasing pattern from the late 19th century to the early 20th century but starts increasing from the middle 20th century. The bathtub pattern in this data set suggests that there are possibly multiple change points.

In summary, the literature on change point detection for the gamma distribution mainly focused on a single change point, and most studies assumed either the shape or scale parameter is fixed, or even known. To address the issues in detecting the multiple changes in both parameters, this study uses the closed-form estimator proposed in Ye and Chen (2017) to approximate the maximum of the gamma log-likelihood. We show that the new estimator proposed by Ye and Chen (2017) approximates the MLE up to a minor stochastic disturbance which depends only on the gamma shape parameter and converges to zero when the gamma shape parameter goes large. Therefore, substituting this new estimator into the gamma log-likelihood function induces a sharp approximation to the maximum of exact gamma log-likelihood. Furthermore, a detailed analysis on the approximation error induces a simple closed form calibration of the approximation which is asymptotically equivalent to the exact maximum. This approximation circumvents the numerical optimization to maximize the gamma log-likelihood function which may fall into pitfalls occasionally. Especially, when computing multiple change points, one has to optimize the gamma log-likelihood functions for different segments from the data sequence. The sufficient statistics of some segments may be nearly ill-posed which results in slow convergence of numerical algorithms. Both theoretical analyses and simulation studies show that the approximation error is very small and almost negligible to the order of log-likelihood. As the resulting procedures are constructed in closed forms, the approximate likelihood can be used to quickly detect the change points while achieving almost the same performance as the exact likelihood.

The rest of this article is organized as follows. Section 2 presents the general framework for the computation of the change point(s). Section 3 discusses the approximate log-likelihood and its calibration. Section 4 presents the implementation of the algorithms to detect change points with both exact and approximate log-likelihoods. Section 5 presents the simulation studies to demonstrate the computation efficiency and estimation accuracy of our approaches. The case studies on real data sets are presented in Section 6. Concluding remarks and future perspectives are given in Section 7.

2. Formulation of change points problem

The probability density function (PDF) of a gamma distribution, denoted as Gamma(λ, k), is given by

$$f(x;\lambda,k)=\frac{x^{k-1}\mathrm{e}^{-x/\lambda}}{\lambda^{k}\Gamma(k)}, x\geq 0, \qquad [1]$$

where k is the shape parameter, λ is the scale parameter, and $\Gamma(\cdot)$ is the gamma function. Given an i.i.d. sample $Y_{1:n}=(Y_1, Y_2, ..., Y_n)$ from Gamma(λ, k), $l(\cdot)$

is the associated log-likelihood function as follows,

$$l(\lambda, k; Y_{1:n}) = n\left[(k-1)\overline{\ln Y} - \bar{Y}/\lambda - \ln\Gamma(k) - k\ln\lambda\right], \quad [2]$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ and $\overline{\ln Y} = \sum_{i=1}^{n} \ln Y_i/n$ are the sufficient statistics for gamma distribution. The MLE of $\boldsymbol{\theta} = (\lambda, k)^T$ is denoted as $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{k})^T$.

Consider a sequence of independent but non-identical gamma random variables $X_{1:t} = (X_1, X_2, ..., X_t)$ where $t \in \mathbb{N}$ is the length of the sequence. Let $\tau_{1:m} = (\tau_1, \tau_2, ..., \tau_m)$ be the unknown change points satisfying $1 < \tau_1 < \tau_2 < ... < \tau_m < t$. We use $X_{1:t}$ with possible change points to discriminate it from the i.i.d. sample $Y_{1:n}$. Here, m is the number of change points which is also unknown. Naturally, m change points divide the data to $m+1$ segments. In the ith segment, $X_{(\tau_{i-1}+1):\tau_i} = (X_{\tau_{i-1}+1}, ..., X_{\tau_i}) \sim \text{Gamma}(\lambda_i, k_i)$ where $1 \le i \le m+1$ with the convention $\tau_0 = 1$ and $\tau_{m+1} = t$. The parameters $\lambda_{1:(m+1)}$ and $k_{1:(m+1)}$ are also unknown. We assume $m \ge 1$ to ensure that there exists at least one change point as we focuses on the computation of the change point(s). We further assume that $\lambda_i \ne \lambda_j$ or $k_i \ne k_j$ if $i \ne j$ for the identifiability of the change points. Therefore, we need to estimate the total number of the change points m, the locations of the change points $\tau_{1:m}$, and the gamma parameters $\lambda_{1:(m+1)}$ and $k_{1:(m+1)}$.

If there is a single change point, *i.e.*, $m = 1$, we can estimate the location of this unique change point τ by searching the maximum of the log-likelihood function as follows,

$$\hat{\tau} = \arg\max_{0<\tau<t} [l(\hat{\lambda}_1, \hat{k}_1; X_{1:\tau}) + l(\hat{\lambda}_2, \hat{k}_2; X_{(\tau+1):t})] \quad [3]$$

where the hat notation denotes that this parameter is an estimate. Testing and estimating a single change point in a sequence are usually carried out together. The log-likelihood ratio test statistic is defined by the maximum of the log-likelihood ratio process as follows.

$$\Lambda^* = \max_{0<\tau<t} \{2 \cdot \omega(\tau) \cdot [l(\hat{\lambda}_1, \hat{k}_1; X_{1:\tau}) + l(\hat{\lambda}_2, \hat{k}_2; X_{(\tau+1):t}) - l(\hat{\lambda}, \hat{k}; X_{1:t})]\} := \max_{0<\tau<t} \Lambda(\tau), \quad [4]$$

where $\omega(\tau) = \tau(t-\tau)/t^2$ is a weight function to ensure the form of asymptotic limit—the maximum norm of a standard two-dimensional Brownian bridge (Csörgö and Horváth 1997). The statistic Λ^* is distribution-free and can then be used to test the null hypothesis. Since $l(\hat{\lambda}, \hat{k}; X_{1:t})$ is a constant as a function of τ, the log-likelihood ratio test directly reports a change point once H_0 is rejected. This procedure is also called testimation of a single change point. It is well-known that the gamma log-likelihood function in Eq. [2] is concave with a unique maximum. However, searching the maximum point, *i.e.*, MLE, does not yield a closed form solution. To estimate λ and k, some iterative optimization algorithms, *e.g.*, the generalized Newton-Raphson method, have to be used. We need to run the iterative algorithm around $2(t-1)$ times to find $\hat{\tau}$ in Eq. [3]. In other words, the time complexity to detect a single change point is typically $O(t)$. The computation cost can be potentially expensive when t is large.

Consider multiple change points. When m is known, to find the locations of the change points, the estimation of multiple change points can be formulated as the follows:

$$(\hat{\tau}_1, \hat{\tau}_2, ..., \hat{\tau}_m) = \arg\max_{0<\hat{\tau}_1<\cdots<\hat{\tau}_m<t} \left[\sum_{i=1}^{m+1} l(\hat{\lambda}_i, \hat{k}_i; X_{(\hat{\tau}_{i-1}+1):\hat{\tau}_i})\right]. \quad [5]$$

Eq. [5] can be regarded as a combinatorial optimization problem. Since there are $\binom{t-1}{m}$ choices for the locations of m change points, brute-force search of the whole parameter space is time-consuming when m is large. Moreover, the number of change points m is usually unknown in most real applications. If m is unknown, estimating the number of the change points can be regarded as a model selection problem. To balance the goodness of fit and the parsimony of the model, some penalties on the model complexity, *i.e.*, a penalty function of m, shall be added into the target function in Eq. [5] when searching for a best model:

$$(\hat{\tau}_1, \hat{\tau}_2, ..., \hat{\tau}_{\hat{m}}) = \arg\min_{0<\hat{\tau}_1<\cdots<\hat{\tau}_{\hat{m}}<t} \left[-2\sum_{i=1}^{\hat{m}+1} l(\hat{\lambda}_i, \hat{k}_i; X_{(\hat{\tau}_{i-1}+1):\hat{\tau}_i}) + \beta\hat{m}\right]. \quad [6]$$

Here, $\beta = 2p$ for Akaike information criterion (AIC) and $\beta = p\ln t$ for Bayesian information criterion (BIC), where $p = 3$ is the number of additional unknown parameters when a change point is added into the gamma model. Since BIC is consistent in change point model selection of the exponential family (Lee 1998; Yao 1988), it is used in this study. The number of possible models with multiple change points is $\sum_{m=1}^{t-1}\binom{t-1}{m} = 2^{t-1} - 2$ which is of exponential order. Exhaustion of all possibilities for change

point estimation proliferates the time complexity to a prohibitive order of $O(2^t)$.

Estimation of multiple change points can be greatly simplified when the target function satisfies Bellman's principle of dynamic programing (DP). Particularly, there are some nice recursive structures in Eq. [5] which can be explored efficiently by DP algorithms to greatly reduce the computation cost from an unfeasible order to a quadratic or even linear order. Quite a few DP algorithms have been developed for multiple change point problem. DP algorithms with $O(t^2)$ time-complexity have been studied in Hawkins (2001), Bai and Perron (2003), and Jackson et al. (2005) for detecting change points in the exponential families and linear regression models. However, algorithms with a quadratic time complexity may be infeasible even for the modern computer when t is very large (Huo and Székely 2016). More recently, Rigaill (2015) and Killick, Fearnhead, and Eckley (2012) proposed two DP algorithms with pruning. Both algorithms achieve a nearly linear computational cost when the distribution of change points satisfies some regularity conditions. In this study, we will extend the pruned exact linear time algorithm proposed by Killick, Fearnhead, and Eckley (2012) which achieves a linear time complexity $O(t)$ when the number of the change points m increases linearly with the data length t, and $O(t^2)$ in the worst case with a fixed number of change points.

3. Approximate likelihood and its calibration

As argued, the lack of a closed-form MLE makes it time-consuming to solve the multiple change point problem in Eq. [6]. To circumvent the possible hindrance of the numerical computation, we use the closed-form estimator $\tilde{\boldsymbol{\theta}} = (\tilde{\lambda}, \tilde{k})$ proposed in Ye and Chen (2017), which is given by

$$\begin{aligned} \tilde{\lambda} &= \overline{Y \ln Y} - \bar{Y} \cdot \overline{\ln Y}, \\ \tilde{k} &= \bar{Y}/\tilde{\lambda}, \end{aligned} \qquad [7]$$

where $\overline{Y \ln Y} = \sum_{i=1}^{n} Y_i \ln Y_i / n$. Basic asymptotic properties of $\tilde{\boldsymbol{\theta}}$ have been investigated in Ye and Chen (2017). In this study, it is found that this new estimator can be regarded as a stochastic approximation to MLE and the approximation error depends only on the shape parameter k and converges to zero rapidly when $k \to \infty$. The details are presented in Appendix.

Let $l^* = l(\hat{\boldsymbol{\theta}})$ and $\tilde{l}^* = l(\tilde{\boldsymbol{\theta}})$ be the log-likelihood values corresponding to the MLE and the estimator proposed by Ye and Chen (2017), respectively. Then $\tilde{l}^*$, the approximate maximum of log-likelihood (AML) is an approximation to l^*, the exact maximum of log-likelihood function (EML). We further show in the Appendix that

$$2(l^* - \tilde{l}^*) \to k(k^2\psi_1^2(k) - \psi_1(k) - 1) \cdot \chi^2(1). \qquad [8]$$

where $\psi_1(\cdot)$ is the trigamma function. The rescaling factor $\sigma_l^2(k) = k \cdot (k^2\psi_1^2(k) - \psi_1(k) - 1) \sim O(1/k) \to 0$ as $k \to \infty$. Although Eq. [8] suggests the approximation error $l^* - \check{l}* \sim O_p(1)$, the relative approximation accuracy of AML is $O_p(1/n)$ since $l^* \sim O_p(n)$. Therefore, the AML $\tilde{l}^*$ is expected to give a satisfactory approximation to the EML l^* and our simulation will further justify this result.

Continuing the analysis on the approximation error of $\tilde{l}^*$, we can calibrate the approximation error and obtain a better approximation which is equivalent to the EML l^* in an asymptotical sense. The AML with calibration (AMLC) is given by

$$\check{l}* = \tilde{l}^* + \frac{n\left[\overline{\ln Y} - \ln \bar{Y} - \psi(\tilde{k}) + \ln \tilde{k}\right]^2}{2\left[\psi_1(\tilde{k}) - 1/\tilde{k}\right]}, \qquad [9]$$

where $\psi(\cdot)$ is the digamma function. This correction yields an asymptotically valid approximation to the EML l^* which means $\check{l}* - l^* \to_p 0$. The detailed discussion can be found in Appendix.

Furthermore, replacing the exact MLE $\hat{\boldsymbol{\theta}}$ in Eq. [4] by $\tilde{\boldsymbol{\theta}}^*$, we obtain an approximate likelihood ratio test (LRT) statistic denoted by $\tilde{\Lambda}^*$. We can further calibrate the approximate LRT statistic $\tilde{\Lambda}^*$ by Eq. [9] to obtain the calibrated approximate LRT statistic $\check{\Lambda}*$ which follows the same asymptotic distribution as Λ^*. However, one shall notice that $\tilde{\Lambda}^*$ differs from Λ^* by $O_P(n^{-1/2})$. Since $\tilde{\boldsymbol{\theta}}$ is a sharp approximation to $\hat{\boldsymbol{\theta}}$, we simply use the null distribution of Λ^* as a surrogate of the null distribution of $\tilde{\Lambda}^*$ which can be complicated. A simulation study is used to further validate the appropriateness of the approximation in Section 5.

4. Algorithm development

The testimation of a single change point can be achieved easily by finding the LRT statistic Λ^* defined in Eq. [4], while the efficient computation of multiple change points requires the DP algorithms. In particular, we introduce a vectorization trick to greatly accelerate the detection of a single change point, and extend the DP algorithm proposed by Killick, Fearnhead, and Eckley (2012) for multiple change points with the vectorization trick.

4.1. Testimation of a single change point

The testimation of one change point seems easy at the first sight. But one shall notice that the maximum in Eq. [4] is taken over $t-2$ possible log-likelihood maximum values which heavily rely on some numerical algorithms. Some loop structures can be used to find the values one by one without any difficulties in programing. But this naive approach leads to a burden when t gets larger especially for R since the loops in R can be rather slow. The key to circumvent the loop issue is the vectorization. The vectorization of algorithms is fundamentally important for their efficient implementation in R; See Eddelbuettel et al. (2011) and Venables and Ripley (2013) for examples.

Particularly, we apply a fast numerical algorithm for computing the exact MLE based on generalized Newton method in Minka (2002). Though there is already a function *fitdistr* in the R package *MASS* (and other alternatives in R) which can compute the MLE for a lot of parametric distributions including gamma distribution, Minka's algorithm is much faster than *fitdistr* for gamma distribution and can even be vectorized to evaluate MLEs of many segments simultaneously. For any point $\tau \in \{2,3,\ldots,t-2\}$, it separates the data sequence $X_{1:t}$ to two segments $X_{1:\tau}$ and $X_{(\tau+1):t}$ with at least two observations. We can compress the information in the two segments by the sufficient statistics of gamma distribution as $(\sum_{i=1}^{\tau} X_i, \sum_{i=1}^{\tau} \ln X_i, \sum_{i=\tau+1}^{t} X_i, \sum_{i=\tau+1}^{t} \ln X_i)$. By varying τ, the sufficient statistics for all segments are summarized by the following cumulative sum sequences of X_i and $\ln X_i$ as

$$\left\{\left(\sum_{i=1}^{\tau} X_i, \sum_{i=1}^{\tau} \ln X_i, \sum_{i=1}^{t} X_i - \sum_{i=1}^{\tau} X_i, \sum_{i=1}^{t} \ln X_i - \sum_{i=1}^{\tau} \ln X_i\right) | \tau = 2,3,\ldots,t-2\right\} \quad [10]$$

Minka's algorithm can handle the multiple pairs of gamma sufficient statistics and return their associated MLEs simultaneously. The log-likelihood ratios can be obtained and maximized accordingly.

Even if the exact LRT statistic can be calculated efficiently by our vectorization scheme, the approximate LRT statistic and its calibration are attractive in avoiding the numerical optimization which occasionally falls into some pits especially when $\ln(\sum_{i=1}^{\tau} X_i) \approx \sum_{i=1}^{\tau} \ln X_i$. The estimator by Ye and Chen (2017) and the further calibration can be easily vectorized for all segments by adding the cumulative sum sequences of the auxiliary statistics $\{(\sum_{i=1}^{\tau} X_i \ln X_i, \sum_{i=\tau+1}^{t} X_i \ln X_i) | \tau = 3,4,\ldots,t-3\}$ to Eq. [10].

4.2. Catching multiple change points with DP algorithm

In this subsection, we present the DP algorithm with EML and AML to optimize Eq. [6]. Our implementation is based on the pruned exact linear time (PELT) algorithm proposed in Killick, Fearnhead, and Eckley (2012). This algorithm has been integrated into the R package *changepoint* developed by Killick and Eckley (2014) for the change point detection in several parametric models, including the normal, Poisson, exponential, and a special case of gamma with prefixed shape parameter. Killick's algorithm can be regarded as an extended version of the optimal partitioning (OP) approach proposed in Jackson et al. (2005).

Given data $X_{1:s}$, denote the minimum in Eq. [6] as $M(s)$. Then the following recursive relation $M(s) = \min_{r<s}[M(r) - 2l^*(X_{(r+1):s}) + \beta]$ holds where $M(0) = -\beta$ (Jackson et al. 2005). Based on this recursive formulation, Eq. [6] can be solved by searching from $s=1$ to $s=t$ greedily. For each s, one need to search from $r=1$ to $r=s-1$. Then, the OP approach is an algorithm with $O(t^2)$ time complexity for Eq. [6] as we need to search over all $t(t-1)/2$ segments in the form $\{X_{r:s} | s=1,\ldots,t, r=1,\ldots,s-1\}$. Killick, Fearnhead, and Eckley (2012) further accelerated the OP by introducing a pruning technique to avoid searching the whole subsequence from $r=1$ to $r=s-1$. Their results showed that only a few points can be the candidate change points if the log-likelihood always increases when a change point is introduced into the sequence. More precisely, if $l^*(X_{r:s}) + l^*(X_{(s+1):u}) \geq l^*(X_{r:u})$ for all $r<s<u$, r will never be the last change point before u when $M(r) - 2l^*(X_{(r+1):s}) \geq M(s)$ holds. Then, for any subsequence $X_{1:u}$ with $u>s$, s will always be the last change point detected. Therefore, we can ignore the point $\{r | 1 \leq r < s, M(r) - 2l^*(X_{(r+1):s}) \geq M(s)\}$ in each iteration of the OP approach.

If we replace the exact log-likelihood by the approximate log-likelihood, the error term may influence the results presented above. The algorithm can still work, but it does not guarantee finding the same solution to Eq. [5]. However, we can achieve a fast computational rate at a small price of approximation. By combining the vectorization trick and the PELT algorithm, we summarize the algorithm for gamma distributed sequence with the approximate log-likelihood in Algorithm 1 which outlines the general idea

on pruning of PELT algorithm when calculating multiple change points.

Algorithm 1. Pseudo-code of PELT algorithm with AML

Input: $X_{1:t} = (X_1, X_2, ..., X_t)$; β: the constant penalty term; h: the minimum segment length;

$\mathcal{R}_{2h} = \{0, h\}; M(0) = -\beta; \mathcal{C}_{2h-1} = \{0\}$.

for $s = h, ..., 2h-1$ **do**

$M(s) = -2\tilde{l}^*(X_{(r+1):s})$.

end for

for $s = 2h, ..., t$ **do**

for $r \in \mathcal{R}_s$ **do** ▷ This inner for loop is vectorized.

Calculate the candidate current optimum $\mathcal{M}(r) = M(r) - 2\tilde{l}^*(X_{(r+1):s}) + \beta$.

end for

Find the current optimum $M(s) = \min_{r\in\mathcal{R}_s}\mathcal{M}(r)$;

Find the current last optimal change point $v = \arg\min_{r\in\mathcal{R}_s}\mathcal{M}(r)$;

Update the change point set with the last optimal change point $\mathcal{C}_s = \mathcal{C}_v \cup \{v\}$;

Update the change point candidate set $\mathcal{R}_{s+1} = \{s\} \cup \{r \in \mathcal{R}_s : M(r) - 2\tilde{l}^*(X_{(r+1):s}) < M(s)\}$.

end for

Output: Change points recorded in $\mathcal{C}_t$.

For the detailed implementation of Algorithm 1 with AML, one should notice that the evaluation of the candidate current optima $\mathcal{M}(r) = M(r) - 2\tilde{l}^*(X_{r+1:s}) + \beta$ for all $r \in \mathcal{R}_s$ can be very fast by the vectorization described in the last subsection. Particularly, $M(r)$ is already given and the vectorized AML formula (or vectorized Minka's algorithm) can evaluate all the approximate log-likelihoods $\tilde{l}^*(X_{r+1:s}), r \in \mathcal{R}_s$ (or exact log-likelihoods $l^*(X_{r+1:s}), r \in \mathcal{R}_s$) in the inner for loop of Algorithm 1 at the same time. Therefore, we can get rid of the inner for loop when implementing the algorithm to detect gamma change points since $\mathcal{M}(r)$ can be obtained without using any loop structure. To ensure that the parameters are estimable, the minimum segment length h needs to be specified to be equal to or larger than three. With the minimum segment length h, it is necessary to initialize the hth to $(2h-1)$ th in $M(i)$ before starting to search the change points.

The DP algorithm with EML can be obtained with some minor modifications on Algorithm 1 after replacing $\tilde{l}^*(X_{r:s})$ with $l^*(X_{r:s})$. In particular, the summary statistics are the cumulative sums of $X_1, X_2, ..., X_t$ and $\ln X_1, \ln X_2, ..., \ln X_t$ which are the sufficient statistics of gamma distribution. Therefore, the minimum segment length can be reduced to two. The simple calculation of the approximate likelihood is replaced by the numerical optimization algorithm in Minka (2002).

Both the exact and approximate log-likelihoods can be integrated into the PELT algorithm to detect the change points. However, the exact MLE requires maximizing the log-likelihood function via some numerical algorithms. The numerical algorithm needs much more computation resources than the closed form solution. Furthermore, the global convergence of numerical algorithms can be a problem when we try to optimize a lot of log-likelihood functions over different segments simultaneously.

5. Simulation study

In this section, extensive simulation studies are carried out to investigate the performance of the EML and the two proposed AML for gamma change point problem. All of the simulations are run by R 3.5.1. Here, we assume that each segment has at least three observations for the estimability of gamma parameters with the closed form estimator proposed by Ye and Chen (2017) and the comparability between EML and AML. We use $(\lambda_1, k_1) \xrightarrow{\tau} (\lambda_2, k_2)$ to denote the event that the parameter vector changes from (λ_1, k_1) to (λ_2, k_2) right after the τth observation. When there are m change points, the following notation is used,

$$(\lambda_1, k_1) \xrightarrow{\tau_1} (\lambda_2, k_2) \xrightarrow{\tau_2} \cdots \xrightarrow{\tau_m} (\lambda_{m+1}, k_{m+1}).$$

5.1. Testimation of a single change point

We first investigate the performance of several approaches to detect a single change point, including the exact LRT, approximate LRT with and without calibration, and AMOC(k) which is the At most one change approach implemented in the package *changepoint*. AMOC(k) only works when the shape parameter does not change and its value is manually supplied as k. The BIC is used as penalty in AMOC(k).

Firstly, we compare the sizes and powers of approximate LRTs with those of alternatives under several different parameter set-ups at the significance level $\alpha = 0.05$. Therefore, the nominal sizes of the tests are all equal to 0.05. The length of the data sequence is given as $t = 50$, 100, 300, and a change point is located at $\tau = t/2$ or $\tau = t/5$.

We consider two different parameter shift set-ups as

$$(\lambda_1, k_1) = (2, 2) \xrightarrow{\tau} (\lambda_2, k_2) = (2 + 0.1h, 2 + 0.1h)$$

and

$$\begin{aligned}(\lambda_1, k_1) &= (1, 1.5) \xrightarrow{\tau} (\lambda_2, k_2) \\ &= (1.5/(1.5 + 0.1h), 1.5 + 0.1h)\end{aligned}$$

where $h = -10, \ldots, -1, 0, 1, \ldots, 10$. In the first set-up both the mean and variance of the sequence will change while in the second set-up only the variance will change. For AMOC(k), the shape parameter k is fixed at $k = 2$ and $k = 1.5$ for the above two scenarios, respectively.

By varying h, we can adjust the signal strength of a change to see its effect on the power of the test, *i.e.*, one minus the type-II error probability, which is given as follows,

$$\text{Power} = \Pr(\text{Reject} \quad H_0 \mid H_0 \quad \text{is false}).$$

Particularly, when $h = 0$, there is no change in the data stream and we would get the Monte Carlo estimate of the actual size of the test, *i.e.*, the type-I error probability, which is given as follows,

$$\text{Size} = \Pr(\text{Reject} \quad H_0 \mid H_0 \quad \text{is ture}).$$

As can be seen from Figures 1 and 2, the power functions of all three LRTs are almost the same in all figures. We conclude that either of two approximate tests can be used as an efficient surrogate for the exact LRT. The power functions of all LRTs exhibit a reasonable trend, *i.e.*, decreasing when $h < 0$ and increasing $h > 0$. A large data length increases the powers as one may expect. When $h = 0$ the actual sizes of LRTs are slightly larger than the nominal size 0.05, but they approach the nominal size when t increases. Since we are using the asymptotical null distribution in our test, the actual sizes look acceptable. One should notice that the results also suggest that the change point in the second parameter shift set-up is considerably harder to detect, especially when the change point is not located in the middle of the data.

To test a single change point is a typical problem with a linear time-complexity $c \cdot O(t)$, where the constant c accounts for the time to compute the maximum log-likelihood for each segment. To compare the runtime of the EML and AML, we let t vary from 10,000 to 100,000, and the change point is located at the middle of the process. The parameter shift set-up of the process is $(2, 5) \xrightarrow{\tau} (2, 10)$ at the change point. The shape parameter k is fixed at 2 which allows us to use AMOC(k) approach in the R package *changepoint* by Killick and Eckley (2014). Their package has been well developed, maintained, and updated in past years. Therefore, their approach can serve as the benchmark to compare the computation efficiency. We apply all four approaches to compute this unique change point with 10 replications for each data length. All approaches have been vectorized to achieve their fastest possible speeds. The averages of the computation times for different data lengths are presented in Figure 3.

From Figure 3, it is easily observed that all these four approaches achieve a linear computational complexity as we may expect. Furthermore, approximate tests and AMOC(k) both equipped with closed forms are much faster than the exact LRT which requires

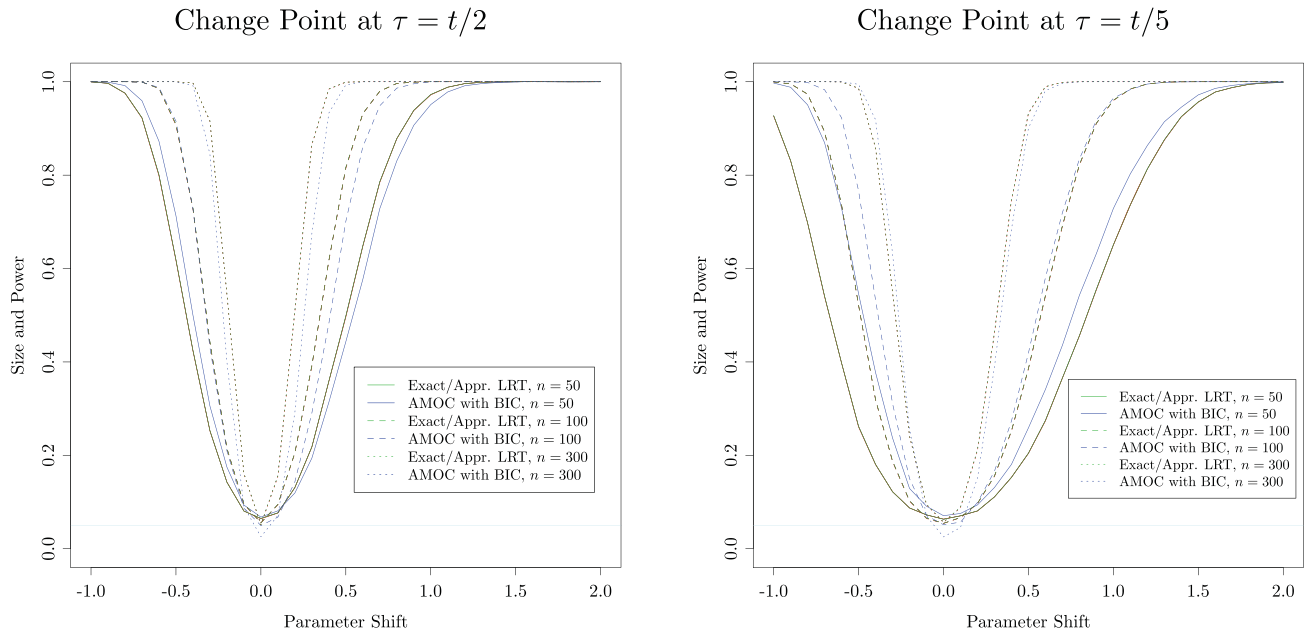

Figure 1. Power functions of LRTs (in green lines) and AMOC($k = 2$) (in blue lines) with $t = 50$ (in solid lines), 100 (in dash lines), and 300 (in dot lines) for the first parameter shift set-up. Note: All three power functions for exact and approximate LRTs are plotted in different colors (black - exact LRT, red - approximate LRT, green - approximate LRT with calibration), but they overlap with each other and only the green one is shown in the plot and the legend.

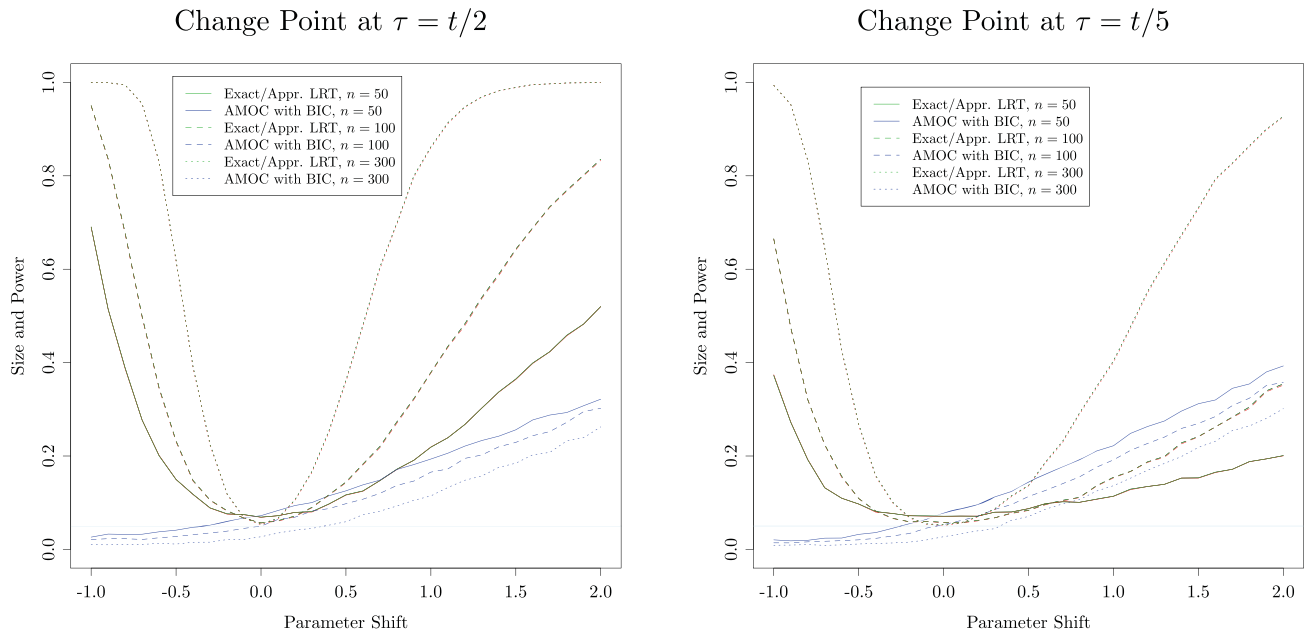

Figure 2. Power functions of LRTs (in green lines) and AMOC($k = 2$) (in blue lines) with $t = 50$ (in solid lines), 100 (in dash lines), and 300 (in dot lines) for the second parameter shift set-up. Note: All three power functions for exact and approximate LRTs are plotted in different colors (black - exact LRT, red - approximate LRT, green - approximate LRT with calibration), but they overlap with each other and only the green one is shown in the plot and the legend.

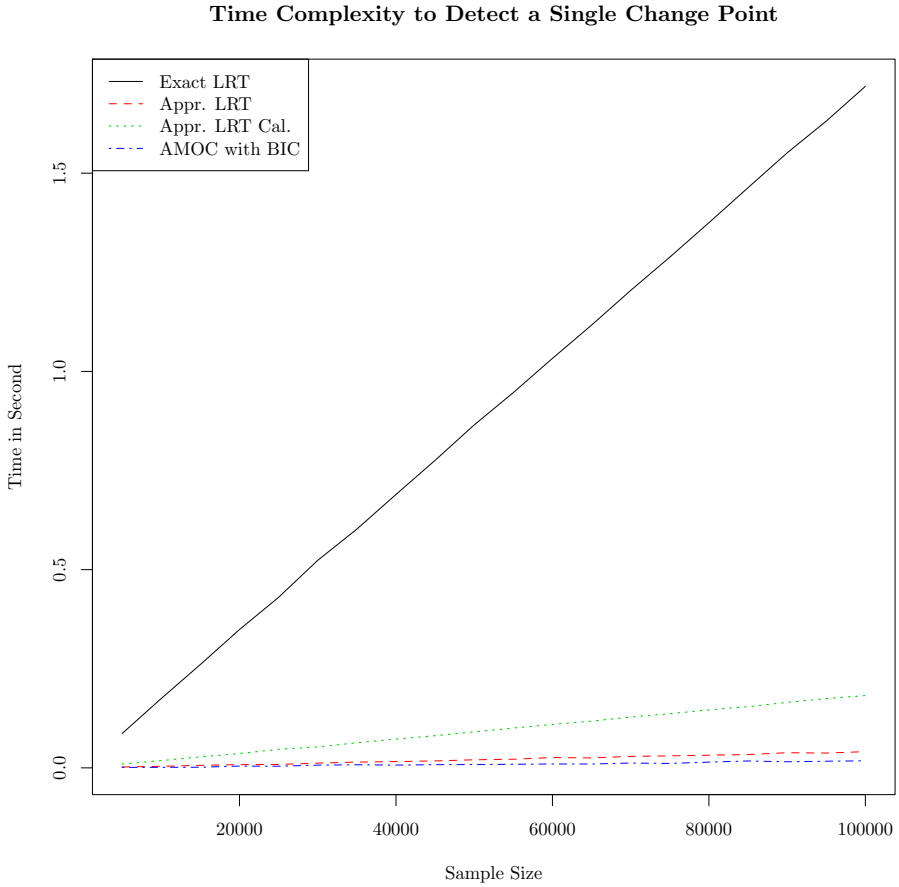

Figure 3. Computation efficiency for the exact LRT (in black solid line), approximate LRT (in red dash line), approximate LRT with calibration (in green dot line), and AMOC(k) (in blue dot dash line).

the numerical optimization. The interesting issue is that AMOC(k) is almost twice as fast as the approximate LRT without calibration. When the shape parameter is given, all the calculations within AMOC(k) are just arithmetic and logarithmic. However, it is necessary to compute the log-gamma function of the shape parameter when evaluating the approximate LRT. After profiling the source codes in R, we find that the computational time spent on the log-gamma function is almost a half of the total computational time of approximate LRT. For the exact LRT and approximate LRT with calibration, the program needs to spend more time on the evaluation of the digamma and trigamma functions. These complicated functions are considerably harder to tackle with and become the major burden for the computation. Overall speaking, the computational rate of approximate LRT without calibration is most satisfactory and cannot be further improved significantly.

5.2. Multiple change points

In this subsection, we will investigate the estimation accuracy and computational efficiency of EML, AML, AMLC for multiple change points. The function *cpt.meanvar* in the R package *changepoint* (Killick and Eckley 2014) can also be used to estimate multiple change points in a gamma sequence by prefixing the shape parameter k. We include this alternative approach as PELT(k). We further consider two nonparametric approaches based on Kolmogorov-Smirnov test (CPM-KS) and Cramér-von-Mises test (CPM-CM) (Ross and Adams 2012). Both of them have been implemented in the function *processStream* in the R package *cpm* (Ross 2015).

5.2.1. Numerical comparison

To examine the case where m increases with t, we first let m increases from 4 to 40, and let the length of $m+1$ segments is generated by

$$(L_1, ..., L_{m+1}) \sim 50 + \text{Multinomial}(50 \cdot (m+1), p_1, ..., p_{m+1}), \quad [11]$$

where $(p_1, ..., p_{m+1})$ is generated from a flat Dirichlet distribution with a constant concentration parameter $\alpha = 1$ which is essentially a uniform distribution over the open standard m-simplex. This randomization setup puts a lower bound (50) on the segment length while allowing the segments to be of significantly different lengths.

Therefore, m change points are given by

$$\tau_j = 50j + \sum_{i=1}^{j} L_i, j = 1, ..., m.$$

The following shift pattern of the parameters is then used

$$\begin{aligned}&(\lambda_1 + U_1, k_1 + V_1) \xrightarrow{\tau_1} (\lambda_2 + U_2, k_2 + V_2) \xrightarrow{\tau_2} \\ &(\lambda_1 + U_3, k_1 + V_3) \xrightarrow{\tau_3} \cdots \\ &\xrightarrow{\tau_{m\bar{1}}} (\lambda_2 + U_m, k_2 + V_m) \xrightarrow{\tau_m} (\lambda_1 + U_{m+1}, k_1 + V_{m+1}),\end{aligned} \quad [12]$$

where U_i and V_i follow an i.i.d. uniform distribution over the interval $(-0.25, 0.25)$. U_i and V_i introduce extraordinary uncertainties to the generated data stream and it can help to examine the robustness of different approaches.

In our simulation study, three combinations of (λ_1, k_1) and (λ_2, k_2) are studied as follows.

$$\begin{aligned}\lambda_1 = 1, \lambda_2 = 3, k_1 = k_2 = 2; \\ \lambda_1 = 2, \lambda_2 = 4, k_1 = 1, k_2 = 3; \\ \lambda_1 = 1, \lambda_2 = 3, k_1 = 3, k_2 = 1.\end{aligned}$$

In the simulation, the shape parameter is prefixed at $k = k_1$ for PELT(k). For each parameter combination, the number of Monte Carlo replication is 2,000. Three criteria are used to evaluate the performance, *i.e.*, the matching rate, the ratio of change point numbers, and the positive detection rate (PDR). The matching rate is defined as

$$\Pr(\tilde{\boldsymbol{\tau}}_{1:\tilde{m}} = \hat{\boldsymbol{\tau}}_{1:\hat{m}}) \text{ and } \Pr(\check{\boldsymbol{\tau}}_{1:\check{m}} = \hat{\boldsymbol{\tau}}_{1:\hat{m}}).$$

The ratio of change point numbers is given as

$$\frac{\hat{m}}{m}, \frac{\tilde{m}}{m}, \text{ and } \frac{\check{m}}{m},$$

and the positive detection rate (PDR) is

$$\begin{aligned}\text{PDR} &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\big[I(\tau_i \text{ is correctly identified})\big] \\ &= \frac{1}{m}\sum_{i=1}^{m} \Pr(\tau_i \text{ is correctly identified}),\end{aligned}$$

where $I(\cdot)$ is an indicator function. An estimate of PDR for one replication in Monte Carlo experiments can be expressed in the following form:

$$\widehat{\text{PDR}} = \frac{1}{m}\sum_{i=1}^{m} \max_{j=1,...,\hat{m}} I(||\tau_i - \hat{\tau}_j|| \le b),$$

where $\hat{\cdot}$ shall be replaced by $\tilde{\cdot}$ and $\check{\cdot}$ for AML and AMLC. In our simulation study, we set $b = 10$. Monte Carlo simulation results on the matching rates, the ratios of change point numbers and the PDR for all three parameter shift set-ups are represented in

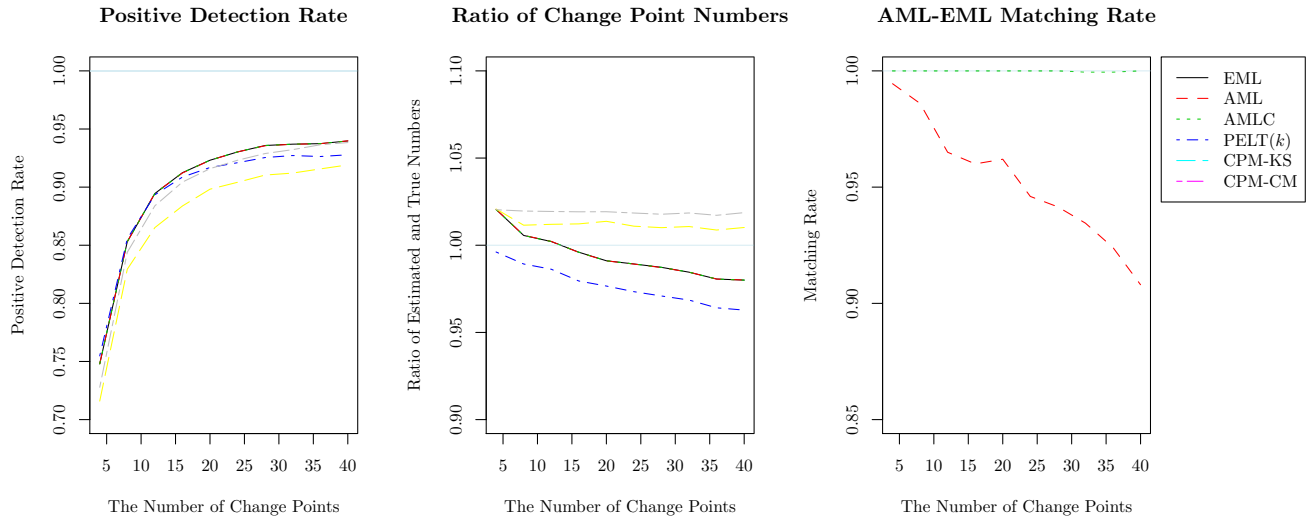

Figure 4. Simulation results of EML (in black solid line), AML (in red dash line), AMLC (in green dot line), PELT(k) (in blue dot dash line), CPM-KS (in yellow long dash line), and CPM-CM (in gray long dot dash line) for the first parameter shift set-up.

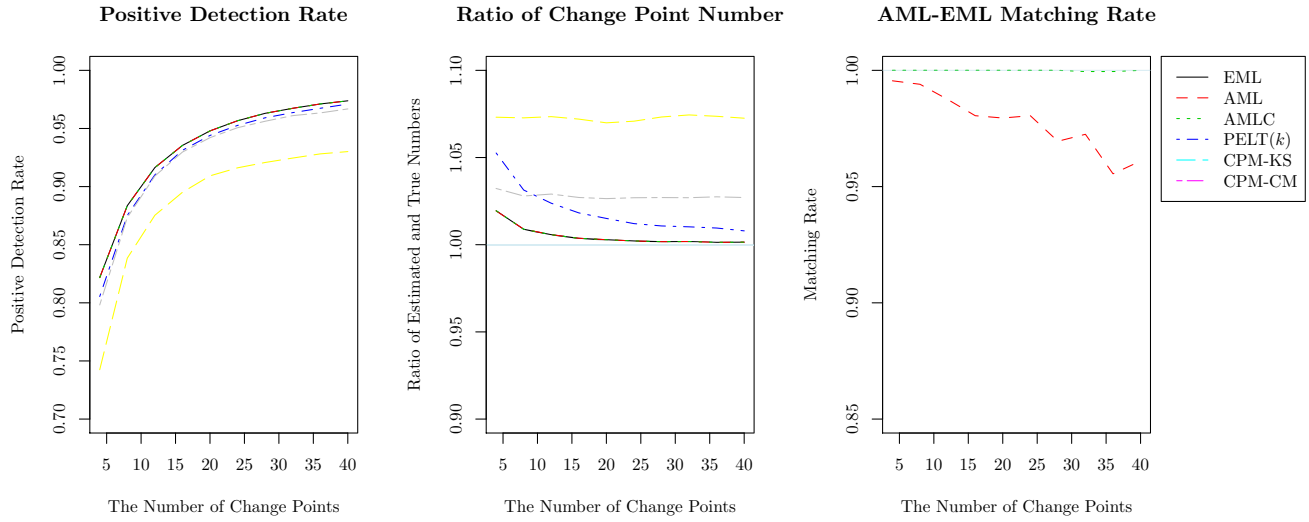

Figure 5. Simulation results of EML (in black solid line), AML (in red dash line), AMLC (in green dot line), PELT(k) (in blue dot dash line), CPM-KS (in yellow long dash line), and CPM-CM (in gray long dot dash line) for the second parameter shift set-up.

Figures 4–6. As can be seen, EML, AML, and AMLC gives almost the same results and their plots overlap with each other. These results again confirm the high approximation accuracy of AML and AMLC to EML. Furthermore, in all three scenarios, EML, AML, and AMLC dominate two non-parametric approaches—CPM-KS and CPM-CM by providing ratios of change point number closer to 1 and higher PDR. This validates the superiority of parametric change point model over the nonparametric one when the model can be correctly-specified. In addition, CPM-KS and CPM-CM provides a consistent ratio of change point numbers in each scenario since they are based on the corresponding nonparametric hypothesis tests which always allows false positives. Some detailed discussions are presented as follows.

Firstly, the right panels of Figures 4–6 show that the matching rates of AML are above 0.8 for all cases. Though there are moderate fluctuations in these three plots, the matching rates of AML all exhibit a decreasing trend when the number of change points increases. This agrees with our anticipation as the change point has the so-called "local" properties and the matching rate is an extremely strict criterion. More change points introduce more uncertainties and make the exact match of EML and AML on all change points become harder. Furthermore, the matching rates of AMLC are almost perfect (≈ 1) which suggests that the proposed calibration can be very beneficial from the perspective of approximation accuracy.

For the first set-up $\lambda_1 = 1, \lambda_2 = 3, k_1 = k_2 = 2$, the middle panel of Figure 4 show that three likelihood

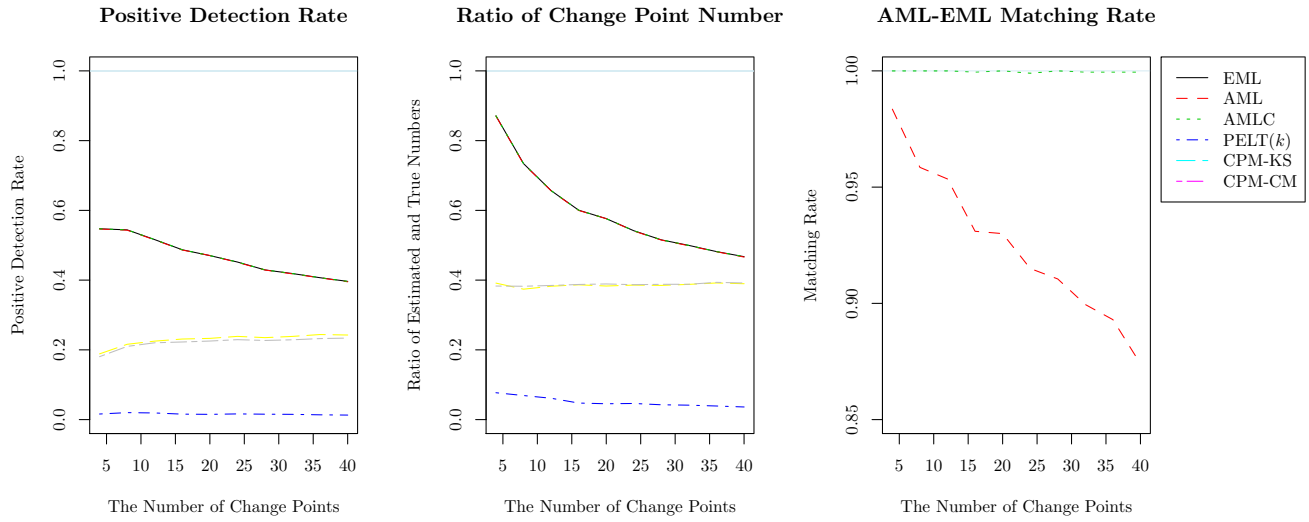

Figure 6. Simulation results of EML (in black solid line), AML (in red dash line), AMLC (in green dot line), PELT(k) (in blue dot dash line), CPM-KS (in yellow long dash line), and CPM-CM (gray long dot dash line) for the third parameter shift set-up.

based approaches overestimate the number of change points m at the beginning while PELT(k) tends to underestimates m a bit. When m increases, the ratio decreases and all approaches underestimate the number of change points. Particularly, the performance of PELT(k) deteriorates most. The underestimation issue with large m can be explained by the shape parameter k. Since k is almost a constant, the penalty term in BIC may be a little too large here. This also explains the reason why the PDR of PELT(k) is even slightly better than EML when m is small (≤ 10) in the left panel of Figure 4. For the second set-up $\lambda_1 = 2, \lambda_2 = 4, k_1 = 1, k_2 = 3$, Figure 5 shows that the performance of all approaches are close in terms of PDR except for CPM-KS. This may be explained by the significant change in both two parameters. It leads to a large shift in both mean and variance of the data stream which makes change points relatively easier to detect. In addition, all approaches overestimate the number of change points but the ratios of our approaches are much closer to 1 especially when m gets larger. For the last set-up $\lambda_1 = 1$, $\lambda_2 = 3, k_1 = 3, k_2 = 1$, Figure 6 suggests that all six approaches perform badly in terms of PDR and underestimate the number of change points seriously. The PDR of PELT(k) are all below 0.1 which means a lazy approach by prefixing k is not capable to handle this case. The PDR of nonparametric approach are merely slighter better than PELT(k) while our likelihood-based approaches identify around half of the change points. It is interesting to see that the performance of likelihood-based approaches deteriorates when m increases. This case is a little tricky as the mean of the process is nearly a constant $k\lambda = 3$. Some better approach can be designed by taking this information into consideration. Generally speaking, three likelihood-based approaches provide most robust results for all three set-ups by outperforming other alternatives in most cases.

5.2.2. Robustness analysis

The results in Section 5.2.1 suggest that, although the AML and AMLC approximate EML well, the estimation of change point can be tricky if the parameters shift in some particular way. In real applications, the underlying model of a change point problem can be even more complicated. The distribution of data stream may be mis-specified and the observations can be serially correlated. The following simulation study further investigates the robustness of our approaches against various parameter shift set-ups, model mis-specification, and autocorrelation.

In the first scenario, we consider independent gamma distributed sequences with the following parameter shift set-up,

$$(1.5, 1.5) \overset{\tau_1}{\to} (2.5, 2.5) \overset{\tau_2}{\to} (3.5, 1.0) \overset{\tau_3}{\to} (2.0, 2.5) \overset{\tau_4}{\to} (0.5, 0.5) \overset{\tau_5}{\to} (1.5, 1.0) \overset{\tau_6}{\to} (0.4, 2.0) \overset{\tau_7}{\to} (0.3, 4.0) \overset{\tau_8}{\to} (4.0, 0.3).$$

All eight change points lead to various combinations of parameter shift patterns. The segment lengths are generated from

$$(L_1, \ldots, L_9) \sim 25 + \text{Multinomial}(75 \cdot 9, p_1, \ldots, p_9),$$

which is similar to Eq. [11]. Therefore, the data length is fixed at $t = 900$. Figure 7 represents a simulated path of the data stream with its mean and variance as functions of the time index. For the rest three scenarios, we generate three different log auto-regressive

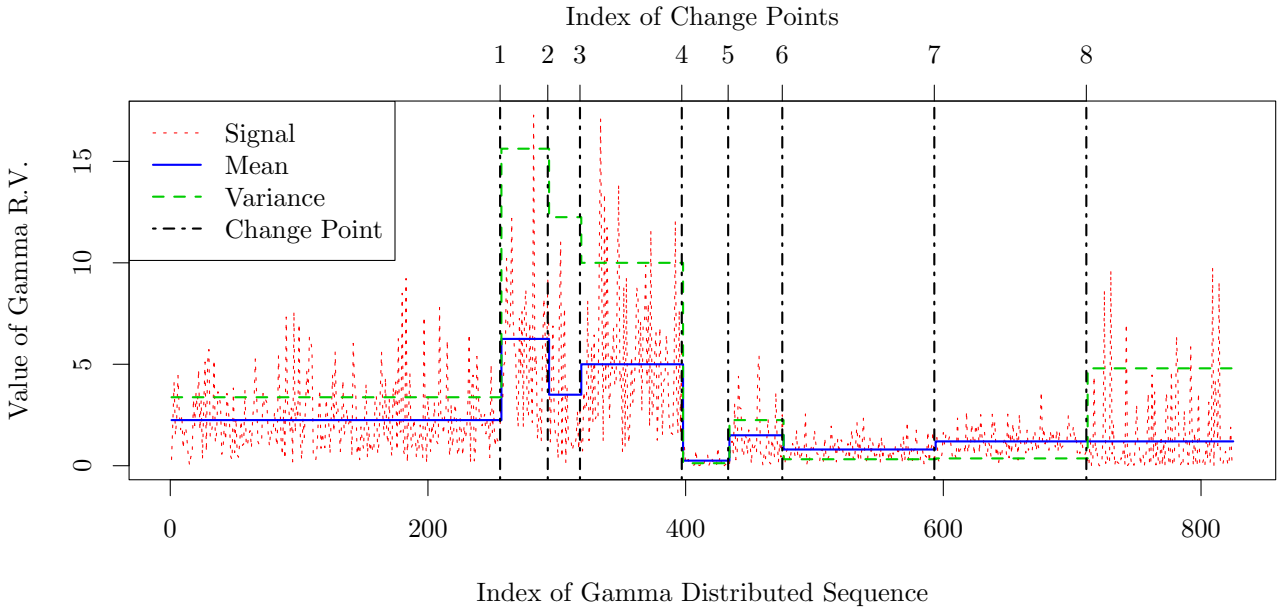

Figure 7. Simulated data (in red dot line) with mean (in blue solid line), variance (in green dash line), and change points (in black dot dash line).

(AR) processes of order 1 with Gaussian innovations. A suitable location-scale transformation is applied to AR(1) processes and the transformed processes are further exponentiated to matching their means and variances with those of the independent gamma distributed sequence in the first scenario. The AR coefficients are $\phi = 0, \pm 0.3$ which correspond to log AR processes with different dependence structures. Particularly, $\phi = 0$ leads to an independent log-normal distributed sequence.

AML, AMLC, EML, PELT(k), CPM-KS, and CPM-CM are implemented in all the four scenarios with 2,000 Monte Carlo replications. We represent the PDR of all approaches for each change point in Figure 8. Here, $k = 2$ in PELT(k) which is selected from a candidate set to achieve a high PDR by simulation.

In general, Figure 8 represents that EML, AML, and AMLC provide almost the same results in all four scenarios. They also achieve best overall PDR in the first scenario (the topleft panel in Figure 8)—independent gamma distributed sequence, which agrees with the results in Section 5.2.1. Particularly, all six approaches can efficiently detect three change points with high PDR, *i.e.*, $(1.5, 1.5) \overset{\tau_1}{\rightarrow} (2.5, 2.5)$, $(2.0, 2.5) \overset{\tau_4}{\rightarrow} (0.5, 0.5) \overset{\tau_5}{\rightarrow} (1.5, 1.0)$. We can find that the shape parameters and scale parameters of these three change points are increasing or decreasing simultaneously. Correspondingly, both the mean and variance of this process change significantly which makes the change point detection easier. This coincides with the results of the second set-up shown in Figure 5. It is found that the ratios (not differences) in means and variances of two adjacent segments play an important role in the efficiency of change point estimation. One can investigate the difference of expected log-likelihoods of two adjacent segments as well as the variance of them, *i.e.*, the signal-noise ratio of log-likelihoods, to further justify this issue. EML, AML, and AMLC also perform well for the last change point $(0.3, 4.0) \overset{\tau_8}{\rightarrow} (4.0, 0.3)$. However, their PDR decrease significantly for the rest four change points but still dominate PELT(k). From Figure 7, one can figure out the reason why the likelihood based approaches perform badly sometimes. The change of the parameters does not necessarily lead to a significant change in the real data. Without enough data, it is hard to detect this kind of change points.

However, the performance of likelihood based approaches deteriorates when the underly distribution of data stream becomes independent log-normal (the topright panel in Figure 8). Though EML, AML, and AMLC still dominate PELT(k), CPM-KS, and CPM-CM provides higher PDR in the second and the third change points, *i.e.*, $(2.5, 2.5) \overset{\tau_2}{\rightarrow} (3.5, 1.0) \overset{\tau_3}{\rightarrow} (2.0, 2.5)$. Since AML and AMLC still outperform other approaches for the last three change points $(1.5, 1.0) \overset{\tau_6}{\rightarrow} (0.4, 2.0) \overset{\tau_7}{\rightarrow} (0.3, 4.0) \overset{\tau_8}{\rightarrow} (4.0, 0.3)$, our approaches are still competitive even if we mis-specify the underlying log-normal distribution as a gamma distribution. However, when the autocorrelation is introduced into the observed sequence (the bottomleft and bottomright panels in Figure 8), the performance of likelihood based approaches become much worse and only dominate others in the last change point $(0.3, 4.0) \overset{\tau_8}{\rightarrow} (4.0, 0.3)$. This shows that AML and

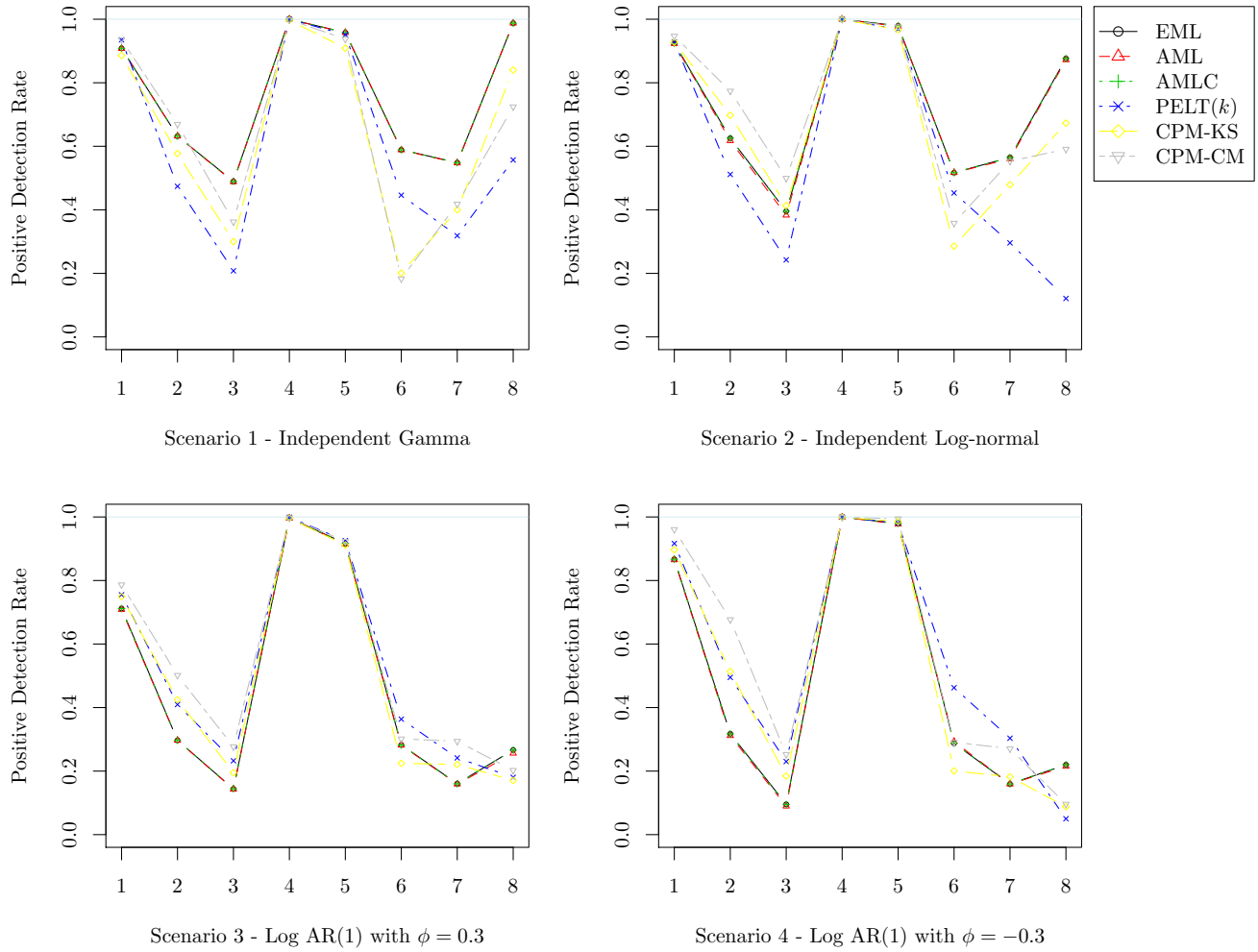

Figure 8. Positive detection rates of EML (in black solid line), AML (in red dash line), AMLC (in green dot line), PELT(k) (in blue dot dash line), CPM-KS (in yellow long dash line), and CPM-CM (in gray long dot dash line) for different data generating mechanisms.

AMLC can't handle even the simplest dependence structure in the data stream. This is expected since any likelihood based approaches have been tailored to specific parametric assumptions.

5.2.3. Computation efficiency

We further study the time complexity of the proposed approaches in two different scenarios of multiple change points, *i.e.*, the increasing number of change points with the increasing data length ($m \sim O(t)$) and the fixed number of change points with the increasing data length ($m \sim O(1)$). We apply EML, AML, AMLC, and PELT(k) to compute the multiple change points in the following two different scenarios. CPM-KS and CPM-CM are excluded since they are way much slower than the rest approaches.

For the first scenario ($m \sim O(t)$), the number of change points m grows linearly with t and each segment in the sequence has a fixed length 100. In particular, m increases from 9 to 199 by 20 and the number of the segment $m+1$ increases from 10 to 200 by 20. The data length t increases from 1,000 to 20,000 by 2,000. The parameter shift set-up of the process is given as

$$(1,2) \overset{100}{\rightarrow} (10,2) \overset{200}{\rightarrow} (1,2) \overset{300}{\rightarrow} \cdots \overset{100m}{\rightarrow} (10,2).$$

For each data length, the replication number is 10 to reduce the uncertainties in the computation.

For the second scenario ($m \sim O(1)$), the number of change points m is fixed at 9 and the number of the segments is 10. The data length t increases from 500 to 10,000 by 500. All the segments are fixed to the same length s which increases from 50 to 1,000 by 50. The parameter shift set-up of the process is given as

$$(1,2) \overset{s}{\rightarrow} (10,2) \overset{2s}{\rightarrow} (1,2) \overset{3s}{\rightarrow} \cdots \overset{9s}{\rightarrow} (10,2).$$

The replication number is still 10.

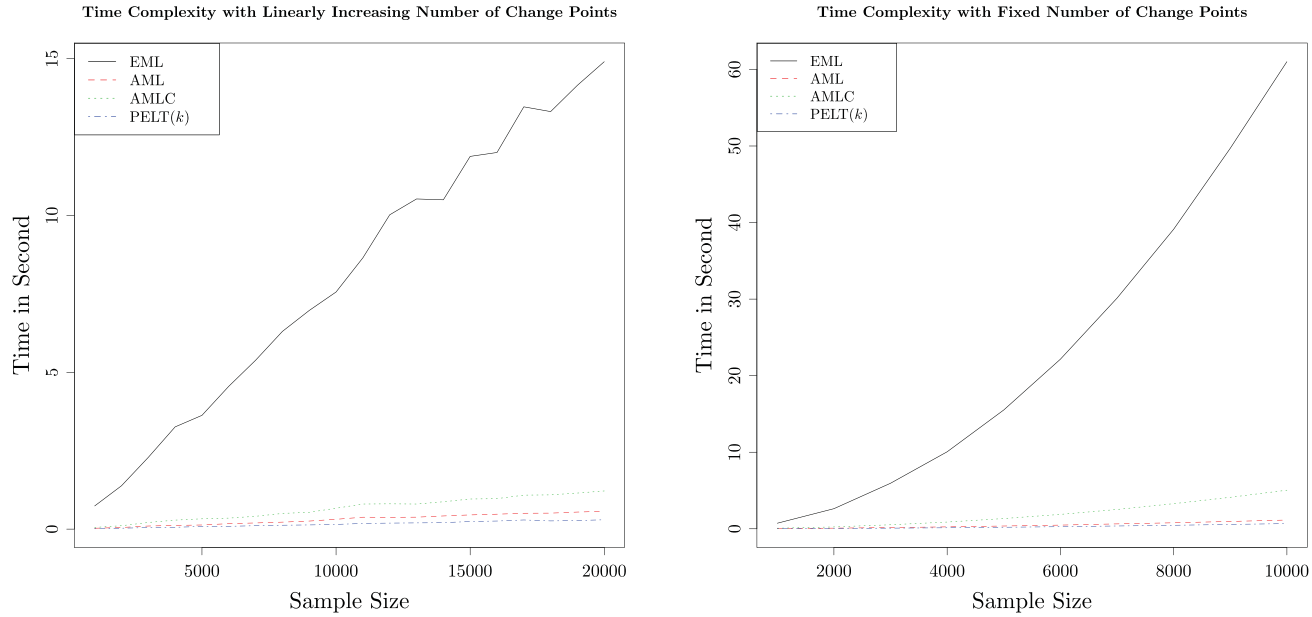

Figure 9. Computation efficiency of EML (in black solid line), AML (in red dash line), AMLC (in greem dot line), and PELT(k) (in blue dot dash line) for multiple change points.

In Figure 9, we represent the computation rates of EML, AML, AMLC, and PELT(k) for the above two different scenarios. As shown in the left panel of Figure 9, when the number of the change points is growing linearly with the data length t, all of these four approaches achieve the linear computational complexity. AML is just a little slower than PELT(k), but this can still be explained by the additional time to compute the log-gamma function. EML is much slower than AML and PELT(k) as one may expect while AMLC is between AML and EML. Since EML gets involved with multiple rounds of updating with the digamma and trigamma functions while AMLC requires only one round to evaluating those complicated functions. The difficulty in computing digamma and trigamma functions may explain the high computational cost of AMLC if it is compared to AML and PELT(k). But it still saves more than half of computational time if one uses it as a surrogate for EML. In general, the computational efficiencies of AML and AMLC are satisfactory for even a very long sequence as they all find the solution in around one second when the data length reaches 20,000. For EML, the computational rate is still acceptable with a little more than three seconds.

For the scenario with a fixed number of change points represented in the right panel of Figure 9, all approaches become much slower. This phenomenon has also been discussed as the worst case with a quadratic computational complexity in Killick, Fearnhead, and Eckley (2012). When the data length reaches 10,000, it takes almost 1 second to find the change points for both AML and PELT(k). AMLC takes around 5 seconds which is still acceptable. However, EML is again the worst one in terms of speed as it spends almost one minute. The computational cost will greatly increase for an $O(t^2)$ algorithm if the length of data stream further increases. Therefore, AML is much more competitive in terms of the computational cost for this scenario and AMLC can also be used. However, EML tends to be too slow for a longer sequence.

6. Case studies

In this section, two real data sets are studied to illustrate and compare different approaches discussed in this article. Particularly, we focus on the time between events data arising from the industrial accident problems. Monitoring the time between accidents is a critical topic in the safety management of the manufacturing production. It is of interest for both the researchers and the practitioners to detect the possible shifts of the historical accident data to support the identification of the root causes of the accidents. Maguire, Pearson, and Wynn (1952) first studied the distribution of the time between two consecutive industrial accidents. They showed that the exponential distribution which is a special case of the gamma distribution provides a satisfactory fit to a lot of data sets arising from industrial accidents. Therefore, it is reasonable to apply our approach to the industrial accident data. Since the first data set, *i.e.*, British coal mine disaster data, has been well studied in a series of papers, we only present a concise analysis on it. The second data set, *i.e.*, American mine disaster data,

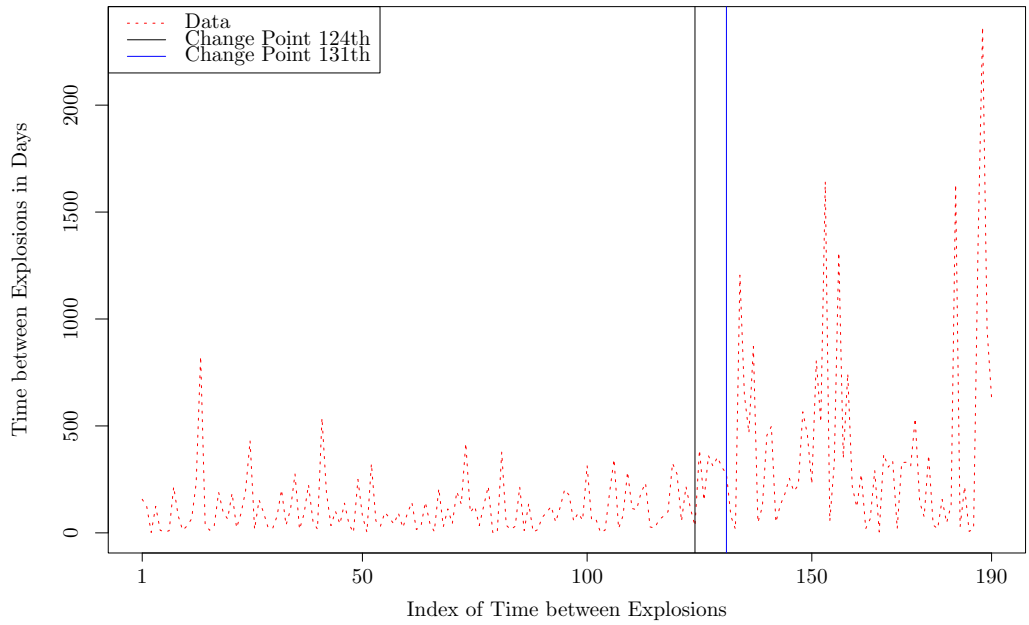

Figure 10. British coal mine data (in red dot line) with two detected change points (in vertical black and blue solid lines).

contains more observations and has never been investigated in the past literature. Therefore, we carry out a detailed study on this date set.

6.1. British coal mine disaster data

British coal mine disaster data is taken from Jarrett (1979) which is a note to the seminal article by Maguire, Pearson, and Wynn (1952). It consists of the time intervals in days between successive explosions involving 10 or more men killed in British coal mines from March 15, 1851 to March 22, 1962, a total of 40,550 days. There are 191 explosions in total with 190 times between explosions recorded. There is a zero in this data set which means there were two accidents happening on the same day. We replace this zero by 1/2 to avoid the numerical difficulty to handle the left-censored data. From the increasing trend in Figure 10, we may conjecture that there exists one change point in this data set.

This data set has been studied by Worsley (1986), Ramanayake (2005), Gan (1998), and Zhang et al. (2007). Both Worsley (1986) and Ramanayake (2005), reported that there exists one change point—the 126th time interval in the year 1890. Zhang et al. (2007) also found that several out-of-control signals are triggered when monitoring this data set with their control chart for gamma distribution. We will continue with the previous work to validate their results and check if there exist any other change points.

We perform the exact LRT, approximate LRT and approximate LRT with calibration based on the gamma distribution on this data set. An LRT based on the exponential distribution implemented in the R package *changepoint* is also performed. All four approaches reject the null hypothesis and suggest that a change point is located right after the 124th observation which is slightly different from the previous results in Worsley (1986) and Ramanayake (2005). This minor difference is reasonable since a change point estimator converges to its true location at a rate $O_p(1)$ (Csörgö and Horváth 1997).

To check the possible multiple change points within this data set, EML, AML, and AMLC are applied to it. We also use the R function *cpt.meanvar* specified for the exponential distribution in the R package *changepoint*. If we use the standard BIC with the minimum segmentation length 3, EML, AML, and AMLC report that there are two change points as the 126th and 131th observations while *cpt.meanvar* identifies two change points as the 124th and 186th observations. As pointed by Zhang and Siegmund (2007), the modified BIC tends to perform better than BIC by accounting for the irregularity of the log-likelihood of a change point model. Therefore, we adopt the modified BIC proposed by Zhang and Siegmund (2007). All four approaches agree that the 124th observation is the only change point and the false alarms are eliminated.

6.2. American mine disaster data

This data set is taken from NIOSH (2013) maintained by the National Institute for Occupational Safety and Health (NIOST). It consists of the records on the mine disasters in America from March 18, 1839 to

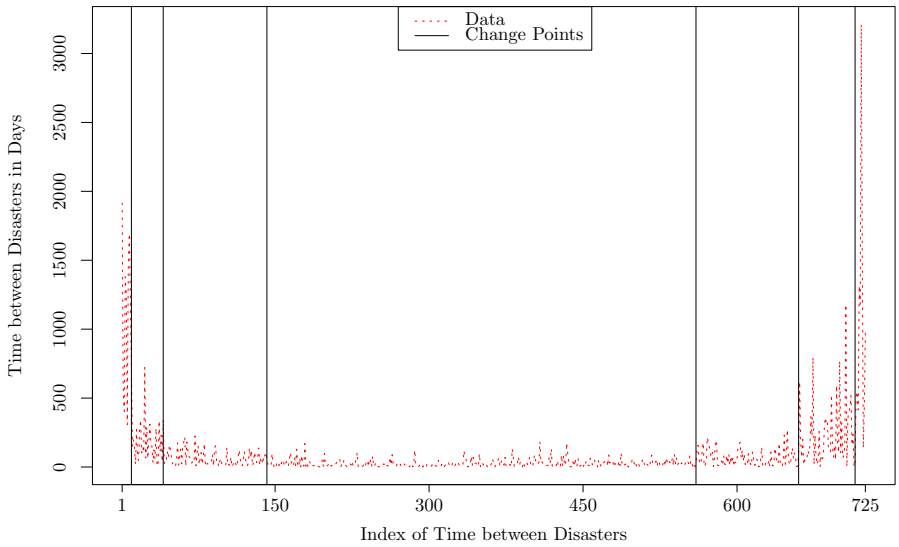

Figure 11. American mine disaster data (in red dot line) with the detect change points (in vertical black solid lines).

April 5, 2010, a total of 62,476 days. There are 726 mine disasters and most of them (623, 85.8%) are coal mine disasters. After taking the difference of 726 dates, we get 725 times between mine disasters and 8 (1.1%) of them are zeros which is replaced by 1/2 in the following analysis.

Unlike British coal mine disaster data which merely consists of the coal mine explosions with more than 10 deaths, American mine disaster data provides a relatively complete historical record. Our approach is used to investigate the development of mining safety in the US. From the line plot on the times between disasters in Figure 11, it is obvious that the times between disasters exhibit a bath-tub curve which alludes to at least two change points in this data set.

Firstly, we apply all LRTs to test the existence of a unique change point. All four tests used in the previous subsection agree that there is a change point right after the 660th time between events occurred in the year 1948. Then, we estimate the locations of possible multiple change points by EML, AML, AMLC, and *cpt.meanvar*. The minimum segmentation length is fixed at 3 and we try both BIC and modified BIC (mBIC) since BIC tends to trigger a few false alarms in change point analysis.

All approaches provide highly consistent results even with different penalties. Specifically, EML with mBIC, AML with mBIC, and AMLC with mBIC report the same and most concise solution as $(10, 41, 142, 560, 660)$. EML with BIC, AML with BIC, AMLC with BIC, and *cpt.meanvar* with mBIC report one more change point as $(10, 41, 142, 560, 660, 715)$. *cpt.meanvar* with BIC gives the longest result as $(10, 81, 179, 333, 560, 660, 715)$. To balance the goodness-of-fit and the parsimony of the final model, we conclude that the final estimates of change points are $(10, 41, 142, 560, 660, 715)$ in the following years—1869, 1884, 1901, 1931, 1948. They divide the whole time horizon (from 1839 to 2013) to seven segments with the length 30, 15, 17, 30, 17, 34, and 31 (in years). We mark these change points in Figures 11 and 12 in vertical black dash lines.

The parameter estimates and p-values of Kolmogrov Smirnov test for all segments are summarized in Table 1. In general, both the exponential distribution and gamma distribution fit each segment reasonably well. An interesting issue is that the shape parameter estimate of gamma distribution in the first segment looks rather high (>3). Furthermore, the shape parameter exhibits a decreasing trend in the first four segments. Then, it just fluctuates around 1.

Identifying the factors which cause these change points is of course our interest as it can provide insights on the safe management of the complex industrial system. Firstly, we represent the cumulative number of mine disasters against the time in Figure 12 with all the change points identified in solid black lines. In addition, as most of the mine disasters are coal mine disasters, the American annual coal production from 1870 to 2013 is represented in Figure 13 (Wikipedia 2018). In general, the first inflection point in the bathtub curve can be explained by the rapid development of American mining industry in the late 19th century. The safety regulations and equipments are left behind during this period. Therefore, the time between mine disasters shows a decreasing trend in the first four time segments. We cannot find the exact coal production before 1870 which means the

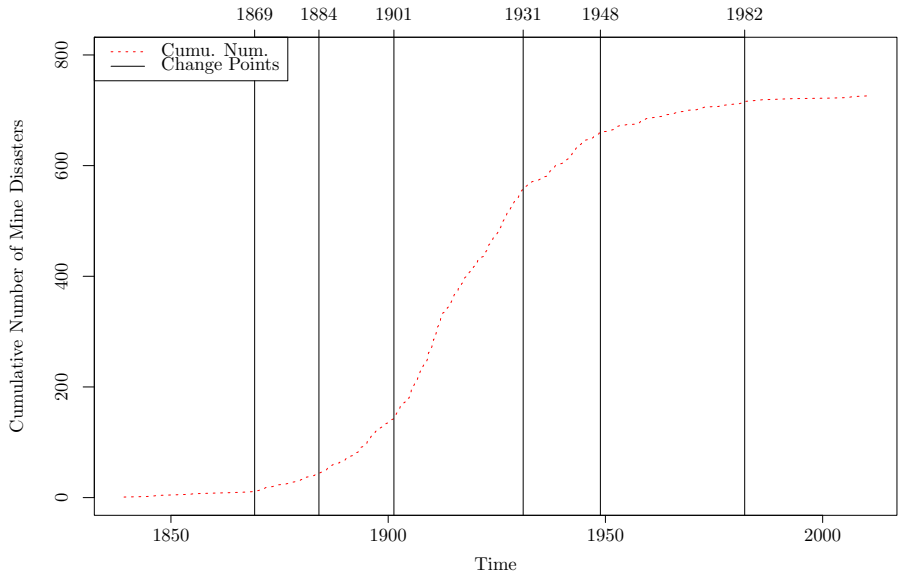

Figure 12. Cumulative number of American mine disasters (in red dot line) against time with the change points detected (in vertical black solid lines).

Table 1. Parameter estimation and goodness of fit for American mine disaster data.

Segments	Gamma EML	Gamma AML	Gamma AMLC	Exponential
1–10	3.377, 325.1 (0.497)	3.527, 311.3 (0.496)	3.370, 325.1 (0.490)	1098 (0.385)
11–41	1.265, 151.7 (0.935)	1.259, 152.4 (0.933)	1.265, 151.7 (0.935)	191.9 (0.867)
42–142	1.142, 57.07 (0.871)	1.131, 57.63 (0.892)	1.142, 57.07 (0.871)	65.16 (0.898)
143–560	0.935, 27.91 (0.632)	0.907, 28.78 (0.775)	0.934, 27.92 (0.647)	26.09 (0.294)
561–660	1.050, 61.42 (0.593)	1.051, 61.34 (0.589)	1.050, 61.42 (0.593)	64.48 (0.766)
661–715	0.901, 242.2 (0.700)	0.907, 240.6 (0.718)	0.901, 242.2 (0.699)	218.2 (0.758)
716–725	1.160, 811.0 (0.950)	1.166, 806.2 (0.953)	1.160, 811.0 (0.950)	940.5 (0.836)

Note: The numbers before the left bracket is the parameter estimates and the numbers in the brackets are the p-value of Kolmogrov Smirnov test given the parameter estimates. The parameter estimates are represented as $\hat{k}, \hat{\lambda}$ for the gamma distribution and only $\hat{\lambda}$ for the exponential distribution.

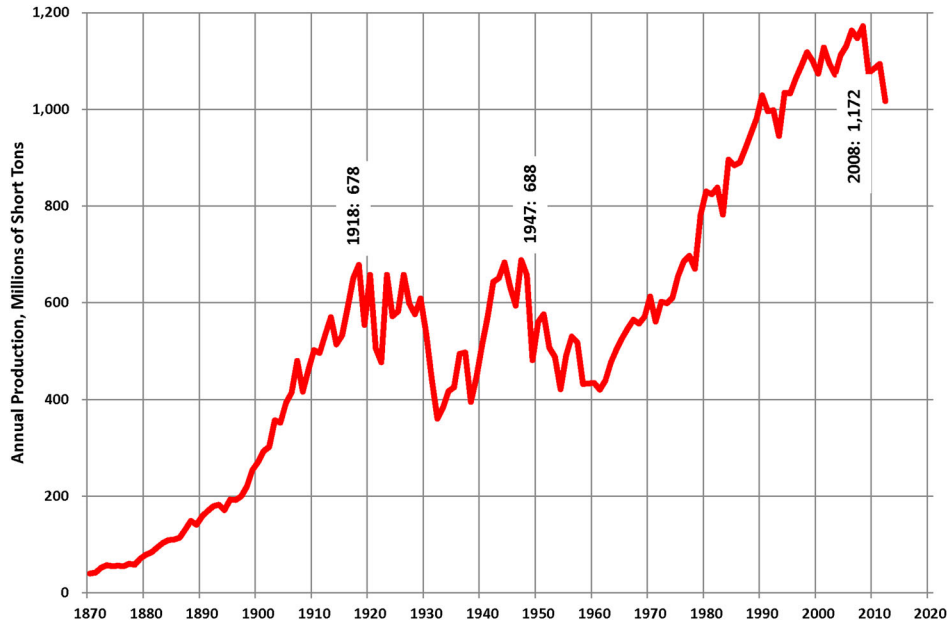

Figure 13. American coal production (the red solid line) from 1870 to 2013. Source: Wikipedia (2018). Figure created by and attributed to Plazak. © 2013 Plazak. This figure is published on Wikipedia under a CC BY SA license at https://en.wikipedia.org/wiki/History_of_coal_mining_in_the_United_States. The figure is reprinted in this article under the guidelines of the CC BY SA license. Disclaimer: Reuse of the figure in this article does not indicate endorsement of the content of the article from the creator or licensor.

historical records are incomplete before this year. As the first change point (1869) is just before 1870, this may imply that the mine disasters were highly under reported during that period which leads to a large shape parameters (>3). The unrecorded issue may also influence the second segment with a shape parameter (>1.2). An interesting issue is that most of the notable labor strikes and events of American coal miners were happened between 1884 and 1921 according to Wikipedia (2018) which is after the second change point (1884) and before the fourth change point (1931). This fact reflects the miners' working environment was extremely bad at that time. The second inflection point in the bathtub curve can be explained by the technological development and the mechanization which greatly reduce the risk of mine disaster. In particular, at the fourth and fifth change points (1931 and 1948), the annual coal productions dropped significantly in Figure 13. This may also contribute to the change points.

7. Conclusions

In this article, we present an efficient computation framework to estimate the locations of change points as well as the associated gamma parameters within a gamma distributed sequence with change points. This approach is based on a closed form approximation to MLE of gamma distribution proposed by Ye and Chen (2017). We first discover the linkage between their estimator and the standard MLE and further derive a closed form calibration of the approximate likelihood which achieves the same asymptotical performance of exact likelihood. Necessary theoretical justifications and extensive simulation studies on its approximation accuracy, estimation efficiency, and robustness are provided. This approach can also be incorporated into the DP algorithm to calculate the locations of multiple change points efficiently. Two real data case studies on industrial accidents are studied thoroughly and deeply to illustrate and compare different approaches to detect the change point(s) in a gamma distributed sequence. Our approaches further dig out some insightful information from these data sets. All the results show that our approaches achieve satisfactory performance for the real problems.

For the future research, we notice that there are still some cases without a nice solution, *i.e.*, the third parameter shift set-up in Figure 6 and serially correlated sequence in Figure 8, in our article. These gaps can be filled with appropriate parametrization of gamma distribution and change point model with a refined dependence structure. A solid theoretical justification on the consistency of the change point estimation with AML and the null distribution of approximate LRT statistic also need to be developed. The estimator proposed by Ye and Chen (2017) can certainly be applied to other problems on gamma distribution, *e.g.*, the fast estimation of the finite mixture of gamma distributions. Furthermore, the magnitude of the industrial accident is another critical factor in the engineering safety management. To detect the changes in time between events as well as the magnitudes of events with some bi-variate models will be an attractive but challenging topic. Finally, American mine disaster data set has not been studied in the literature and it is of interest for us to explore more details on it.

Funding

Ye was supported by the Natural Science Foundation of Jiangsu Province under grant BK20180232.

About the authors

Dr. Xiao received the B.S. degree in Statistics from University of Science and Technology of China in 2011 and the PhD degree in systems engineering and engineering management from City University of Hong Kong. He is currently a Lecture in Statistics with School of Fundamental Sciences, Massey University, New Zealand. His research interests include reliability and quality control, industrial statistics, and spatial temporal modeling.

Dr. Chen received the B.E. degree in industrial engineering from Shanghai Jiao Tong University, China, in 2013, and the Ph.D. degree in industrial systems engineering and management from the National University of Singapore in 2017. He is currently an Assistant Professor with Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands. His research interests include data analysis, reliability engineering and statistical inference.

Dr. Ye received the joint B.E. degree in material science and engineering and economics from Tsinghua University, Beijing, China in 2008, and the Ph.D. degree in industrial and systems engineering from the National University of Singapore in 2012. He is currently an Assistant Professor with the Department of Industrial Systems Engineering and Management, National University of Singapore. His research interests include reliability engineering, complex systems modeling, and industrial statistics.

Dr. Tsui is the Chair Professor of Industrial Engineering in the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong. He is a Fellow of ASQ. His research interests include Quality Engineering, Process Control and Surveillance, System Informatics, Data Mining, Health Informatics, Bioinformatics, Infectious Disease Modeling, Logistics and Supply Chain Management.

ORCID

Xun Xiao http://orcid.org/0000-0002-8780-5471

References

Abramowitz, M., and I. A. Stegun, eds. 1964. *Handbook of mathematical functions: With formulas, graphs, and mathematical tables*. New York: Dover.

Amiri, A., and S. Allahyari. 2012. Change point estimation methods for control chart postsignal diagnostics: A literature review. *Quality and Reliability Engineering International* 28 (7):673–85. doi: 10.1002/qre.1266.

Andrews, D. W. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61 (4):821–56. doi: 10.2307/2951764.

Bai, J., and P. Perron. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66 (1):47–78. doi: 10.2307/2998540.

Bai, J., and P. Perron. 2003. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18 (1):1–22.

Brodsky, E., and B. S. Darkhovsky. 1993. *Nonparametric methods in change-point problems*. Dordrecht: Kluwer.

Chen, J., and A. K. Gupta. 2000. *Parametric statistical change point analysis*. New York: Birkhauser.

Csörgö, M., and L. Horváth. 1997. *Limit theorems in change-point analysis*. Vol. 18. New York: John Wiley & Sons Inc.

Diaz, J. 1982. Bayesian detection of a change of scale parameter in sequences of independent gamma random variables. *Journal of Econometrics* 19 (1):23–9. doi: 10.1016/0304-4076(82)90049-5.

Eddelbuettel, D., R. François, J. Allaire, J. Chambers, D. Bates, and K. Ushey. 2011. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* 40 (8):1–18. doi: 10.18637/jss.v040.i08.

Gan, F. 1998. Designs of one-and two-sided exponential EWMA charts. *Journal of Quality Technology* 30 (1): 55–69. doi: 10.1080/00224065.1998.11979819.

Hawkins, D. M. 2001. Fitting multiple change-point models to data. *Computational Statistics & Data Analysis* 37 (3): 323–41. doi: 10.1016/S0167-9473(00)00068-2.

Hawkins, D. M., P. Qiu, and C. W. Kang. 2003. The changepoint model for statistical process control. *Journal of Quality Technology* 35 (4):355–66. doi: 10.1080/00224065.2003.11980233.

Hinkley, D. V. 1970. Inference about the change-point in a sequence of random variables. *Biometrika* 57 (1):1–17. doi: 10.1093/biomet/57.1.1.

Hinkley, D. V., and E. A. Hinkley. 1970. Inference about the change-point in a sequence of binomial variables. *Biometrika* 57 (3):477–88. doi: 10.1093/biomet/57.3.477.

Hsu, D. 1979. Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association* 74 (365): 31–40. doi: 10.1080/01621459.1979.10481604.

Huang, W., L. Shu, W. Jiang, and K.-L. Tsui. 2013. Evaluation of run-length distribution for CUSUM charts under gamma distributions. *IIE Transactions* 45 (9): 981–94. doi: 10.1080/0740817X.2012.705455.

Huo, X., and G. J. Székely. 2016. Fast computing for distance covariance. *Technometrics* 58 (4):435–47. doi: 10.1080/00401706.2015.1054435.

Jackson, B., J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. 2005. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters* 12 (2): 105–8. doi: 10.1109/LSP.2001.838216.

Jarrett, R. 1979. A note on the intervals between coal-mining disasters. *Biometrika* 66 (1):191–3. doi: 10.1093/biomet/66.1.191.

Killick, R., and I. Eckley. 2014. changepoint: An R package for changepoint analysis. *Journal of Statistical Software* 58 (3):1–19. doi: 10.18637/jss.v058.i03.

Killick, R., P. Fearnhead, and I. Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* 107 (500):1590–8. doi: 10.1080/01621459.2012.737745.

Lee, C.-B. 1998. Bayesian analysis of a change-point in exponential families with applications. *Computational Statistics & Data Analysis* 27 (2):195–208. doi: 10.1016/S0167-9473(98)00009-7.

Maguire, B. A., E. Pearson, and A. Wynn. 1952. The time intervals between industrial accidents. *Biometrika* 39 (1–2):168–80. doi: 10.1093/biomet/39.1-2.168.

Meeker, W. Q., and L. A. Escobar. 1998. *Statistical methods for reliability data*. New York: John Wiley & Sons.

Minka, T. P. 2002. Estimating a gamma distribution. Tech. Rep, Microsoft Research, Cambridge, UK. http://research.microsoft.com/en-us/um/people/minka/papers/minka-gamma.pdf

NIOSH. 2013. All mining disasters: 1839 to present. Accessed November 11, 2016. http://www.cdc.gov/niosh/mining/statistics/content/allminingdisasters.html

Ramanayake, A. 2005. Tests for a change point in the shape parameter of gamma random variables. *Communications in Statistics - Theory and Methods* 33 (4):821–33. doi: 10.1081/STA-120028728.

Rigaill, G. 2015. A pruned dynamic programming algorithm to recover the best segmentations with 1 to $K_{\max}$ change-points. *Journal de la Société Française de Statistique* 156 (4): 180–205.

Ross, G. J. 2015. Parametric and nonparametric sequential change detection in R: The cpm package. *Journal of Statistical Software* 66 (3):1–20.

Ross, G. J., and N. M. Adams. 2012. Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology* 44 (2):102–16. doi: 10.1080/00224065.2012.11917887.

Tan, C.-C., B.-Q. Miao, and X.-C. Zhou. 2013. Statistical inference for the shape parameter change-point estimator in negative associated gamma distribution. *Journal of Inequalities and Applications* 2013 (1):1–11. doi: 10.1186/1029-242X-2013-161.

Vaseghi, S. V. 2008. *Advanced digital signal processing and noise reduction*. New York: John Wiley & Sons.

Venables, W. N., and B. D. Ripley. 2013. *Modern applied statistics with S-PLUS*. New York: Springer Science & Business Media.

Wikipedia. 2018. History of coal mining in the United States—Wikipedia, the free encyclopedia. Accessed November 17, 2018. https://en.wikipedia.org/wiki/History_of_coal_mining_in_the_United_States

Woodall, W. H., and D. C. Montgomery. 1999. Research issues and ideas in statistical process control. *Journal of Quality Technology* 31 (4):376–86. doi: 10.1080/00224065.1999.11979944.

Worsley, K. 1986. Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* 73 (1):91–104. doi: 10.1093/biomet/73.1.91.

Yao, Y.-C. 1988. Estimating the number of change-points via Schwarz's criterion. *Statistics & Probability Letters* 6 (3):181–9. doi: 10.1016/0167-7152(88)90118-6.

Ye, Z.-S., and N. Chen. 2017. Closed-form estimators for the gamma distribution derived from likelihood equations. *The American Statistician* 71 (2):177–81. doi: 10.1080/00031305.2016.1209129.

Zhang, C., M. Xie, J. Liu, and T. Goh. 2007. A control chart for the gamma distribution as a model of time between events. *International Journal of Production Research* 45 (23):5649–66. doi: 10.1080/00207540701325082.

Zhang, N. R., and D. O. Siegmund. 2007. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63 (1):22–32. doi: 10.1111/j.1541-0420.2006.00662.x.

Zou, C., P. Qiu, and D. Hawkins. 2009. Nonparametric control chart for monitoring profiles using change point formulation and adaptive smoothing. *Statistica Sinica* 19 (3):1337–57.

Zou, C., G. Yin, L. Feng, and Z. Wang. 2014. Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics* 42 (3):970–1002. doi: 10.1214/14-AOS1210.

Zou, C., Y. Zhang, and Z. Wang. 2006. A control chart based on a change-point model for monitoring linear profiles. *IIE Transactions* 38 (12):1093–103. doi: 10.1080/07408170600728913.

Appendix

$\tilde{\boldsymbol{\theta}}$ as a stochastic approximation to $\hat{\boldsymbol{\theta}}$

Taking the partial derivatives of Eq. [2], the gamma score equations are given as,

$$S(\lambda,k)=\begin{pmatrix} n\bar{Y}/\lambda^2-nk/\lambda \\ n\overline{\ln Y}-n\psi(k)-n\ln\lambda \end{pmatrix}=0. \qquad [A1]$$

The solution to Eq. [A1] is the MLE $\hat{\boldsymbol{\theta}}$.

The estimator $\tilde{\boldsymbol{\theta}}$ in Ye and Chen (2017) is very interesting because it closely mimics the MLE $\hat{\boldsymbol{\theta}}$. In fact, the MSE of $\tilde{\boldsymbol{\theta}}$ is almost equal to that of $\hat{\boldsymbol{\theta}}$ in their simulation study. However, they showed that $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ have quite different asymptotic covariance matrices. We extend a bit their discussions by showing that $\tilde{\boldsymbol{\theta}}$ can be regarded as a stochastic approximation to $\hat{\boldsymbol{\theta}}$. The error depends only on the shape parameter k and converges to zero rapidly when $k\to\infty$.

Substituting $\tilde{\boldsymbol{\theta}}=(\tilde{\lambda},\tilde{k})$ into the gamma score functions in Eq. [A1], we obtain the partial derivatives of the gamma log-likelihood evaluated at the new estimator as

$$\begin{aligned} S(\tilde{\lambda},\tilde{k}) &= \begin{pmatrix} n\bar{Y}/\tilde{\lambda}^2-n\tilde{k}/\tilde{\lambda} \\ n\overline{\ln Y}-n\psi(\tilde{k})-n\ln\tilde{\lambda} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ n(\overline{\ln Y}-\psi(\tilde{k})-\ln\tilde{\lambda}) \end{pmatrix}. \end{aligned} \qquad [A2]$$

The first term in Eq. [A2] simply implies that the estimator $\tilde{\boldsymbol{\theta}}$ is located on the curve defined by $k\lambda=\bar{Y}$ just like $\hat{\boldsymbol{\theta}}$ while the second term contains the information on the gap between these two estimators. A detailed probabilistic analysis on the second term $e(\tilde{\lambda},\tilde{k})=\overline{\ln Y}-\psi(k)-\ln\tilde{\lambda}$ in Eq. [A2] with rescaling will shed some light on the relationship between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$. First of all, it is easy to notice that

$$\begin{aligned} e(\tilde{\lambda},\tilde{k}) &= \overline{\ln Y}-\psi(\tilde{k})-\ln\tilde{\lambda}\to\mathbb{E}\ln Y-\psi(k)-\ln\lambda \\ &= 0,\ \text{a.s.}\ . \end{aligned}$$

This result is not amazing since both $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are consistent estimators of gamma parameters. Furthermore, by the central limit theorem and the delta method, we notice that

$$\sqrt{n}\cdot e(\tilde{\lambda},\tilde{k})=\sqrt{n}(\overline{\ln Y}-\psi(\tilde{k})-\ln\tilde{\lambda})\to N(0,\sigma_e^2).$$

where σ_e^2 is the asymptotical variance to be determined. Therefore, the estimator $\tilde{\boldsymbol{\theta}}$ can be regarded as a stochastic approximation to the MLE $\hat{\boldsymbol{\theta}}$ and the approximation accuracy is controlled by the asymptotic variance term σ_e^2 since the value of the score functions at the new estimator will close to zero if σ_e^2 is small enough. After some cumbersome calculations based on the delta method and an asymptotical expansion of trigamma function (Abramowitz and Stegun 1964), we find that

$$\sigma_e^2(k)=(k^2\psi_1^2(k)-\psi_1(k)-1)(k\psi_1(k)-1)\sim O(k^{-3}). \qquad [A3]$$

As one can see from Eq. [A3], the asymptotic variance only depends on the shape parameter k. More importantly, we show that $\sigma_e^2(k)\sim O(k^{-3})$ converges to zero rapidly. Figure A1 outlines the asymptotic variance $\sigma_e^2(k)$ as well as the simulated variances with sample sizes $n=30$, 100, 300 against the shape parameter k. The asymptotic variance denoted by the solid black line clearly dominates the simulated variances which suggest that the new estimator will be close to the MLE even for small sample sizes when k is not too small. Even if k is small, since the asymptotical variance is calculated after rescaling by $\sqrt{n}$ and $\sigma_e^2(k)<1$ for $k>0.2427$, the difference in estimates before rescaling will be much smaller provided a moderate n.

Asymptotical analysis on $l^*-\tilde{l}^*$ and the calibration

The above analysis only gives a rough idea on the gap between $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$. Actually, we quantify their gap more accurately. Since both $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ satisfy the first order moment condition $k\lambda=\bar{Y}$, the gap between these two estimators can be explored through the profile log-likelihood by substituting $\lambda=\bar{Y}/k$ into the log-likelihood function Eq. [2] as

$$l_k(k)=n\left[k(\overline{\ln Y}-\ln\bar{Y}-1)-\overline{\ln Y}-\ln\Gamma(k)+k\ln k\right].$$

which is a univariate function of k. We immediately know that $l_k(k)$ achieves its maximum at the MLE $k=\hat{k}$. Furthermore, its first order and second order derivatives are given as

$$\begin{aligned} l_k'(k) &= n\left(\overline{\ln Y}-\ln\bar{Y}-\psi(k)+\ln k\right), \\ l_k''(k) &= n\left(1/k-\psi_1(k)\right), \end{aligned}$$

with $l_k'(\hat{k})=0$. Since both $l_k(k)$ and $l_k'(k)$ are differentiable on the interval $(0,\infty)$, by the mean value theorem there exists κ_0 bewteen $\hat{k}$ and $\tilde{k}$ such that

$$l_k'(\tilde{k})=l_k'(\tilde{k})-l_k'(\hat{k})=l_k''(\kappa_0)(\tilde{k}-\hat{k}). \qquad [A4]$$

We can then estimate the gap between $\tilde{k}$ and $\hat{k}$ without estimating the MLE $\hat{k}$ by plugging $\tilde{k}$ into Eq. [A4] as

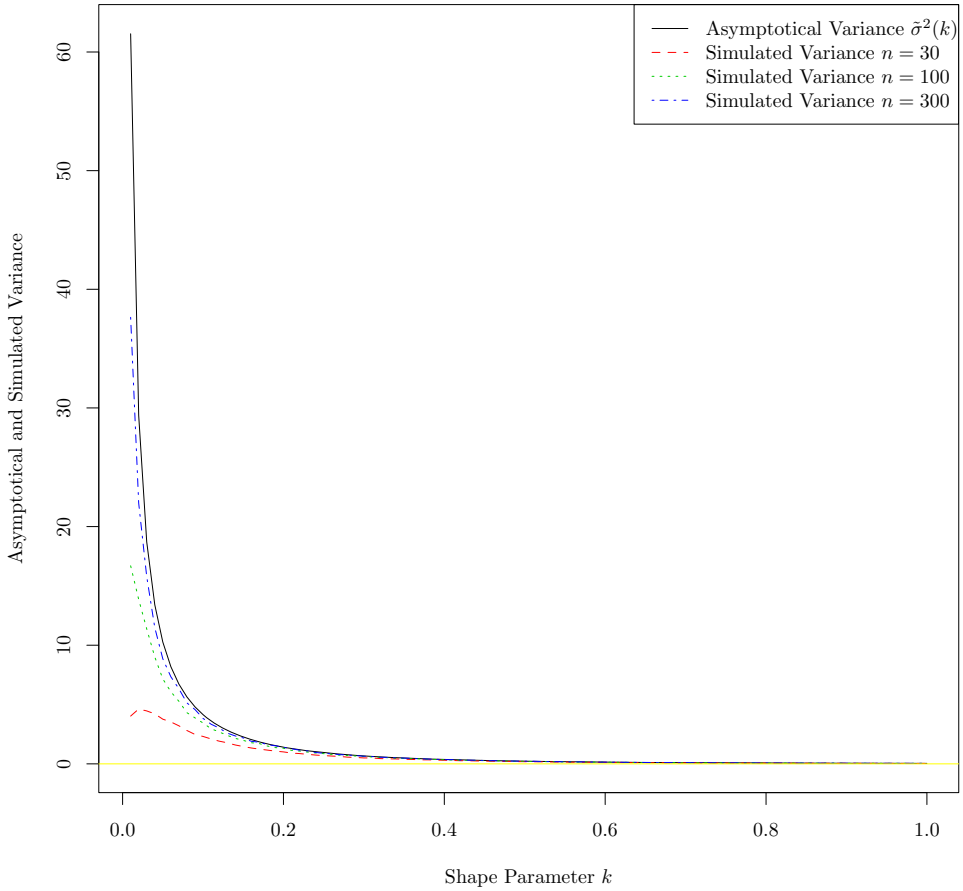

Figure A1. The asymptotic variance $\sigma_e^2(k)$ versus the shape parameter k. The asymptotic variance (in black solid line) is directly calculated from Eq. [A3]. The simulated variances for sample sizes 30 (in red dash line), 100 (in green dot line), and 300 (in blue dot dash line) are calculated based on 10,000 Monte Carlo repetitions from $k = 0.01$ to $k = 1$ by 0.01.

$$\tilde{k} - \hat{k} \approx \hat{e}_k = \frac{l'_k(\tilde{k})}{l''_k(\tilde{k})} = \frac{\overline{\ln Y} - \ln \bar{Y} - \psi(\tilde{k}) + \ln \tilde{k}}{1/\tilde{k} - \psi_1(\tilde{k})}.$$

The gap between $\tilde{\lambda}$ and $\hat{\lambda}$ can then be estimated by subtracting the estimated error term $\hat{e}_k$ from $\tilde{k}$ and invoking the first order moment condition $\mathbb{E}[Y] = k\lambda$ as follows.

$$\tilde{\lambda} - \hat{\lambda} \approx \hat{e}_\lambda = \tilde{\lambda} - \frac{\bar{Y}}{\tilde{k} - \hat{e}_k} = -\frac{\tilde{\lambda}\hat{e}_k}{\tilde{k} - \hat{e}_k}.$$

Since both two estimators satisfy the first order moment condition, we can replace the gamma log-likelihood by the profile log-likelihood as

$$\tilde{l}^* = l_k(\tilde{k}), \quad l^* = l_k(\hat{k}).$$

By the second order Taylor expansion of $l_k(k)$ at $k = \tilde{k}$, we have

$$l^* - \tilde{l}^* = l_k(\hat{k}) - l_k(\tilde{k}) \approx l'_k(\tilde{k})(\hat{k} - \tilde{k}) + \frac{l''_k(\tilde{k})}{2}(\hat{k} - \tilde{k})^2.$$

Furthermore, the difference between AML and EML can be approximated by plugging in $\tilde{k} - \hat{k} \approx \hat{e}_k = l'_k(\tilde{k})/l''_k(\tilde{k})$ as follows:

$$\begin{aligned} l^* - \tilde{l}^* &\approx l'_k(\tilde{k}) \cdot (-\hat{e}_k) + \frac{l''_k(\tilde{k})}{2} \cdot (-\hat{e}_k)^2 \\ &= -\frac{\left[l_k(\tilde{k})\right]^2}{l''_k(\tilde{k})} + \frac{\left[l_k(\tilde{k})\right]^2}{2l''_k(\tilde{k})} = -\frac{\left[l_k(\tilde{k})\right]^2}{2l''_k(\tilde{k})} \end{aligned} \quad [A5]$$

$$= \frac{n\left[\overline{\ln Y} - \ln \bar{Y} - \psi(\tilde{k}) + \ln \tilde{k}\right]^2}{2\left[\psi_1(\tilde{k}) - 1/\tilde{k}\right]} = \frac{\left[\sqrt{n}e(\tilde{\lambda}, \tilde{k})\right]^2}{2\left[\psi_1(\tilde{k}) - 1/\tilde{k}\right]}. \quad [A6]$$

where the last equality follows by substituting $\tilde{k} = \bar{Y}/\tilde{\lambda}$ into $\ln \tilde{k}$. The calibration of AML is then given by adding the error term in Eq. [A6] to the AML $\tilde{l}^*$ as

$$l* = \tilde{l}^* + \frac{n\left[\overline{\ln Y} - \ln \bar{Y} - \psi(\tilde{k}) + \ln \tilde{k}\right]^2}{2\left[\psi_1(\tilde{k}) - 1/\tilde{k}\right]}.$$

By Slutsky's theorem, it is easy to show that after a suitable rescaling the asymptotical distribution of $l^* - \tilde{l}^*$ follows a $\chi^2(1)$ distribution as

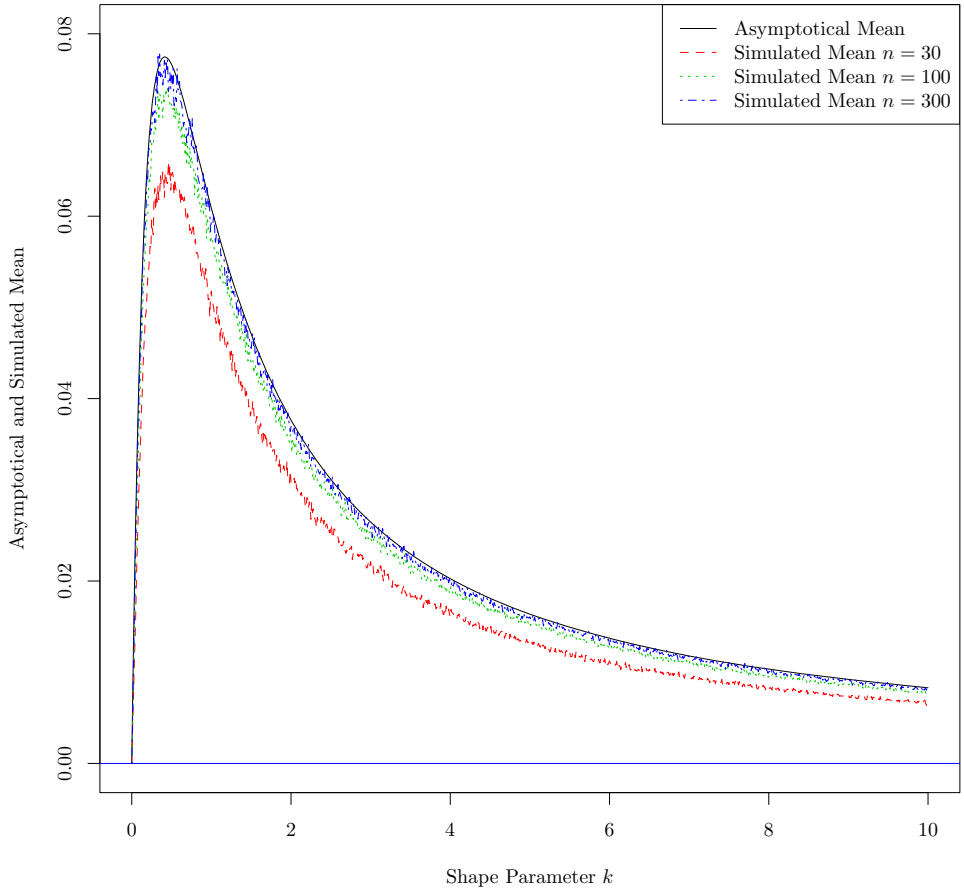

Figure A2. The asymptotic mean of $l^* - \tilde{l}^*$, *i.e.*, $\sigma_l^2(k)$, versus the shape parameter k. The asymptotic mean (in black solid line) is directly calculated from Eq. [8]. The simulated means for sample sizes 30 (in red dash line), 100 (in green dot line), and 300 (in blue dot dash line) are calculated based on 10,000 Monte Carlo repetitions from $k = 0.01$ to $k = 10$ by 0.01.

$$2(l^* - \tilde{l}^*) \approx \frac{\sigma_e^2(k)\left[\sqrt{n}e(\tilde{\lambda}, \tilde{k})/\sigma_e(k)\right]^2}{\psi_1(\tilde{k}) - 1/\tilde{k}} \to k(k^2\psi_1^2(k) - \psi_1(k) - 1) \cdot \chi^2(1). \quad [A7]$$

$\sigma_l^2(k) = k(k^2\psi_1^2(k) - \psi_1(k) - 1)$ shown in Figure A2 is bounded when $k \to 0$ and $k \to \infty$. The maximum of the asymptotical mean of the error term in AML, *i.e.*, $\max_{k>0}\sigma_l^2(k) \approx 0.0775$ is attained around $k = 0.4154$. Figure A2 also depicts the Monte Carlo simulated mean of the error term of AML. We see that the simulated means are bounded by the asymptotical mean and approach the asymptotical one when the sample size n increases.

The approximation error of AML is controlled by the shape parameter k only. The above investigations reveal that the AML is near to the EML for any possible values of the gamma parameters up to a minor stochastic disturbance. Moreover, the error term is free of the sample size n. Therefore, we establish that $l^* - \tilde{l}^* = O(1/k) \cdot O_p(1)$. Since $\tilde{\boldsymbol{\theta}}$ is only an $O_p(1/\sqrt{n})$ approximation to the MLE, these two estimates may differ a lot when the sample size is small. As the log-likelihood function is $O_p(n)$, the relative accuracy of the proposed approximation is $O_p(1/n)$ which is higher than that of the estimates. Furthermore, this additional term as a scaled χ^2 random variable with degree of freedom one can be regarded as the price of the approximation. Estimating the parameter by MLE needs only the sufficient statistics $(\bar{Y}, \overline{\ln Y})$ while estimating the parameters by the closed form estimator $\tilde{\boldsymbol{\theta}}$ needs one more statistic $\overline{Y \ln Y}$. Similarly, since the calibrated log-likelihood is asymptotically equivalent to the exact one, $\tilde{l}*$ approximate l^* with a higher order naturally and the approximation error of AMLC is negligible as $l^* - \tilde{l}* \to_p 0$.