Detecting Performance Degradation of Software-Intensive Systems in the Presence of Trends and Long-Range Dependence

Alexey Artemov
Lomonosov Moscow State University
Complex Systems Modeling Laboratory,
27-1 Lomonosovsky Ave., Moscow 119991, Russia
Yandex Data Factory,
16 Leo Tolstoy St., Moscow 119021, Russia,
Email: artemov@physics.msu.ru

Evgeny Burnaev
Skolkovo Institute of Science and Technology,
3 Skolkovo Innovation Center, Moscow, 143026, Russia
Institute for Information Transmission Problems,
19 Bolshoy Karetny Lane, Moscow 127994, Russia,
Email: e.burnaev@skoltech.ru

***Abstract*—As contemporary software-intensive systems reach increasingly large scale, it is imperative that failure detection schemes be developed to help prevent costly system downtimes. A promising direction towards the construction of such schemes is the exploitation of easily available measurements of system performance characteristics such as average number of processed requests and queue size per unit of time. In this work, we investigate a holistic methodology for detection of abrupt changes in time series data in the presence of quasi-seasonal trends and long-range dependence with a focus on failure detection in computer systems. We propose a trend estimation method enjoying optimality properties in the presence of long-range dependent noise to estimate what is considered "normal" system behaviour. To detect change-points and anomalies, we develop an approach based on the ensembles of "weak" detectors. We demonstrate the performance of the proposed change-point detection scheme using an artificial dataset, the publicly available Abilene dataset as well as the proprietary geoinformation system dataset.**

I. Introduction

The last decade has witnessed the emergence of a novel type of high-tech systems: the software-intensive systems [1]. The latter[1] include digital communication systems, internet systems (including devices, data transfer networks and services), call centers, automated power grids, intellectual transport systems, electronic trading platforms and many others. The obvious requirement for such systems is the effective, reliable and uninterrupted operation. However, recent studies of large-scale software-intensive systems indicate quite the opposite state of affairs: due to their sheer scale[2] "software and hardware failures will be the norm rather than the exception" [3]. According to the research, the dominant cause of costly and dangerous system failures are the software failures which makes software "the most problematic element of large-scale systems" [2].

Among the efforts undertaken in order to improve system reliability a major role is played by failure detection which aims to identify failures based on the analysis of data collected during the system operation. Examples of such data include the average number of processed requests and the queue size per time unit, the volume of transferred traffic, the number of rejected queries, etc. During both unexpected events (such as network equipment failures and network attacks) and scheduled occasions (e. g. data center maintenance and system software upgrades) the data experience abrupt deviations from the target state. The goal then is to detect sudden changes (referred to as anomalies or disorders) in the flow of the observed data. The detection is to be performed online; within the online (sequential) setting, as long as the behavior of the observations is consistent with the target state, one is content to let the process continue. If the state changes, then one is interested in detecting the change as rapidly as possible. Problems concerned with constructing efficient procedures for detecting changes in observed stochastic processes are known in the literature as change-point detection problems [4].

In the present work, we investigate the change-point detection problem for localization and diagnosis of anomalies in large-scale software-intensive systems in the presence of quasi-periodic trends and long-range dependence. The key step in the change-point detection approach is the specification of what is "normal" and "abnormal" state. This problem represents a challenge due to a number of reasons. First, the systems we consider here experience anthropogenic "nearly periodic" load variations that are difficult to model due to a complex load shape and its random variations over time. An example of quasi-periodic time series we investigate in this work is shown in Fig. 1; they reflect weekly and daily load profiles for several internet services.

[1] defined in ISO/IEC/IEEE 42010:2011 as systems where "software contributes essential infuences to the design, construction, deployment, and evolution of the system as a whole"

[2] Expressed in "number of lines of code; number of people employing the system for different purposes; amount of data stored, accessed, manipulated, and refined; number of connections and interdependencies among software components; and number of hardware elements" [2].

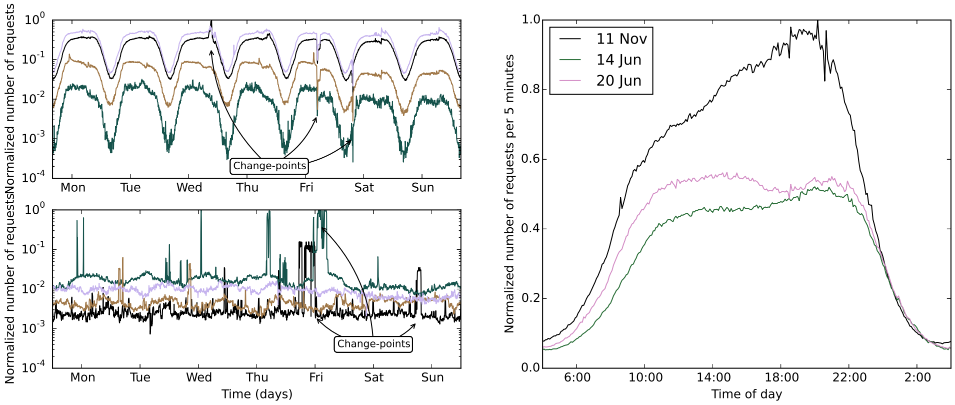


Fig. 1: Top-left: weekly load profile of a geoinfomation system at Yandex along with several change-points. Right: daily load of a system at Yandex aggregated over consecutive 5-minute intervals for three different days in 2014: Saturday, 15th June, Friday, 20th June, and Tuesday, 11th November. Note the change of the load profile from weekday to weekend and throughout the year. Bottom-left: weekly load shape of the traffic in the Abilene network (1008 measurements), for the period of 14–21 June, 2004, along with several change-points.

The second essential property of data flows in large-scale computer systems is that long-term correlations are present in these quantities, i. e., they are statistically self-similar [5]. As self-similarity (also referred to as *long-range dependence* or *LRD*) has significant impact on queueing performance and represents the dominant cause of load "bursts", the model should be able to efficiently capture it.

A natural approach to change-point detection involves utilization of statistical detection procedures such as the CUSUM procedure [6], the control charts procedure [7], etc., as they possess certain efficiency properties. In turn, for these procedures to be implemented, the change-point model (the model of the "normal" and the "abnormal" signals) must be specified. The latter often cannot be specified accurately; as a consequence, even theoretically optimal procedures suffer significant degradation in change-point performance.

Finally, a considerable difficulty is caused by the large scale of contemporary software-intensive systems. For instance, volume of the dataset measured at Yandex[3] reaches hundreds of thousands of characteristics, while other authors report software systems consisting of up to tens of thousand nodes [3]. As the cost of manual model selection for each individual observed signal might be unacceptable, one should consider an automatic approach to model learning.

In this paper, we present an optimal method for signal estimation and an efficient procedure for change-point and anomaly detection in the presence of quasi-periodic trends and long-range dependence. We use our theoretical results regarding the structure of the optimal filter to construct a practical trend estimation algorithm. Using the estimate, we develop the change-point detection algorithm based on the ensemble of "weak" detectors to improve change-point detection performance when the standard assumptions regarding the change-point model are violated.

We briefly describe existing change-point detection approaches as well as some of the conventional filtering techniques in Sect. II. In Sect. III, we specify our time series model and propose the model estimation algorithm. In Sect. IV, we consider the particular change-point detection problem for our model and develop the ensemble-based change-point detection method. Sect. V presents the evaluation results for a simulated and two real-world datasets: a publicly available Abilene network dataset and a proprietary Yandex dataset.

II. Related Work

A vast body of research covers the problem of failure detection in computer systems, and efficient detection algorithms have been developed for anomaly detection in computer networks, data stream networks, etc, see, e. g., [8], [9] and references therein. In these applications, the change-point detection problem is investigated for the case of stationary random series, which is a well-studied setting (See [4] for a bird's eye review).

Process stationarity assumption is rather restrictive for practice since in many applications the observed process is non-stationary. While no specific assumptions about the structure of the observed process are made, purely data-driven approaches such as principal component analysis (PCA) and its modifications are often taken under consideration [9]. PCA and the subspace methods classify the observed data into "normal" and "abnormal" subspaces and have proven themselves efficient in anomaly detection applications [8], [10], [9].

Change-point detection approaches mentioned above are difficult to apply directly to our problem. On the one hand, the observed data in our system are non-stationary; on the other hand, these data are characterized by trends and LRD noise that make PCA and the subspace methods ineffective.

[3] Yandex is one of the largest internet companies in Europe, operating Russia's most popular search engine and its most visited website, see http://company.yandex.com.

A body of research covers a vast number of trend modeling and estimation approaches, such as multiple exponential smoothing [11], autoregressive models [12], decomposition methods [13], parametric and nonparametric regression [14]. Neither of the approaches incorporates an explicit model of LRD; consequently, efficient trend estimation in the presence of LRD cannot be achieved. On the contrary, our trend extraction approach relies on an explicit model of LRD signal and yields theoretically efficient estimates.

III. Trend Estimation in the Presence of Long-Range Dependence

A. LRD and the Fractional Brownian Motion

Long-range dependence is a phenomenon shared by many natural and technical systems. It relates to the rate of decay of statistical dependence of points with increasing time interval. In relation to software-intensive systems, LRD may be qualified as the presence of "burstiness" across an extremely wide range of time scales [5]. During the last decades, the fractional Brownian motion has been established as the standard model for LRD signals.

The fractional Brownian motion (fBm) was introduced by Kolmogorov in connection with his works on the theory of turbulence [15] and later was constructively defined by Mandelbrot [16]. In what follows, we adopt the notation from [17]. A standard fBm $B^H = (B_t^H)_{0 \leqslant t \leqslant T}$ with Hurst exponent $H \in (0,1)$ on $[0,T]$ is a Gaussian process with continuous trajectories, $B_0^H = 0$, $\mathbf{E}B_t^H = 0$, $\mathbf{E}B_s^H B_t^H = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right)$. When $H = \frac{1}{2}$, the process B^H is a standard Brownian motion but in the case $H \neq \frac{1}{2}$ the process B^H is not a semimartingale. In many applications, process B^H is used for modeling of time series with very chaotic movements (the case $H < 1/2$) and with a relatively smooth behavior (the case $H \geqslant 1/2$).

B. The Specification of the Theoretical Filter

Let the observed continuous-time process $X = (X_t)_{0 \leqslant t \leqslant T}$ satisfy the relation

$$X_t = \sum_{i=0}^{n} \theta_i \varphi_i(t) + \sigma B_t^H, \tag{1}$$

where $\{\varphi_i(t)\}_{i=0}^n$ is a dictionary of differentiable functions on $[0,T]$, $B^H = (B_t^H)_{0 \leqslant t \leqslant T}$ is the standard fBm on $[0,T]$ with a known Hurst index H, and the variance $\sigma > 0$ is assumed to be known. The unknown parameters $\{\theta_i\}_{i=0}^n$ must be estimated using the observations $\{X_s, 0 \leqslant s \leqslant t\}$ available up to time t.

In [17], theoretical results regarding the structure of the optimal filter in (1) for the general dictionary of functions $\{\varphi_i(t)\}_{i=0}^n$ were obtained for the case of (a) the maximum likelihood estimate and (b) the Bayesian estimate. For the purpose of the current work, we use the maximum likelihood (ML) filter to estimate a smooth trend against the LRD noise. We assume that:

- the dictionary consists of power functions: $\varphi_i(t) = t^i, i = 0, \ldots, 3$, allowing to estimate the polynomial trend $f(t)$;
- the value of the Hurst exponent H is known (in practice, H can be estimated from the observations using such approaches as introduced in [18], [19]);
- the value of the variance σ is known (in fact, the filter from [17] does not depend on the variance, see below).

The ML estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ for the drift parameter $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_3)$ is given by

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \boldsymbol{R}_H^{-1}(t)\boldsymbol{\psi}_t^H, \tag{2}$$

where $\boldsymbol{R}_H(t) = (\boldsymbol{R}_H(t))_{ij}$ and $\boldsymbol{\psi}_t^H = ((\boldsymbol{\psi}_t^H)_0, \ldots, (\boldsymbol{\psi}_t^H)_3)$ are defined by $(\boldsymbol{R}_H(t))_{ij} = \alpha_H(i,j)t^{i+j-2H}$ and $(\boldsymbol{\psi}_t^H)_i = \beta_H(i)\int\limits_0^t s^{i-1} dM_s^H$, where $\lambda_H = 2H\frac{\Gamma(3-2H)\Gamma(1/2+H)}{\Gamma(3/2-H)}$, $\alpha_H(i,j) = \lambda_H^{-1}\beta_H(i)\beta_H(j)\frac{2-2H}{i+j-2H}$, $\beta_H(i) = i^{\frac{2-2H+i-1}{2-2H}}\frac{\Gamma(3-2H)}{\Gamma(3-2H+i-1)}\frac{\Gamma(3/2-H+i-1)}{\Gamma(3/2-H)}$, $i,j = 0, \ldots, n$, and $M^H = (M_t^H)_{0 \leqslant t \leqslant T}$ is a martingale defined by $M_t^H \equiv \kappa_H^{-1}\int\limits_0^t s^{1/2-H}(t-s)^{1/2-H}dX_s$, $\kappa_H = 2H\Gamma(3/2-H)\Gamma(1/2+H)$.

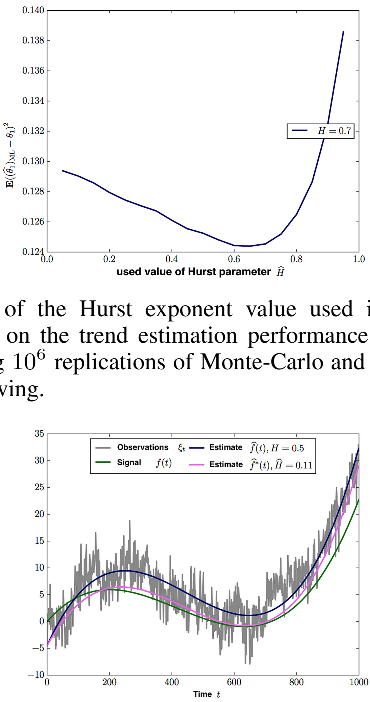
C. The Trend Estimation Algorithm with LRD Correction

The algorithm assumes the observations are taken according to the model

$$X_t = f(t) + \eta^H(t), \qquad t \geqslant 0, \tag{3}$$

where the trend $f(t)$ is some smooth function observed in the LRD noise $\eta^H(t)$. Taking advantage of the smoothness of the trend $f(t)$, we approximate it using some finite-order polynomial $\sum_{i=0}^n \theta_i(t-t_0)^i$ in the neighbourhood of any $t_0 > 0$. We model $\eta^H(t)$ using the fractional Gaussian noise (fGn) Z_t^H with some (unknown but nonrandom) variance $\sigma(t)$ and Hurst exponent H: $\eta^H(t) = \sigma(t)Z_t^H$. Given the noisy observations $\{(X_k, t_k)\}_{k=1}^{\ell}$, the goal is to estimate the expected value $f(t) = \mathbf{E}X_t$ for any $t \geqslant 0$. The following algorithm provides a solution to this problem.

1) Consider an interval $[a,b]$ and select observations window $W(a,b) = \{(X_k, t_k) : a \leqslant t_k \leqslant b\}$.
2) Compute the estimate $\widehat{f}_{[a,b]}(t)$ of the trend $f(t)$ for $a \leqslant t \leqslant b$:
   a) Assuming a cubic polynomial model for the observations
   $$X_k = \sum_{i=0}^{3} \theta_i(t_k - t_0)^i + \sigma Z_k^H, \tag{4}$$
   where $(X_k, t_k) \in W(a,b)$, $t_0 = (a+b)/2$, σ is assumed to be constant, and $H = \frac{1}{2}$, estimate the value of $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_3)$ using the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ described in Sect. III-B.
   b) Compute the trend estimate on $[a,b]$ using the relation $\widehat{f}_{[a,b]}(t) = \sum_{i=0}^3 (\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})_i(t-t_0)^i$ for each $t \in [a,b]$.
   c) Compute the variance estimate $\widehat{\sigma}$ as the sample variance of residuals $\{X_k - \widehat{f}_{[a,b]}(t_k) \mid t_k \in [a,b]\}$.



(a) Influence of the Hurst exponent value used in the algorithm in Sect. III-C on the trend estimation performance. The results are obtained using 10^6 replications of Monte-Carlo and rescaled to $[0, 1]$ for better viewing.

(b) The effect of the correction in 2a–2c on the trend estimation accuracy. The estimate $\widehat{f}(t)$ was obtained using $\widehat{H} = 0.5$, while the corrected estimate $\widehat{f}^*(t)$ was obtained using $\widehat{H} = 0.11$ (true $H = 0.1$).

Fig. 2: Correction employed in the algorithm in Sect. III-C and its effect on the trend extraction performance.

   d) Compute the estimate of the Hurst exponent $\widehat{H}$ using an approach from [19] and the standardized residuals $\{(X_k - \widehat{f}_{[a,b]}(t_k))/\widehat{\sigma} \mid t_k \in [a, b]\}$.
   e) Using the Hurst exponent estimate $\widehat{H}$, compute corrected trend and variance estimates in a)–c).

3) We use the sliding window $[a, b] = [a, a + \Delta]$ with sufficiently large Δ and obtain $n_{[a,b]}(t) = |A(t)|$ local corrected estimates $\widehat{f}_{[a,b]}(t)$ for each $t \geqslant 0$, where $A(t) = \{(a, b) \mid t \in [a, b]\}$. To obtain the final estimate $\widehat{f}(t)$ we average the corrected estimates using the relation $\widehat{f}(t) = \frac{1}{n_{[a,b]}(t)} \sum\limits_{(a,b)\in A(t)} \widehat{f}_{[a,b]}(t)$.

The two-step procedure for computing the estimate $\widehat{f}(t)$ is necessary since in practice the Hurst exponent is unknown but important constant strongly influencing an estimation performance, see Fig. 2a. By applying the correction in the algorithm steps 2a–2c we achieve better trend estimation accuracy compared to a generic approach with $H = \frac{1}{2}$, see Fig. 2b.

IV. Change-point Detection in the Presence of Trends and Long-Range Dependence

A. The Change-point Model

We consider the following change-point model for the noise $\eta^H(t)$ in (4):

$$\eta^H(t) = \mu \mathbb{1}_{[\theta,\theta+\Delta t]}(t) + \sigma Z_t^H, \qquad t \geqslant 0, \tag{5}$$

where θ is an unknown time of a change, μ is an unknown change magnitude, σ is an unknown (non-random) variance, and Z_t^H is the fGn. The characteristic duration Δt of the considered change is short; hence the change represents a local deviation in the values of the observed series, see Fig. 6a.

To detect the change, we introduce a residual process $R = (R_t)_{t\geqslant 0}$

$$R_t = \sigma^{-1}(X_t - \widehat{X}_t), \qquad t \geqslant 0, \tag{6}$$

where X_t is the signal with a known variance σ observed in (4) and $\widehat{X}_t$ is an estimate of X_t obtained via filtering algorithm described in Sect. III-C. In absence of a change, R_t is an approximately zero-mean process with unit variance, however, in presence of a change, neither of these properties holds. Note that R_t is a natural estimate for Z_t^H and σR_t is a natural estimate for the noise component $\eta^H(t)$. We use the process R_t in Sect. IV-B to detect the change.

B. The Ensemble-based Change-point Detection Procedure

The standard assumptions regarding the change-point model state that pre- and post-change distributions are Gaussian i.i.d. with different (yet known) parameters [7], [6], [4]. These assumptions are heavily violated in our case due to (a) the approximation error introduced by substitution of the real trend $f(t)$ with a locally cubic trend, (b) the estimation error introduced by the estimation algorithm in Sect. III-C, (c) the unknown change signature, and (d) the modeling errors due to interpreting noise in the real signal as the fBm. Moreover, the absence of accurate detection procedures for LRD signals makes the change-point detection performance low when "classical" change-point detection methods are used.

Let $\Pi_1, \ldots, \Pi_n$ denote n change-point detection procedures, such as the cumulative sum (CUSUM) procedure [6] based on the process $T = (T_t)_{t\geqslant 0}$:

$$T_t = \max(0, T_{t-1} + \zeta_t), \quad T_0 = 0, \quad t \geqslant 0, \tag{7}$$

where $\zeta_t = \log(f_0(X_t)/f_\infty(X_t))$ is the log-likelihood ratio, and $f_\infty(\cdot)$ and $f_0(\cdot)$ are one-dimentional pre- and post-change distributions, respectively. Each procedure Π_k prescribes to stop observations at time τ_k which is the first hitting time of some process $S^k = (S_t^k)_{t\geqslant 0}$ to a level $h_k > 0$: $\tau_k = \inf\{t \geqslant 0 : S_t^k \geqslant h_k\}$. We further consider a set of *signals* $\left\{s^k = (s_t^k)_{t\geqslant 0}\right\}_{k=1}^{n}$ defined by $s_t^k = S_t^k / h_k, t \geqslant 0$. We call the procedure A an *ensemble* if its stopping time τ_{A} is defined as the first hitting time of some process $a = (a_t)_{t\geqslant 0}$ to a specified level $h_{\mathrm{A}} > 0$: $\tau_{\mathrm{A}} = \inf\{t \geqslant 0 : a_t \geqslant h_{\mathrm{A}}\}$, where

$$a_t = \psi(\boldsymbol{\lambda}; \mathbf{S}_t^1, \ldots, \mathbf{S}_t^n), \tag{8}$$

$\boldsymbol{\lambda} \in \mathbb{R}^d$ $(d \geqslant n)$ and $\mathbf{S}_t^k = \{s_s^k, 0 \leqslant s \leqslant t\}$ is the history of the signal $s^k = (s_t^k)_{t\geqslant 0}$ up to the time t, $k = 1, \ldots, n$.

Each ensemble is completely defined by the choice of the "aggregation function" $\psi(\cdot)$. In this work, we consider a *logistic regression-based* classifier for which the aggregation function could be written as

$$a_t = \psi_{\mathrm{LOG}-p}(\boldsymbol{\lambda}; \mathbf{S}_t^1, \dots, \mathbf{S}_t^n) = \sigma\Big(\sum_{j=0}^{p}\sum_{k=1}^{n} \lambda_{kj} s_{t-j}^k - \lambda_0\Big), \tag{9}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the logistic function. The value a_t can be interpreted as a posterior probability of a change-point given the observations history $\mathbf{X}_t = \{X_s, 0 \leqslant s \leqslant t\}$ up to the moment t. Note that for this ensemble the threshold h_{A} must be chosen to belong to the interval $(0,1)$ [20].

C. Learning Ensemble Parameters

Ensemble parameters $\boldsymbol{\lambda} \in \mathbb{R}^d$ can be *learned* to optimize a certain performance measure. Let $\mathcal{X}^\ell = \{(X^i, Y^i)\}_{i=1}^\ell$ be the labeled data where each point $(X^i, Y^i) \in \mathcal{X}^\ell$ is a pair, its first component $X^i = (X_t^i)_{0 \leqslant t \leqslant T}$ being a sample path of the observations, and its label $Y^i = (Y_t^i)_{0 \leqslant t \leqslant T}$ being an "abnormal" state indicator: $Y_t^i = \mathbb{1}_{\mathcal{T}_0^i}(t)$. Let T_∞^i and T_0^i be the durations of "normal" and "abnormal" states $\mathcal{T}_\infty^i$ and $\mathcal{T}_0^i$ for each point $(X^i, Y^i), i = 1, \dots, \ell$, respectively. We formulate the problem of learning the parameters $\boldsymbol{\lambda} \in \mathbb{R}^d$ of an ensemble as an optimization problem $\mathbf{F}(\mathrm{A}) \to \inf\limits_{\boldsymbol{\lambda} \in \mathbb{R}^d}$ for the Average Relative Error Rate measure

$$\mathbf{F}(\mathrm{A}) = c_\infty \mathbf{E}_\infty \left[\frac{\int \mathbb{1}_{\{a_t \geqslant h_{\mathrm{A}}\}}(t) \mathbb{1}_{\mathcal{T}_\infty}(t) dt}{\int \mathbb{1}_{\mathcal{T}_\infty}(t) dt} \right] + c_0 \mathbf{E}_0 \left[\frac{\int \mathbb{1}_{\{a_t < h_{\mathrm{A}}\}}(t) \mathbb{1}_{\mathcal{T}_0}(t) dt}{\int \mathbb{1}_{\mathcal{T}_0}(t) dt} \right], \tag{10}$$

where c_∞ and c_0 are the costs of false alarm and false silence, respectively. As $\mathbf{F}(\mathrm{A})$ is a non-differentiable function and cannot be optimized using standard approaches, we introduce its empirical approximation $\widehat{\mathbf{F}}_{\mathrm{D}}(\mathrm{A})$ defined by

$$\widehat{\mathbf{F}}_{\mathrm{D}}(\mathrm{A}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left\{ \frac{c_\infty}{T_\infty^i} \sum_{t \in \mathcal{T}_\infty^i} \sigma(a_t - h_{\mathrm{A}}) + \frac{c_0}{T_0^i} \sum_{t \in \mathcal{T}_0^i} \sigma(h_{\mathrm{A}} - a_t), \right\} \tag{11}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the logistic function. Note now that the function $\widehat{\mathbf{F}}_{\mathrm{D}}(\mathrm{A})$ is differentiable w. r. t. the ensemble parameters $\boldsymbol{\lambda} \in \mathbb{R}^d$ and can therefore be optimized using standard methods.

V. Performance Evaluation

A. Evaluation Datasets

We study the performance of filtering and change-point detection algorithms on two artificial datasets ARTIFICIAL-EASY and ARTIFICIAL-HARD and on two real-world datasets: the publicly available Abilene network dataset and on the proprietary Yandex dataset.

Artificial datasets consist of one-week samples of artificial data $\{(X_k, t_k)\}_{k=1}^K$, $K = 2016$, measured at consecutive 5-minute intervals according to the model $X_k = f(t_k) + \eta^H(t_k)$, where $f(t_k) = A \sin(2\pi t_k / T)$ with $A = 1.5, T = 288$, and $\eta^H(t)$ is the LRD noise process. To model the change-point in the artificial data, for each replication of the sample we generate the LRD noise $\eta^H(t)$ according to the model in (5) with $\sigma = 1$, a random change-point time: $\theta \sim U(T, 6T)$, a random change-point duration: $\Delta t \sim U(5, 100)$, and $Z^H = (Z_t^H)_{t \geqslant 0}$ formed as a discrete approximation of the fGn process with $H = 0.95$. For ARTIFICIAL-EASY, we set the change-point magnitude $\mu = 5$, and for ARTIFICIAL-HARD, change-point magnitude is set to $\mu = 3$. Despite this seemingly large magnitude, as we show below, the change-points we generated are remarkably hard to detect, due to the presence of seasonal trends and LRD noise, see Fig. 3 (left). We generated 1000 independent replications of the sample for training the ensemble and another 1000 for testing. We denote these dataset $\mathcal{X}_{\mathrm{TRAIN}}^\ell$ and $\mathcal{X}_{\mathrm{TEST}}^\ell$, where $\ell = 1000$, respectively.

The Abilene dataset[4] describes network load in the Abilene network in terms of the amount of traffic transmitted between network endpoints during consecutive 5-minute intervals. The data is available for the period of March 1, 2004 to September 10, 2004, and consists of 132 different time series describing traffic transmitted between 12 different network nodes located in 12 different locations across the USA. An example of Abilene data is shown in Fig. 1, bottom-left, for 4 different pairs of endpoints for a particular measurement period. The Abilene dataset is frequently used for evaluation of anomaly detection methods due to its complex structure and presence of both short-lived and long-lived anomalies [10], [9].

The Yandex dataset consists of time series describing the performance of a geoinformation system at Yandex. Each time series is sampled at consecutive 5-minute intervals and it represents the total number of requests processed by the system. An example of Yandex time series is shown in Fig. 1 (top-left) and in Fig. 6a (right) along with labels displaying the anomalies subject to detection.

B. Evaluated Procedures

We train the ensemble using five "weak" detectors: the cumulative sum detector, the Shiryaev-Roberts detector, the Shewhart detector, the changepoint detector, and the posterior probability process detector (for details, refer to [20], Sect. 2).

We empirically compare the performance of our ensemble-based procedure to that of several well-studied approaches, specifically, threshold-based procedure, CUSUM procedure, and the subspace method. The threshold-based procedure EWMA-THRESHOLD uses EWMA to estimate the mean $\widehat{\mu}_t$ and variance $\widehat{\sigma}_t^2$ of the time series X_t, obtains the residuals $R_t = (X_t - \widehat{\mu}_t)/\widehat{\sigma}_t$, and calculates the fraction of the residual points within the time window $[t - \Delta, t]$ located above the threshold h. The stopping time for raising the alarm is defined as $\tau_{\mathrm{THR}} = \inf\{k \geqslant 1 : S_k \geqslant h_{\mathrm{THR}}\}$

[4] See http://www.cs.utexas.edu/~yzhang/research/AbileneTM.

where $S_k = \sum_{i=k-\Delta}^{k} \mathbb{1}_{\{R_i \geqslant h\}}(i)$. The threshold h_{THR}, the per-point threshold h and the window size Δ are algorithm parameters; we only report results regarding the calibrated values of these parameters which result in best performance of the procedure. The EWMA-CUSUM procedure replaces the statistic $S = (S_t)_{t \geqslant 0}$ defined above with the CUSUM statistic T_t defined in (7). The densities $f_\infty(\cdot)$ and $f_0(\cdot)$ are assumed to be normal with unit variances and means $\mu_\infty = 0$ and $\mu_0 = \mu_\infty + \delta$, respectively. The parameter δ is selected to obtain the best performance on training set in terms of the area under the precision-recall curve. The subspace method PCA is closely related to the singular spectrum analysis (SSA) approach and subspace methods from the literature [10], [9], [21]. In the PCA procedure, a decomposition of the time series $X = (X_t)_{t \geqslant 0}$ is obtained using the SSA procedure, and the component $\mathbf{X}_t^{\text{RES}}$ living in the residual subspace is considered. The statistic $P = (P_t)_{t \geqslant 0}$ of the procedure is the norm of the residual component: $P_t = \|\mathbf{X}_t^{\text{RES}}\|$. We note that the subspace method benefits greatly from pretraining on historic data. To exploit this advantage, we supplied the SSA procedure with a week of historic data to obtain a better decomposition. We call this procedure PCA-PRETRAINING. Note that no other procedure receives any additional input when trained.

C. Trend Approximation Accuracy

We first compare the trend extraction accuracy on the dataset ARTIFICIAL-EASY. We use the relative root mean squared forecast error $\text{RRMSE}(X_t, \widehat{X}_t) = \sqrt{\frac{1}{K}\sum_{t=1}^{K}(X_t - \widehat{X}_t)^2/X_t^2}$, to evaluate forecasting performance. Table I presents trend extraction accuracy on two tasks: trend approximation and one-point-ahead forecasting. Trend approximation accuracy $\text{RRMSE}(f(t), \widehat{f}(t))$ measures how closely the extracted trend follows the true trend $f(t)$. One-point-ahead forecasting accuracy estimates how well an algorithm predicts incoming new data X_t given the observed values $\{X_k, k < t\}$. Our study shows that our approach produces significantly more accurate estimates than EWMA. An example of trend approximation is presented in Fig. 3 for the artificial dataset and for the Abilene dataset. We note that our approach yields a smooth approximation and allows for more robust anomaly isolation, while EWMA follows the data more closely.

TABLE I: Trend extraction accuracy for the artificial dataset in terms of RRMSE (%) for EWMA, PCA and our approach.

METHOD	TREND	ONE-POINT-AHEAD
	APPROXIMATION	FORECASTING
EWMA	7.84	7.34
PCA	8.96	5.65
PCA-PRETRAINING	5.58	3.80
OURS	5.72	3.06

D. Change-point Detection Performance Measures

To evaluate the change-point detection performance, we use two performance measures. The first measure is the Precision-Recall Curve, which is a standard performance measure in the area of machine learning. The second measure is the Average Relative Error Rate curve proposed in (10)–(11). Before discussing the obtained results, we briefly explain how these performance measures are computed. Suppose that a procedure Π is defined by a statistic $S = (S_t)_{t \geqslant 0}$. When computed on a test instance $(X^i, Y^i) \in \mathcal{X}_{\text{TEST}}^{\ell}$, procedure Π generates a trajectory $\{S_1^i, \ldots, S_l^i\}$ and for some specified threshold $h_\Pi > 0$ produces M_Π segments $\left\{[t_{a_m}, t_{b_m}]\right\}_{m=1}^{M_\Pi}$ such that $\forall t \in [t_{a_m}, t_{b_m}] \quad S_t^i \geqslant h_\Pi$. We declare the detection $[t_{a_m}, t_{b_m}]$ true positive if it intersects with the "abnormal" segment, i. e. if $[\theta, \theta + \Delta t] \cap [t_{a_m}, t_{b_m}] \neq \emptyset$. If, on the other hand, this intersection is empty (the statistic signals outside the interval $[\theta, \theta + \Delta t]$), then the detection is declared false positive. The

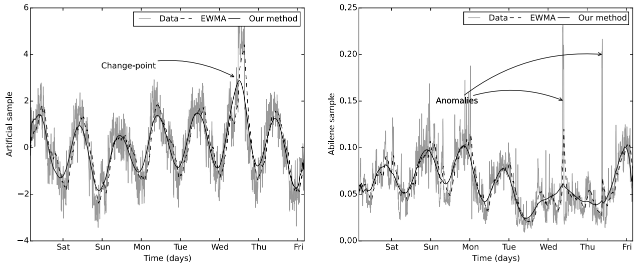
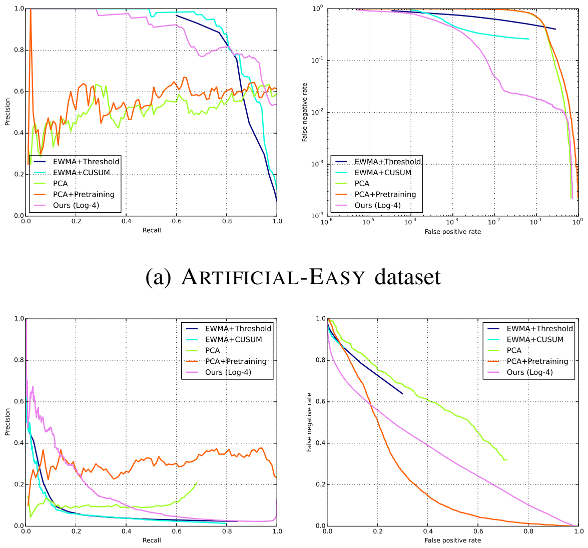


Fig. 3: Example data from the ARTIFICIAL-EASY dataset (left) and the Abilene dataset (right) and trend extraction results obtained using EWMA and our approach. Marked are the labeled anomalies.



(a) ARTIFICIAL-EASY dataset

(b) ARTIFICIAL-HARD dataset

Fig. 4: Empirical comparison of change detection performance for EWMA-based approaches, PCA-based approaches, and our approach. Left: Precision-Recall curves. Right: Average Relative Error Rate curves.

Precision-Recall Curve is plotted by varying the threshold h_Π. The Average Relative Error Rate curve is a plot of Average False Positive Rate $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{c_\infty}{T_\infty^i}\sum_{t\in\mathcal{T}_\infty^i}\mathbb{1}_{\{S_t^i\geqslant h_\Pi\}}(t)$ versus Average False Negative Rate $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{c_0}{T_0^i}\sum_{t\in\mathcal{T}_0^i}\mathbb{1}_{\{S_t^i< h_\Pi\}}(t)$. Average Relative Error Rate can be thought of as a segmentation rather than classification measure.

E. Results

For the ARTIFICIAL-EASY data, our approach is outperformed only by the optimal CUSUM procedure by a little margin when measured in terms of AUC, see Fig. 4a, left. On ARTIFICIAL-HARD, our approach outperforms all other methods in equal conditions. However, adding more data to PCA to improve decomposition accuracy makes it the best on this task, see Fig. 4b, left. Our approach also yields the most accurate segmentations, as can be seen on both Fig. 4a, right, and Fig. 4b, left, meaning both lower average false silence and lower average false alarm durations.

We conclude that our approach significantly outperforms the rival algorithms in terms of the precision-recall characteristic. The reason for this increase in change-point detection performance is the high correlation between the true change-points and the proposed detections, as can be seen in Fig. 5, right. We note, however, that due to the complex nature of both artificial datasets, many change-points are difficult to detect.

Trend extraction results for the two real-world datasets are presented in Fig. 6a for EWMA and our approach, and in Fig. 6b for PCA and our approach. As can be seen from these figures, our filtering approach would result in residuals which violate the change-point model in (5) to a lesser extent; the ensemble then further should improve detection performance because it optimizes (10) on the residual data. PCA-based approach performs generally comparable to our approach (and even outperforms it in case of pretraining); however, PCA requires retraining which is computationally very expensive when performed online on a large number of time series. Our filtering approach is advantageous in that it may be implemented online via a simple linear filter. More results are in Fig. 7, where change-point detection results using the logistic regression-based ensemble are presented for both Yandex and Abilene data. We conclude that our approach is effective for both artificial and real data and can readily be applied for anomaly detection in a multitude of environments.

VI. CONCLUSION

We investigated change-point detection in the presence of quasi-seasonal trends and long-range dependent noise with an application to fault detection in software-intensive systems. We proposed an effective trend estimation algorithm based on the theoretically optimal filter and a practical change-point detection procedure based on the ensemble of “weak” detectors. An empirical study of the change-point detection procedure shows that it significantly ourperforms the standard EWMA and PCA-based algorithms when the conventional assumptions about the change-point model are violated.

VII. ACKNOWLEDGEMENTS

The research, presented in Section V of this paper, was supported by the RFBR grants 16-01-00576 A and 16-29-09649 ofi_m; the research, presented in other sections, was conducted in IITP RAS and supported solely by the Russian Science Foundation grant (project 14-50-00150).

REFERENCES

[1] ISO/IEC/IEEE: Systems and software engineering – architecture description. ISO/IEC/IEEE 42010:2011(E) (Revision of ISO/IEC 42010:2007 and IEEE Std 1471-2000) (1 2011) 1 –46

[2] Northrop, L., Feiler, P., Gabriel, R.P., Goodenough, J., Linger, R., Longstaff, T., Kazman, R., Klein, M., Schmidt, D., Sullivan, K., et al.: Ultra-large-scale systems: The software challenge of the future. Technical report, DTIC Document (2006)

[3] Yigitbasi, N., Gallet, M., Kondo, D., Iosup, A., Epema, D.: Analysis and modeling of time-correlated failures in large-scale distributed systems. Proceedings - IEEE/ACM International Workshop on Grid Computing (2010) 65–72

[4] Polunchenko, A.S., Tartakovsky, A.G.: State-of-the-Art in Sequential Change-Point Detection. Methodology and Computing in Applied Probability **14**(3) (2012) 649–684

[5] Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V.: On the self-similar nature of Ethernet traffic (extended version) (1994)

[6] Page, E.: Continuous inspection schemes. Biometrika **41**(1) (1954) 100–115

[7] Shewhart, W.A.: Economic control of quality of manufactured product (1931)

[8] Pham, Duc-Son and Venkatesh, Svetha and Lazarescu, Mihai and Budhaditya, S.: Anomaly detection in large-scale data stream networks. Data Mining and Knowledge Discovery **28**(1) (2014) 145–189

[9] Casas, P., Vaton, S., Fillatre, L., Nikiforov, I.: Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements. Computer Networks **54**(11) (2010) 1750–1766

[10] Lakhina, A., Crovella, M., Diot, C.: Diagnosing network-wide traffic anomalies. ACM SIGCOMM Computer Communication Review **34**(4) (2004) 219

[11] Winters, P.R.: Forecasting sales by exponentially weighted moving averages. Management Science **6**(3) (1960) 324–342

[12] Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.C.: New capabilities and methods of the x-12-arima seasonal-adjustment program. Journal of Business & Economic Statistics **16**(2) (1998) 127–152

[13] Hodrick, R.J., Prescott, E.C.: Postwar us business cycles: an empirical investigation. Journal of Money, credit, and Banking (1997) 1–16

[14] Artemov, A.V., Burnaev, E.V., Lokot, A.S.: Nonparametric decomposition of quasi-periodic time series for change-point detection. In: Eighth International Conference on Machine Vision, International Society for Optics and Photonics (2015) 987520–987520

[15] Kolmogorov, A.N.: The wiener spiral and some other interesting curves in hilbert space. In: Dokl. Akad. Nauk SSSR. Volume 26. (1940) 115–118

[16] Mandelbrot, B.B., Van Ness, J.W.: Fractional Brownian Motions, Fractional Noises and Applications (1968)

[17] Artemov, A.V., Burnaev, E.V.: Optimal estimation of a signal perturbed by a fractional brownian noise. Theory Probab. Appl. **60**(1) (2016) 126–134

[18] Kirichenko, L., Radivilova, T., Deineko, Z.: Comparative analysis for estimating of the hurst exponent for stationary and nonstationary time series. Information Technologies & Knowledge **5**(1) (2011) 371–388

[19] Hardstone, R., Poil, S.S., Schiavone, G., Jansen, R., Nikulin, V.V., Mansvelder, H.D., Linkenkaer-Hansen, K.: Detrended fluctuation analysis: a scale-free view on neuronal oscillations. Scale-free Dynamics and Critical Phenomena in Cortical Activity (2012) 75

[20] Artemov, A.V., Burnaev, E.V.: Ensembles of detectors for online detection of transient changes. In: Eighth International Conference on Machine Vision, International Society for Optics and Photonics (2015) 98751Z–98751Z

[21] Vautard, R., Yiou, P., Ghil, M.: Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. Physica D: Nonlinear Phenomena **58**(1) (1992) 95–126

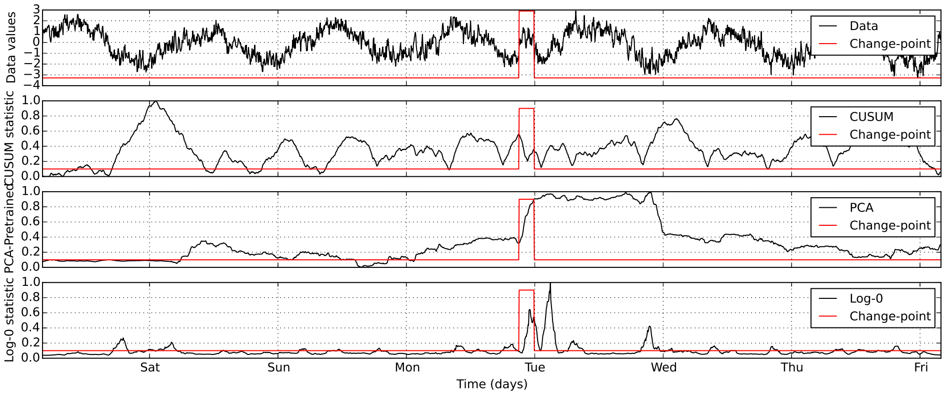
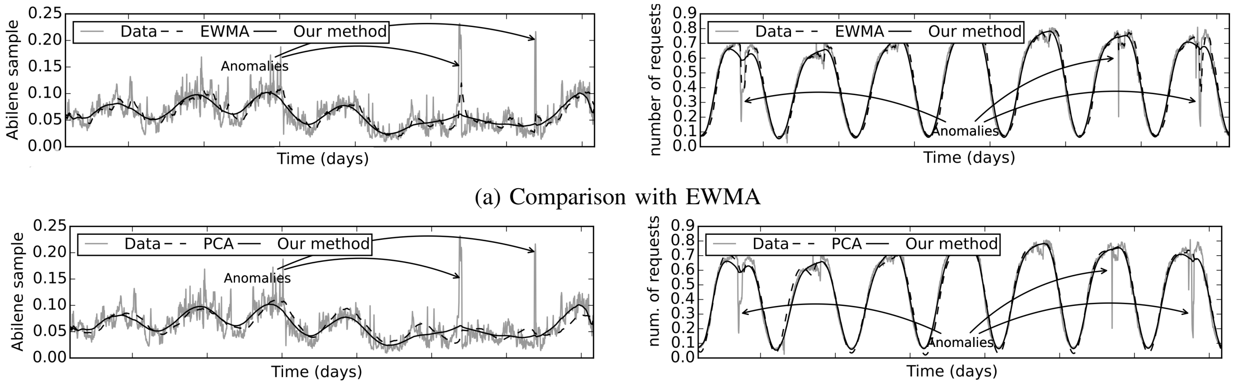


Fig. 5: Comparison of CUSUM, PCA-PRETRAINED and LOG-0 ensemble trajectories for the ARTIFICIAL-HARD dataset. Top: a sample of artificial data with the change-point indicator. Upper middle: sample path of the CUSUM statistic. Note no correlation between the two. Lower middle: sample path of the PCA-PRETRAINED statistic. Note the longer duration of the large values of the statistic. Lower: sample path of the logistic regression-based classifier statistic along with the change-point indicator. Note the strong correlation between the two. All statistics have been rescaled to $[0, 1]$ for better viewing experience.



(a) Comparison with EWMA

(b) Comparison with PCA

Fig. 6: Example data from the Abilene dataset (left) and the Yandex dataset (right), and the trend extraction results obtained using competing methods and our approach. Marked are the detected anomalies.

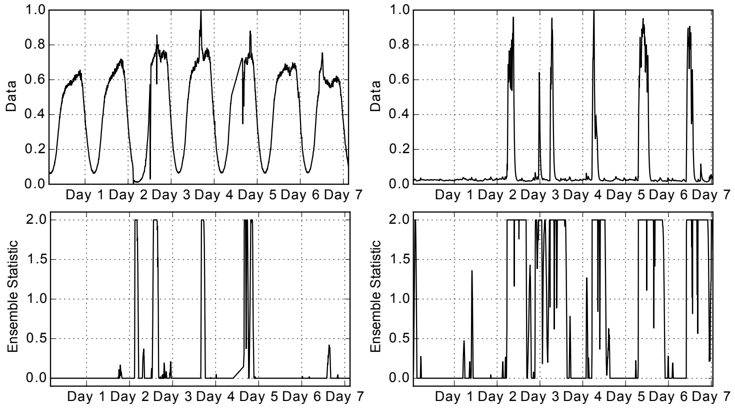


Fig. 7: Change-point detection results for two different cases: Yandex, short-lived change, and Abilene, short-lived change (left to right). From top to bottom are shown: the source data X_t and the logistic regression-based ensemble statistic a_t.