Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss

Cesare Alippi,[1,2] **Giacomo Boracchi,**[1] **Diego Carrera,**[1] **Manuel Roveri**[1]
[1]Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, Milano, Italy
[2]Università della Svizzera Italiana, Lugano, Switzerland
{firstname.lastname@polimi.it}

Abstract

We address the problem of detecting changes in multivariate datastreams, and we investigate the intrinsic difficulty that change-detection methods have to face when the data dimension scales. In particular, we consider a general approach where changes are detected by comparing the distribution of the log-likelihood of the datastream over different time windows. Despite the fact that this approach constitutes the frame of several change-detection methods, its effectiveness when data dimension scales has never been investigated, which is indeed the goal of our paper.

We show that the magnitude of the change can be naturally measured by the symmetric Kullback-Leibler divergence between the pre- and post-change distributions, and that the detectability of a change of a given magnitude worsens when the data dimension increases. This problem, which we refer to as *detectability loss*, is due to the linear relationship between the variance of the log-likelihood and the data dimension. We analytically derive the detectability loss on Gaussian-distributed datastreams, and empirically demonstrate that this problem holds also on real-world datasets and that can be harmful even at low data-dimensions (say, 10).

1 Introduction

Change detection, namely the problem of detecting changes in probability distribution of a process generating a datastream, has been widely investigated on scalar (i.e. univariate) data. Perhaps, the reason beyond the univariate assumption is that change-detection tests (CDTs) were originally developed for quality-control applications [Basseville *et al.*, 1993], and much fewer works address the problem of detecting changes in multivariate datastreams.

A straightforward extension to the multivariate case would be to independently inspect each component of the datastream with a scalar CDT [Tartakovsky *et al.*, 2006], but this does not clearly provide a truly multivariate solution, e.g., it is unable to detect changes affecting the correlation among the data components. A common, truly multivariate approach consists in computing the log-likelihood of the datastream and compare the distribution of the log-likelihood over different time windows (Section 2). In practice, computing the log-likelihood is an effective way to reduce the multivariate change-detection problem to a univariate one, thus easily addressable by any scalar CDT. Several CDTs for multivariate datastreams pursue this approach, and compute the log-likelihood with respect to a model fitted to a training set of stationary data: [Kuncheva, 2013] uses Gaussian mixtures, [Krempl, 2011; Dyer *et al.*, 2014] use nonparametric density models. Other CDTs have been designed upon specific multivariate statistics [Schilling, 1986; Agarwal, 2005; Lung-Yut-Fong *et al.*, 2011; Wang and Chen, 2002; Ditzler and Polikar, 2011; Nguyen *et al.*, 2014]. In the classification literature, where changes in the distribution are referred to as concept-drift [Gama *et al.*, 2014], changes are typically detected by monitoring the scalar sequence of classification errors over time [Gama *et al.*, 2004; Alippi *et al.*, 2013; Bifet and Gavalda, 2007; Ross *et al.*, 2012].

Even though this problem is of utmost relevance in datastream mining, no theoretical or experimental study investigate how the data dimension d impacts on the change detectability. In Section 3, we consider change-detection problems in $\mathbb{R}^d$ and investigate how d affects the detectability of a change when monitoring the log-likelihood of the datastream. In this respect, we show that the symmetric Kullback-Leibler divergence (sKL) between pre-change and post-change distributions is an appropriate measure of the *change magnitude*, and we introduce the *Signal-to-Noise Ratio of the change* (SNR) to quantitatively assess the change detectability when monitoring the log-likelihood.

Then, we show that the detectability of changes having a given magnitude progressively reduces when d increases. We refer to this phenomenon as *detectability loss*, and we analytically demonstrate that, in case of Gaussian random variables, the change detectability is upperbounded by a function that decays as $1/d$. We demonstrate that detectability loss occurs also in non Gaussian cases as far as data components are independent, and we show that it affects also real-world datasets, which we approximate by Gaussian mixtures in our empirical analysis (Section 4). Most importantly, detectability loss is not a consequence of density-estimation problems, as it holds either when data distribution is estimated from training samples or known. Our results indicate that detectability loss is a potentially harmful also at reasonably low-dimensions (e.g., 10) and not only in Big-Data scenarios.

2 Monitoring the Log-Likelihood

2.1 The Change Model

We assume that, in stationary conditions, the datastream $\{\mathbf{x}(t), t = 1, \dots\}$ contains independent and identically distributed (i.i.d.) random vectors $\mathbf{x}(t) \in \mathbb{R}^d$, drawn from a random variable $\mathcal{X}$ having probability-density-function (pdf) ϕ_0, that for simplicity we assume continuous, strictly positive and bounded. Here, t denotes the time instant, bold letters indicate column vectors, and $'$ is the matrix transpose operator.

For the sake of simplicity, we consider permanent changes $\phi_0 \to \phi_1$ affecting the expectation and/or correlation of $\mathcal{X}$:

$$\mathbf{x}(t) \sim \begin{cases} \phi_0 & t < \tau \\ \phi_1 & t \geq \tau \end{cases} , \text{ where } \phi_1(\mathbf{x}) = \phi_0(Q\mathbf{x} + \mathbf{v}) \, , \quad (1)$$

where τ is the unknown change point, $\mathbf{v} \in \mathbb{R}^d$ changes the location ϕ_0, and $Q \in O(d) \subset \mathbb{R}^{d \times d}$ is an orthogonal matrix that modifies the correlation among the components of $\mathbf{x}$. This rather general change-model requires a truly multivariate monitoring scheme: changes affecting only the correlation among components of $\mathbf{x}$ cannot be perceived by analyzing each component individually, or by extracting straightforward features (such as the norm) out of vectors $\mathbf{x}(t)$[1].

2.2 The Considered Change-Detection Approach

We consider the popular change-detection approach that consists in monitoring the log-likelihood of $\mathbf{x}(t)$ with respect to ϕ_0 [Kuncheva, 2013; Song *et al.*, 2007; Sullivan and Woodall, 2000]:

$$\mathcal{L}(\mathbf{x}(t)) = \log(\phi_0(\mathbf{x}(t))) \, , \; \forall t \, . \quad (2)$$

We denote by $L = \{\mathcal{L}(\mathbf{x}(t)), t = 1, \dots, \}$ the sequence of log-likelihood values, and observe that in stationary conditions, L contains i.i.d. data drawn from a scalar random variable. When $\mathcal{X}$ undergoes a change, the distribution of $\mathcal{L}(\cdot)$ is also expected to change. Thus, changes $\phi_0 \to \phi_1$ can be detected by comparing the distribution of $\mathcal{L}(\cdot)$ over W_P and W_R, two non-overlapping windows of L, where W_P refers to past data (that we assume are generated from ϕ_0), and W_R refers to most recent ones (that are possibly generated from ϕ_1). In practice, a suitable test statistic $\mathcal{T}(W_P, W_R)$, such as the t-statistic, Kolmogorov-Smirnov or Lepage [Lepage, 1971], is computed to compare W_P and W_R. In an hypothesis testing framework, this corresponds to formulating a test having as null hypothesis "*samples in W_P and W_R are from the same distribution*". When $\mathcal{T}(W_P, W_R) > h$ we can safely consider that the log-likelihood values over W_P and W_R are from two different distributions, indicating indeed a change in $\mathcal{X}$. The threshold $h > 0$ controls the test significance.

There are two important aspects to be considered about this change-detection approach. First, that comparing data on different windows is not a genuine sequential monitoring scheme. However, this mechanism is at the core of several online change-detection methods [Kuncheva, 2013; Song *et al.*, 2007; Bifet and Gavalda, 2007; Ross *et al.*, 2011]. Moreover, the power of the test $\mathcal{T}(W_P, W_R) > h$, namely the probability of rejecting the null hypothesis when the alternative holds, indicates the effectiveness of the test statistic $\mathcal{T}$ when the same is used in sequential-monitoring techniques. Second, that ϕ_0 in (2) is often unknown and has to be preliminarily estimated from a training set of stationary data. Then, ϕ_0 is simply replaced by its estimate $\widehat{\phi}_0$. In practice, it is fairly reasonable to assume a training set of stationary data is given, while it is often unrealistic to assume ϕ_1 is known, since the datastream might change unpredictably.

3 Theoretical Analysis

The section sheds light on the relationship between change detectability and d. To this purpose, we introduce: *i*) a measure of the *change magnitude*, and *ii*) an indicator that quantitatively assesses *change detectability*, namely how difficult is to detect a change when monitoring $\mathcal{L}(\cdot)$ as described in Section 2.2. Afterward, we can study the influence of d on the change detectability provided that changes $\phi_0 \to \phi_1$ have a constant magnitude.

3.1 Change Magnitude

The magnitude of $\phi_0 \to \phi_1$ can be naturally measured by the symmetric Kullback-Leibler divergence between ϕ_0 and ϕ_1 (also known as Jeffreys divergence):

$$\begin{aligned} \text{sKL}(\phi_0, \phi_1) &:= \text{KL}(\phi_0, \phi_1) + \text{KL}(\phi_1, \phi_0) \\ &= \int_{\mathbb{R}^d} \log \frac{\phi_0(\mathbf{x})}{\phi_1(\mathbf{x})} \phi_0(\mathbf{x}) d\mathbf{x} + \int_{\mathbb{R}^d} \log \frac{\phi_1(\mathbf{x})}{\phi_0(\mathbf{x})} \phi_1(\mathbf{x}) d\mathbf{x} \, . \end{aligned} \quad (3)$$

This choice is supported by the Stein's Lemma [Cover and Thomas, 2012], which states that $\text{KL}(\phi_0, \phi_1)$ yields an upper-bound for the power of parametric hypothesis tests that determine whether a given sample population is generated from ϕ_0 (null hypothesis) or ϕ_1 (alternative hypothesis). In practice, large values of $\text{sKL}(\phi_0, \phi_1)$ indicate changes that are very apparent, since hypothesis tests designed to detect either $\phi_0 \to \phi_1$ or $\phi_1 \to \phi_0$ can be very powerful.

3.2 Change Detectability

We define the following indicator to quantitatively assess the detectability of a change when monitoring $\mathcal{L}(\cdot)$.

Definition 1. *The* signal-to-noise ratio *(SNR) of the change $\phi_0 \to \phi_1$ is defined as:*

$$\text{SNR}(\phi_0 \to \phi_1) := \frac{\left(\underset{\mathbf{x} \sim \phi_0}{E}[\mathcal{L}(\mathbf{x})] - \underset{\mathbf{x} \sim \phi_1}{E}[\mathcal{L}(\mathbf{x})] \right)^2}{\underset{\mathbf{x} \sim \phi_0}{\text{var}}[\mathcal{L}(\mathbf{x})] + \underset{\mathbf{x} \sim \phi_1}{\text{var}}[\mathcal{L}(\mathbf{x})]} \, , \quad (4)$$

where $\text{var}[\cdot]$ *denotes the variance of a random variable.*

In particular, $\text{SNR}(\phi_0 \to \phi_1)$ measures the extent to which $\phi_0 \to \phi_1$ is detectable by monitoring the expectation of $\mathcal{L}(\cdot)$. In fact, the numerator of (4) corresponds to the shift introduced by $\phi_0 \to \phi_1$ in the expectation of $\mathcal{L}(\cdot)$ (i.e., the relevant information, the *signal*) which is easy/difficult to detect

[1] We do not consider changes affecting data dispersion as these can be detected by monitoring the Euclidean norm of $\mathbf{x}(t)$.

relatively to its random fluctuations (i.e., the *noise*), which are assessed in the denominator of (4). Note that, if we replace the expectations and the variances in (4) by their sample estimators, we obtain that $\text{SNR}(\phi_0 \to \phi_1)$ corresponds – up to a scaling factor – to the square statistic of a Welch's t-test [Welch, 1947], that detects changes in the expectation of two sample populations. This is another argument supporting the use of $\text{SNR}(\phi_0 \to \phi_1)$ as a measure of change detectability.

The following proposition relates the change magnitude $\text{sKL}(\phi_0, \phi_1)$ with the numerator of (4).

Proposition 1. *Let us consider a change $\phi_0 \to \phi_1$ such that*

$$\phi_1(\mathbf{x}) = \phi_0(Q\mathbf{x} + \mathbf{v}) \tag{5}$$

where $Q \in \mathbb{R}^{d\times d}$ is orthogonal and $\mathbf{v} \in \mathbb{R}^d$. Then, it holds:

$$\text{sKL}(\phi_0, \phi_1) \geq \underset{\mathbf{x}\sim\phi_0}{E}[\mathcal{L}(\mathbf{x})] - \underset{\mathbf{x}\sim\phi_1}{E}[\mathcal{L}(\mathbf{x})] \tag{6}$$

Proof. From the definition of $\text{sKL}(\phi_0, \phi_1)$ in (3) it follows

$$\begin{aligned}\text{sKL}(\phi_0, \phi_1) = & \underset{\mathbf{x}\sim\phi_0}{E}[\log(\phi_0(\mathbf{x}))] - \underset{\mathbf{x}\sim\phi_0}{E}[\log(\phi_1(\mathbf{x}))] + \\ & + \underset{\mathbf{x}\sim\phi_1}{E}[\log(\phi_1(\mathbf{x}))] - \underset{\mathbf{x}\sim\phi_1}{E}[\log(\phi_0(\mathbf{x}))] \,.\end{aligned}$$

Since $\mathcal{L}(\cdot) = \log(\phi_0(\cdot))$, (6) holds if and only if

$$\underset{\mathbf{x}\sim\phi_1}{E}[\log(\phi_1(\mathbf{x}))] \geq \underset{\mathbf{x}\sim\phi_0}{E}[\log(\phi_1(\mathbf{x}))]. \tag{7}$$

From (5) it follows that $\phi_0(\mathbf{x}) = \phi_1(Q'(\mathbf{x} - \mathbf{v}))$, thus, by replacing the mathematical expectations with their integral expressions, (7) becomes

$$\int \log(\phi_1(\mathbf{x}))\,\phi_1(\mathbf{x})d\mathbf{x} \geq \int \log(\phi_1(\mathbf{x}))\,\phi_1(Q'(\mathbf{x} - \mathbf{v}))d\mathbf{x} \tag{8}$$

Let us define $\mathbf{y} = Q'(\mathbf{x} - \mathbf{v})$, then $\mathbf{x} = Q\mathbf{y} + \mathbf{v}$ and $d\mathbf{x} = |\text{det}(Q)|d\mathbf{y} = d\mathbf{y}$, since Q is orthogonal. Using this change of variables in the second summand of (8) we obtain

$$\int \log(\phi_1(\mathbf{x}))\,\phi_1(\mathbf{x})d\mathbf{x} \geq \int \log(\phi_1(Q\mathbf{y} + \mathbf{v}))\,\phi_1(\mathbf{y})d\mathbf{y}. \tag{9}$$

Finally, defining $\phi_2(\mathbf{y}) := \phi_1(Q\mathbf{y} + \mathbf{v})$ turns (9) into

$$\int \log(\phi_1(\mathbf{x}))\,\phi_1(\mathbf{x})d\mathbf{x} - \int \log(\phi_2(\mathbf{y}))\,\phi_1(\mathbf{y}))d\mathbf{y} \geq 0, \tag{10}$$

which holds since the left-hand-side of (10) is $\text{KL}(\phi_1, \phi_2)$. □

3.3 Detectability Loss

It is now possible to investigate the intrinsic challenge of change-detection problems when data dimension increases. In particular, we study how the change detectability (i.e., $\text{SNR}(\phi_0 \to \phi_1)$) varies when d increases and changes $\phi_0 \to \phi_1$ preserve constant magnitude (i.e., $\text{sKL}(\phi_0, \phi_1) = \textit{const}$). Unfortunately, since there are no general expressions for the variance of $\mathcal{L}(\cdot)$, we have to assume a specific distribution for ϕ_0 to carry out any analytical development. As a relevant example, we consider Gaussian random variables, which enable a simple expression of $\mathcal{L}(\cdot)$. The following theorem demonstrates the *detectability loss* for Gaussian distributions, namely that $\text{SNR}(\phi_0 \to \phi_1)$ decays as d increases.

Theorem 1. *Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ be a d-dimensional Gaussian pdf and $\phi_1 = \phi_0(Q\mathbf{x} + \mathbf{v})$, where $Q \in \mathbb{R}^{d\times d}$ is orthogonal and $\mathbf{v} \in \mathbb{R}^d$. Then, it holds*

$$\text{SNR}(\phi_0 \to \phi_1) \leq \frac{C}{d} \tag{11}$$

where the constant C depends only on $\text{sKL}(\phi_0, \phi_1)$.

Proof. Basic algebra leads to the following expression for $\mathcal{L}(\mathbf{x})$ when $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$:

$$\mathcal{L}(\mathbf{x}) = -\frac{1}{2}\log\left((2\pi)^d \text{det}(\Sigma_0)\right) - \frac{1}{2}(\mathbf{x} - \mu_0)'\Sigma_0^{-1}(\mathbf{x} - \mu_0)\,. \tag{12}$$

The first term in the right-hand-side of (12) is constant, while the second term is distributed as a chi-squared having d degrees of freedom. Therefore,

$$\underset{\mathbf{x}\sim\phi_0}{\text{var}}[\mathcal{L}(\mathbf{x})] = \text{var}\left[-\frac{1}{2}\chi^2(d)\right] = \frac{d}{2}\,. \tag{13}$$

Then, from the definition of $\text{SNR}(\phi_0 \to \phi_1)$ in (4) and Proposition 1, it follows that

$$\begin{aligned}\text{SNR}(\phi_0 \to \phi_1) \leq & \frac{\text{sKL}(\phi_0, \phi_1)^2}{\underset{\mathbf{x}\sim\phi_0}{\text{var}}[\mathcal{L}(\mathbf{x})] + \underset{\mathbf{x}\sim\phi_1}{\text{var}}[\mathcal{L}(\mathbf{x})]} \leq \frac{\text{sKL}(\phi_0, \phi_1)^2}{\underset{\mathbf{x}\sim\phi_0}{\text{var}}[\mathcal{L}(\mathbf{x})]} \\ & = \frac{\text{sKL}(\phi_0, \phi_1)^2}{d/2} = \frac{C}{d}\,.\end{aligned}$$

□

Theorem 1 shows detectability loss for Gaussian distributions. In fact, when d increases and $\text{sKL}(\phi_0, \phi_1)$ remains constant, $\text{SNR}(\phi_0 \to \phi_1)$ is upper-bounded by a function that monotonically decays as $1/d$. The decaying trend of $\text{SNR}(\phi_0 \to \phi_1)$ indicates that detecting changes becomes more difficult when d increases. Moreover, the decaying rate does not depend on $\text{sKL}(\phi_0, \phi_1)$, thus this problem equally affects all possible changes $\phi_0 \to \phi_1$ defined as in (1), disregarding their magnitude.

3.4 Discussion

First of all, let us remark that Theorem 1 implicates detectability loss only when $\text{sKL}(\phi_0, \phi_1)$ is kept constant. Assuming constant change magnitude is necessary to correctly investigate the influence of the sole data dimension d on the change detectability. In fact, when the change magnitude increases with d, changes might become even easier to detect as d grows. This is what experiments in [Zimek *et al.*, 2012](Section 2.1) show, where outliers[2] become easier to detect when d increases. However, in that experiment, the change-detection problem becomes easier as d increases, since each component of $\mathbf{x}$ carries additional information about the change, thus increases $\text{sKL}(\phi_0, \phi_1)$.

Detectability loss can be also proved when ϕ_0 is non Gaussian, as far as its components are independent. In fact, if

[2] Even though similar techniques can be sometimes used for both change-detection and anomaly-detection, the two problems are intrinsically different, since the former aims at recognizing process changes, while the latter at identifying spurious data.

$\phi_0(\mathbf{x}) = \prod_{i=0}^{d} \phi_0^{(i)}(x^{(i)})$, where $(\cdot)^{(i)}$ denotes either the marginal of a pdf or the component of a vector, it follows

$$\underset{\mathbf{x}\sim\phi_0}{\text{var}}[\mathcal{L}(\mathbf{x})] = \sum_{i=0}^{d} \underset{\mathbf{x}\sim\phi_0}{\text{var}} \left[\log\left(\phi_0^{(i)}(x^{(i)})\right)\right] , \quad (14)$$

since $\log(\phi_0^{(i)}(x^{(i)}))$ are independent. Clearly, (14) increases with d, since its summands are positive. Thus, also in this case, the upperbound of $\text{SNR}(\phi_0 \to \phi_1)$ decays with d when $\text{sKL}(\phi_0, \phi_1)$ is kept constant.

Remarkably, detectability loss does not depend on how the change $\phi_0 \to \phi_1$ affects $\mathcal{X}$. Our results hold, for instance, when either $\phi_0 \to \phi_1$ affects all the components of $\mathcal{X}$ or some of them remain irrelevant for change-detection purposes. Moreover, detectability loss occurs independently of the specific change-detection method used on the log-likelihood (e.g. sequential analysis, or window comparison), as our results concern $\text{SNR}(\phi_0 \to \phi_1)$ only.

In the next section we show that detectability loss affects also real-world change-detection problems. To this purpose, we design a rigorous empirical analysis to show that the power of customary hypothesis tests actually decreases with d when data are non Gaussian and possibly dependent.

4 Empirical Analysis

Our empirical analysis has been designed to address the following goals: *i*) showing that $\text{SNR}(\phi_0 \to \phi_1)$, which is the underpinning element of our theoretical result, is a suitable measure of change detectability. In particular, we show that the power of hypothesis tests able to detect both changes in mean and in variance of $\mathcal{L}(\cdot)$ also decays. *ii*) Showing that detectability loss is not due to density-estimation problems, but it becomes a more serious issue when ϕ_0 is estimated from training data. *iii*) Showing that detectability loss occurs also in Gaussian mixtures, and *iv*) showing that detectability loss occurs also on high-dimensional real-world datasets, which are far from being Gaussian or having independent components. We address the first two points in Section 4.1, while the third and fourth ones in Sections 4.2 and 4.3, respectively.

In our experiments, the change-detection performance is assessed by numerically computing the power of two customary hypothesis tests, namely the Lepage [Lepage, 1971] and the one-sided t-test[3] on data windows W_P and W_R which contains 500 data each. As we discussed in Section 3.2, the t-statistic on the log-likelihood is closely related to $\text{SNR}(\phi_0 \to \phi_1)$, while the Lepage is a nonparametric statistic that detects both location and scale changes[4]. To compute the power, we set h to guarantee a significance level[5] $\alpha = 0.05$. Following the procedure in Appendix, we synthetically introduce changes $\phi_0 \to \phi_1$ having $\text{sKL}(\phi_0, \phi_1) = 1$ which, in the univariate Gaussian case, corresponds to $\mathbf{v}$ equals to the standard deviation of ϕ_0.

4.1 Gaussian Datastreams

We generate Gaussian datastreams having dimension $d \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ and, for each value of d, we prepare 10000 runs, with $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $\phi_1 = \mathcal{N}(\mu_1, \Sigma_1)$. The parameters $\mu_0 \in \mathbb{R}^d$ and $\Sigma_0 \in \mathbb{R}^{d\times d}$ have been randomly generated, while $\mu_1 \in \mathbb{R}^d$ and $\Sigma_1 \in \mathbb{R}^{d\times d}$ have been set to yield $\text{sKL}(\phi_0, \phi_1) = 1$ (see Appendix). In each run we generate 1000 samples: $\{\mathbf{x}(t),\ t = 1, \dots, 500\}$ from ϕ_0, and $\{\mathbf{x}(t),\ t = 501, \dots, 1000\}$ from ϕ_1. Then, we compute the datastream $L = \{\mathcal{L}(\mathbf{x}(t)),\ t = 1, \dots, 1000\}$, and define $W_P = \{\mathcal{L}(\mathbf{x}(t)), t = 1, \dots, 500\}$ and $W_R = \{\mathcal{L}(\mathbf{x}(t)), t = 501, \dots, 1000\}$.

We repeat the same experiment replacing ϕ_0 with its estimate $\widehat{\phi}_0(\mathbf{x})$, where $\widehat{\mu}_0$ and $\widehat{\Sigma}_0$ are computed using the sample estimators over an additional training set TR whose size grows linearly with d, i.e. $\#TR = 100 \cdot d$. We denote by $\widehat{L} = \{\widehat{\mathcal{L}}(\mathbf{x}(t)), t = 1, \dots, 1000\}$ the sequence of estimated log-likelihood values. Finally, we repeat the whole experiments keeping $\#TR = 100$ for any value of d, and we denote by $\widehat{L}_{100}$ the corresponding sequence of log-likelihood values.

Figure 1(a) shows that the power of both the Lepage and one-sided t-test substantially decrease when d increases. This result is coherent with our theoretical analysis of Section 3, and confirms that $\text{SNR}(\phi_0 \to \phi_1)$ is a suitable measure of change detectability. While it is not surprising that the power of the t-test decays, given its connection with the $\text{SNR}(\phi_0 \to \phi_1)$, it is remarkable that the power of the Lepage test also decays, as this fact indicates that it becomes more difficult to detect both changes in the mean and in the dispersion of L. The decaying power of both tests indicates that the corresponding test statistics decrease with d, which imply larger detection delays when using this statistics in sequential monitoring schemes.

Note that detectability loss is not due to density-estimation issues, but rather to the fact that the change-detection problem becomes intrinsically more challenging, as it occurs in the ideal case where ϕ_0 is known (solid lines). When $\mathcal{L}$ is computed from an estimated $\widehat{\phi}_0$ (dashed and dotted lines), the problem becomes even more severe, and worsens when the number of training data does not grow with d (dotted lines).

4.2 Gaussian mixtures

We now consider ϕ_0 and ϕ_1 as Gaussian mixtures, to prove that detectability loss occurs also when datastreams are generated/approximated by more general distribution models. Mimicking the proof of Theorem 1, we show that when d increases and $\text{sKL}(\phi_0, \phi_1)$ is kept constant, the upper-bound of $\text{SNR}(\phi_0 \to \phi_1)$ decreases. To this purpose, it is enough to show that $\underset{\mathbf{x}\sim\phi_0}{\text{var}}[\mathcal{L}(\mathbf{x})]$ increases with d.

The pdf of a mixture of k Gaussians is

$$\phi_0(\mathbf{x}) = \sum_{i=1}^{k} \lambda_{0,i} \mathcal{N}(\mu_{0,i}, \Sigma_{0,i})(\mathbf{x}) = \\ = \sum_{i=1}^{k} \frac{\lambda_{0,i}}{(2\pi)^{d/2}\text{det}(\Sigma_{0,i})^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{0,i})'\Sigma_{0,i}^{-1}(\mathbf{x}-\mu_{0,i})}, \quad (15)$$

[3] We can assume that $\phi_0 \to \phi_1$ decreases the expectation of $\mathcal{L}$ since $\underset{\mathbf{x}\sim\phi_0}{E}[\log(\phi_0(\mathbf{x}))] - \underset{\mathbf{x}\sim\phi_1}{E}[\log(\phi_0(\mathbf{x}))] \geq 0$ follows from (7).

[4] The Lepage statistic is defined as the sum of the squares of the Mann-Whitney and Mood statistics, see also [Ross *et al.*, 2011].

[5] The value of h for the Lepage test is given by the asymptotic approximation of the statistic in [Lepage, 1971].

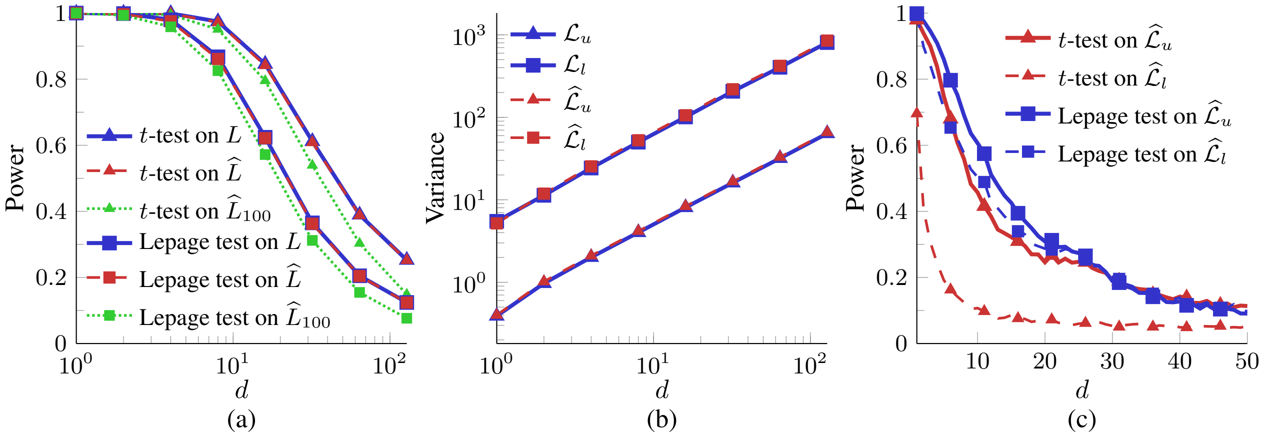

Figure 1: (a) Power of the Lepage and one-sided t-test empirically computed on sequences generated as in Section 4.1. Detectability loss clearly emerges when the log-likelihood is computed using ϕ_0 (denoted by L) or its estimates fitted on $100 \cdot d$ samples ($\widehat{L}$) or from 100 samples ($\widehat{L}_{100}$). (b) The sample variance of $\mathcal{L}_u(\cdot)$ (16) and $\mathcal{L}_l(\cdot)$ (17) computed as in Section 4.2. As in the Gaussian case, both these variances grow linearly with d and similar results hold when using $\widehat{\phi}_0$, which is estimate from $200 \cdot d$ training data. (c) The power of both Lepage and one-sided t-test indicate detectability loss on the Particle dataset, which is approximated by a mixture of 2 Gaussians, using both $\widehat{\mathcal{L}}_u$ (16) and $\widehat{\mathcal{L}}_l$ (17). Using $\widehat{\mathcal{L}}_u$ guarantees better performance than $\widehat{\mathcal{L}}_l$ because this latter yields a larger variance, as shown in (b). We achieve similar results on the Wine dataset, that was approximated by a mixture of 4 Gaussians. Results have not been reported due to space limitations.

where $\lambda_{0,i} > 0$ is the weight of the i-th Gaussian $\mathcal{N}(\mu_{0,i}, \Sigma_{0,i})$. Unfortunately, the log-likelihood $\mathcal{L}(\mathbf{x})$ of a Gaussian mixture does not admit an expression similar to (12) and two approximations are typically used to avoid severe numerical issues when $d \gg 1$.

The first approximation consists in considering only the Gaussian of the mixture yielding the largest likelihood, as in [Kuncheva, 2013] i.e.,

$$\mathcal{L}_u(\mathbf{x}) = -\frac{k\lambda_{0,i^*}}{2}\Big(\log\left((2\pi)^d \det(\Sigma_{0,i^*})\right) + \\ + (\mathbf{x}-\mu_{0,i^*})'\Sigma_{0,i^*}^{-1}(\mathbf{x}-\mu_{0,i^*})\Big) \tag{16}$$

where i^* is defined as

$$i^* = \underset{i=1,\dots,k}{\operatorname{argmax}} \left(\frac{\lambda_{0,i}}{\det(\Sigma_{0,i})^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{0,i})'\Sigma_{0,i}^{-1}(\mathbf{x}-\mu_{0,i})} \right).$$

The second approximation we consider is:

$$\mathcal{L}_l(\mathbf{x}) = -\frac{1}{2}\sum_{i=1}^{k} \lambda_{0,i}\Big(\log\left((2\pi)^d \det(\Sigma_{0,i})\right) + \\ + (\mathbf{x}-\mu_{0,i})'\Sigma_{0,i}^{-1}(\mathbf{x}-\mu_{0,i})\Big), \tag{17}$$

that is a lower bound of $\mathcal{L}(\cdot)$ due to the Jensen inequality.

We consider the same values of d as in Section 4.1 and, for each of these, we generate 1000 datastreams of 500 data drawn from a Gaussian mixture ϕ_0. We set $k = 2$ and $\lambda_{0,1} = \lambda_{0,2} = 0.5$, while the parameters $\mu_{0,1}$, $\mu_{0,2}$, $\Sigma_{0,1}$, $\Sigma_{0,2}$ are randomly generated. We then compute the sample variance of both $\mathcal{L}_u$ and $\mathcal{L}_l$ over each datastream and report their average in Figure 1(b). As in Section 4.1, we repeat this experiment by preliminarily estimating $\widehat{\phi}_0$ from a training set containing $200 \cdot d$ additional samples, then we compute $\widehat{\mathcal{L}}_u$ and $\widehat{\mathcal{L}}_l$.

Figure 1(b) shows that the variances of $\mathcal{L}_u$ and $\mathcal{L}_l$ grow linearly with respect to d, as in the Gaussian case (13). This indicates that detectability loss occurs also when $\mathcal{X}$ is generated by a simple bimodal distribution and, most importantly, also when using $\mathcal{L}_u$ or $\mathcal{L}_l$ that are traditionally adopted when fitting Gaussian mixtures. As in Section 4.1, we see that log-likelihoods $\widehat{\mathcal{L}}_u$ and $\widehat{\mathcal{L}}_l$ computed with respect to fitted models follow the same trend. We further observe that $\mathcal{L}_l$ exhibits a much larger variance than $\mathcal{L}_u$, thus we expect this to achieve lower change-detection performance than $\mathcal{L}_u$.

4.3 Real-World Data

To investigate detectability loss in real-world datasets, we design a change-detection problem on the *Wine Quality Dataset* [Cortez *et al.*, 2009] and the *MiniBooNE Particle Dataset* [Roe *et al.*, 2005] from the UCI repository [Lichman, 2013]. The Wine dataset has 12 dimensions: 11 corresponding to numerical results of laboratory analysis (such as density, Ph, residual sugar), and one corresponding to a final grade (from 0 to 10) for each different wine. We consider the vectors of laboratory analysis of all white wines having a grade above 6, resulting in a 11-dimensional dataset containing 3258 data. The Particle dataset contains numerical measurements from a physical experiment designed to distinguish electron from muon neutrinos. Each sample has 50-dimensions and we considered only data from muon class, yielding 93108 data.

In either datasets we have to estimate ϕ_0 for both introducing changes having constant magnitude and computing the log-likelihood. We adopt Gaussian mixtures and estimate k by 5-fold cross validation over the whole datasets, obtaining

$k = 4$ and $k = 2$ for Wine and Particle dataset, respectively.

We process each dataset as follows. Let us denote by D the dataset dimension and for each value of $d = 1, \dots, D$ we consider only d components of our dataset that are randomly selected. We then generate a d-dimensional training set of $200 \cdot d$ samples and a test set of 1000 samples (datastream), which are extracted by a bootstrap procedure without replacement. The second half of the datastream is perturbed by the change $\widetilde{\phi}_0 \to \widetilde{\phi}_1$, which is defined by fitting at first $\widetilde{\phi}_0$ on the whole d-dimensional dataset, and then computing $\widetilde{\phi}_1$ according to the procedure in Appendix. Then, we estimate $\widehat{\phi}_0$ from the training set and we compute $\mathcal{T}(\widehat{W}_P, \widehat{W}_R)$, where $\widehat{W}_P$, $\widehat{W}_R$ are defined as in Section 4.1. This procedure is repeated 5000 times to estimate the test power numerically. Note that the number of Gaussians in both $\widetilde{\phi}_0$ and $\widehat{\phi}_0$ is the value of k estimated from the whole D-dimensional dataset, and that $\widetilde{\phi}_0$ is by no means used for change-detection purposes.

Figure 1(c) reports the power of both Lepage and one-sided t-tests on the Particle dataset, considering $\widehat{\mathcal{L}}_u$ (16) and $\widehat{\mathcal{L}}_l$ (17) as approximated expressions of the likelihoods. The power of both tests is monotonically decreasing, indicating an increasing difficulty in detecting a change among $\widehat{W}_P$ and $\widehat{W}_R$ when d grows. This result is in agreement with the claim of Theorem 1 and the results in the previous sections. The Lepage test here turns to be more powerful than the t-test and this indicates that it is important to monitor also the dispersion of $\mathcal{L}(\cdot)$ when using Gaussian mixtures, where $\mathcal{L}(\cdot)$ can be multimodal. Moreover, the decaying power of the Lepage test indicates that, as in Section 4.1, monitoring both expectation and dispersion of $\mathcal{L}(\cdot)$ does not prevent the detectability loss. Figure 1(c) indicates that $\widehat{\mathcal{L}}_u(\cdot)$ guarantees superior performance than $\widehat{\mathcal{L}}_l(\cdot)$ since this has lower variance than $\widehat{\mathcal{L}}_u(\cdot)$. This fact also underlines the importance of considering the variance of $\mathcal{L}(\cdot)$ in measures of change detectability, as in (4). Experiments on Wine dataset, which is approximated by a more sophisticated distribution ($k = 4$), confirms detectability loss, but the results have not been reported due to space limitations.

We finally remark that have set a change magnitude ($\text{sKL}(\phi_0, \phi_1) = 1$) that is quite customary in change-detection experiments, as in the univariate Gaussian case this corresponds to setting $\mathbf{v}$ equals to the standard deviation of ϕ_0. Therefore, since in our experiments the power of both tests is almost halved when $d \approx 10$, we can conclude that detectability loss is not only a Big-Data issue.

5 Conclusions

We provide the first rigorous study of the challenges that change-detection methods have to face when data dimension scales. Our theoretical and empirical analyses reveal that the popular approach of monitoring the log-likelihood of a multivariate datastream suffers detectability loss when data dimension increases. Remarkably, detectability loss is not a consequence of density-estimation errors – even though these further reduce detectability – but it rather refers to an intrinsic limitation of this change-detection approach. Our theoretical results demonstrate that detectability loss occurs independently on the specific statistical tool used to monitor the log-likelihood and does not depend on the number of input components affected by the change. Our empirical analysis, which is rigorously performed by keeping the change-magnitude constant when scaling data-dimension, confirms detectability loss also on real-world datastreams. Ongoing works concern extending this study to other change-detection approaches and to other families of distributions.

Appendix: Generating Changes of Constant Magnitude

Here we describe a procedure to select, given ϕ_0, an orthogonal matrix $Q \in \mathbb{R}^{d \times d}$ and a vector $\mathbf{v} \in \mathbb{R}^d$ such that $\phi_1 = \phi_0(Q\mathbf{x} + \mathbf{v})$ guarantees $\text{sKL}(\phi_0, \phi_1) = 1$ in the case of Gaussian pdfs, and $\text{sKL}(\phi_0, \phi_1) \approx 1$ for arbitrary distributions. Extensions to different values of $\text{sKL}(\phi_0, \phi_1)$ are straightforward. Since $\phi_1(\mathbf{x}) = \phi_0(Q\mathbf{x} + \mathbf{v})$, we formulate the problem as generating at first a rotation matrix Q such that

$$0 < \text{sKL}(\phi_0(\cdot), \phi_0(Q\cdot)) < 1$$

and then defining the translation vector $\mathbf{v}$ to adjust ϕ_1 such that $\text{sKL}(\phi_0, \phi_1)$ reaches (or approaches) 1.

We proceed as follows: we randomly define a rotation axis $\mathbf{r}$, and a sequence of rotations matrices $\{Q_j\}_j$ around $\mathbf{r}$, where the rotation angles monotonically decrease toward 0 (thus Q_j tends to the identity matrix as $j \to \infty$). Then, we set Q_{j^*} as the largest rotation yielding a sKL < 1, namely

$$j^* = \min\{j \,:\, \text{sKL}(\phi_0(\cdot), \phi_0(Q_j\cdot)) < 1\}. \tag{18}$$

When ϕ_0 is continuous and bounded (as in case of Gaussian mixtures) it can be easily proved that such a j^* exists.

In the case of Gaussian pdfs, when $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$, $\text{sKL}(\phi_0, \phi_1)$ admits a closed-form expression:

$$\begin{aligned}\text{sKL}(\phi_0, \phi_1) = \frac{1}{2}\Bigg[&\mathbf{v}'\Sigma_0^{-1}\mathbf{v} + \mathbf{v}'Q\Sigma_0^{-1}Q'\mathbf{v} + \\ &+ 2\mathbf{v}'\Sigma_0^{-1}(I - Q)\mu_0 + 2\mathbf{v}'Q\Sigma_0^{-1}(Q' - I)\mu_0 + \\ &+ \text{Tr}(Q'\Sigma_0^{-1}Q\Sigma_0) + \text{Tr}(\Sigma_0^{-1}Q'\Sigma_0 Q) - 2d + \\ &+ 2\mu_0'(I - Q')\Sigma_0^{-1}(I - Q)\mu_0\Bigg],\end{aligned} \tag{19}$$

and $\text{sKL}(\phi_0(\cdot), \phi_0(Q_j\cdot))$ can be exactly computed to solve (18). When there are no similar expressions for $\text{sKL}(\phi_0, \phi_1)$ this has to be computed via Monte Carlo simulations.

After having set the rotation matrix Q, we randomly generate a unit-vector $\mathbf{u}$ as in [Alippi, 2014] and determine a suitable translation along the line $\mathbf{v} = \rho\mathbf{u}$, where $\rho > 0$, to achieve $\text{sKL}(\phi_0, \phi_1) = 1$. Again, the closed-form expression (19) allows to directly compute the exact value of ρ by substituting $\mathbf{v} = \rho\mathbf{u}$ into (19). This yields a quadratic equation in ρ, whose positive solution ρ^* provides $\mathbf{v} = \rho^*\mathbf{u}$ that leads to $\text{sKL}(\phi_0, \phi_1) = 1$. When the are no analytical expressions for $\text{sKL}(\phi_0, \phi_1)$, we generate an increasing sequence $\{\rho_n\}_n$ such that $\rho_0 = 0$ and $\rho_n \to \infty$ as $n \to \infty$, and set

$$n^* = \max\{n \,:\, \text{sKL}(\phi_0(\cdot), \phi_0(Q \cdot + \rho_n\mathbf{u})) < 1\}, \tag{20}$$

where $\mathrm{sKL}(\phi_0(\cdot), \phi_0(Q \cdot + \rho_n \mathbf{u}))$ is computed by Monte Carlo simulations. After having solved (20), we determine ρ^* via linear interpolation of $[\rho_{n^*}, \rho_{n^*+1}]$ on the corresponding values of the sKL. In this case, we can only guarantee $\mathrm{sKL}(\phi_0, \phi_1) \approx 1$ with an accuracy that can be improved by increasing the resolution of $\{\rho_n\}_n$.

References

[Agarwal, 2005] Deepak Agarwal. An empirical bayes approach to detect anomalies in dynamic multidimensional arrays. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, 2005.

[Alippi *et al.*, 2013] Cesare Alippi, Giacomo Boracchi, and Manuel Roveri. Just-in-time classifiers for recurrent concepts. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4), 2013.

[Alippi, 2014] Cesare Alippi. *Intelligence for Embedded Systems, a Methodological Approach*. Springer, 2014.

[Basseville *et al.*, 1993] Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*. Prentice Hall Englewood Cliffs, 1993.

[Bifet and Gavalda, 2007] Albert Bifet and Ricard Gavalda. Learning from time-changing data with adaptive windowing. In *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2007.

[Cortez *et al.*, 2009] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 2009.

[Cover and Thomas, 2012] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[Ditzler and Polikar, 2011] Gregory Ditzler and Robi Polikar. Hellinger distance based drift detection for nonstationary environments. In *Proceedings of IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments (CIDUE)*, 2011.

[Dyer *et al.*, 2014] Karl B Dyer, Robert Capo, and Robi Polikar. Compose: A semisupervised learning framework for initially labeled nonstationary streaming data. *IEEE Transactions on Neural Networks and Learning Systems*, 25(1), 2014.

[Gama *et al.*, 2004] Joao Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with drift detection. In *Proceedings of Brazilian Symposium on Artificial Intelligence (SBIA)*, 2004.

[Gama *et al.*, 2014] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46(4), 2014.

[Krempl, 2011] Georg Krempl. The algorithm APT to classify in concurrence of latency and drift. In *Advances in Intelligent Data Analysis X (IDA)*, pages 222–233, 2011.

[Kuncheva, 2013] Ludmila I Kuncheva. Change detection in streaming multivariate data using likelihood detectors. *IEEE Transactions on Knowledge and Data Engineering*, 25(5), 2013.

[Lepage, 1971] Yves Lepage. A combination of wilcoxon's and ansari-bradley's statistics. *Biometrika*, 58(1), 1971.

[Lichman, 2013] M. Lichman. UCI machine learning repository, 2013.

[Lung-Yut-Fong *et al.*, 2011] Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, and Olivier Cappé. Robust changepoint detection based on multivariate rank statistics. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[Nguyen *et al.*, 2014] Tuan Duong Nguyen, Marthinus Christoffel Du Plessis, Takafumi Kanamori, and Masashi Sugiyama. Constrained least-squares density-difference estimation. *IEICE Transactions on Information and Systems*, 97(7), 2014.

[Roe *et al.*, 2005] Byron P Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2), 2005.

[Ross *et al.*, 2011] Gordon J Ross, Dimitris K Tasoulis, and Niall M Adams. Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53(4), 2011.

[Ross *et al.*, 2012] Gordon J. Ross, Niall M. Adams, Dimitris K. Tasoulis, and David J. Hand. Exponentially weighted moving average charts for detecting concept drift. *Pattern Recognition Letters*, 33(2), 2012.

[Schilling, 1986] Mark F Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395), 1986.

[Song *et al.*, 2007] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. Statistical change detection for multi-dimensional data. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[Sullivan and Woodall, 2000] Joe H Sullivan and William H Woodall. Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations. *IIE Transactions*, 32(6), 2000.

[Tartakovsky *et al.*, 2006] Alexander G Tartakovsky, Boris L Rozovskii, Rudolf B Blazek, and Hongjoong Kim. 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(9), 2006.

[Wang and Chen, 2002] Tai-Yue Wang and Long-Hui Chen. Mean shifts detection and classification in multivariate process: A neural-fuzzy approach. *Journal of Intelligent Manufacturing*, 13(3), 2002.

[Welch, 1947] Bernard L Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2), 1947.

[Zimek *et al.*, 2012] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5), 2012.