Abstract

We investigate the problem of distributed monitoring of large-scale data streams where an undesired event may occur at some unknown time and affect only a few unknown data streams. We propose to develop scalable global monitoring schemes by parallel running local detection procedures and by combining these local procedures together to make a global decision based on SUM-shrinkage techniques.

Problem Formulation and Existing Research

Problem formulation

Online Monitoring independent large-scale data streams:

Data Stream 1: \( X_1, X_1, \ldots \)

Data Stream 2: \( X_2, X_2, \ldots \)

Data Stream \( K: X_K, X_K, \ldots \)

At some unknown time \( \nu \), an event occurs and affects a few unknown data streams in the sense of changing the distributions of \( X_k \)'s from \( \mathcal{N}(0, 1) \) to \( \mathcal{N}(\mu, 1) \), while \( \mu_k \) may or may not be known.

Objective: Detect the true change time \( \nu \) as soon as possible. Mathematically, find stopping time \( T \) to minimize the worst-case detection delay proposed by Lorden (1971):

\[
E_{\mathcal{H}_0} \max_{0 \leq t \leq \infty} E^T_f(t) = \sup_{0 \leq t < \infty} \mathbb{E}^T_f(t),
\]

subject to the global false alarm constraint:

\[
E^T_f(T) = \gamma.
\]

Applications: Industrial quality control, signal detection, bio-surveillance (CDC BioSense) etc.

Challenges:

- **Time domain**: Repeatedly test hypothesis of \( H_0 : \nu = \infty \) (no change) against \( H_1 : \nu = 1, 2, \ldots \), (a change occurs) at each and every time step \( n \) when new data arrives.

- **Spatial domain**: We do not know which subset of data streams is affected, and the post-change parameter \( \mu_k \) might also be unknown. If \( r \) out of \( K \) data streams are affected, there are \( \binom{K}{r} \) possible combinations.

Existing Research

- Tartakovsky and Veeravalli (2008) and Mei (2010): Assume the post-change parameter \( \mu_k \) is completely specified if affected.

- Xie and Siegmund (2013) proposed a semi-Bayesian scheme by assuming the proportions:

\[
\tilde{T}_{\mathcal{G}}(n, \mu_0) = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq K} \sum_{i=1}^{K} \log(1 - p_i) \right\},
\]

\[
p_i = \exp \left\{ \max \left\{ 0, 1 - \sum_{j=1}^{K} \frac{X_{ij}}{\phi} \right\} \right\},
\]

where \( p_i \) is the fraction of affected data streams.

Our Proposed Methodology

We propose two main components:

- **Local detection statistics \( W_k,n \)'s** that can efficiently detect local change at \( K \) local sensor up to time \( n \).

- **A SUM-shrinkage method** that combines the local detection statistics \( W_k,n \)'s suitably.

Let us postpone the discussion of \( W_k,n \)'s and focus on the SUM-shrinkage method first, which is motivated by parallel computing in the censoring sensor networks.

Non-Homogeneous Sensors with Known Post-Change Distributions

The \( W_k,n \)'s are chosen as the well-known local CUSUM statistics (Page 1954):

\[
W_k,n = \max \left\{ W_k,n-1 + \Delta_k, 0 \right\},
\]

for \( n \geq 1 \) and \( W_0 = 0 \) for \( k = 1, \ldots, K \).

The choice of \( \Delta_k \)'s: If sensors are homogeneous, a “good” choice is \( \Delta_k = \rho \Delta \), for \( k = 1, \ldots, K \) and constant \( \Delta \geq 0 \), where \( \Delta = \sum_{k=1}^{K} \log \Delta_k \).

For the two-sided test in multi-dimensional case: The local detection statistics for \( r \)th data stream is:

\[
W_r^{(r)} = \max \left\{ W_r^{(r)}, W_r^{(r)} \right\},
\]

where \( W_r^{(r)} \) and \( W_r^{(r)} \) are applied to detect positive and negative mean shifts, respectively. The estimate of \( \mu_k \) is defined in the similar form as in (6). The three SUM-shrinkage methods can then be applied to combine this new local detection statistics \( W_k,n \)'s together.

Reference:

**Theorem 1.** For any given post-change hypothesis set \( \{\delta_1, \ldots, \delta_K\} \subseteq \mathbb{R} \), subject to the false alarm constraint (1), as \( \gamma \) goes to \( \infty \), the hard thresholding scheme \( \mathcal{N}_{W_0}(\nu, \delta) \) asymptotically minimizes \( \mathcal{E}_{\mathcal{H}_0}(\nu, \delta) \) (up to the first-order). The conclusion also holds for the soft thresholding scheme \( \mathcal{N}_{W_0}(\nu, \delta) \) and the combined thresholding scheme \( \mathcal{N}_{\text{combined}}(\nu) \) when the occurring event affects at most \( r \) data streams.

Homogeneous Sensors with Unknown Post-Change Distributions

Challenge: Determine the local detection statistics \( W_k,n \)'s properly when the post-change mean \( \mu_k \) is unknown.


A technical assumption: \( \mu_k \geq \rho \), where \( \rho \) is the smallest mean shift that is meaningful in practice.

Let \( \hat{\mu}_k \) replace the unknown \( \mu_k \) by its estimate from the past observed data in (5). At time \( n \), we can produce a candidate post-change time \( \hat{T} = (0, 1, \ldots, n) \) and thus \( \mu_k \) is estimated by \( \hat{\mu}_k = \hat{T}_k \).

Procedure:

- Define \( \delta \) as the largest \( 0 \leq \delta \leq \nu + 1 \) such that \( W_0 = 0 \) and denote by \( T_{\mathcal{H}_0} = \nu + \delta \) and \( S_{\mathcal{H}_0} = \sum_{i=1}^{\nu+\delta} X_i \) the total number and the summation of observations \( X_i \)'s between the candidate post-change time \( \hat{T} \) and step time \( n+1 \).

- A Neyman-type estimator of \( \mu_k \) at time \( n \):

\[
\hat{\mu}_k = \max \left\{ \rho, \frac{S_k - X_k}{\phi} \right\}.
\]

- Let \( S_k = T_{\nu,n} = X_k = 0 \), and \( \rho_k = \rho \). For all \( n \geq 1 \),

\[
W_k,n = \max \left\{ W_k,n-1 + \hat{\mu}_k, 0 \right\},
\]

where the \( S_k \) and \( T_{\nu,n} \) in (6) has the recursive formula:

Reference:

**Proposition 1.** Suppose that the delay effects \( \delta_k \)'s satisfy the following post-change hypothesis set :\n
\[
\Delta = \{ \delta_k: \delta_k \leq \infty \} \text{ or } \delta_k < \infty \text{ and } \min_{\delta_k \in \Delta} \delta_k = 0 \}
\]

where \( \gamma \) is the false alarm constraint in (1).

Simulation Results

Table 1: A comparison of detection delays when the change is instantaneous and the post-change mean \( \mu_k \) is only affected. Results are based on 200 Monte Carlo simulations.

<table>
<thead>
<tr>
<th>Data stream</th>
<th>Largest standard error</th>
<th>Actual delay</th>
<th>False alarm rate</th>
<th>Normalized delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
</tr>
<tr>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
<td>( W_k^{(r)} )</td>
</tr>
</tbody>
</table>

**References (39 total publications)**


