Detecting Changes in the Scale of Dependent Gaussian Processes: a Large Deviations Approach

Julia Kuhn[1 2], Wendy Ellens[1], and Michel Mandjes[1]

[1] University of Amsterdam, The Netherlands
[2] University of Queensland, Australia

Abstract. This paper devises new hypothesis tests for detecting changes in the *scale* of interdependent and serially correlated data streams, i.e, proportional changes of the mean and (co-)variance. Such procedures are of great importance in various networking contexts, since they enable automatic detection of changes, e.g. in the network load. Assuming the underlying structure is Gaussian, we compute the log-likelihood ratio test statistic, either as a function of the observations themselves or as a function of the innovations (i.e., a sequence of i.i.d. Gaussians, to be extracted from the observations). An alarm is raised if the test statistic exceeds a certain threshold. Based on large deviations techniques, we demonstrate how the threshold is chosen such that the ratio of false alarms is kept at a predefined (low) level. Numerical experiments validate the procedure, and demonstrate the merits of a multidimensional detection approach (over multiple one-dimensional tests). Also a detailed comparison between the observations-based approach and the innovations-based approach is provided.

1 Introduction

Statistical change point detection is an important tool in network control, and has been widely applied in e.g. intrusion detection systems, [17,18,19], and overload detection [13]. In order to enable the network operator to adequately respond to persistent changes in the (inherently random) observations, the main task is to detect persistent changes as quickly as possible while keeping the number of false alarms at a predefined low level (for instance 5%).

Traditionally, in the change point detection literature the main focus has been on detecting a change in the *mean* value corresponding to a sequence of *independent*, *one-dimensional* observations [4,8,16]. However, in many situations this setting is far from adequate. In the first place, in practice there is typically positive correlation between subsequent data points [20]. Moreover, single data points often consist of multidimensional records, rather than one-dimensional values. In addition, in the context of communication networks, an increase in the number of active users tends to be not reflected by a change in the mean only, but rather as a *change in scale* – a change in both the mean and (proportionally) the corresponding variance. Therefore, to only focus on the detection

of mean shifts neglects an additional indicator that a change has taken place [2, Ex. 4.1.9].

Motivated by the above considerations, a number of procedures have been proposed that allow for data streams to be either serially correlated [15] or multidimensional [8]. In [19] a detection method for testing serially correlated and multidimensional data streams is presented but the multiple data streams are assumed to be independent. The more general setting of dependent multidimensional data streams is covered in [2], where testing against a change in mean or variance is considered separately. In the current paper, we develop techniques different from those of [2] for the detection of changes in scale in multidimensional and serially correlated sequences, which allow the network operator to limit the false alarm rate to a level of her choice. We focus on *Gaussian* sequences $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$, where sets of observations have a multivariate Normal distribution. Gaussian time series are popular for modeling network traffic, see e.g. [1] and [12, Part A].

Let us consider the illustrative example of a link of a communication network. If the bandwidth consumed by different users is i.i.d., then the mean and variance of the total bandwidth consumption are both proportional to the number of users. As a consequence, a change in the number of users can be considered as a change of scale, in the sense defined above: the mean and variance exhibit the same relative (i.e., percentage-wise) change. When measuring not only at a single link but at various points in the network, then more information is available, potentially facilitating earlier detection or a lower risk of false alarms. In this case – apart from *serial correlation* (correlation over time) – also *cross-correlation* between data sequences generated by different sensors has to be taken into account, because the same traffic may be captured by several sensors. For large traffic aggregates, the Gaussianity is justified by central-limit type of arguments. Based on the above, we conclude that the setup considered in this paper can be used to detect changes in load, caused by, for instance, a (legal) increase in the number of users, or a DDoS (distributed denial of service) attack.

At the methodological level, the testing procedure we propose is a sequential hypothesis test, in line with the popular CUSUM algorithm [14]. Our procedure monitors a likelihood ratio test statistic, and raises an alarm as soon as it exceeds some predefined threshold. The question arises how this threshold should be chosen so as to ensure that the number of false alarms does not exceed a given (low) level. In case the observations are one-dimensional and independent, the test's false alarm performance can be assessed using a functional central limit theorem to establish the convergence of the test statistic to a Brownian motion [16]. Alternatively, since a false alarm is required to be a rare event, a limit expression for the false alarm probability can be derived using large deviations theory (concerned with the asymptotic behavior of rare event probabilities), see e.g. [10] and [5, Ch. VI.E]. Choosing the smallest threshold that satisfies the predefined level of false alarms ensures that an alarm is raised quickly once a change has occurred.

In [11] we extended the large deviations (LD) approach to detect a change in mean in serially correlated (one-dimensional) autoregressive moving average (ARMA) processes. The main objective of the present paper is to further extend such a LD approach to make it applicable to detect changes in scale in multi-dimensional correlated data. Furthermore, we compare the method of testing the sequence of *observations* $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ themselves, with an innovations based approach, where first the observations are transformed into i.i.d. *innovations*, and then change point detection tests are performed on the innovations, see e.g. [2]. Our testing procedure differs from the one considered in [2], which uses the so-called *local approach* [2, 4.2.3] to determine the threshold for testing an i.i.d. sequence, whereas in this paper the false alarm probabilities are evaluated in the large deviations regime to choose the threshold. For the innovations based approach we impose the weak assumption [3, 5.7.1] that the process be linear and invertible, while for the observations based approach we need additional assumptions on the underlying correlation structure. We validate the proposed tests in a series of numerical experiments, which (i) study the trade-off between the detection ratio and the corresponding delay, (ii) assess the gain of multidimensional testing procedures (over multiple one-dimensional tests), and (iii) provide a systematic comparison between (A) the observations-based and (B) the innovations-based method.

This paper is organized as follows. In the next section we explain the change in scale and the set-up of our LD-based hypothesis test in greater detail. In Section 3 we compute the log-likelihood ratio test statistics for the observations and the innovations based detection approach, before we derive the threshold functions in Section 4. The results of the numerical evaluation are presented in Section 5. We conclude in Section 6.

2 Detection Procedure for a Change in Scale

We are concerned with testing a stationary multidimensional Gaussian sequence $(\boldsymbol{X}_t)$ against a change in scale, where after the change both the mean and variance are multiplied by some constant c. Each $\boldsymbol{X}_t$ is a d-dimensional column vector consisting of the measurements of d different 'sensors' at (discrete) time t. In this section we explain the general detection procedure; a more detailed description for the case of a change in scale can be found in the following two sections.

To detect a change in the traffic streams, we monitor windows of size $n \in \mathbb{N}$, i.e., at time t the n most recent observations (in the sequel denoted by $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$) are tested in order to decide whether a change has occurred at some point $k \in \{1, \ldots, n\}$. In other words, we consider the hypotheses:

H_0**:** *No change has occurred within the window.*
H_1**:** *A change occurred at some point within the window.*

Thus, the alternative hypothesis is essentially the union of hypotheses:

$H_1(k)$**:** *A change in scale occurred exactly at time k, for a specific $k \in \{1, \ldots, n\}$.*

It will turn out to be convenient to express the *change point* k via the window size n, that is, we write $k = n\beta + 1$, where (throughout the paper) $\beta \in \mathscr{B} = \{0/n, 1/n, \ldots, (n-1)/n\}$.

To set up the testing procedure, we may consider (A) testing the observations directly, or (B) testing the extracted independent sequence of *innovations* – denoted by $(\boldsymbol{\varepsilon}_t)$ and defined in Section 3. We list some of the benefits and drawbacks of both approaches in Table 1; the details are explained in the sequel. Our method for testing a window of size n can be summarized as follows.

(i) The log-likelihood ratio (LLR) test statistic $\mathscr{L}_{n\beta}(\cdot)$ for testing H_0 against the simple alternative hypothesis $H_1(n\beta + 1)$ is computed as either (A) $\mathscr{L}^{\boldsymbol{X}}_{n\beta}(\boldsymbol{X})$ (when considering the observations) or (B) $\mathscr{L}^{\boldsymbol{\varepsilon}}_{n\beta}(\boldsymbol{\varepsilon})$ (when considering the innovations). The two approaches are equivalent under H_0.

(ii) Based on large deviations theory, the threshold $b(\beta)$ is obtained as (A) $b_{\boldsymbol{X}}(\beta)$ or (B) $b_{\boldsymbol{\varepsilon}}(\beta)$; it is a function of β such that for any value of β asymptotically (for large n) the probability of raising a false alarm is kept at level α.

(iii) In line with [5, Ch. VI.E, Eqn. (43)] we reject H_0 ("raise an alarm")

(A) as soon as

$$\max_{\beta\in\mathscr{B}}\left(\frac{1}{n}\mathscr{L}^{\boldsymbol{X}}_{n\beta}(\boldsymbol{X}) - b_{\boldsymbol{X}}(\beta)\right) := \max_{\beta\in\mathscr{B}}\left(\frac{1}{n}\log\frac{g_{n\beta}(\boldsymbol{X})}{f_n(\boldsymbol{X})} - b_{\boldsymbol{X}}(\beta)\right) > 0, \qquad (1)$$

where f_n and $g_{n\beta}$ are the joint densities of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ under H_0 and $H_1(n\beta + 1)$ respectively.

(B) Accordingly, for the innovations based approach, we raise an alarm when

$$\max_{\beta\in\mathscr{B}}\left(\frac{1}{n}\mathscr{L}^{\boldsymbol{\varepsilon}}_{n\beta}(\boldsymbol{\varepsilon}) - b_{\boldsymbol{\varepsilon}}(\beta)\right) > 0. \qquad (2)$$

We explain steps (i) and (ii) in greater detail in Sections 3 and 4.

3 Computation of the Log-Likelihood Ratio Test Statistic

We now formulate the null hypothesis and the alternative hypothesis for the case of a change in scale in terms of an appropriate test statistic, for (A) the observations based approach, and (B) the innovations based approach. Approach (A) can be used to detect a change point in a stationary Gaussian process, for approach (B) we restrict our exposition to linear processes[3], which allows for the rich class of *vector autoregressive moving average* (VARMA) processes. In both cases we may assume, without loss of generality, that the pre-change process has mean vector $\mathbf{0}$ (we may subtract the original mean vector to achieve this).

(A) For the observations based approach, to compute the LLR, we consider the n observations within each window jointly. The joint distribution of $\boldsymbol{X} := (\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_n^{\mathrm{T}})^{\mathrm{T}}$ under H_0 is $\mathcal{N}_{nd}(\mathbf{0}, \Sigma)$, a Gaussian distribution of dimension nd. We write the covariance matrix Σ of the joint observations

[3] Generalization may be possible using Wold's decomposition theorem.

Table 1: Characteristics of (A) the observations and (B) the innovations based approach

(A)	(B)
Suitable test statistic for changes in mean and variance but also in coefficients	Suitable for detecting changes in mean or variance
Computationally expensive	Recursive computation of LLR and reduced dimensionality
How to define the threshold function in the multidimensional case is unclear, unless there is no shift in mean or data streams are independent	We can compute the threshold function for the change in scale explicitly
The process does not need to be invertible	Requires invertibility
The observations are well-defined test statistics	Since innovations are defined in terms of past observations, initial conditions are required

as a block Toeplitz matrix of the individual autocovariance matrices $\Gamma_h = \mathrm{Cov}\,(\boldsymbol{X}_t, \boldsymbol{X}_{t-h})$.
Now we can formulate H_0 and H_1 more specifically. For all $\beta \in \mathscr{B}$ we want to test

$$H_0 : \boldsymbol{X} \sim \mathcal{N}_{dn}\,(\boldsymbol{0}, \Sigma) \quad \text{vs.} \quad H_1(n\beta+1) : \boldsymbol{X} \sim \mathcal{N}_{dn}\,(\boldsymbol{\nu}, T)\,,$$

where

$$\boldsymbol{\nu} = \left(\boldsymbol{0}^T \ldots, \boldsymbol{0}^T, \bar{\boldsymbol{\nu}}^T, \ldots, \bar{\boldsymbol{\nu}}^T\right)^T, \quad T = \left(\begin{array}{c|c} \Sigma^{(dn\beta)} & 0 \\ \hline 0 & c \cdot \Sigma^{(dn(1-\beta))} \end{array}\right),$$

with $\bar{\boldsymbol{\nu}} = c\boldsymbol{\mu} - \boldsymbol{\mu}$, $\boldsymbol{\mu}$ denoting the mean vector before centering, and where m in $\Sigma^{(m)}$ denotes the dimension of the matrix. For method (A), we assume that the sequence before $n\beta + 1$ is independent of the sequence afterward. This assumption enables computations, and is reasonable if a change has taken place, and the cause of the change is 'external' (as in the examples mentioned in the introduction).
The LLR for testing $\boldsymbol{X} \sim \mathcal{N}_{nd}\,(\boldsymbol{0}, \Sigma)$ against the simple alternative hypothesis $\boldsymbol{X} \sim \mathcal{N}_{nd}\,(\boldsymbol{\nu}, T)$ can be computed as

$$\mathscr{L}_n^{\boldsymbol{X}}(\boldsymbol{X}) = \frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|T| + \frac{1}{2}\boldsymbol{X}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{X} - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\nu})^{\mathrm{T}}T^{-1}(\boldsymbol{X}-\boldsymbol{\nu}).$$

Filling in $\boldsymbol{\nu}$, Σ, T, the LLR for testing against a change in scale at a specific point $n\beta + 1$ becomes

$$\begin{aligned}\mathscr{L}_{n\beta}^{\boldsymbol{X}}(\boldsymbol{X}) = &-\frac{1}{2}dn(1-\beta)\log c + \frac{1}{2}\check{\boldsymbol{X}}^{\mathrm{T}}\left(\Sigma^{(dn(1-\beta))}\right)^{-1}\check{\boldsymbol{X}} \\ &-\frac{1}{2c}\left(\check{\boldsymbol{X}} - \boldsymbol{\nu}^{(dn(1-\beta))}\right)^{\mathrm{T}}\left(\Sigma^{(dn(1-\beta))}\right)^{-1}\left(\check{\boldsymbol{X}} - \boldsymbol{\nu}^{(dn(1-\beta))}\right),\end{aligned} \tag{3}$$

where $\check{\boldsymbol{X}} := (\boldsymbol{X}_{n\beta+1}^{\mathrm{T}}, \ldots, \boldsymbol{X}_n^{\mathrm{T}})^{\mathrm{T}}$.

(B) For the innovations based approach we need to impose further assumptions (see also Table 1). We focus on *linear* processes, i.e., we assume that $\boldsymbol{X}_t$ can be modeled as

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \Psi_j \boldsymbol{Z}_{t-j} =: \Psi(L)\boldsymbol{Z}_t, \tag{4}$$

(L denoting the lag operator: $L\boldsymbol{Z}_t := \boldsymbol{Z}_{t-1}$), with uncorrelated error terms $\boldsymbol{Z}_t \sim \mathcal{N}_d(\boldsymbol{0}, \Omega)$, and where the Ψ_j form an absolutely summable sequence of coefficient matrices [3].

We further need to assume that the process be *invertible*, i.e., that i.i.d. the sequence of *innovations*

$$\boldsymbol{\varepsilon}_t := \boldsymbol{X}_t - \mathbb{E}(\boldsymbol{X}_t \,|\, \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_1) \tag{5}$$

can be extracted as a well-defined function of present and past observations (lie in their closed linear span). If $\boldsymbol{X}_t$ is given by a VARMA(p,q) process

$$\boldsymbol{X}_t = \sum_{i=1}^{p} A_i \boldsymbol{X}_{t-i} + \sum_{j=1}^{q} B_j \boldsymbol{Z}_{t-j} + \boldsymbol{Z}_t,$$

then a well-known sufficient condition for invertibility is that $|B(z)|$ has no roots on the unit circle, where $B(z) = \boldsymbol{I} + \sum_{j=1}^{q} B_j z^j$ denotes the MA-polynomial [3].

Given such an invertibility assumption holds, a proportional change in the covariance matrix of the observations (i.e. covariances are inflated by c) can be detected as a proportional change in the covariance matrix of the innovations, as it is known [3, Eqn. (11.1.13)] that under H_0 the autocovariances of $\boldsymbol{X}_t$ are given by $\Gamma_h = \sum_j \Psi_j \Omega \Psi_{j-h}^{\mathrm{T}}$. It has been shown in [2] that (for VARMA processes) the sequence of innovations is a sufficient statistic for detecting a change in the mean value.

Then, defining $\theta = \Psi(L)^{-1}\bar{\boldsymbol{\nu}}$, the above hypotheses can equivalently be formulated as

$$H_0 : \boldsymbol{\varepsilon}_t \sim \mathcal{N}_d(\boldsymbol{0}, \Omega),\ t = 1, \ldots, n \quad \text{vs.} \quad H_1(n\beta+1) : \begin{cases} \boldsymbol{\varepsilon}_t \sim \mathcal{N}_d(\boldsymbol{0}, \Omega), & t \le n\beta, \\ \boldsymbol{\varepsilon}_t \sim \mathcal{N}_d(\boldsymbol{\theta}, c\Omega), & t > n\beta. \end{cases}$$

Since the innovations are independent, the LLR $\mathscr{L}_{n\beta}^{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$ for testing H_0 against $H_1(n\beta+1)$ can be expressed as the sum of the LLRs at time $t > n\beta$ (since the LLR is zero for $t \le n\beta$). Therefore, $\mathscr{L}_{n\beta}^{\boldsymbol{\varepsilon}}$ becomes

$$\mathscr{L}_{n\beta}^{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) = \sum_{t=n\beta+1}^{n} \frac{1}{2}\log\frac{1}{c^d} + \frac{1}{2}\boldsymbol{\varepsilon}_t^{\mathrm{T}}\Omega^{-1}\boldsymbol{\varepsilon}_t - \frac{1}{2c}(\boldsymbol{\varepsilon}_t - \theta)^T \Omega^{-1}(\boldsymbol{\varepsilon}_t - \theta). \tag{6}$$

Note that in this case we can compute the LLR for each new window recursively (for details, see the literature on CUSUM, e.g., [6]). On the other hand, in practice the true innovations after the change points can only be estimated as the recursion (5) requires initial conditions. The effect is minor if the order of the process is small (see Section 5).

The LLR test statistics obtained for approach (A) and (B) are compared with the associated threshold functions as derived in the next section.

4 Derivation of the Threshold Function

In this section we show how to obtain the threshold function as $b_{\boldsymbol{X}}(\beta)$ for the observations based or $b_{\boldsymbol{\varepsilon}}(\beta)$ for the innovations based approach. We first outline the main idea behind the derivation of the threshold function for both approaches (therefore, the subscripts $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ are omitted).

Let $\mathbb{P}_0, \mathbb{E}_0$ denote probability and expectation under H_0. When testing H_0 against $H_1(n\beta+1)$ for any fixed $\beta \in \mathscr{B}$, the probability of a type I error is given by $\mathbb{P}_0(\mathscr{L}_{n\beta}(\cdot)/n > b(\beta))$. Since we wish this probability to be *small*, it certainly holds that $b(\beta) > \mathbb{E}_0\mathscr{L}_{n\beta}(\cdot)/n$, so that we are indeed concerned with a rare event. LD theory suggests that for fixed β the false alarm probability can be approximated by

$$\mathbb{P}_0(\frac{1}{n}\mathscr{L}_{n\beta}(\cdot) > b(\beta)) \approx \exp\Big(-n\mathscr{I}(b(\beta))\Big),$$

where $\mathscr{I}$ denotes a function specified below. Recall that we wish the false alarm probability on the left hand side to be kept at a small level α. This suggests to pick the threshold function b such that it satisfies

$$\alpha = \exp\Big(-n\mathscr{I}(b(\beta))\Big) \tag{7}$$

for all $\beta \in \mathscr{B}$. This choice entails that raising a false alarm is essentially equally likely irrespective of the supposed location of the change point within the window.

Now let us make the above more rigorous. The *limiting logarithmic moment generating function* $\Lambda(\lambda)$ associated with the distribution of the LLR is defined as

$$\Lambda(\lambda) := \lim_{n\to\infty} \frac{1}{n} \log M_n(\lambda) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_0\left(e^{\lambda\mathscr{L}_{n\beta}(\cdot)}\right); \tag{8}$$

we assume for now that this function exists and is finite for every $\lambda \in \mathbb{R}$. Define $\mathscr{I}$ as the Fenchel-Legendre transform of $\Lambda(\lambda)$, that is,

$$\mathscr{I}(b(\beta)) = \sup_{\lambda\in\mathbb{R}} (\lambda b(\beta) - \Lambda(\lambda)). \tag{9}$$

Provided that Λ exists and is finite, by the Gärtner-Ellis theorem [5,9], it holds that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0(\mathscr{L}_{n\beta}(\cdot) > nb(\beta)) = -\mathscr{I}(b(\beta)).$$

In accordance with the idea expressed in (7), we choose the threshold function $b(\beta)$ such that it satisfies

$$-\mathscr{I}(b(\beta)) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0\left(\frac{1}{n}\mathscr{L}_{n\beta}(\cdot) - b(\beta) > 0\right) = -\gamma \tag{10}$$

for some positive $\gamma = -1/n \log \alpha$, across all $\beta \in \mathscr{B}$. Asymptotically, as $n \to \infty$, the probability of raising a false alarm within the window is then kept at level α.

To be able to obtain $b(\beta)$ from (10), we need to compute the limiting log-moment generating function $\Lambda(\lambda)$ in more explicit terms (this way we also check that it indeed exists and is finite for all λ).

(A) In Section 3 of [11] we outlined how to compute the moment generating function $M_n(\lambda)$ for testing $\boldsymbol{X} \sim \mathcal{N}_{nd}(\mathbf{0}, \Sigma)$ against $\boldsymbol{X} \sim \mathcal{N}_{nd}(\boldsymbol{\nu}, T)$ (for arbitrary $\boldsymbol{\nu}, \Sigma, T$):

$$M_n(\lambda) = \left(\frac{|\Sigma|}{|T|}\right)^{\lambda/2} \frac{1}{|\lambda T^{-1}\Sigma + (1-\lambda) I_{dn}|^{1/2}} \times \exp\left(-\frac{\lambda}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu} + \frac{\lambda^2}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\left(\lambda T^{-1} + (1-\lambda)\Sigma^{-1}\right)^{-1} T^{-1}\boldsymbol{\nu}\right).$$

Filling in the specific $\boldsymbol{\nu}, \Sigma, T$ for testing against a change in scale, this expression reduces to

$$M_{n\beta}(\lambda) = c^{-\lambda dn(1-\beta)/2}\left(\frac{\lambda}{c} + 1 - \lambda\right)^{-dn(1-\beta)/2} \times \exp\left(\bar{\boldsymbol{\nu}}^{\mathrm{T}} s_{n\beta}\bar{\boldsymbol{\nu}} \frac{\lambda^2 - \lambda}{2(\lambda + c - \lambda c)}\right),$$

where $s_{n\beta}$ denotes the sum of all d dimensional covariance matrices within the lower right $dn(1-\beta) \times dn(1-\beta)$ dimensional block matrix in Σ^{-1}.
Using the expression we obtained for $M_n(\lambda)$, the limiting log-moment generating function as defined in (8) becomes

$$\Lambda(\lambda) = -\frac{1}{2}\lambda d(1-\beta)\log(c) - \frac{1}{2}d(1-\beta)\log\left(\frac{\lambda}{c} + 1 - \lambda\right) + \lim_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{\nu}}^{\mathrm{T}} s_{n\beta}\bar{\boldsymbol{\nu}}\frac{\lambda^2 - \lambda}{2(\lambda + c - \lambda c)}.$$

We can evaluate the limit in the specific cases (i) $\boldsymbol{X}_t$ can be modeled as d independent ARMA processes

$$X_{it} = Z_{it} + \sum_{j=1}^{p} a_{ij} X_{i,t-j} + \sum_{j=1}^{q} b_{ij} Z_{i,t-j},$$

(i.e., the d monitored traffic streams are independent), or (ii) there is no shift in mean, i.e. $\bar{\boldsymbol{\nu}} = \mathbf{0}$. The latter may happen, for example, if the number of users stays constant while the variance of their load changes (e.g. due to application changes).

(i) In the first case, the autocovariance matrices Γ_h are diagonal, and thus the expression $\bar{\boldsymbol{\nu}}^T s_{n\beta}\bar{\boldsymbol{\nu}}$ reduces to $\sum_{i=1}^{d} \bar{\nu}_i^2 t_{i,n\beta}$, where $\bar{\nu}_i$ is the size of the mean shift of X_{it}, and $t_{i,n\beta}$ denotes the sum of the entries of the lower right $n(1-\beta) \times n(1-\beta)$-dimensional block matrix of Σ_i^{-1}, the inverse covariance matrix of X_{it}. From [11, Lemma 1] we have

$$\lim_{n\to\infty}\frac{t_{i,n\beta}}{n(1-\beta)} = \left(\frac{1 - \sum_{j=1}^{p} a_{ij}}{\sigma_i\left(1 + \sum_{j=1}^{q} b_{ij}\right)}\right)^2 =: \tau_i,$$

and hence, the limiting log-moment generating function exists and is finite. The threshold $b_X(\beta)$ can then be evaluated by putting the resulting rate function

$$\sup_{\lambda}\left\{\lambda b_X(\beta)+\frac{1}{2}(1-\beta)\left[\lambda d\log c+d\log\left(\frac{\lambda}{c}+1-\lambda\right)-\frac{\lambda^2-\lambda}{\lambda+c-\lambda c}\sum_{i=1}^{d}\bar{v}_i^2\tau_i\right]\right\}$$

equal to γ. Defining $\eta=-d\,(1-c)^2/2\sum_{i=1}^{d}\bar{v}_i^2\tau_i$, we compute the optimizing λ to be

$$\frac{c}{1-c}\left[\left(\eta+\sqrt{\eta^2+c-d+1+\frac{4c\eta}{1-c}\left(\frac{b(\beta)}{1-\beta}+\frac{1}{2}\log c\right)}\right)^{-1}-1\right]. \tag{11}$$

The threshold function $b_X(\beta)$ can be evaluated using standard numerical procedures.

(ii) If there is no shift in mean, then $M_n(\lambda)$ does not depend on $s_{n\beta}$. Hence the limiting log-moment generating function always exists, and $b_X(\beta)$ follows from

$$\gamma=\mathscr{I}(b_X(\beta))=\sup_{\lambda}\left(\lambda b_X(\beta)+\frac{1}{2}d(1-\beta)\left[\lambda\log c+\log\left(\frac{\lambda}{c}+1-\lambda\right)\right]\right).$$

The optimizing λ is

$$-\left(\frac{d(1-\beta)}{2b_X(\beta)+d(1-\beta)\log c}+\frac{c}{1-c}\right).$$

(B) When using the innovations based approach, we may make use of the fact that innovations are independent, in which case the LLR can be written as a sum of the form $\sum_{t=n\beta+1}^{n}s_t$ as given in (6). It follows that $\Lambda(\lambda)$ exists as a finite number:

$$\Lambda(\lambda)=\lim_{n\to\infty}\frac{1}{n}\log\left[\mathbb{E}_0\exp(\lambda s_1)\right]^{n(1-\beta)}=(1-\beta)\log\mathbb{E}_0\exp(\lambda s_1).$$

The threshold can be found from putting

$$\sup_{\lambda}\left[\lambda b_{\varepsilon}(\beta)+\frac{1}{2}(1-\beta)\left(\lambda d\log c+d\log\left(\frac{\lambda}{c}+1-\lambda\right)-\frac{\lambda^2-\lambda}{\lambda+c-\lambda c}\theta^T\Omega^{-1}\theta\right)\right] \tag{12}$$

equal to γ.
The optimizing λ is similar to (11) (replace η by $-d\,(1-c)^2/2\theta^T\Omega^{-1}\theta$).

As expected both approaches yield the same threshold function in case there is no shift in mean. We now know how to compute the LLR and the threshold function either using the observations or the innovations based approach. In the next section we evaluate the performance of the resulting detection procedures (1) and (2) respectively.

5 Numerical Evaluation

In this section we summarize the results of our numerical experimentation, carried out with MATLAB. We investigate the performance of detection methods (A) and (B) with respect to the false alarm rate and the detection delay, when testing vector autoregressive (VAR) processes against a change in scale.

We begin in Section 5.1 with an illustrative example which outlines how the testing methods (A) and (B) could be applied in practice. Then, in Section 5.2, we explain how the performance measures, false alarm rate and detection delay, are evaluated. Finally, in Section 5.3, we demonstrate the potential gain from using multidimensional detection procedures by comparing the multidimensional procedure to the corresponding one-dimensional procedure that tests each data stream individually.

5.1 On-line Detection

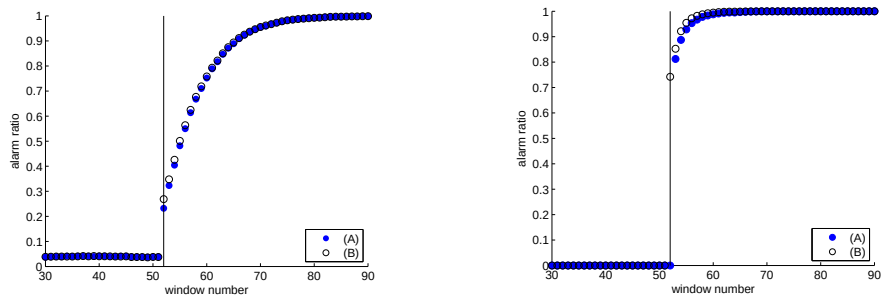
Let us first explain how to apply the detection methods set up in this paper for on-line detection of changes in scale in multidimensional Gaussian processes. We assume that one new observation arrives at a time, and the n most recent observations are being tested against a change with scaling factor c. As an illustrative example, we run the following procedure.

- We simulate a VAR(1) process of length N according to

$$\boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + \boldsymbol{Z}_t, \tag{13}$$

  where $\boldsymbol{Z}_t$ is Gaussian white noise with $\boldsymbol{Z}_t \sim \mathcal{N}(\boldsymbol{0}, \Omega)$ for $t = 1, \ldots, k-1$, $0 < k < N$, and $\boldsymbol{Z}_t \sim \mathcal{N}(\boldsymbol{\theta}, c\Omega)$ afterward.
- We consider windows of size $n < k$, adding one new observation at a time while deleting the oldest.
- In order to test whether a change in scale with scaling factor c has occurred in a particular window, we determine whether (A) criterion (1) holds true if the LLR is computed as a function of observations, or (B) criterion (2) holds true if the LLR is expressed as a function of innovations. In the latter case, the innovations are extracted as $\boldsymbol{X}_t - A\boldsymbol{X}_{t-1}$ *for all* t, and thus, the assumed independence between pre- and post-change observations is neglected. We do so to account for the fact that in practice the true value of $\boldsymbol{\varepsilon}_k$ is not known as it depends on unknown initial values.
- We repeat the above steps $15,000$ times, and divide the total number of alarms raised for each window by $15,000$ so as to obtain the alarm ratio for each window.

Two examples are presented in Fig. 1. It can be seen that the false alarm rate (the ratio of alarms before the change point as indicated by the vertical line) is indeed kept at a low level, whereas the alarm rate increases gradually to 1 after the change has occurred. It is not surprising that the detection ratio depends



(a) Mean shift $\bar{\boldsymbol{\nu}} = (0, 0, 0)^T$ (b) Mean shift $\bar{\boldsymbol{\nu}} = (2, 2, 2)^T$

Fig. 1: Alarm ratios obtained when testing a three-dimensional AR(1) sequence of observations, simulated according to (13) with diagonal coefficient matrix A with diagonal entries 0.5 and diagonal input variance matrix Ω with diagonal entries 1, against a change in scale with $c = 2$, $\alpha = 0.01$, window size $n = 50$. The first window containing the change is indicated by a vertical line.

on the position of the change point within the window – the more observations have been affected by the change, the easier the change can be detected.

The figure shows that method (B) results in a slightly higher detection rate than method (A). This may be due to the fact that in the test set-up for approach (A) we neglected the dependence between $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}$ and $\boldsymbol{X}_k, \ldots, \boldsymbol{X}_N$ under H_0.

As expected, we also see that if $\bar{\boldsymbol{\nu}} \neq \mathbf{0}$, i.e., if there is a change in the mean value also, then both false alarm rate and detection rate improve; the shift in mean is an additional indicator that a change has occurred (for a formal proof of this intuitive result, see [2, Ex. 4.1.9]). In the following, we focus on the worst-case setting $\bar{\boldsymbol{\nu}} = \mathbf{0}$ when evaluating the performance measures, false alarm ratio and the detection delay, in the next section.

5.2 Performance Measures

To evaluate the *false alarm rate*, we perform the above experiment; however, instead of shifting windows along a series of length $N > n$, we now consider a single window of observations that all correspond to H_0. Then every alarm that is raised in 15,000 runs is a false alarm, and hence, the number of change points detected on average gives an estimate for the false alarm rate. The significance level is set to $\alpha \in \{0.01, 0.05\}$, and we pick $c = 2$ (as no change is simulated, the choice of c has little impact on the test results).

In order to evaluate the *detection delay*, we simulate a VAR(1) sequence where the first 49 observations correspond to H_0 while all later observations have been affected by the change. We test windows of size 50, at each point in time adding one new observation and dropping the oldest (thus, in window i only i out of 50 observations have been affected by the change). The procedure is stopped as

soon as the change has been recognized, i.e., when the first alarm was raised. We then take the number of the first window for which this happened, averaged over $30,000$ runs (to obtain an estimate for the *average run length* under H_1, i.e. the number of decisions that have to be taken before the change is detected), and subtract one to obtain the detection delay.

The results of these experiments, where data streams are tested *jointly*, are presented in Table 2 for a number of two-dimensional examples (next to the results from testing the streams separately as explained in Section 5.3). It can be seen that – as expected – the outcome of the experiments is similar for methods (A) and (B), and the false alarm rate is generally close to the significance level α as desired. Table 2 also shows that the detection delay is small, and provides quantitative insight into the the trade-off between the false alarm rate and the detection delay: It suffices if 22% of the observations have been affected by the change when $\alpha = 0.01$ while less than 12% need to be affected when $\alpha = 0.05$.

These and similar examples suggest that the test performance is affected neither by the sign (positive or negative) nor by the magnitude of the correlation induced by Ω because the change size is relative to the size of the covariances if $\bar{\nu} = 0$. (The effect of the shift size $\bar{\nu}$ has been investigated in [11] for the case of a change in mean only.) A higher correlation via A on the other hand seems to have a positive effect on the delay – the effect of a change is enhanced due to the cross correlation.

5.3 A Case for Multidimensional Testing Procedures

In this section we demonstrate the merits of multidimensional detection procedures. In general, the signature of a change in scale is stronger when it affects $d > 1$ data streams simultaneously. In fact, in case the d tested data streams are independent, and the detection probability for each of them is p, then the detection probability when testing the d streams simultaneously is $1 - (1 - p)^d$. For example, if the detection probability for one data stream is 0.8, then the detection probability for testing three i.i.d. data streams simultaneously is 0.992. As a consequence, the multidimensional procedure outperforms a procedure that tests one of the individual data streams.

The more interesting question is whether the multidimensional procedure (testing data streams jointly) performs better than a one-dimensional approach where each of the d data streams is tested *separately* but an alarm is raised as soon as a change has been detected in *any* of the streams. In the latter case the significance level is corrected using the (conservative) Bonferroni method [7], that is, it is put to α/d for each one-dimensional testing procedure.

The main conclusion we draw from the results presented in Table 2 is that indeed the multidimensional detection procedure outperforms the method of separate testing of data streams in terms of false alarm rate and detection delay, even if the sequences are independent. However, it should be noted that this benefit comes at the cost of a longer computation time.

Furthermore, it can be seen that testing the data streams separately results in a considerably larger false alarm rate as soon as the data streams are mutually

dependent via the coefficient matrix A; due to the increased correlation, the process $\boldsymbol{X}_t$ makes larger jumps, but the separate testing does not account for this. It is surprising that the performance in terms of detection delay is good when streams are tested separately, but this may be explained by the high false alarm rate.

Cross-correlations in the covariance matrix of the innovations process on the other hand have a negative impact on the detection delay when testing the streams separately, whereas the false alarm rate remains low. This is because the fluctuations of the process $\boldsymbol{X}_t$ are of smaller magnitude if the error terms $\boldsymbol{Z}_{it}$ are cross-correlated. (In the example given in the table, $\boldsymbol{Z}_t$ is generated as $\boldsymbol{Z}_t = \Omega^{1/2}\boldsymbol{Y}_t$, where the two components of $\boldsymbol{Y}_t$ are independent standard Normals. Therefore, $\boldsymbol{Z}_{1t} = \boldsymbol{Y}_{1t}$ and $\boldsymbol{Z}_{2t} = 0.5\boldsymbol{Y}_{1t} + 0.866\boldsymbol{Y}_{2t}$. This way it can be seen that jumps of $\boldsymbol{Z}_t$ are more moderate than when there is no cross-correlation in Ω.)

6 Conclusion

In this paper we explained how to set up a testing procedure for detecting a change in scale within multidimensional serially correlated Gaussian processes, and found appropriate threshold functions. In the networking context, this type of change may occur for instance as a change in scale in correlated traffic streams due to an increase in the number of users, or due to an attack on the network.

We applied the testing procedure to (A) the sequence of observations and (B) the sequence of innovations. We listed benefits and drawbacks of each approach, and saw that both performed well in numerical experiments. We also demonstrated the supremacy of multidimensional detection procedures – compared to one-dimensional testing methods – for detecting changes that affect multiple data streams simultaneously, even if the data streams are independent.

A number of interesting questions arise. For example, can we quantify the advantage of approach (B) over (A) in terms of running time? Can we compute the threshold function in more general cases? How can we generalize the LD testing procedure, for example, to detect changes in processes that are not purely indeterministic, or to detect different types of changes, such as changes in correlation structure? We hope to address these questions in future research.

Acknowledgements. Julia Kuhn is supported by Australian Research Council (ARC) grant DP130100156. The authors thank Yoni Nazarathy for many helpful comments.

Table 2: False alarm rates and detection delays obtained from testing two-dimensional VAR(1) sequences, using (A) the observations-based approach and (B) the innovations-based approach, with $c = 2$, window size $n = 50$, mean zero. Streams are tested jointly with significance level α, and separately (ignoring interdependence) with significance level $\alpha/2$. In the latter case an alarm is raised as soon as a change point is found in any of the d streams. The standard error is given in parentheses.

Example	α	Testing	False alarm rate		Delay	
			(A)	(B)	(A)	(B)
$A = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, $\Omega = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix}$	0.01	separately	0.007 (0.0006)	0.007 (0.0006)	14.278 (0.075)	14.139 (0.075)
		jointly	0.008 (0.0007)	0.007 (0.0007)	10.510 (0.058)	10.289 (0.058)
	0.05	separately	0.031 (0.0014)	0.032 (0.0015)	7.998 (0.050)	7.818 (0.050)
		jointly	0.038 (0.0016)	0.038 (0.0016)	5.992 (0.040)	5.802 (0.040)
$A = \begin{pmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{pmatrix}$, $\Omega = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix}$	0.01	separately	0.397 (0.0040)	0.374 (0.0040)	3.264 (0.036)	3.438 (0.037)
		jointly	0.008 (0.0007)	0.007 (0.0007)	7.384 (0.055)	6.970 (0.054)
	0.05	separately	0.552 (0.0041)	0.529 (0.0041)	1.527 (0.022)	1.625 (0.023)
		jointly	0.038 (0.0016)	0.038 (0.0016)	4.105 (0.037)	3.768 (0.036)
$A = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, $\Omega = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$	0.01	separately	0.006 (0.0006)	0.006 (0.0006)	15.502 (0.082)	15.340 (0.082)
		jointly	0.008 (0.0007)	0.007 (0.0007)	10.509 (0.058)	10.289 (0.058)
	0.05	separately	0.031 (0.0014)	0.031 (0.0014)	8.782 (0.055)	8.634 (0.055)
		jointly	0.038 (0.0016)	0.038 (0.0016)	5.992 (0.040)	5.802 (0.040)
$A = \begin{pmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{pmatrix}$, $\Omega = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$	0.01	separately	0.515 (0.0041)	0.485 (0.0041)	2.674 (0.035)	2.919 (0.037)
		jointly	0.008 (0.0007)	0.007 (0.0007)	7.458 (0.055)	7.023 (0.055)
	0.05	separately	0.640 (0.0039)	0.610 (0.0040)	1.295 (0.023)	1.428 (0.022)
		jointly	0.038 (0.0016)	0.038 (0.0016)	4.146 (0.037)	3.796 (0.036)

References

1. R. ADDIE, P. MANNERSALO, and I. NORROS (2002). Most probable paths and performance formulae for buffers with Gaussian input traffic. *European Transactions on Telecommunications*, 13, pp. 183–196.
2. M. BASSEVILLE and I. NIKIFOROV (1993). *Detection of Abrupt Changes: Theory and Application* (Vol. 104). Prentice Hall, Englewood Cliffs, NJ, USA.
3. P. BROCKWELL and R. DAVIS (1987). *Time Series: Theory and Methods*. Springer, Berlin, Germany.
4. B. E. BRODSKY and B. S. DARKHOVSKY (1993). *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, The Netherlands.
5. J. BUCKLEW (1990). *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York, NY, USA.
6. C. CALLEGARI, A. COLUCCIA, A. D'ALCONZO, W. ELLENS, S. GIORDANO, M. MANDJES, M. PAGANO, T. PEPE, F. RICCIATO, and P. ŻURANIEWSKI (2013). A methodological overview on anomaly detection. In: E. Biersack, C. Callegari, and M. Matijasevic (Eds.), *Data Traffic Monitoring and Analysis*, Springer, Berlin, Germany, pp. 148–183.
7. G. CASELLA and R. L. BERGER (1990). *Statistical Inference* (Vol. 70). Duxbury Press, Belmont, CA, USA.
8. J. CHEN and A. GUPTA (2012). *Parametric Statistical Change Point Analysis: with Applications to Genetics, Medicine, and Finance*. Springer, Berlin, Germany.
9. A. DEMBO and O. ZEITOUNI (1998). *Large Deviations Techniques and Applications* (2nd Edition). Springer, New York, NY, USA.
10. J. DESHAYES and D. PICARD (1986). Off-line statistical analysis of change-point models using non parametric and likelihood methods. In: M. Thoma and A. Wyner, (Eds.), *Detection of Abrupt Changes in Signals and Dynamical Systems*. Springer, Berlin, Germany, pp. 103–168.
11. W. ELLENS, J. KUHN, M. MANDJES, and P. ŻURANIEWSKI (2013). Changepoint detection for dependent Gaussian sequences. Submitted, `arXiv:1307.0938`.
12. M. MANDJES (2007). *Large Deviations for Gaussian Queues*. John Wiley & Sons, Chichester, UK.
13. M. MANDJES and P. ŻURANIEWSKI (2011). $M/G/\infty$ transience, and its applications to overload detection. *Performance Evaluation*, 68, pp. 507–527.
14. E. PAGE (1954). Continuous inspection scheme. *Biometrika*, 41, pp. 100–115.
15. M. ROBBINS, C. GALLAGHER, R. LUND, and A. AUE (2011). Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32, pp. 498–511.
16. D. SIEGMUND (1985). *Sequential Analysis*. Springer, New York, NY, USA.
17. A. SPEROTTO, M. MANDJES, R. SADRE, P. T. DE BOER, and A. PRAS (2012). Autonomic parameter tuning of anomaly-based IDSs: an SSH case study. *IEEE Transactions on Network and Service Management*, 9, pp. 128–141.
18. A. G. TARTAKOVSKY, B. L. ROZOVSKII, R. B. BLAZEK, and H. KIM (2006). A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54, pp. 3372–3382.
19. A. G. TARTAKOVSKY and V. VEERAVALLI (2004). Change-point detection in multichannel and distributed systems. In: N. Mukhopadhyay, S. Datta, and S. Chattopadhyay, (Eds.), *Applied Sequential Methodologies: Real-World Examples with Data Analysis*. Marcel Dekker, NY, USA, pp. 339–370.
20. M. WILSON (2006). A historical view of network traffic models. Unpublished survey paper. See `http://www.arl.wustl.edu/~mlw2/classpubs/traffic_models/`