# Robust Sequential Change-Point Detection by Convex Optimization

Yang Cao and Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
{caoyang, yao.xie}@gatech.edu

Abstract—We address the computational challenge of finding the robust sequential change-point detection procedures when the pre- and post-change distributions are not completely specified. Earlier works [1], [2] establish the general conditions for robust procedures which include finding a pair of least favorable distributions (LFDs). However, in the multi-dimensional setting, it is hard to find such LFDs computationally. We present a method based on convex optimization that addresses this issue when the distributions are Gaussian with unknown parameters from prespecified uncertainty sets. We also establish theoretical properties of our robust procedures, and numerical examples demonstrate their good performance<sup>1</sup>.

## I. INTRODUCTION

Sequential detection of an abrupt change has wide applications such as statistical quality control and network security monitoring. In the classic settings, one obtains a sequence of observations of a signal of which the distribution changes at some unknown point in time, referred to as the "change-point". The goal is to detect the change as quickly as possible, subject to the false-alarm constraint. With the ever growing complexity of systems and enlarging number of sensors to monitor the systems, multi-sensor change-point detection has become a quite important subject (see, [3], [4] and [5]).

Classic sequential change-point detection assumes that the distributions before and/or after the change-point are completely specified (e.g., the classic CUSUM for one-sensor [6], [7] and Shiryaev-Roberts procedure [7], [8]). Under this setting, CUSUM is optimal (see, e.g., [9]). However, CUSUM procedure is known to be sensitive to the misspecified distributions [10].

Robust detector dates back to Huber's seminal work [11]. Subsequent follow-up work considers robust detector such as [13]. Huber considers the class of all symmetric densities that satisfies the so-called  $\epsilon$ -contamination model, with symmetric but unknown contaminations on the nominal distributions. Robust sequential change detection based on the above framework was considered in [14]. The more recent contributions [1], [2] introduce a so-called Joint Stochastic Boundedness (JSB), under which one can identify a pair of least favorable distributions (LFDs) from the uncertainty classes such that the CUSUM procedure designed for the LFDs is optimal for the robust problem in the minimax sense. However, in the multi-dimensional setting, there remains the computational challenge to establish robust sequential detection procedures or to find the LFDs.

Closed-form LFDs can only be found for a few special one-dimensional cases (e.g, [15] and [16]). Moreover, the JSB condition in [2] is defined on the real line; direct extension of JSB to multi-dimensional setting becomes quite restrictive even in very simple cases, illustrated in the following example. Consider two bivariate normal distributions. Assume that  $\Sigma$  is a positive-definite matrix in  $\mathbb{R}^{2\times 2}$ , and we would like to detect a possible transition from  $\mathcal{P}_0 = \{\mathcal{N}(0,\Sigma)\}$ , to a family of distributions  $\mathcal{P}_1 = \{\mathbb{P} \mid \mathbb{P} = \mathcal{N}(\mu_1,\Sigma), \|\mu_1 - (10,10)^T\|_2 \le 1, \mu_1 \in \mathbb{R}^2\}$ . In this case, it is impossible to find a distribution in  $\mathcal{P}_1$  that is stochastically larger than any other distribution in  $\mathcal{P}_1$  due the following Lemma 1 which satisfies the JSB condition (also see Fig. 1 for the illustration).

**Lemma 1** (Theorem 5 in [17] ). Let  $X \sim \mathcal{N}(\mu, \Sigma)$  and  $X' \sim \mathcal{N}(\mu', \Sigma')$  be n-dimensional normally distributed random vectors. Then X' is stochastically larger than X if and only if  $\mu'^{(i)} \geq \mu^{(i)}$ , for all  $1 \leq i \leq n$  and  $\Sigma = \Sigma'$ , where  $\mu^{(i)}$  denotes the ith entry of  $\mu$ .

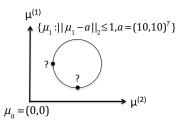


Fig. 1: It is impossible to find a point on the circle of which all the entries are larger than those of other points on the circle.

In this paper, we present a method of establishing the robust procedure by solving a convex optimization program. Given convex sets for parameters under the null and the alternative distributions, instead of identifying LFDs, we find a pair of least favorable parameters such that the Hellinger distance between the corresponding distributions from uncertainty sets are minimized. In this paper, we consider detecting the change in the mean vector and covariance matrix of a multivariate normal distribution, and hence, we may restrict our attention to linear and quadratic "detectors" (the methodology can be generalized to other sub-Gaussian distributions, see, e.g., [18]). Then a CUSUM procedure is defined for the pair of parameters solved from the optimization problem. We analyze the theoretical properties of our procedure. Note that since we use parametric models and represent uncertainty as "uncertainty sets" for the parameters, this is different from the previous work that identifies LFDs where the

<sup>&</sup>lt;sup>1</sup>Proofs to theorems can be found in the arXiv version of this paper: arXiv:1701.06952.

uncertainty class is represented by a set of probability functions. A benefit of our approach is that it leads to computationally more efficient methods.

Our approach is motivated by the recent work using convex optimization for hypothesis testing [18], [19]. The difference of these approaches from our work is that they consider finite time-horizon sequential change-point detection problem and treat it as multiple hypothesis tests. Since for each time t, there are kpossible change-point locations, for a fixed time horizon t < Tthere are a finite number of hypotheses. Then one may design a test such that the probability of error for each of the hypothesis is uniformly controlled and the total probability of error is less than a given level  $\epsilon$ . This approach is not convenient for infinite horizon setting considered in the usual sequential change-point detection problem. In this paper, we essentially proposal another approach for using the framework in [18], [19] for infinitely horizon sequential hypothesis testing, where one can determine the Average Run Length (ARL) and Expected Detection Delay (EDD) analytically.

#### II. FORMULATION

# A. General setup

Assume that we observe a sequence of observations  $\{\xi_i\}_{i=1}^{\infty}$  that take values in  $\mathcal{X}$ . Denote  $\mathcal{P}(\mathcal{X})$  as the set of all the probability distributions on  $\mathcal{X}$  and assume that there are two known distributions  $\nu_0, \nu_1 \in \mathcal{P}(\mathcal{X})$ . If there is no change, the observations are drawn i.i.d. from distribution  $\nu_0$ . The probability and expectation in this case are denoted by  $\mathbb{P}_{\infty}^{\nu_0}$  and  $\mathbb{E}_{\infty}^{\nu_0}$ , respectively. Alternatively, the i.i.d. observations  $\xi_i \sim \nu_0$  for  $i=1,\ldots,\kappa-1$ , and at some  $\mathit{unknown}$  change-point  $\kappa$ , the distributions of the observations switch abruptly to  $\nu_1$ , namely,  $\xi_i \sim \nu_1$  for  $i=\kappa,\kappa+1,\ldots$  The observations are independent conditioned on the change-point  $\kappa$ . The probability and expectation in this case are denoted by  $\mathbb{P}_{\kappa}^{\nu_0,\nu_1}$  and  $\mathbb{E}_{\kappa}^{\nu_0,\nu_1}$ , respectively. In particular,  $\kappa=0$  denotes an immediate change occurring at the initial time.

A sequential change detection procedure is characterized by a stopping time T with respect to the observation sequence. To evaluate the performance of the detection procedure T, two performance measures are widely used: the average run length (ARL) and the expected detection delay (EDD). There are three commonly used mathematical formulations about ARL and EDD: Lorden's worst-case formulation in [7], Pollak's average worst-case formulation in [20] and the Bayesian formulation in [8]. In this paper, we adopt the Lorden's formulation, where the worst-case EDD of a detection procedure T is defined as follows:

$$WDD(T) = \sup_{k \ge 1} \operatorname{esssup} \mathbb{E}_k^{\nu_0, \nu_1} \left[ (T - k + 1)^+ \mid \mathcal{F}_{k-1} \right], \quad (1)$$

where  $(x)^+ = \max(x,0)$ . The quantity in (1) is called the worst-case EDD as a result of the two supreme appearing in (1). The first supreme means that the detection delay is taken over all possible locations of the change-point k and the second essential supreme means that the detection delay is taken over all possible realizations of the observations before the change-point k. ARL can be interpreted as the mean time between two false alarms, denoted by  $\mathbb{E}_{\infty}^{\nu_0}[T]$ . In practice, one usually

fixes a lower bound  $\gamma$  for the ARL and denotes  $C(\gamma)$  as the set of stopping times with ARL larger than  $\gamma>0$ , in other words,  $C(\gamma)=\{T:\mathbb{E}_{\infty}^{\nu_0}[T]\geq\gamma\}$ . Then, our goal is to solve the following problem:

$$\min_{T \in C(\gamma)} WDD(T). \tag{2}$$

In [7] and [9], it has been proven that the cumulative sum (CUSUM) procedure [6] is both the asymptotically optimal solution as  $\gamma \to \infty$  and the exact optimal solution to (2) for any given  $\gamma > 0$ . Hence, in the following, we will focus on CUSUM-type procedures.

Now we consider the case when  $\nu_0$  and  $\nu_1$  are not specified exactly but belong to two classes of distributions  $\mathcal{P}_0, \mathcal{P}_1 \in \mathcal{P}(\mathcal{X})$ , respectively (such definitions have been considered in [2]). Denote  $C(\mathcal{P}_0, \gamma) = \{T : \mathbb{E}_{\infty}^{\nu_0}[T] \geq \gamma, \forall \nu_0 \in \mathcal{P}_0\}$  as the set of all candidate stopping times whose ARL is lower bounded by  $\gamma$ . Then our goal is to solve the following robust version of (2):

$$\min_{T \in C(\mathcal{P}_0, \gamma)} \sup_{\nu_0 \in \mathcal{P}_0, \nu_1 \in \mathcal{P}_1} WDD(T). \tag{3}$$

In the following, we specify the uncertainty set for parametric distributions, by assuming *convex uncertainty sets* for the parameters. This is a versatile formulation which finds useful in many practical situations (see, e.g., [21]).

*Mean change:* Assume that we observe a sequence of d-dimensional multivariate normal distribution with a known covariance matrix that does change. At some time  $\kappa$ , the mean vector switches from  $\mu_0, \mu_0 \in \mathcal{M}_0$  to  $\mu_1, \mu_1 \in \mathcal{M}_1$ , where  $\mathcal{M}_0$  and  $\mathcal{M}_1$  are two known convex sets in  $\mathbb{R}^d$  that are user-specified beforehand. The observations are independent conditioned on the change-point  $\kappa$ . Mathematically, we formulate the problem as the following hypothesis testing problem:

$$H_{0}: \quad \xi_{i} \sim \mathcal{N}(\mu_{0}, \Sigma), \mu_{0} \in \mathcal{M}_{0}, i = 1, 2, ...$$

$$H_{1}: \quad \xi_{i} \sim \mathcal{N}(\mu_{0}, \Sigma), \mu_{0} \in \mathcal{M}_{0}, i = 1, 2, ..., \kappa,$$

$$\xi_{i} \sim \mathcal{N}(\mu_{1}, \Sigma), \mu_{1} \in \mathcal{M}_{1}, i = \kappa + 1, \kappa + 2, ...,$$
(4)

where  $\Sigma$  is the known positive definite covariance matrix. Here, the mean vector  $\mu_0$  and  $\mu_1$  can be any element in the convex sets  $\mathcal{M}_0$  and  $\mathcal{M}_1$ , respectively. For example, in the context of quality control,  $\mathcal{M}_0$  can be defined as the set of all the allowable mean vectors if the system is in-control and  $\mathcal{M}_1$  denotes the set of all the possible mean vectors if the system is out-of-control. Our goal is to identify the occurrence of the change as fast as possible subject to the false alarm constraints.

Covariance matrix change: Similarly, we may come up with a formulation when both the mean and the covariance matrix of the observations change. Assume a sequence of d-dimensional multivariate normal observations. At some time  $\kappa$ , the mean vector changes from  $\mu_0, \mu_0 \in \mathcal{M}_0$  to  $\mu_1, \mu_1 \in \mathcal{M}_1$  and the covariance matrix changes from  $\Theta_0, \Theta_0 \in \mathcal{U}_0$  to  $\Theta_1, \Theta_1 \in \mathcal{U}_1$ , where  $\mathcal{M}_0$  and  $\mathcal{M}_1$  are two known convex sets in  $\mathbb{R}^d$ ,  $\mathcal{U}_0$  and  $\mathcal{U}_1$  are two known convex sets in  $\mathbb{S}^d_+$ , which are user-specified beforehand. We may formulate this problem as a hypothesis test similar to above.

Even if the formulation for the covariance case looks similar to the formulation (4), here the problem is much more difficult than (4). For instance, a natural approach is to use sample mean and sample covariance matrices from the in-control and out-of-control data (there usually are these training data available in certain form) as the parameters before and after the change when designing the procedures. Then the uncertainty sets represents the estimation "precision", which depend on the sample size and how the estimators are constructed. Mean vectors can usually be estimated up to good precision. However, it is much harder to estimate high-dimensional covariance matrix accurately (see, e.g, [22], [23], and [24]). Fortunately, most of the existing methods can guarantee that the true covariance matrix belongs to a convex set in  $\mathbb{S}^d_+$ , which enables us to reasonably construct uncertainty sets for covariance matrices.

## III. MAIN RESULTS

# A. Robust procedure for detecting mean change

For the robust version for mean shift detection (4), we consider a CUSUM-type procedure. CUSUM procedure needs specified likelihood ratio for two *singleton* pre-change and post-change distributions. Here, we solve a convex optimization problem to identify an appropriate pairs of parameters for the pre-change and post-change distributions, and use them to form the CUSUM procedure.

Let  $\mathcal{P}_0 = \{\mathcal{N}(\mu_0, \Sigma), \mu \in \mathcal{M}_0\}$  and  $\mathcal{P}_1 = \{\mathcal{N}(\mu_1, \Sigma), \mu \in \mathcal{M}_1\}$ . Specifically, denote  $(\mu_0^*, \mu_1^*)$  as the solution to the following convex optimization problem:

$$(\mu_0^*, \mu_1^*) = \underset{\mu_0 \in \mathcal{M}_0, \mu_1 \in \mathcal{M}_1}{\arg \min} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1).$$
 (5)

In other words,  $\mu_0^*$  and  $\mu_1^*$  are two points in  $\mathcal{M}_0$  and  $\mathcal{M}_1$  with the minimal Mahalanobis distance.

Our detection procedure is given as follows:

$$T_1 = \inf \left\{ t > 0 : \max_{1 \le k \le t} \sum_{i=k}^t \frac{1}{2} L^*(\xi_i) \ge b \right\},$$
 (6)

where  $L^*$  denotes the likelihood ratio between  $\nu_1^* \sim \mathcal{N}(\mu_1^*, \Sigma)$  and  $\nu_0^* \sim \mathcal{N}(\mu_0^*, \Sigma)$ . The threshold b is chosen such that  $\mathbb{E}_{\infty}^{\nu_0}[T_1] \geq \gamma$  for all  $\nu_0 \in \mathcal{P}_0$  and a prescribed lower bound  $\gamma$  for ARL. We can show the following relationship between  $\gamma$  and b, which offers a guideline about how to determine b given any  $\gamma$ .

**Theorem 1** (ARL). For any  $\nu_0 \in \mathcal{P}_0$ , for the detection procedure  $T_1$  defined in (6), we have that  $\mathbb{E}_{\infty}^{\nu_0}[T_1] \geq \gamma$  as long as

$$b \ge \log \gamma + \log \frac{\epsilon^*}{1 - \epsilon^*},\tag{7}$$

where

$$\epsilon^* = \exp(-\frac{1}{8}(\mu_0^* - \mu_1^*)^T \Sigma^{-1}(\mu_0^* - \mu_1^*)). \tag{8}$$

**Remark 1.** When  $\mathcal{P}_0 = \{\nu_0\}$  and  $\mathcal{P}_1 = \{\nu_1\}$  are two singletons,  $T_1$  is just the classic CUSUM procedure and the classic analysis tells us that if  $b \geq \log \gamma$  then  $\mathbb{E}_{\infty}^{\nu_0}[T_1] \geq \gamma$ . The additional second term  $\log(\epsilon^*/(1-\epsilon^*))$  in (7) can be seen as a cost for the uncertainty. Specifically,  $\epsilon^*$  is the upper bound for the Type-I and

Type-II error for the one sample composite hypothesis testing problem:  $H_0: \xi \sim \nu_0, \nu_0 \in \mathcal{P}_0$  versus  $H_1: \xi \sim \nu_1, \nu_1 \in \mathcal{P}_1$ .

Next, we prove an upper bound for the worst-case detection delay as the threshold b goes to infinity. In the following, let o(1) be a vanishing term as  $\gamma \to \infty$ .

**Theorem 2** (EDD). For any  $\nu_0 \in \mathcal{P}_0$  and  $\nu_1 \in \mathcal{P}_1$ , for the detection procedure  $T_1$  defined in (6), as  $b \to \infty$ , we have that

$$WDD(T_1) \le \frac{b}{1 - \epsilon^*} (1 + o(1)),$$

where  $\epsilon^*$  is defined in (8) and o(1) is a vanishing term as  $b \to \infty$ . Therefore, as  $\gamma \to \infty$ , we can have both  $\mathbb{E}_{\infty}^{\nu_0}[T_1] \ge \gamma$  and

$$WDD(T_1) \le \frac{\log \gamma}{1 - \epsilon^*} (1 + o(1)),$$

where  $\epsilon^*$  is defined in (8)

Remark 2. Note that  $1-\epsilon^*$  is just the Hellinger distance between the two multivariate normal distributions found by solving the convex optimization problem:  $\mathcal{N}(\mu_0^*, \Sigma)$  and  $\mathcal{N}(\mu_1^*, \Sigma)$ . When  $\mathcal{P}_0 = \{\nu_0\}$  and  $\mathcal{P}_1 = \{\nu_1\}$  are two singletons, the classic analysis tells that the  $WDD(T_1)$  is asymptotically upper bounded by 2b/I, where I is the Kullback-Leibler(KL) divergence between pre-change and post-change distributions. The Hellinger distance plays a similar role with the KL divergence as the denominator in Lorden's work [7]. Since KL divergence is known to be bounded below by Hellinger distance, our upper bound is a little bit looser. This can also be seen as the cost for uncertainty.

**Remark 3.** Define that  $\bar{\nu}_0$  and  $\bar{\nu}_1$  are true pre-change and postchange distributions. Since we can interpret the robust detection procedure  $T_1$  as a repeated one-sided sequential probability ratio test (SPRT) between  $\nu_0^* = \mathcal{N}(\mu_0^*, \Sigma)$  and  $\nu_1^* = \mathcal{N}(\mu_1^*, \Sigma)$ , we in fact can obtain that the WDD of  $T_1$  is asymptotically upper bounded by  $2b/(KL(\bar{\nu}_1||\nu_0^*) - KL(\bar{\nu}_1||\nu_1^*))$ . As stated in the seminal work [2], compared with the optimal CUSUM procedure between  $\bar{\nu}_0$  and  $\bar{\nu}_1$ ,  $WDD(T_1)$  is asymptotically larger by a factor no more than  $KL(\bar{\nu}_0 || \bar{\nu}_1)/(KL(\bar{\nu}_1 || \nu_0^*) - KL(\bar{\nu}_1 || \nu_1^*))$ . Furthermore, as a consequence of theorem 2, for any two true pre-change and post-change distributions  $\bar{\nu}_0$  and  $\bar{\nu}_1$ , we have that  $WDD(T_1)$  is asymptotically larger by a factor no more than  $KL(\bar{\nu}_0||\bar{\nu}_1)/[2(1-\epsilon^*)]$ . When the Mahalanobis distance between  $\mathcal{M}_0$  and  $\mathcal{M}_1$  increases,  $\epsilon^*$  in (8) becomes smaller and then factor above decreases, which means that our procedure moves closer to the optimal one. This is consistent with our intuition that one can detect the change more easily when the change is more obvious.

# B. Robust procedure for detecting covariance change

Next, consider the case when both the mean vector and the covariance matrix of a multivariate normal distribution change and they belong to some uncertainty sets. In this case, we may consider linear and quadratic detectors, parameterized by vector h and matrix H defined below, as suggested in [18]. We include the original derivation from [18] below.

First we define the cost function, which can be viewed as exponential loss function which relates to the type-I and type-II error in the test (in the fixed sample size scenario). Let  $\|\cdot\|$  denote the spectral norm and  $\|\cdot\|_F$  the Frobenius norm, respectively. Let  $\mathcal U$  be a convex compact set contained in the interior of the cone  $S^d_+$  of positive semidefinite  $d\times d$  matrices in the space  $S^d_+$  of symmetric  $d\times d$  matrices. Let  $\Theta_*\in S^d_+$  be such that  $\Theta_*\succeq \Theta$  for all  $\Theta\in \mathcal U$ , and let  $\delta\in [0,2]$  be such that

$$\|\Theta^{1/2}\Theta_*^{-1/2} - I_d\| \le \delta, \quad \forall \Theta \in \mathcal{U}. \tag{9}$$

Let  $\mathcal{Z}$  be a nonempty convex compact subset of the set  $\mathcal{Z}^+ = \{Z \in S^{d+1}_+ : Z_{d+1,d+1} = 1\}$ , and let

$$\phi_{\mathcal{Z}}(Y) \triangleq \max_{Z \in \mathcal{Z}} \operatorname{Tr}(ZY) \tag{10}$$

be the support function of  $\mathcal{Z}$ ; this function is used in the following definition of  $\Phi_{\mathcal{Z}}$ . These specify the closed convex set

$$\mathcal{H} = \mathcal{H}^{\beta}$$

$$:= \{ (h, H) \in \mathbb{R}^d \times \mathbb{S}^d : -\beta \Theta_*^{-1} \leq H \leq \beta \Theta_*^{-1} \},$$
(11)

and the function  $\Phi_{\mathcal{Z}}: \mathcal{H} \times \mathcal{U} \to \mathbb{R}$ ,

$$\begin{split} & \Phi_{\mathcal{Z}}(h, H; \Theta) = \\ & - \frac{1}{2} \log \text{Det}(I - \Theta_{*}^{1/2} H \Theta_{*}^{1/2}) + \frac{1}{2} \text{Tr}([\Theta - \Theta_{*}] H) \\ & + \frac{\delta(2 + \delta)}{2(1 - \|\Theta_{*}^{1/2} H \Theta_{*}^{1/2}\|)} \|\Theta_{*}^{1/2} H \Theta_{*}^{1/2}\|_{F}^{2} \\ & + \frac{1}{2} \phi_{\mathcal{Z}} \left( \left[ \left[ \frac{H}{h^{T}} \frac{h}{h} \right] + [H, h]^{T} [\Theta_{*}^{-1} - H]^{-1} [H, h] \right] \right). \end{split}$$
(12)

Then, we have that  $\Phi_{\mathcal{Z}}$  is continuous on its domain, convex in  $(h, H) \in \mathcal{H}$  and concave in  $\Theta \in \mathcal{U}$ .

Next, we specify the uncertainty sets for the pre-change and post-change multivariate normal distributions. Given two collections of data as above:  $(\mathcal{U}_{\chi}, \Theta_*^{(\chi)}, \beta_{\chi}, \mathcal{Z}_{\chi}), \chi = 0, 1$ , we define that

$$\mathcal{G}_{\chi} = \{ N(\mu, \Theta) : \Theta \in \mathcal{U}_{\chi}$$

$$\exists u : \mu = [u; 1], [u; 1][u; 1]^{T} \in \mathcal{Z}_{\chi} \}, \ \chi = 0, 1.$$
(13)

Now to solve for the quadratic detector (h, H), which will be applied on each individual samples and then used to construct the CUSUM recursion, we consider the convex-concave saddle point problem

$$SV = \min_{(h,H)\in\mathcal{H}_0\cap\mathcal{H}_1} \max_{\Theta_0\in\mathcal{U}_0,\Theta_1\in\mathcal{U}_1} \underbrace{\frac{1}{2} \left[\Phi_{\mathcal{Z}_0}(-h,-H;\Theta_0) + \Phi_{\mathcal{Z}_1}(h,H;\Theta_1)\right]}_{\Phi(h,H;\Theta_0,\Theta_1)}.$$
 (14)

A saddle point  $(H_*,h_*;\Theta_0^*,\Theta_1^*)$  in this problem does exist, which corresponds to the parameters of the quadratic detector and the picked worst-case parameters, which can be solved using a semi-definite program (SDP) solver. We obtain the following quadratic detector

$$\phi^{*}(\xi) = \frac{1}{2} \xi^{T} H_{*} \xi + h_{*}^{T} \xi + \frac{1}{2} \left[ \Phi_{\mathcal{Z}_{0}}(-h_{*}, -H_{*}; \Theta_{0}^{*}) - \Phi_{\mathcal{Z}_{1}}(h_{*}, H_{*}; \Theta_{1}^{*}) \right],$$
(15)

Given above (which is pre-solved before we have seen any data), now given a sequence of data, we may evaluate  $\phi^*$  in (15) for

each sample and define our detection procedure as follows:

$$T_2 = \inf \left\{ t > 0 : \max_{1 \le k \le t} \sum_{i=k}^t (-\phi^*(\xi_i)) \ge b \right\},$$
 (16)

where b is a prescribed threshold.

**Corollary 1** (ARL). For any  $\nu_0 \in \mathcal{G}_0$ , for the detection procedure  $T_2$  defined in (16), we have that  $\mathbb{E}_{\infty}^{\nu_0}[T_2] \geq \gamma$  as long as

$$b \ge \log \gamma + \log \frac{\epsilon^*}{1 - \epsilon^*},$$

where

$$\epsilon^* = \exp(\mathcal{SV}) \tag{17}$$

and SV is defined in (14).

**Corollary 2** (EDD). For any  $\nu_0 \in \mathcal{G}_0$  and  $\nu_1 \in \mathcal{G}_1$ , for the detection procedure  $T_2$  defined in (16), as  $b \to \infty$ , we have that

$$WDD(T_2) \le \frac{b}{1 - \epsilon^*} (1 + o(1)),$$

where  $\epsilon^*$  is defined in (17) and o(1) is a vanishing term as  $b \to \infty$ . Therefore, as  $\gamma \to \infty$ , we can have both  $\mathbb{E}_{\infty}^{\nu_0}[T_1] \ge \gamma$  and

$$WDD(T_2) \le \frac{\log \gamma}{1 - \epsilon^*} (1 + o(1)),$$

where  $\epsilon^*$  is defined in (17) and o(1) is a vanishing term as  $\gamma \to \infty$ .

### IV. NUMERICAL EXAMPLES

In this section, we compare our procedures numerically with the corresponding classic CUSUM procedure. In all the following experiments, we set the dimension d=30 and choose b's such that the ARL of  $T_1$  and  $T_{\rm CUSUM}$  are both 5000. The classic CUSUM procedure are formed using randomly chosen pre-change and post-change distributions from the uncertainty sets. In the following, we denote 1 as an all-one vector. The comparison of the numerical example is shown in Table I.

**TABLE I:** Comparison of Robust and Original CUSUM. Results are obtained from 500 Monte Carlo trials. The standard deviation is in the bracket.

	mean-shift, $\ell_1$ set	mean-shift, $\ell_2$ set
$T_1$	7.6 (2.3)	10.3 (2.9)
$T_{ m CUSUM}$	32.2 (30.1)	32.1 (31.0)
	cov-shift, set 1	cov-shift, set 2 (larger)
$T_1$	9.10 (4.21)	2.06 (0.33)
$T_{ m CUSUM}$	8.28 (5.10)	10.28 (9.22)

## A. Mean change detection

Assume  $\mathcal{M}_0=\{0\}$  and  $\Sigma=I$  in (4). In the first example, set  $\mathcal{M}_1=\{x\in\mathbb{R}^d:\|x-\mathbf{1}\|_1\leq 27\}$  in (4). We run 1000 experiments and for each run we choose a mean vector  $\mu$  whose entries are random from [0.1,0.5], then generate the post-change observations from  $\mathcal{N}(\mu,I)$ . For classic CUSUM, we specify the pre-change distribution as  $\mathcal{N}(0,I)$  and the post-change distribution as  $\mathcal{N}(1,I)$ . Then we obtain 1000 simulated detection delays of  $T_1$  and  $T_{\text{CUSUM}}$ . The mean and standard

deviation of detection delay of  $T_1$  are 7.6 and 2.3, and those of  $T_{\rm CUSUM}$  is 32.2 and 30.1, respectively. In this case,  $T_1$  performs much better than  $T_{\rm CUSUM}$  since it is difficult to choose a good post-change distribution in  $\mathcal{M}_1$  that is close to the true post-change distribution.

In the second example, the only difference between the second and the first example is that we replace the norm in  $\mathcal{M}_1$  from  $\ell_1$  to  $\ell_2$ . Set  $\mathcal{M}_1 = \{x \in \mathbb{R}^d : \|x-\mathbf{1}\|_2^2 \leq 27\}$  in (4). We run 1000 experiments, and for each run we choose a mean vector  $\mu$  whose entries are random from [0.1,0.5], then generate the post-change observations from  $\mathcal{N}(\mu,I)$ . For classic CUSUM, we specify the pre-change distribution to be  $\mathcal{N}(0,I)$ , and the post-change distribution to be  $\mathcal{N}(1,I)$ . Then we obtain 1000 simulated detection delays of  $T_1$  and  $T_{\text{CUSUM}}$ . The mean and standard deviation of detection delay of  $T_1$  is 10.3 and 2.9, and those of  $T_{\text{CUSUM}}$  is 32.1 and 31.0, respectively. In this case,  $T_1$  again performs much better than  $T_{\text{CUSUM}}$ .

# B. covariance matrix change detection

Consider  $\mathcal{M}_0 = \mathcal{M}_1 = \{0\}$  and  $\mathcal{U}_0 = \{I\}$ . In the first example, we set  $U_1 = \{I + \sigma V, \sigma \in [0.5, 1]\}$ , where V is a known matrix with diagonal entries  $V_{i,i} = 0, i = 1, ..., d$  and off-diagonal entries  $V_{i,j} = \exp(-(i-j)^2), i, j = 1, \dots, d, i \neq j$ . We run 500 experiments and for each run we randomly choose  $\sigma \in [0.5, 1]$  and then generate the post-change observations from  $\mathcal{N}(0, I + \sigma V)$ . For classic CUSUM, we specify the prechange distribution as  $\mathcal{N}(0,I)$  and the post-change distribution as  $\mathcal{N}(0, I + 0.75V)$ . Then we obtain 500 experiments for  $T_2$ and  $T_{\rm CUSUM}$ . The mean and standard deviation of detection delay of  $T_2$  is 9.10 and 4.21, and those of  $T_{\rm CUSUM}$  is 8.28and 5.10. In this case, there is no obvious difference between the two detection procedures, which means that  $T_2$  performs almost as well as classical CUSUM procedure. The reason is that the set  $\mathcal{U}_1$  is so small that the cost for mis-specified model is not large.

In the second example, consider the case with larger uncertainty sets:  $\mathcal{U}_1 = \{\Theta \in \mathbb{S}^d_+ : \|\Theta\|_2 \leq 0.5\}$ . Again, we run 500 experiments and for each run we randomly choose a  $\Sigma \in \mathcal{U}_1$  and generate the post-change observations from  $\mathcal{N}(0,\Sigma)$ . For classic CUSUM, we randomly choose a matrix in  $\mathcal{U}_1$  as the covariance matrix of its post-change normal distribution. Then, we obtain the detection delays of  $T_2$  and  $T_{\text{CUSUM}}$ . The mean and standard deviation of detection delay of  $T_2$  is 2.06 and 0.33, and those of  $T_{\text{CUSUM}}$  is 10.28 and 9.22. In this case,  $T_2$  outperforms  $T_{\text{CUSUM}}$  since  $\mathcal{U}_1$  is a large convex set and cost for a misspecified model is greater. Note that for the above two choices of  $\mathcal{U}_1$ , (14) can be solved by first removing the inner maximum since the maximum is achieved at the boundary of  $\mathcal{U}_1$ . Then solving saddle point is equivalent to solving a convex optimization.

## V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose robust detection procedures for detecting the change for mean vectors and covariance matrices, when they belong to some convex uncertainty sets. The proposed procedures are similar to classic CUSUM procedure, and the task is to determine appropriate pre-change and post-change distributions by convex optimization, which can be done efficient in both high dimensional cases. Future work includes generalizing the approach to Pollak's average worst case formulation in [20] and the Bayesian formulation in [8]. Ongoing work also includes generalizing the current framework to non-Gaussian distributions utilizing the results for sub-Gaussian distribution in [18], [19].

#### REFERENCES

- V. V. Veeravalli, T. Basar, and H. V. Poor, "Minimax robust decentralized detection," *Information Theory, IEEE Transactions on*, vol. 40, no. 1, pp. 35–40, 1994.
- [2] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, "Minimax robust quickest change detection," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1604–1614, 2011.
- [3] A. G. Tartakovsky and V. V. Veeravalli, "An efficient sequential procedure for detecting changes in multichannel and distributed systems," in *Informa*tion Fusion, 2002. Proceedings of the Fifth International Conference on, vol. 1, pp. 41–48, IEEE, 2002.
- [4] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.
- [5] Y. Xie, D. Siegmund, et al., "Sequential multi-sensor change-point detection," The Annals of Statistics, vol. 41, no. 2, pp. 670–692, 2013.
- [6] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.
- [7] G. Lorden, "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, pp. 1897–1908, 1971.
- [8] A. N. Shiryaev, "On optimum methods in quickest detection problems," Theory of Probability & Its Applications, vol. 8, no. 1, pp. 22–46, 1963.
- [9] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, pp. 1379–1387, 1986.
- [10] Z. G. Stoumbos, M. R. Reynolds Jr, T. P. Ryan, and W. H. Woodall, "The state of statistical process control as we proceed into the 21st century," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 992– 998, 2000.
- [11] P. J. Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.
- [12] S. A. Kassam and J. B. Thomas, "Asymptotically robust detection of a known signal in contaminated non-gaussian noise," *IEEE Trans. Info. Theory*, vol. IT-22, pp. 22–26, Jan 1976.
- [13] G. V. Moustakides, "Robust detection of signals: A large deviation approach," IEEE Trans. Info. Theory, vol. IT-31, pp. 822–825, Nov. 1985.
- [14] R. Crow and S. Schwartz, "On robust quickest detection procedures," in *Information Theory*, 1994. Proceedings., 1994 IEEE International Symposium on, p. 258, IEEE, 1994.
- [15] P. J. Huber, "Robust statistics. 1981."
- [16] B. C. Levy, Principles of signal detection and parameter estimation. Springer Science & Business Media, 2008.
- [17] A. Müller, "Stochastic ordering of multivariate normal distributions," Annals of the Institute of Statistical Mathematics, vol. 53, no. 3, pp. 567–575, 2001
- [18] V. Guigues, A. Juditsky, A. Nemirovski, Y. Cao, and Y. Xie, "Change detection via affine and quadratic detectors," arXiv preprint arXiv:1608.00524, 2016.
- [19] A. Goldenshluger, A. Juditsky, A. Nemirovski, et al., "Hypothesis testing by convex optimization," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 1645–1712, 2015.
- [20] M. Pollak, "Optimal detection of a change in distribution," *The Annals of Statistics*, pp. 206–227, 1985.
- [21] L. E. A. Ben-Tal and A. Nemirovski, Robust Optimization. Princeton University Press, 2009.
- [22] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.
- [23] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu, et al., "High-dimensional covariance estimation by minimizing 11-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.
- [24] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 4, pp. 603–680, 2013.
- [25] D. Siegmund, Sequential analysis: tests and confidence intervals. Springer, 1985.

#### APPENDIX A

In this appendix, we prove the main results. In the following, we denote  $\mathbb{E}_{\xi \sim \nu}[f(\xi)]$  as the expected value of  $f(\xi)$  when  $\xi$  follows some distribution  $\nu$ .

*Proof of Theorem 1.* Define that  $\phi^* \triangleq -\frac{1}{2}L^*$ . From Theorem 2.1 in [19], we have that

$$\mathbb{E}_{\xi \sim \nu_0}[\exp(-\phi^*(\xi))] \le \epsilon^*, \quad \forall \nu_0 \in \mathcal{P}_0, \tag{18}$$

$$\mathbb{E}_{\xi \sim \nu_1}[\exp(\phi^*(\xi))] \le \epsilon^*, \quad \forall \nu_1 \in \mathcal{P}_1, \tag{19}$$

where  $\epsilon^*$  is the solution to the equation

$$\mathbb{E}_{\xi \sim \nu_0^*} [\exp(-\phi^*(\xi))] = \mathbb{E}_{\xi \sim \nu_1^*} [\exp(\phi^*(\xi))],$$

or equivalently, it is defined in (8).

Define a stopping time  $T=\inf\{t>0: \sum_{i=1}^t -\phi^*(\xi_t)>b\}$ , then  $T_1$  in (6) is the same procedure as T and the arguments about T are also true for  $T_1$ . Following the definition of T, for any m>0, we have that

$$\mathbb{P}_{\infty}^{\nu_{0}}(T \leq m) \leq \mathbb{P}_{\infty}^{\nu_{0}} \left( \bigcup_{k=1}^{m} \left\{ \sum_{i=1}^{k} -\phi^{*}(\xi_{i}) > b \right\} \right) \\
\leq \sum_{k=1}^{m} \mathbb{P}_{\infty}^{\nu_{0}} \left( \sum_{i=1}^{k} -\phi^{*}(\xi_{i}) > b \right) \\
= \sum_{k=1}^{m} \mathbb{P}_{\infty}^{\nu_{0}} \left( \sum_{i=1}^{k} \left( -\phi^{*}(\xi_{i}) - \frac{b}{k} \right) > 0 \right).$$
(20)

Fix m and k, we define that  $\widetilde{\phi}^* = \phi^* + b/k$  and then we use Chernoff inequality and inequality (18) to obtain that

$$\mathbb{P}_{\xi \sim \nu}(-\widetilde{\phi}^*(\xi) > 0) \le \frac{\mathbb{E}_{\xi \sim \nu}[\exp(-\widetilde{\phi}^*(\xi))]}{1}$$

$$\le \exp(-\frac{b}{k})\epsilon^*, \ \forall \nu \in \mathcal{P}_0.$$
(21)

Under  $H_0$ ,  $\xi_i \sim \nu_0 \in \mathcal{P}_0, i=1,\ldots,m$  and  $\xi_i$ s are independent. If we apply the shifted detector  $\widetilde{\phi}^*$  on the independent variables  $\xi_1, \xi_2, \ldots, \xi_k$ , from the result for k-repeated observations (Section 2.4 in [19]), we can have that

$$\mathbb{P}_{\infty}^{\nu_0} \left( \sum_{i=1}^k \left( -\phi^*(\xi_i) - \frac{b}{k} \right) > 0 \right) \le \left( \exp\left( -\frac{b}{k} \right) \epsilon^* \right)^k.$$

Then, we have that

$$\mathbb{P}_{\infty}^{\nu_0}(T \le m) \le \sum_{k=1}^m \left(\exp\left(-\frac{b}{k}\right)\epsilon^*\right)^k$$

$$= \sum_{k=1}^m \exp\left(-b\right)\left(\epsilon^*\right)^k,$$

$$= \exp(-b) \cdot \frac{\epsilon^* - (\epsilon^*)^{m+1}}{1 - \epsilon^*}.$$
(22)

Letting m go to infinity, we have that

$$\mathbb{P}^{\nu_0}_{\infty}(T<\infty) = \exp(-b) \cdot \frac{\epsilon^*}{(1-\epsilon^*)}.$$

Applying Theorem 2 in [7], we have that

$$\mathbb{E}_{\infty}^{\nu_0}(T) \ge \frac{1}{\mathbb{P}_{\infty}^{\nu_0}(T < \infty)} = \exp(b) \cdot \frac{1 - \epsilon^*}{\epsilon^*},$$

which concludes our result.

Proof of Theorem 2. Similar with the proof for Theorem 1, we define that  $\phi^* = -\frac{1}{2}L^*$ ,  $S_t = \sum_{i=1}^t -\phi^*(\xi_t)$  and a stopping time  $T = \inf\{t > 0: S_t > b\}$ . Then T is the same as  $T_1$ . Noticing that under  $\mathbb{P}_0^{\nu_0,\nu_1}$ ,  $\xi_1,\xi_2,\ldots$  is a sequence of i.i.d random variables following some distribution  $\nu_1 \in \mathcal{P}_1$ , the well known Wald's equality (e.g, [25]) shows that

$$\mathbb{E}_0^{\nu_0,\nu_1}[T] = \frac{\mathbb{E}_0^{\nu_0,\nu_1}[S_T]}{\mathbb{E}_{\xi_1 \sim \nu_1}[-\phi_*(\xi_1)]} = \frac{b + \mathbb{E}_0^{\nu_0,\nu_1}[S_T - b]}{\mathbb{E}_{\xi_1 \sim \nu_1}[-\phi_*(\xi_1)]}$$

where  $\mathbb{E}_0^{\nu_0,\nu_1}[S_T-b]$  is the expected overshoot above the decision boundary.

Combining (19) and the fact that for any  $x \in \mathbb{R}$ ,  $-x \ge 1 - \exp(x)$ , we have that

$$\mathbb{E}_{\xi_1 \sim \nu_1}[-\phi_*(\xi_1)] \ge 1 - \mathbb{E}_{\xi_1 \sim \nu_1}[\exp(\phi_*(\xi_1))] \ge 1 - \epsilon^*.$$

To estimate the overshoot, we apply (8.18) and (8.50) in [25] to show that as  $b \to \infty$ , the following limit holds,

$$\mathbb{E}_0^{\nu_0,\nu_1}[S_T - b] \to \frac{\mathbb{E}_{\xi_1 \sim \nu_1}[\phi^*(\xi_1)^2]}{2\mathbb{E}_{\xi_1 \sim \nu_1}[\phi^*(\xi_1)]} - \sum_{n=1}^{\infty} \frac{\mathbb{E}_0^{\nu_0,\nu_1}[S_n^-]}{n},$$

where  $x^- = -\min(x, 0)$ .

By the assumption made in the statement, we have that for some M>0,  $\mathbb{E}_{\xi_1\sim\nu_1}[\phi_*^2(\xi_1)]\leq M$ . Therefore, as  $b\to\infty$ , we have that  $\mathbb{E}_0^{\nu_0,\nu_1}[S_T-b]=o(b)$ . Combing the Theorem 2 in [7], we conclude the result.

*Proof of Corollary 1 and 2.* When  $\phi^*$  is obtained from (15), from the Proposition 4.1 in [18], we have that

$$\mathbb{E}_{\xi \sim \nu_0}[\exp(-\phi^*(\xi))] \le \epsilon^*, \quad \forall \nu_0 \in \mathcal{G}_0, \tag{23}$$

$$\mathbb{E}_{\xi \sim \nu_1}[\exp(\phi^*(\xi))] \le \epsilon^*, \quad \forall \nu_1 \in \mathcal{G}_1, \tag{24}$$

where  $\epsilon^*$  is defined in (17). Then, following the same proof routine as Theorem 1 and 2, we conclude the results.

6