THE UNIVERSITY of WISCONSIN MADISON

Industrial and Systems Engineering

Online monitoring of big data streams --- roadmap and recent advances

Dr. Kaibo Liu

Associate Professor
Department of Industrial and Systems Engineering
University of Wisconsin-Madison

Associate Director
UW-Madison IoT Systems Research Center

Date: 4/25/2022

Background

- **Associate Professor** 2019-now, Department of industrial and Systems Engineering, UW-Madison
- **Associate Director** 2019-now, UW-Madison IoT systems research center
- **Assistant Professor** 2013-2019, Department of industrial and Systems Engineering, UW-Madison
- **Ph.D.** 2013, Industrial Engineering (Minor: Machine Learning), Georgia Institute of Technology
- **M.S.** 2011, Statistics, Georgia Institute of Technology
- **B.S.** 2009, Industrial Engineering and Engineering Management, Hong Kong University of Science and Technology

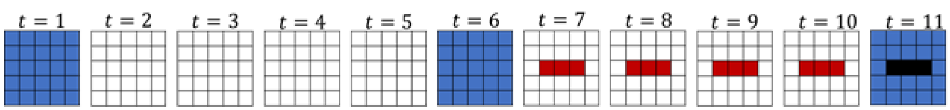
Outline

- **Introduction**
  - What is the problem?
- Research Topics
  - An adaptive sampling strategy for online high-dimensional process monitoring (2015)
  - A Nonparametric Adaptive Sampling Strategy for Online Monitoring of Big Data Streams (2018)
  - Online Nonparametric Monitoring of Heterogeneous Data Streams with Partial Observations based on Thompson Sampling (2022)
- Summary

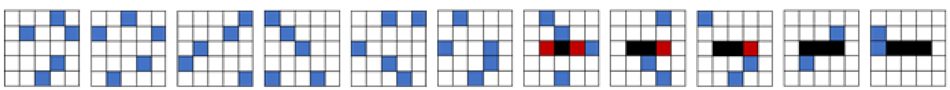
What is the problem?
Motivation & applications

- **Goal**: Develop a systematic and scalable adaptive monitoring and sampling strategy that enables us to actively select the partial “observable” data to maximize the change detection capability of the whole system subject to the resource constraints

- **Innovative idea:**



(a) Conventional sampling strategy over the temporal domain



(b) Dynamic sampling strategy over the spatial domain

exploration

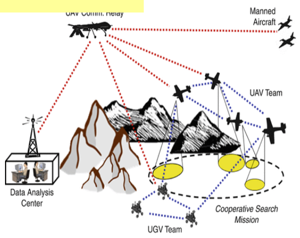
exploitation

blue: sampled data streams

red: anomaly regions

black: overlapping

UAV surveillance

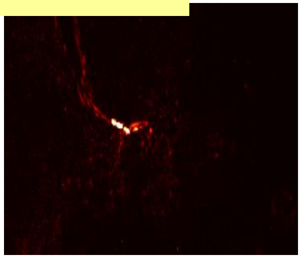


Limited number of devices

Manufacturing process

Limited energy/cost or bandwidth

Solar flare detection



Limited transmission/processing time

Outline

- Introduction
  - What is the problem?
- **Research Topics**
  - An adaptive sampling strategy for online high-dimensional process monitoring (2015)
  - A Nonparametric Adaptive Sampling Strategy for Online Monitoring of Big Data Streams (2018)
  - Online Nonparametric Monitoring of Heterogeneous Data Streams with Partial Observations based on Thompson Sampling (2022)
- Summary

An adaptive sampling strategy for online high-dimensional process monitoring

Liu, K., Mei, Y., and Shi, J. (2015), "An adaptive sampling strategy for online high-dimensional process monitoring", *Technometrics*, 57, 3, 305-319.

Problem Formulation & Assumption

- m physical variables $M = \{1, \ldots, m\}$ and q $(q \leq m)$ sampling resources in a system.
- When the process is in-control:
  - Each variable $k \sim N(0,1)$
- At some unknown time ν:
  - Mean shift occurs at certain variables and will affect an unknown subset of data streams
  - Each variable $k \sim N(u_k, 1)$
- Samples over time are independent of each other
- **Goal: Based on dynamic observations in real time, actively decide which data streams to observe at the next time** for quick detection of anomaly event while still maintaining a system-wide false alarm rate.

Algorithm Illustration

- Five variables and two observable
  - Red: observable; White: unobservable
- $t = 0$: Create a local statistic for each variable and initiate $W_{k,0} = 0$ for all k
- $t = t_1$: Obtain the measurement based on current layout (Step 1)

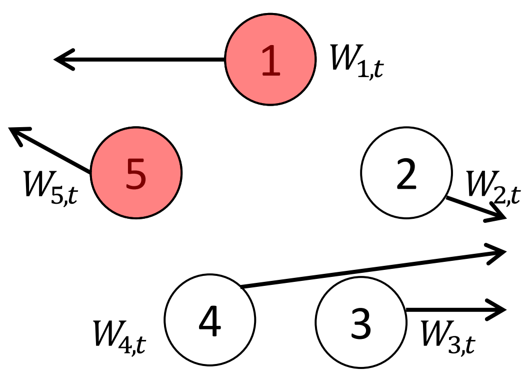
Step 2:

Update $W_{k,t}$ based on the observation at t_1 by CUSUM:

$$W_{k,t}^{(1)} = max\left(W_{k,t-1}^{(1)} + u_{min}X_{k,t} - \frac{u_{min}^2}{2}, 0\right)$$

$$W_{k,t}^{(2)} = max\left(W_{k,t-1}^{(2)} - \boxed{u_{min}}X_{k,t} - \frac{u_{min}^2}{2}, 0\right)$$

interested-smallest magnitude of shift for detection



Update $W_{k,t}$ with a small increment:

$$W_{k,t}^{(1)} = W_{k,t-1}^{(1)} + \Delta$$

$$W_{k,t}^{(2)} = W_{k,t-1}^{(2)} + \boxed{\Delta}$$

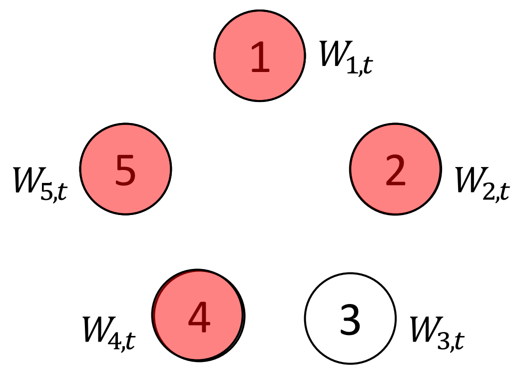
Compensation coefficient

$$W_{k,t} = max\,(W_{k,t}^{(1)}, W_{k,t}^{(2)})$$

t	X_1	X_2	X_3	X_4	X_5	$W_{1,t}$	$W_{2,t}$	$W_{3,t}$	$W_{4,t}$	$W_{5,t}$	Monitoring statistics	Updated s
1	0.0301	N/A	N/A	N/A	0.0033	0	0.1	0.1	0.1	0		

Algorithm Illustration

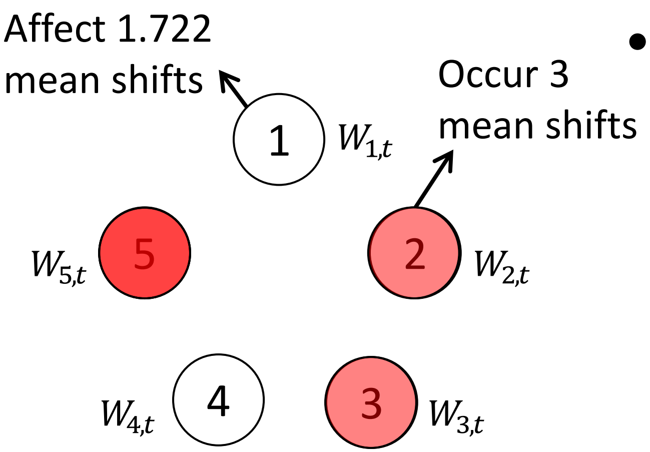
- Calculate the sum of largest r local statistics as the monitoring statistics (Step 3)
  - Engineering domain knowledge: Change only affects a small subset of variables
- Update the sampling layout onto the variables with largest local statistics (Step 4)



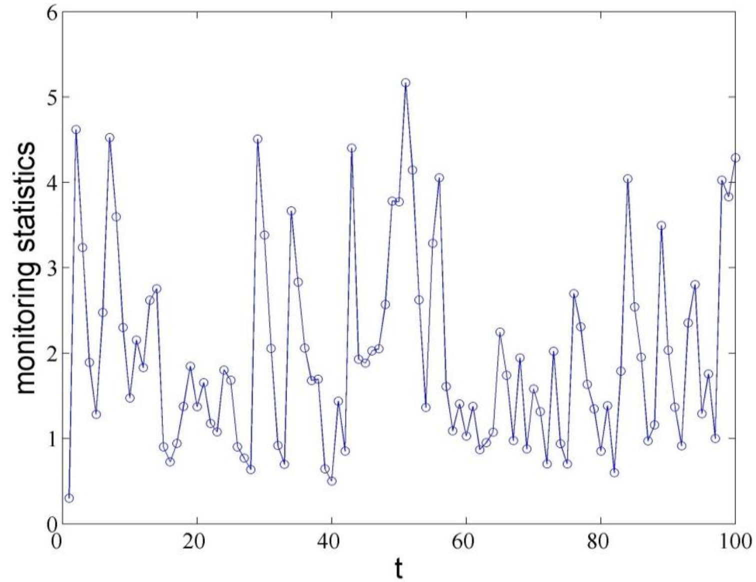
$W_{2,t}$	0.1
$W_{4,t}$	0.1
$W_{3,t}$	0.1
$W_{1,t}$	0
$W_{5,t}$	0

t	X_1	X_2	X_3	X_4	X_5	$W_{1,t}$	$W_{2,t}$	$W_{3,t}$	$W_{4,t}$	$W_{5,t}$	Monitoring statistics	Updated s
1	0.0301	N/A	N/A	N/A	0.0033	0	0.1	0.1	0.1	0	0.3	{2,4}

Illustration (process in-control) $t = 101$

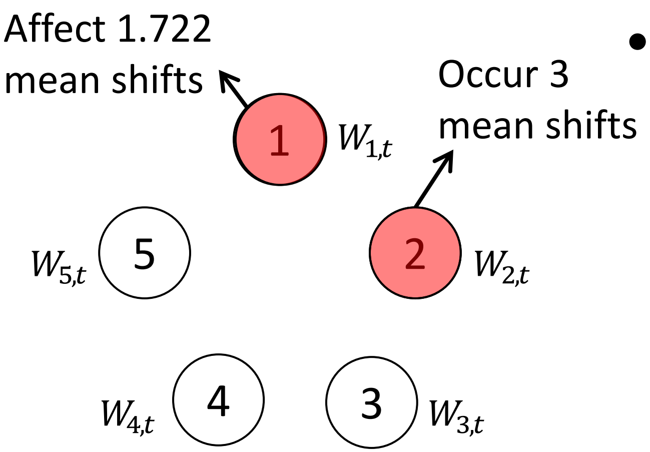


- Interested to detect $u_{min} = 1.5$ mean shift

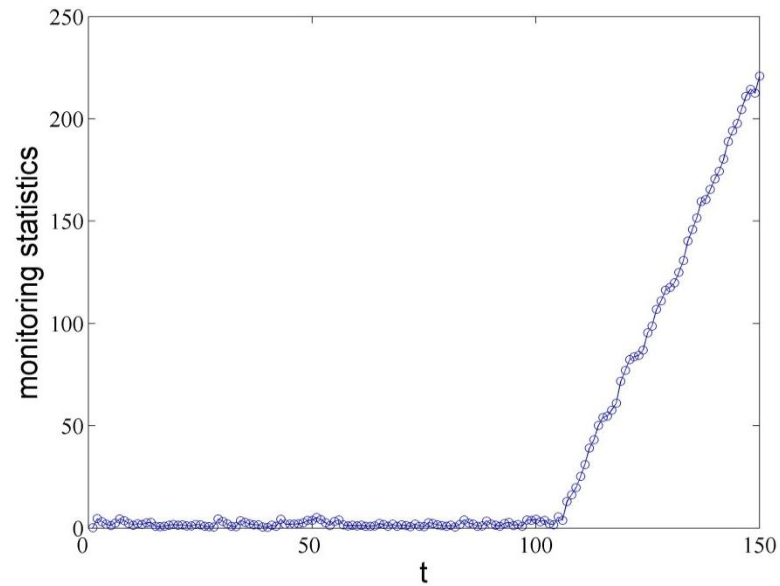


t	X_1	X_2	X_3	X_4	X_5	$W_{1,t}$	$W_{2,t}$	$W_{3,t}$	$W_{4,t}$	$W_{5,t}$	Monitoring statistics	Updated s
1	0.0301	N/A	N/A	N/A	0.0033	0	0.1	0.1	0.1	0	0.3	{2,4}
2	N/A	-2.4866	N/A	-1.8268	N/A	0.1000	2.7049	0.2000	1.7151	0.1000	4.6201	{2,4}
…												
101	N/A	-0.6248	N/A	N/A	0.6495	0.4000	0	0.5000	0.4000	2.2731	3.1731	{3,5}

Illustration (process out-of-control) $t = 150$



- Interested to detect $u_{min} = 1.5$ mean shift



t	X_1	X_2	X_3	X_4	X_5	$W_{1,t}$	$W_{2,t}$	$W_{3,t}$	$W_{4,t}$	$W_{5,t}$	Monitoring statistics	Updated s
1	0.0301	N/A	N/A	N/A	0.0033	0	0.1	0.1	0.1	0	0.3	{2,4}
2	N/A	-2.4866	N/A	-1.8268	N/A	0.1000	2.7049	0.2000	1.7151	0.1000	4.6201	{2,4}
…												
101	N/A	-0.6248	N/A	N/A	0.6495	0.4000	0	0.5000	0.4000	2.2731	3.1731	{3,5}
…												
150	2.8230	4.1308	N/A	N/A	N/A	68.4773	147.6343	4.7718	4.6768	4.6899	220.8834	{1,2}

TRAS algorithm

- Three major steps:
  - how to construct local statistics:
    - *If* $(k \in S)$, $W_{k,t}^{(1)} = max\left(W_{k,t-1}^{(1)} + \boxed{u_{min}} X_{k,t} - \frac{u_{min}^2}{2}, 0\right)$ *and* $W_{k,t}^{(2)} = max\left(W_{k,t-1}^{(2)} - u_{min} X_{k,t} - \frac{u_{min}^2}{2}, 0\right)$;
    - *Otherwise* $(k \notin S)$, $W_{k,t}^{(1)} = W_{k,t-1}^{(1)} + \Delta$ *and* $W_{k,t}^{(2)} = W_{k,t-1}^{(2)} + \boxed{\Delta}.$
    - *Let* $W_{k,t} = max\,(W_{k,t}^{(1)}, W_{k,t}^{(2)})$.
  - when to indicate process is out-of-control:
    - $N_{top,r}(d) = inf\{t \geq 1: \boxed{\sum_{k=1}^{r}} W_{(k),t} \geq d\}$.
  - how to update sampling layout :
    - Denote the index of the decreasing order statistics $W_{(k),t}$ as $l_{(k)}$
    - $S = \{l_{(1)}, \dots, l_{(q)}\}$

interested-smallest magnitude of shift for detection

Incremental parameter

Engineering domain knowledge: Change only affects a small subset of sensors

Properties

Property 1: Assume $\rho_k = u_{min}\, u_k - \frac{u_{min}^2}{2} - \Delta \leq 0$ for any $k \in M$. Denote U by those $k \in M$ such that the variable k will never be observed again after some finite time t_0. Then as $d \to \infty$, $P(U = \emptyset) \to 1$, where $\emptyset$ represents the empty set.

– Resampling each variable with infinite number of times (i.e. all variables will not be left unattended)

Proof by contradiction:

- Denote $M = \{1, \dots, m\}$. Suppose sampling resources will never be redistributed to the set of variables $U \neq \emptyset$ after time t.

$$\textit{Lemma 1}: W_{k',t'} \geq W_{k,t'}, \text{ for } \forall t' > t, \forall k' \in M \backslash U, \text{ and } \forall k \in U.$$

- There must be a series of $X_{k',t'}$ such that either $\sum_{t'=1}^{\infty}\left(u_{min}X_{k',t'} - \frac{u_{min}^2}{2}\right) \geq \Delta t'$ or $\sum_{t'=1}^{\infty}\left(-u_{min}X_{k',t'} - \frac{u_{min}^2}{2}\right) \geq \Delta t'$

- $$lim_{t'\to\infty} P\left(\sum_{t'=1}^{\infty}\left(u_{min}X_{k',t'} - \frac{u_{min}^2}{2}\right) \geq \Delta t'\right) + lim_{t'\to\infty} P\left(\sum_{t'=1}^{\infty}\left(-u_{min}X_{k',t'} - \frac{u_{min}^2}{2}\right) \geq \Delta t'\right) = 0$$

$\Rightarrow$ $U = \emptyset$

Contradictory

Properties & Proofs

Property 2: Let $B = \{k \in M: \rho_k = u_{min}u_k - \frac{u_{min}^2}{2} - \Delta > 0\}$. Then, for any finite time t, once the variable $k \in B$ is observed at time t, there is a nonzero probability such that this variable will be kept observing at all the future time, as $d \rightarrow \infty$.

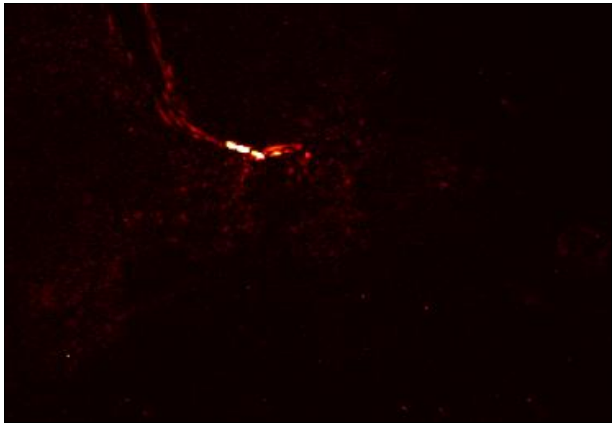
– Sampling resources will eventually stick to the anomaly regions: localize the anomaly event.

Proof :

- Considering the variable k, where $k \in S_t$ at time t and $k \in B$. Then, $W_{k,t} \geq W_{k',t}$ for $\forall k' \notin S_t$. Define $Y_{k,n} = X_{k,t+n} - \frac{u_{min}}{2} - \frac{\Delta}{u_{min}}$ and $H_{k,n} = \sum_{i=1}^{n} Y_{k,i}$. $\{H_{k,n}: n \geq 0\}$ refers to the Guassian random walk process, where $H_{k,0} = 0$.
- $P(G = 0) = P(H_{k,n} \geq 0, \forall n \geq 0) = \sqrt{2}\delta_k exp\{\frac{\delta_k}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\zeta\left(\frac{1}{2}-r\right)}{r!(2r+1)} \left(-\frac{\delta_k^2}{2}\right)^r\}$ (Chang and Peres, 1997), where $G = min\{H_{k,n}: n \geq 0\}$, $\delta_k = u_k - \frac{u_{min}}{2} - \frac{\Delta}{u_{min}}$ and $\zeta(\cdot)$ is the Riemann zeta function. (An increasing function as δ_k gets larger).
- According to property 1, sampling layout will not stick to the variables in $M \backslash B$, and thus they must be redistributed to the variables in B at some time.

Case Study – Solar Flare Detection

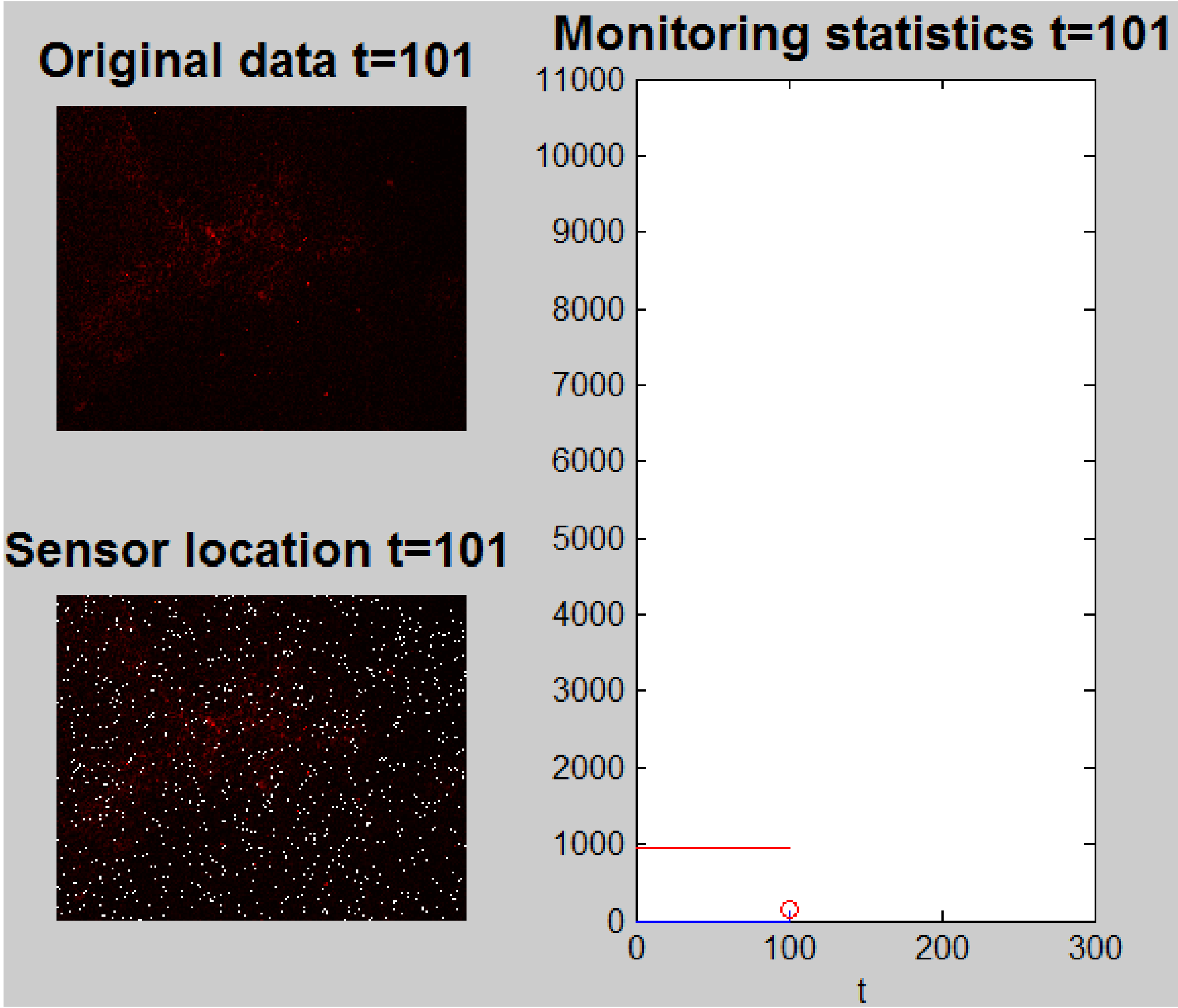

Solar dynamic observatory


Example of Solar flare

source: NASA

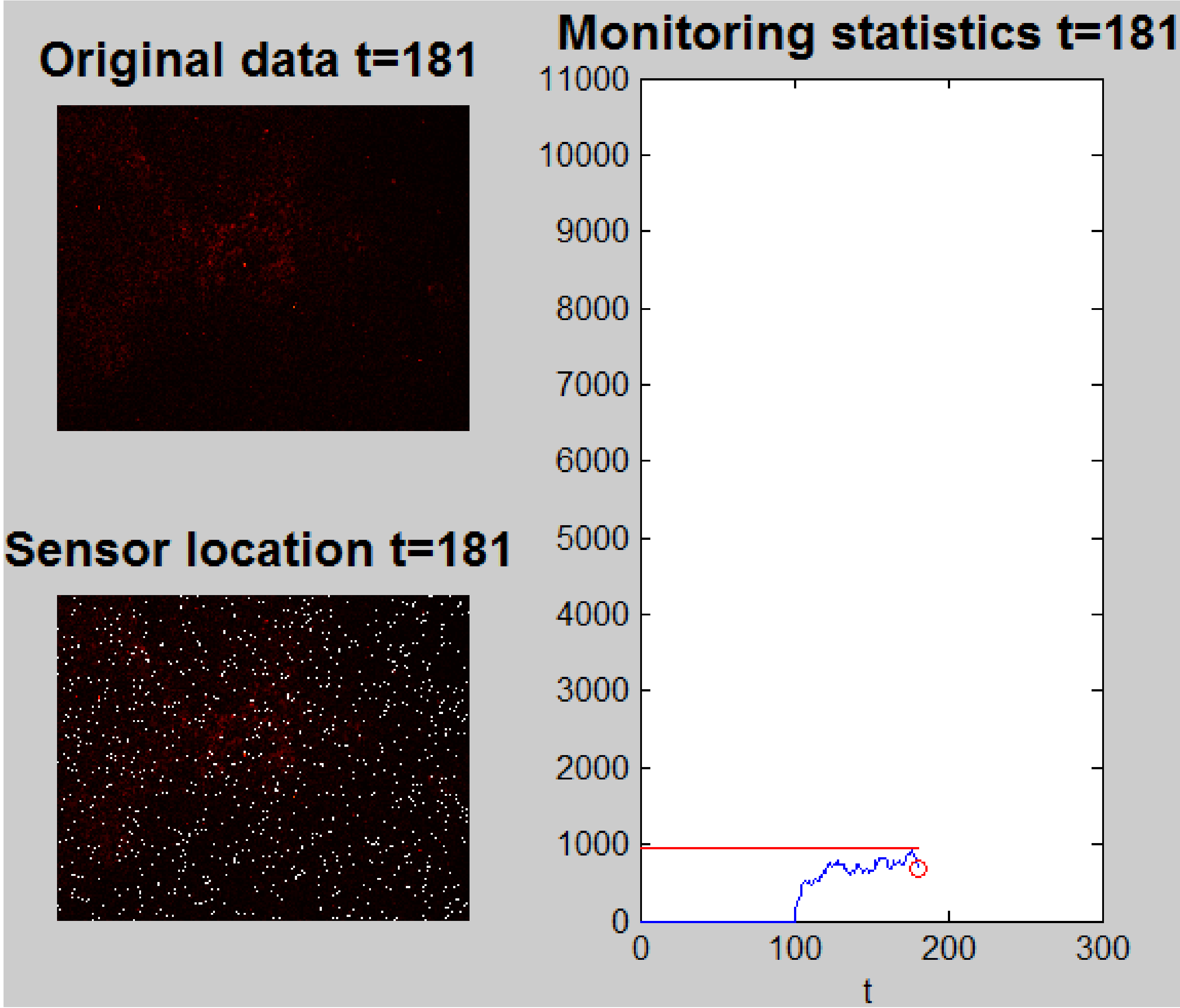
- A solar flare is a sudden, transient, and intense variation in brightness
- High-dimensional: each frame has $232 \times 292 = 67744$ pixels
- Large volume: 130Mbps, acquires $\sim$ 11TB of data each day
- Goal: real-time detect abnormal solar flares from partial streaming data
  - Assume only 2000 out of 67744 pixels are available
  - Accounts for only 2.95% partial information

Result

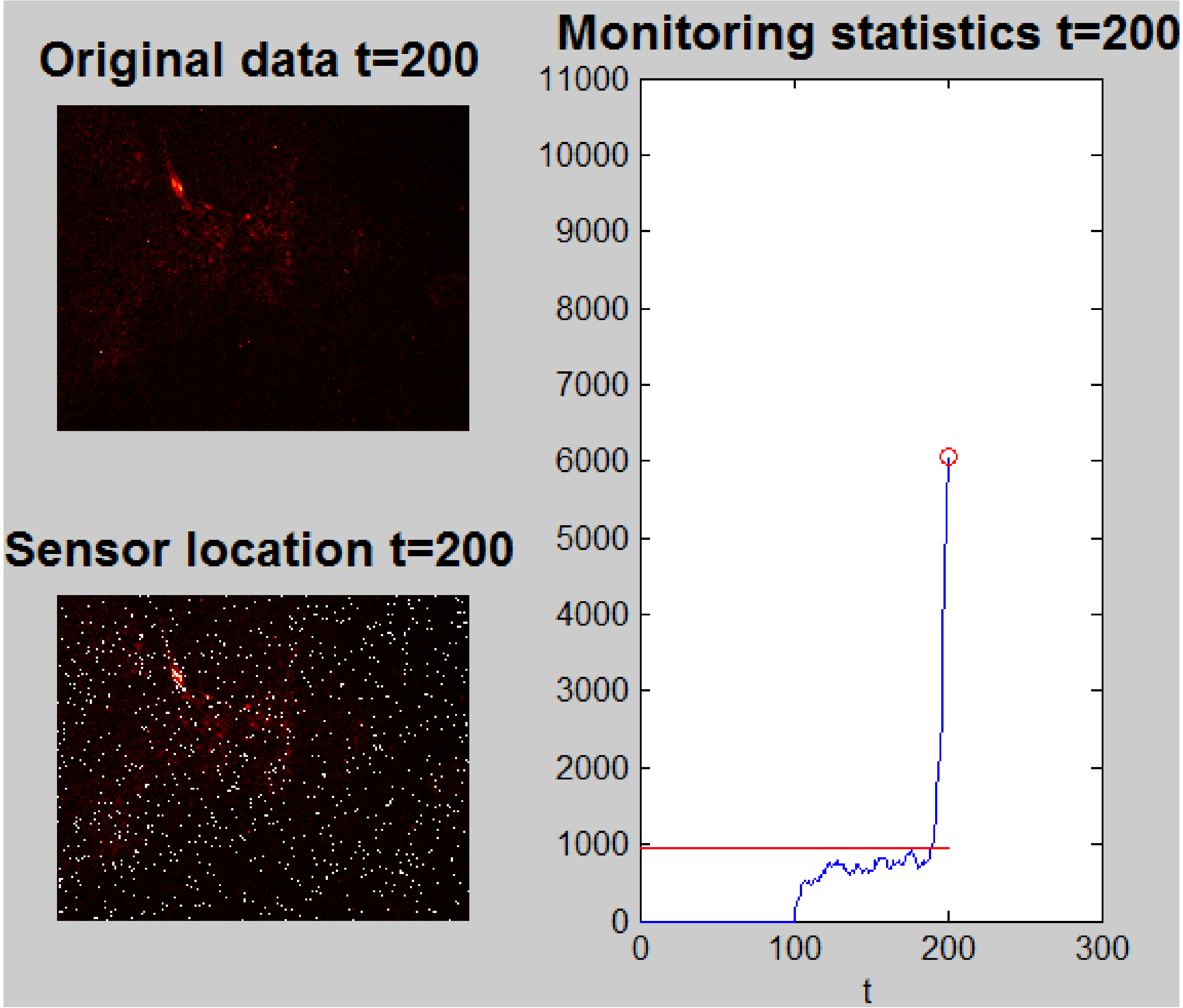


Original data t=101
Sensor location t=101
Monitoring statistics t=101
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
t

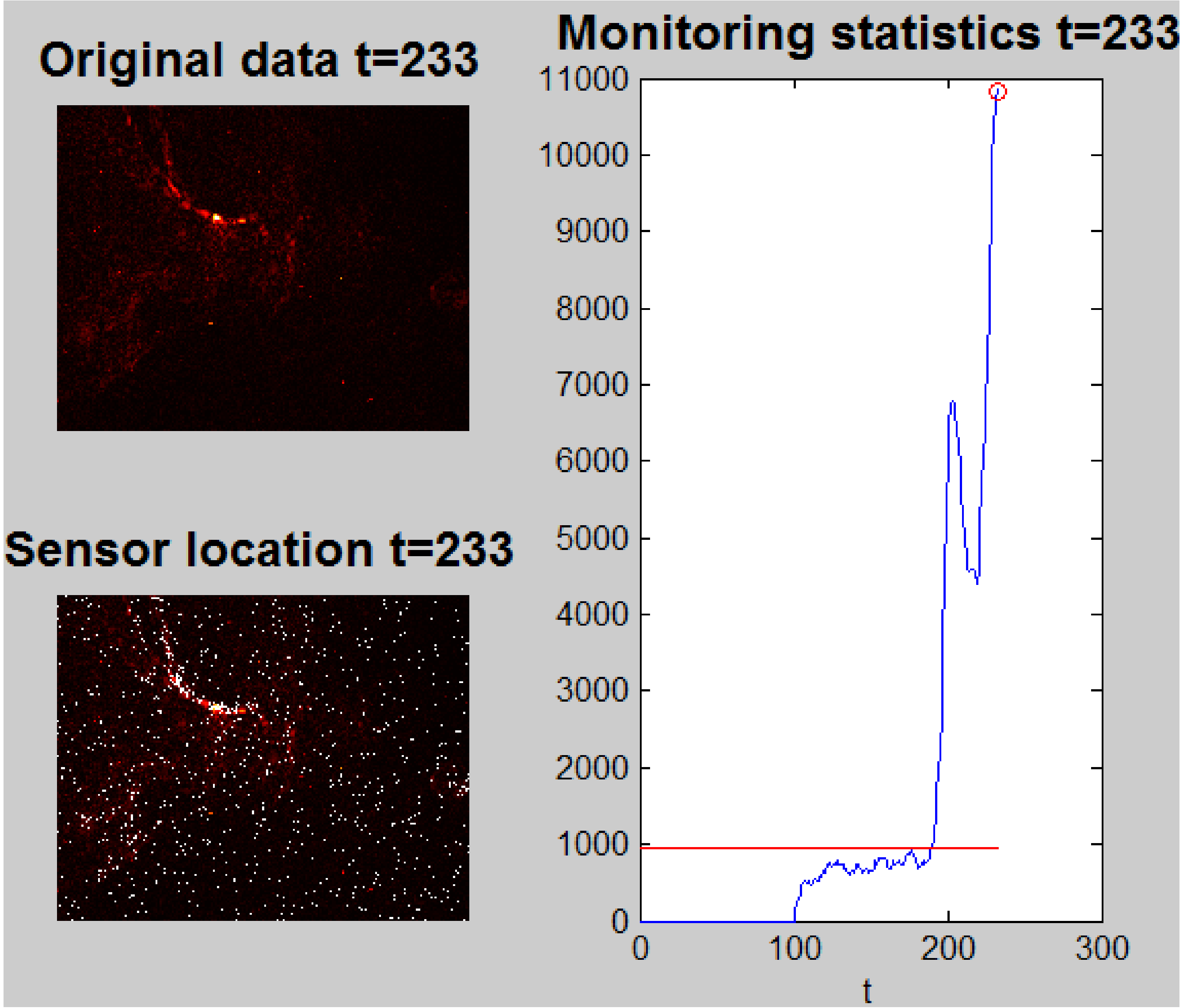
Result


Original data t=181
Sensor location t=181
Monitoring statistics t=181
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
t

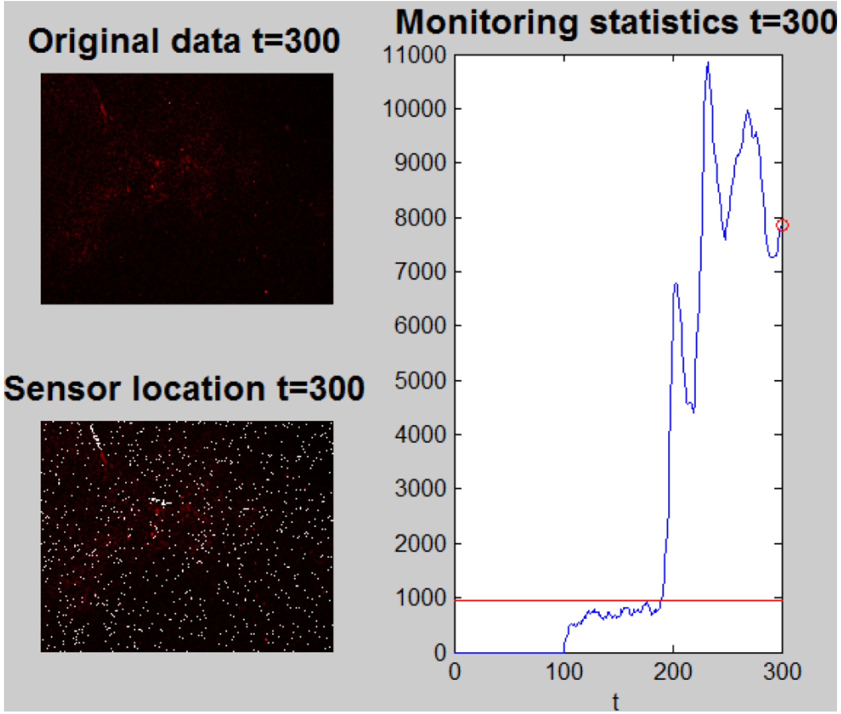
Result


Original data t=200
Sensor location t=200
Monitoring statistics t=200
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
t

Result


Original data t=233
Sensor location t=233
Monitoring statistics t=233
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
t

Result Comparison



	Our method, base on 2000 pixels	Xie et al. (2013), based on 67744 pixels
First solar flare	$t = 190$	$t = 191$
Second solar flare	$t = 221$	$t = 217$
Algorithm type	Efficient (recursive)	Inefficient (no recursive)
Real-time Monitoring	Yes	No

Summary of the proposed TARS sampling strategy

- A systematic adaptive sampling strategy is proposed for real-time monitoring of Big Data streams with dynamically selected partial information.
- **Scalability**: **linear** in the number of data streams
- **Adaptability**:
  - Quickly detect a wide range of possible changes **with no prior knowledge** of the potential anomaly events by adaptively adjusting to the event locations;
  - **Actively select the data streams** to observe from the whole streaming data to maximize the sensitivity for anomaly detection with consideration of resource constraints.
- **Limited in the normality assumption**

A Nonparametric Adaptive Sampling Strategy for Online Monitoring of Big Data Streams

Xian, X., Wang, A., and **Liu, K.** (2018), “A Nonparametric Adaptive Sampling Strategy for Online Monitoring of Big Data Streams”, *Technometrics*, 60, 1, 14-25. (This paper received the Best Student Poster award in Quality, Statistics, and Reliability Section of INFORMS, 2016; This paper is selected for presentation in the Technometrics invited session in the 2018 INFORMS conference)

Objective & Problem formulation

Objective

– Propose an adaptive nonparametric monitoring scheme with only partial observations available.

Problem formulation and assumptions

– Measurement of the p variables at time t:

$$\boldsymbol{X}(t) = \left(X_1(t), X_2(t), \dots, X_p(t)\right)$$

– Only r $(r < p)$ out of p variables **can be "observable"** at each time t.

– When the process is in-control:
- All variables are exchangeable (not necessarily normal!)
- In-control mean of $\boldsymbol{X}(t)$ is $\mathbf{0} = (0, 0, \cdots, 0)'$.

– At some unknown time τ,
- Unknown mean of $\boldsymbol{X}(t)$: $\boldsymbol{\mu}' \neq \mathbf{0}$.
- The number of affected variables is unknown.

Nonparametric Anti-rank based Sampling strategy (NAS)

- **Anti-rank**

Variables and anti-ranks	$\widetilde{\boldsymbol{X}}(t) = \left(X_1(t), X_2(t), \dots, X_p(t), 0\right)$ $\tilde{X}_{B_1(t)}(t) \leq \tilde{X}_{B_2(t)}(t) \leq \cdots \leq \tilde{X}_{B_{p+1}(t)}(t)$.

Illustration purpose, more complicated rank can be considered

Theorem. Let ν_F be the probability measure defined by the joint distribution function $F(\boldsymbol{x})$ of the p measurements. If $\nu_F(O) > 0$ for any open set $O \in R^p$, which includes the origin in its closure and has positive Lebesgue area, then the distribution of the anti-rank indicator $\boldsymbol{\xi}(t)$ is different from its in-control distribution when the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_p = 0$ is violated by the shift in the mean vector of the process, where $\boldsymbol{\mu} = \left(\mu_1, \mu_2, \cdots, \mu_p\right)$ is the mean of $\boldsymbol{X}(t)$.

Can be extended to other anti-ranks to facilitate distributional shift.

Nonparametric Anti-rank based Sampling strategy (NAS)

- **Example in partial observations**

$\boldsymbol{X}(t)$	3.2	-0.5	-2.8	NA	-1.6	NA	0
$\boldsymbol{\xi}(t)$	0	0	λ	Δ	0	Δ	0

Challenge: $\boldsymbol{\xi}(t)$ not well-defined!

$\boldsymbol{X}(t)$	3.2	0.5	2.8	NA	1.6	NA	0
$\boldsymbol{\xi}(t)$	0	0	0	Δ	0	Δ	λ_0

The compensation coefficient Δ is a pseudo observation that represents the likelihood of each unobservable variable taking the first anti-rank.

- **Generalized anti-rank indicator**

$$\xi_j(t) = \begin{cases} \lambda, \text{observable and } B_1(t) = j \\ 0, \text{observable but } B_1(t) \neq j \\ \Delta, \text{unobservable} \end{cases}, j = 1,2,\cdots p.$$

$$\xi_{p+1}(t) = \begin{cases} \lambda_0, B_1(t) = p+1 \\ 0, B_1(t) \neq p+1 \end{cases}$$

$\boldsymbol{g} = (g_1, g_2, \dots, g_{p+1}), g_j = \mathbb{P}(B_1(t) = j)$

long run probability of anti-ranks.

$$\mathbb{E}(\boldsymbol{\xi}) = \boldsymbol{g} \Leftrightarrow$$

(1) Offline parameter settings for λ, Δ, λ_0

(2) Online monitoring

Three major steps of the NAS algorithm

1. Construct CUSUM statistics

Local statistics, weighted sum of $\boldsymbol{\xi}(t)$

Behaves like the expected IC $\boldsymbol{S}(t)$

$$\begin{cases} \boldsymbol{S}(t) = 0, \quad \boldsymbol{R}(t) = 0 & \text{if } C_t \le k, \\ \boxed{\boldsymbol{S}(t)} = \big(\boldsymbol{S}(t-1) + \boldsymbol{\xi}(t)\big)(C_t - k)/C_t & \\ \boxed{\boldsymbol{R}(t)} = (\boldsymbol{R}(t-1) + \boldsymbol{g})(C_t - k)/C_t & \text{if } C_t > k. \end{cases}$$

$C_t = (\boldsymbol{S}(t-1) - \boldsymbol{R}(t-1) + \boldsymbol{\xi}(t) - \boldsymbol{g})' \cdot diag\left(\frac{1}{R_1(t-1)+g_1}, \dots, \frac{1}{R_{p+1}(t-1)+g_{p+1}}\right) \cdot$
$(\boldsymbol{S}(t-1) - \boldsymbol{R}(t-1) + \boldsymbol{\xi}(t) - \boldsymbol{g})$, where $\boldsymbol{S}(0) = \boldsymbol{R}(0) = \boldsymbol{0}$, k is a constant.

2. Stopping time

Measures the difference between $\boldsymbol{S}(t)$ and $\boldsymbol{R}(t)$

$$y_t = \big(\boldsymbol{S}(t) - \boldsymbol{R}(t)\big)' diag\left(\frac{1}{R_1(t)}, \dots \frac{1}{R_{p+1}(t)}\right)\big(\boldsymbol{S}(t) - \boldsymbol{R}(t)\big).$$

Stop the monitoring process when $y_t > h$.

3. Sampling strategy

Observe data streams with the largest local statistics $\boldsymbol{S}(t)$

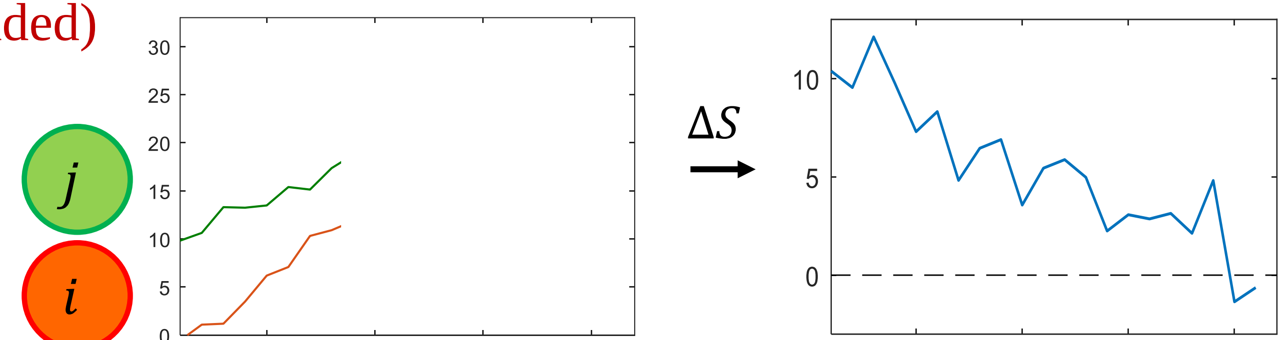
Denote $j_{(l),t}$ to be the variable index of the decreasing order statistics of $\big(S_1(t), \cdots, S_p(t)\big)$, observe $\{j_{(1),t}, \cdots, j_{(r),t}\}$ at time $t+1$.

Properties of the NAS algorithm

IC Property: Assume that $\varDelta > \frac{\lambda}{r}$. Let U denote those variable $i \in \mathcal{P}$ that can never be observed after some finite time t_0, i.e., $\exists t_0$ such that $U = \bigcap_{t=t_0}^{+\infty} \mathcal{U}(t)$. As $h \to \infty$, $P(U = \emptyset) \to 1$, where $\emptyset$ represents the empty set.

$\varDelta$: compensation coefficient, λ: OC penalty, r: number of observable variables.

- Redistributed to each variable with infinite number of times (i.e. no variables will be left unattended)



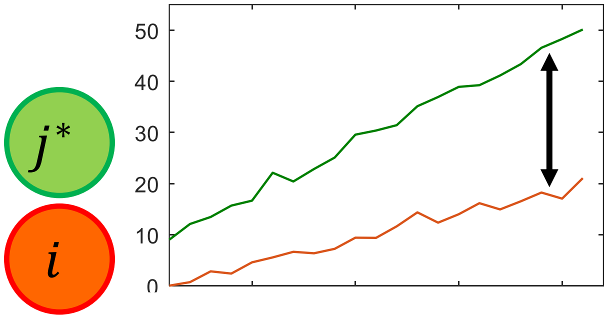
Proof:

- *Lemma 1*: $S_i(t) \leq S_j(t)$, for all $t \geq t_0$, $j \in \mathcal{P} \backslash U$ and $i \in U$.
- $S_j(t_l) - S_i(t_l) \leq S_j(t_0) - S_i(t_0) + \sum_{m=1}^{l} \left(\xi_j(t_m) - \varDelta \right)$.
- $\sum_{m=1}^{l} \left(\xi_j(t_m) - \varDelta \right)$ is a general random walk with mean $\mathbb{E}\xi_j(t_m) - \varDelta = \frac{\lambda}{r} - \varDelta < 0, \Rightarrow S_j(t_l) < S_i(t_l)$. Contradiction!

Properties of the NAS algorithm

OC Property: Suppose after time t_0, there is a mean shift. Let D be a set of out-of-control variables in the sense that $D = \left\{j: \mathbb{P}\big(\xi_j(t) = \lambda | j \in \mathcal{O}(t)\big) > \frac{\Delta}{\lambda}\right\}$. Then once $j^* \in D$ is observed at time t, there is a nonzero probability that variable j^* will always be observed forever, i.e., $j^* \in \mathcal{O}(\tau)$ for $\forall \tau \geq t$.

- Monitoring resources will eventually stick to the variables with large mean shifts.



Proof: Denote $j_{(1),t}, \ldots, j_{(p),t}$ as the variable indices such that $S_{j_{(1),t}}(t) \geq S_{j_{(2),t}}(t) \geq \cdots \geq S_{j_{(p),t}}(t)$. Define the difference between the increments of $S_j(t)$ on variable j^* and $i_{(r+1),t+n}$ at time $t+n$ to be $Z_{j^*,n}$.

- $Z_{j^*,n} > \xi_{j^*}(t+n) - \Delta, P\left(S^{(1)}_{j^*,t+N} - S^{(1)}_{(r+1),t+N} > 0 \text{ for any } N > 0\right) > 0$

Simulation results

Simulation settings

- Based on 5000 replications, $ARL_0 = 370$.
- Different magnitudes of mean shifts.
- Different choices of number of observable variables r.

Competing algorithms

Monitoring normal data

Out-of-control average run length. Assume that $p = 6$ variables follow an independent $N(0,1)$ distribution. One randomly selected variable has a shift with magnitude of δ.

		NAS	TRAS	RS	QH03
$r = 2$	$\delta = 1.0$	24.65 (0.26)	17.57 (0.18)	41.45 (0.75)	12.55 (0.15)
	$\delta = 2.0$	11.78 (0.09)	5.48 (0.03)	20.53 (0.31)	4.41 (0.03)
	$\delta = 3.0$	7.60 (0.06)	3.72 (0.02)	17.18 (0.24)	3.24 (0.01)
$r = 3$		**NAS**	**TRAS**	**RS**	**QH03**
	$\delta = 1.0$	22.00 (0.22)	16.00 (0.15)	38.95 (0.56)	12.55 (0.15)
	$\delta = 2.0$	7.89 (0.06)	4.92 (0.03)	15.30 (0.17)	4.41 (0.03)
	$\delta = 3.0$	5.72 (0.03)	3.42 (0.01)	12.01 (0.12)	3.24 (0.01)
$r = 4$		**NAS**	**TRAS**	**RS**	**QH03**
	$\delta = 1.0$	18.67 (0.21)	15.56 (0.15)	26.65 (0.38)	12.55 (0.15)
	$\delta = 2.0$	5.91 (0.05)	4.81 (0.03)	10.98 (0.10)	4.41 (0.03)
	$\delta = 3.0$	4.09 (0.02)	3.35 (0.01)	8.39 (0.06)	3.24 (0.01)

Monitoring exponential data

Out-of-control average run length. Assume that $p = 6$, variables follow an independent exponential distribution. One randomly selected variable has a shift with magnitude of δ.

		NAS	TRAS	RS	QH03
$r = 2$	$\delta = 1.0$	14.97 (0.12)	30.17 (0.28)	24.97 (0.38)	8.52 (0.11)
	$\delta = 2.0$	8.32 (0.08)	9.22 (0.03)	19.10 (0.27)	3.62 (0.03)
	$\delta = 3.0$	7.33 (0.07)	6.90 (0.02)	17.12 (0.23)	2.54 (0.02)
$r = 3$		NAS	TRAS	RS	QH03
	$\delta = 1.0$	13.40 (0.14)	29.21 (0.26)	21.58 (0.29)	8.52 (0.11)
	$\delta = 2.0$	6.73 (0.05)	8.65 (0.03)	12.69 (0.15)	3.62 (0.03)
	$\delta = 3.0$	5.39 (0.03)	5.67 (0.02)	10.62 (0.12)	2.54 (0.02)
$r = 4$		NAS	TRAS	RS	QH03
	$\delta = 1.0$	12.72 (0.12)	28.65 (0.25)	16.31 (0.18)	8.52 (0.11)
	$\delta = 2.0$	5.55 (0.04)	7.39 (0.03)	8.68 (0.07)	3.62 (0.03)
	$\delta = 3.0$	4.21 (0.02)	4.59 (0.02)	7.38 (0.06)	2.54 (0.02)

Monitoring multinomial data

Out-of-control average run length. Assume that $p = 10$, variables follow a $\mathrm{MN}(100; (P_1, \ldots, P_{10}))$ distribution where $P_i = 0.1$. One randomly selected variable has a shift that $P_i = \frac{1}{10}(1 + \delta)$.

		NAS	TRAS	RS	QH03
$r = 3$	$\delta = 20\%$	48.09 (0.83)	61.00 (0.73)	68.19 (1.80)	15.53 (0.28)
	$\delta = 50\%$	9.75 (0.14)	10.60 (0.06)	20.05 (0.39)	4.21 (0.04)
	$\delta = 100\%$	4.83 (0.03)	5.60 (0.02)	12.99 (0.21)	2.21 (0.01)
$r = 5$		**NAS**	**TRAS**	**RS**	**QH03**
	$\delta = 20\%$	43.74 (0.61)	50.99 (0.61)	48.42 (0.98)	15.53 (0.28)
	$\delta = 50\%$	8.30 (0.08)	8.82 (0.05)	11.49 (0.16)	4.21 (0.04)
	$\delta = 100\%$	4.14 (0.02)	4.53 (0.01)	6.40 (0.07)	2.21 (0.01)
$r = 7$		**NAS**	**TRAS**	**RS**	**QH03**
	$\delta = 20\%$	32.92 (0.53)	50.12 (0.62)	39.79 (0.71)	15.53 (0.28)
	$\delta = 50\%$	5.98 (0.07)	8.50 (0.05)	7.86 (0.09)	4.21 (0.04)
	$\delta = 100\%$	2.93 (0.02)	4.34 (0.01)	4.15 (0.04)	2.21 (0.01)

Summary of the nonparametric big data monitoring research

- The online NAS algorithm is proposed for real-time monitoring **exchangeable distributions,** in the cases that only **partial observation** of data streams is available.
- Two properties of this algorithm with theoretical proofs are investigated.
- **Still limited to the homogeneous assumption.**

Online Nonparametric Monitoring of Heterogeneous Data Streams with Partial Observations based on Thompson Sampling

Ye, H., Xian, X., Cheng, J. C., Hable, B., Shannon, R. W., Elyaderani, M. K. and **Liu, K**. (2022), "Online Nonparametric Monitoring of Heterogeneous Data Streams with Partial Observations based on Thompson Sampling", *IISE Transactions*, accepted. (This paper received the Best Student Paper Finalist award in the QCRE Section of Industrial and Systems Engineering Research Conference (ISERC), 2020).

Objective & Problem formulation

- **Objective**
  - Propose a **nonparametric** framework for monitoring **heterogeneous** data streams with only **partial observations** available.

- **Problem formulation and assumptions**
  - Measurement of the M data streams at time t:
$$\boldsymbol{X}(t) = \left(X_1(t), X_2(t), \dots, X_M(t)\right)$$
  - Only m $(m \leq M)$ out of M data streams are "observable" at each time t.
  - When the process is in-control:
    - Data streams are heterogeneous
    - In-control mean of $\boldsymbol{X}(t)$ is $\boldsymbol{\mu} = \mathbf{0}$ and standard deviation is **1**
  - When the process is out-of-control:
    - Unknown mean of $\boldsymbol{X}(t)$: $\boldsymbol{\mu}' \neq \mathbf{0}$ at unknown time
    - The number of affected data streams is unknown.

First antirank indicator with partial observations

$\boldsymbol{X}(t)$	3.2	NA	-2.8	NA	-1.6	NA
$\xi(t)$	0	?	?	?	0	?

Challenge: $\xi(t)$ not well-defined!

Hierarchical Bayesian structure

$\boldsymbol{\theta}(t)$: true time-varying parameter

$$\xi(t) \sim Cat\big(\boldsymbol{\theta}(t)\big)$$

$\boldsymbol{X}(t)$

New observations

$$\boldsymbol{\theta}(t) | \xi(t) \sim Dir\big(\boldsymbol{\alpha}(t)\big),$$
$$\alpha_j(t) = \alpha_j(t-1) + \boxed{\mathbb{1}\{\xi(t) = 1\}}$$

$$\boldsymbol{\theta}(t) \sim Dir\big(\boldsymbol{\alpha}(t-1)\big)$$

$\boldsymbol{\alpha}(t-1)$: prior concentration parameter

$$\omega_j(t) = \begin{cases} \sum_{k \in O(t)} g_k, & \text{observable and } B_1(t) = j \\ 0, & \text{observable but } B_1(t) \neq j \\ g_j, & \text{unobservable} \end{cases}, j = 1,2,\cdots,M.$$

First antirank indicator with partial observations (con't)

$$\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}(t-1) + \boldsymbol{\omega}(t)$$

The $\boldsymbol{\theta}(t)$ can be estimated by

$$\widehat{\boldsymbol{\theta}}(t) = \frac{\boldsymbol{\alpha}(t)}{\sum_{j=1}^{M} \alpha_j(t)} = \frac{\boldsymbol{g} + \sum_{i=1}^{t} \widehat{\boldsymbol{\omega}}(i)}{t+1}$$

The first antirank indicator $\boldsymbol{\eta}(t)$ is constructed as follows:

$$\eta_j(t) = \begin{cases} \sum_{k \in \mathcal{O}(t)} \hat{\theta}_k(t), & \text{observable and } B_1(t) = j \\ 0, & \text{observable but } B_1(t) \neq j \\ \hat{\theta}_j(t), & \text{unobservable} \end{cases}, j = 1, 2, \cdots, M.$$

AiTS algorithm

1. Construct CUSUM statistics

Local statistics based on observations

Behaves like the expectation of $\boldsymbol{S}_t^{(1)}$

$$\begin{cases} \boldsymbol{S}_t^{(1)} = 0, \qquad \boldsymbol{S}_t^{(2)} = 0 & \text{if } C_t \le k, \\ \boxed{\boldsymbol{S}_t^{(1)}} = \left(\boldsymbol{S}_{t-1}^{(1)} + \boldsymbol{\eta}(t)\right)(C_t - k)/C_t & \\ \boxed{\boldsymbol{S}_t^{(2)}} = \left(\boldsymbol{S}_{t-1}^{(2)} + \boldsymbol{g}\right)(C_t - k)/C_t & \text{if } C_t > k. \end{cases}$$

$$C_t = \left(\boldsymbol{S}_{t-1}^{(1)} - \boldsymbol{S}_{t-1}^{(2)} + \boldsymbol{\eta}(t) - \boldsymbol{g}\right)' \cdot diag\left(\frac{1}{S_{t-1,1}^{(2)} + g_1}, \dots, \frac{1}{S_{t-1,M}^{(2)} + g_M}\right) \cdot \left(\boldsymbol{S}_{t-1}^{(1)} - \boldsymbol{S}_{t-1}^{(2)} + \boldsymbol{\eta}(t) - \boldsymbol{g}\right)$$

where $\boldsymbol{S}_0^{(1)} = \boldsymbol{S}_0^{(2)} = \boldsymbol{0}$, k is a constant.

2. Stopping time

$$y_t = \left(\boldsymbol{S}_t^{(1)} - \boldsymbol{S}_t^{(2)}\right)' diag\left(\frac{1}{S_{1,t}^{(2)}}, \dots \frac{1}{S_{M,t}^{(2)}}\right)\left(\boldsymbol{S}_t^{(1)} - \boldsymbol{S}_t^{(2)}\right).$$

Measures the difference between $\boldsymbol{S}_t^{(1)}$ and $\boldsymbol{S}_t^{(2)}$

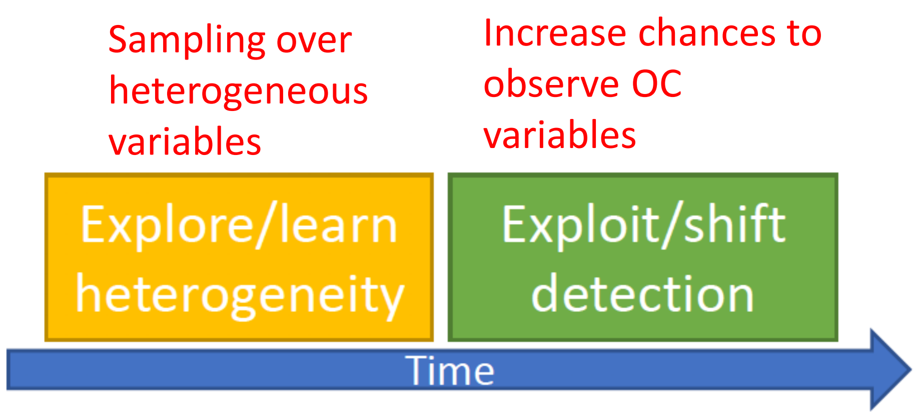
Stop the monitoring process when $y_t > h$.

AiTS algorithm (con't)

3. Sampling strategy

1) Draw a sample $\boldsymbol{Y} \sim Dir(N_1, N_2, \cdots, N_M)$, where $N_j = \frac{s_{j,t}^{(1)}}{\sum_{k=1}^{M} s_{k,t}^{(1)}}$ for $j = 1, 2, \cdots, M$;

2) Let $j_{(l)}$ be the data stream index of the decreasing order of $(Y_1, Y_2, \cdots, Y_M)$ such that $Y_{j_{(1)}} \geq Y_{j_{(2)}} \geq \cdots \geq Y_{j_{(M)}}$. Then, observe $\{j_{(1)}, j_{(2)}, \cdots, j_{(m)}\}$ at time $t + 1$.



Main Theorem

Theorem: Under the null hypothesis $H_0^{(1)}$: $\mu_1 = \mu_2 = \cdots = \mu_M$, suppose that $k = 0$, given $\widehat{\boldsymbol{\theta}}(0) = \boldsymbol{\eta}(0) = \boldsymbol{g}$, if all data streams are independent, then $\mathbb{E}\left(\widehat{\boldsymbol{\theta}}(t)\right) = \mathbb{E}(\boldsymbol{\eta}(t)) = \boldsymbol{g}$ under the AiTS algorithm.

Proof: by induction.

- *Lemma 1*: $\mathbb{E}(\boldsymbol{\omega}(t)) = \boldsymbol{g}$ under $H_0^{(1)}$ given $\mathbb{E}(\boldsymbol{\eta}(t-1)) = \boldsymbol{g}$
- *Unbiasedness*: $\mathbb{E}(\boldsymbol{\alpha}(t)) = (t+1)\boldsymbol{g} \Rightarrow \mathbb{E}\left(\widehat{\boldsymbol{\theta}}(t)\right) = \boldsymbol{g}$
- Sampling property: $\mathbb{E}(\boldsymbol{\eta}(t)) = \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{\eta}(t)|\widehat{\boldsymbol{\theta}}(t)\right)\right) = \boldsymbol{g}$

Take-home message: with partial observations,

- $\hat{\theta}(t)$ is an accurate estimator of the underlying process
- Ensure the validity the antirank-based CUSUM framework and explain why we can effectively handle the heterogeneity among data streams

Simulation results

- **Competing algorithms**
  - The NAS algorithm (Xian *et al.*, 2018)

    *Assuming **partial observation** and **exchangeable***
  - The RS method

    *Assuming **random sampling** strategy*
  - The QH01 method (Qiu and Hawkins 2001)

    *Assuming **full** observations available*

- **Simulation settings**
  - $M = 6$, based on 10000 replications, $ARL_0 = 370$
  - Different magnitudes of mean shifts (δ =1, 2, 3)
  - Different choices of m (m = 2, 3, 4)

Simulation results (Case 1)

Parameters:

- AiTS method: k = 0.1
- NAS method: k = 0.2, Δ=0.13
- RS method: k = 0.1
- QH01 method: k = 0.05

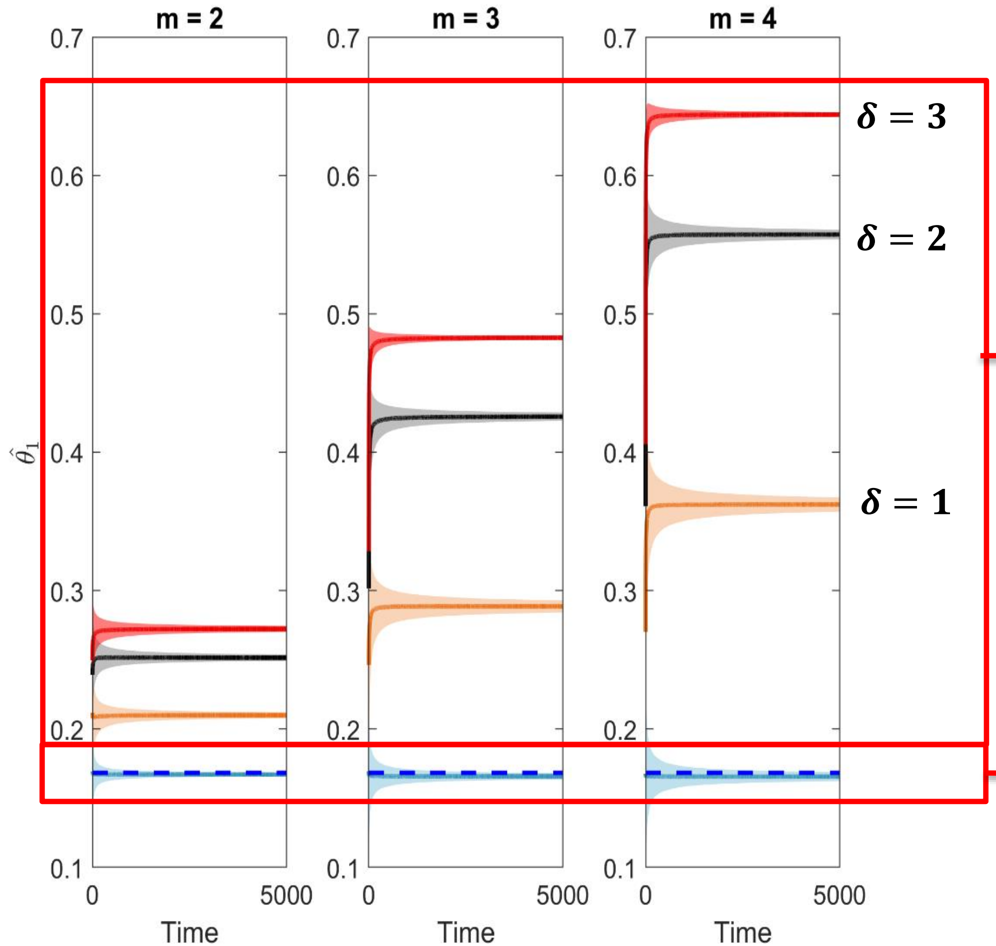
Half data streams are standard normal
Half data streams are POI(3)

$$\boxed{\boldsymbol{g} = [0.169, 0.169, 0.169, 0.164, 0.164, 0.164]}$$

Low-level heterogeneity

		AiTS	NAS	RS	QH01
$m = 2$	$\delta = 1.0$	31.91 (0.50)	39.66 (0.29)	62.33 (1.04)	9.90 (0.09)
	$\delta = 2.0$	11.85 (0.12)	16.00 (0.09)	21.42 (0.28)	3.81 (0.02)
	$\delta = 3.0$	9.06 (0.08)	12.69 (0.06)	15.49 (0.18)	2.81 (0.01)
$m = 3$	$\delta = 1.0$	17.53 (0.19)	26.47 (0.19)	23.34 (0.27)	9.90 (0.09)
	$\delta = 2.0$	6.11 (0.04)	9.18 (0.04)	8.89 (0.06)	3.81 (0.02)
	$\delta = 3.0$	4.52 (0.02)	6.80 (0.02)	6.89 (0.04)	2.81 (0.01)
$m = 4$	$\delta = 1.0$	12.90 (0.12)	20.87 (0.15)	16.26 (0.15)	9.90 (0.09)
	$\delta = 2.0$	4.82 (0.02)	6.86 (0.03)	6.32 (0.04)	3.81 (0.02)
	$\delta = 3.0$	3.63 (0.01)	5.06 (0.01)	4.85 (0.02)	2.81 (0.01)

Bayesian estimation when the shift happens at the first data stream



- Deviations from g_1 imply the capture of mean shift
- A larger m or δ leads to a larger $\hat{\theta}_1(t)$

- $\mathbb{E}\left(\hat{\theta}_1(t)\right) = g_1$
- A narrower confidence band as time goes

Simulation results (Case 2)

Parameters:

- AiTS method: k = 0.1
- NAS method: k = 0.2, Δ=0.12
- RS method: k = 0.1
- QH01 method: k = 0.05

$\boldsymbol{g} = [0.180, 0.127, 0.148, 0.181, 0.185, 0.179]$

Medium-level heterogeneity

- standard normal
- t(3)
- Exponential(1)
- χ^2_{10}
- POI(3)
- Binomial(10, 0.9)

		AiTS	NAS	RS	QH01
$m = 2$	$\delta = 1.0$	42.75 (0.63)	199.99 (2.14)	130.13 (2.33)	11.25 (0.10)
	$\delta = 2.0$	13.96 (0.13)	84.64 (1.22)	35.62 (0.55)	3.89 (0.02)
	$\delta = 3.0$	10.89 (0.09)	61.76 (0.99)	25.00 (0.34)	2.92 (0.01)
$m = 3$	$\delta = 1.0$	18.16 (0.18)	38.93 (0.38)	25.17 (0.30)	11.25 (0.10)
	$\delta = 2.0$	6.71 (0.04)	11.48 (0.06)	9.53 (0.07)	3.89 (0.02)
	$\delta = 3.0$	5.22 (0.02)	8.74 (0.03)	7.53 (0.05)	2.92 (0.01)
$m = 4$	$\delta = 1.0$	12.74 (0.12)	25.79 (0.23)	16.20 (0.16)	11.25 (0.10)
	$\delta = 2.0$	4.72 (0.03)	7.54 (0.03)	6.47 (0.04)	3.89 (0.02)
	$\delta = 3.0$	3.63 (0.01)	5.62 (0.02)	4.96 (0.02)	2.92 (0.01)

Simulation results (Case 3)

Parameters:

- AiTS method: k = 0.1
- NAS method: k = 0.2, Δ=0.10
- RS method: k = 0.1
- QH01 method: k = 0.05

$$\boldsymbol{X}(t) \sim MVN\left(\begin{bmatrix}0\\0\\0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.8 & 0.4 & 0.3 & 0.2 & 0.1\\0.8 & 1 & 0.7 & 0.4 & 0.3 & 0.2\\0.4 & 0.7 & 1 & 0.6 & 0.4 & 0.3\\0.3 & 0.4 & 0.6 & 1 & 0.5 & 0.4\\0.2 & 0.3 & 0.4 & 0.5 & 1 & 0.4\\0.1 & 0.2 & 0.3 & 0.4 & 0.4 & 1\end{bmatrix}\right)$$

$$\boxed{\boldsymbol{g} = [0.191, 0.111, 0.140, 0.154, 0.186, 0.218]}$$

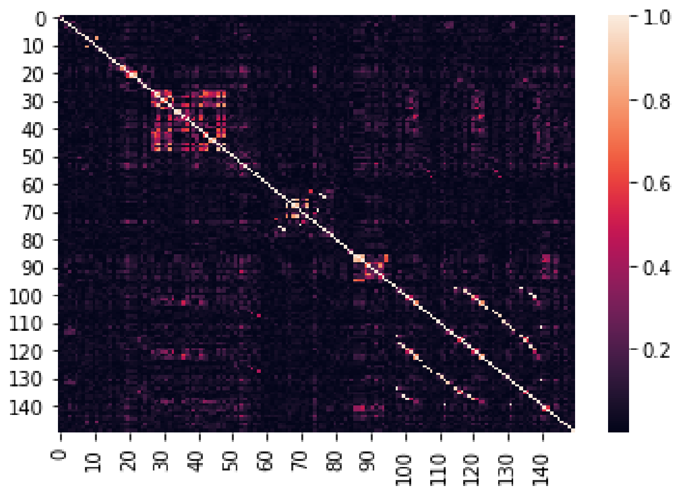
High-level heterogeneity

		AiTS	NAS	RS	QH01
$m = 2$	$\delta = 1.0$	45.77 (0.82)	>370	330.09 (7.59)	7.17 (0.06)
	$\delta = 2.0$	14.19 (0.12)	159.91 (1.52)	89.56 (2.23)	3.14 (0.01)
	$\delta = 3.0$	11.84 (0.09)	146.59 (1.50)	54.87 (1.19)	2.59 (0.01)
$m = 3$	$\delta = 1.0$	13.93 (0.14)	45.40 (0.79)	24.34 (0.32)	7.17 (0.06)
	$\delta = 2.0$	5.96 (0.03)	14.94 (0.32)	9.75 (0.07)	3.14 (0.01)
	$\delta = 3.0$	5.07 (0.02)	11.54 (0.05)	8.30 (0.05)	2.59 (0.01)
$m = 4$	$\delta = 1.0$	9.93 (0.08)	18.65 (0.13)	13.42 (0.12)	7.17 (0.06)
	$\delta = 2.0$	4.39 (0.02)	7.15 (0.03)	6.17 (0.04)	3.14 (0.01)
	$\delta = 3.0$	3.78 (0.01)	6.02 (0.02)	5.16 (0.02)	2.59 (0.01)

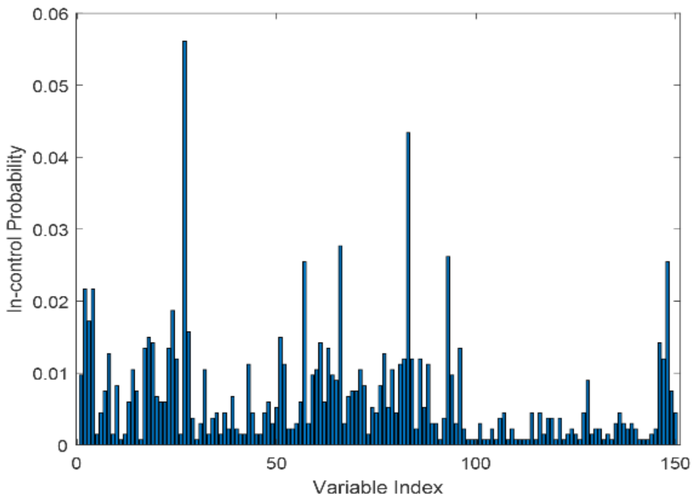
Case study: semiconductor process

- 1336 normal samples and 150 data streams
- m = 30 (20%) observable data streams
- The in-control ARL is set to be 370

High heterogeneous



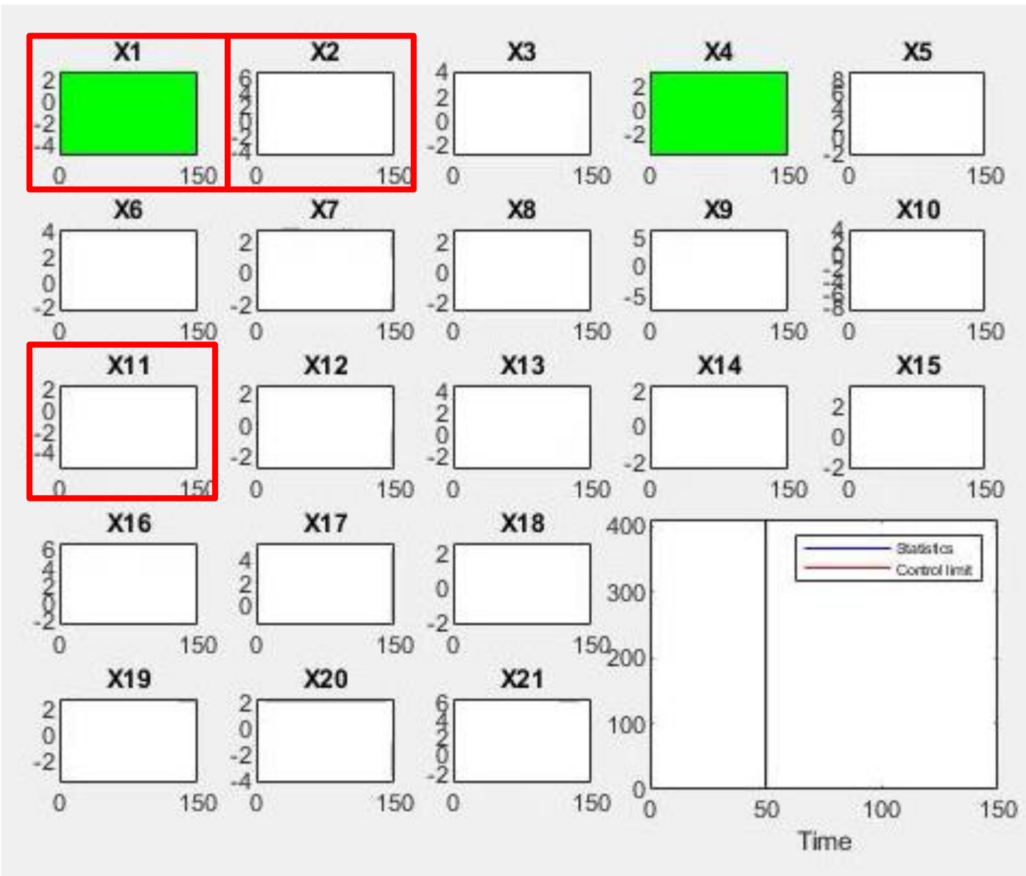
Correlation among 150 data streams



Estimated parameter $\boldsymbol{g}$ of the wafer data

Demo of AiTS method

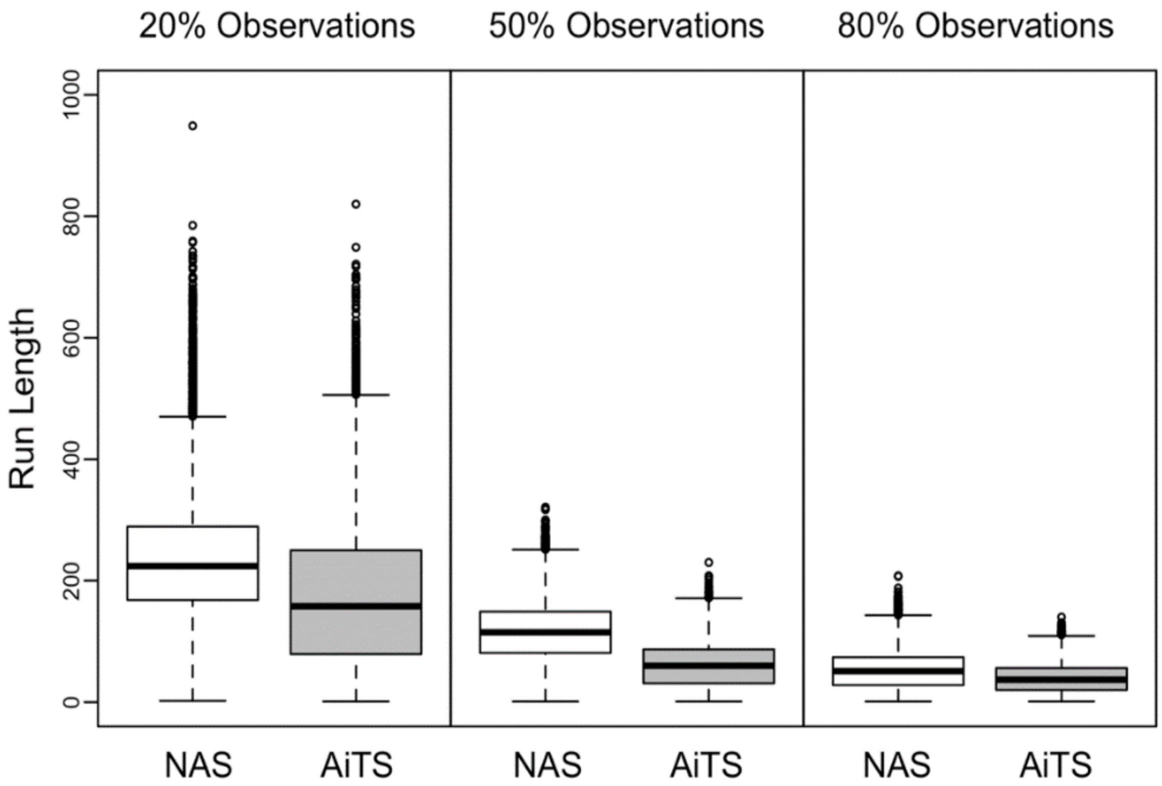
Observed Unobserved Out-of-control



The percentage of time being sampled

OC variables	Before shift	After shift
X1	30%	39%
X2	6%	77%
X11	26%	65%

Case study results



Conclusion

- The AiTS algorithm is proposed for real-time monitoring **heterogeneous data streams** with only **partial observation** of data streams.
- Theoretical justifications of this algorithm are investigated.
- Simulation and case study reveal the capability of AiTS to capture the heterogeneity and to quickly detect a wide range of possible mean shifts.

Outline

- Introduction
  - What is the problem?
- Research Topics
  - An adaptive sampling strategy for online high-dimensional process monitoring (2015)
  - A Nonparametric Adaptive Sampling Strategy for Online Monitoring of Big Data Streams (2018)
  - Online Nonparametric Monitoring of Heterogeneous Data Streams with Partial Observations based on Thompson Sampling (2022)
- **Summary**

Sensor Measurement and Monitoring Strategy

- A Top-r based Adaptive Sampling Strategy: Online monitor normally distributed big data streams in the context of limited resources
- A Nonparametric Adaptive Sampling Strategy: Online monitor non-normal (exchangeable) big data streams in the context of limited resources
- Online Nonparametric Monitoring of Heterogeneous Data Streams: Online monitor arbitrarily distributed big data streams in the context of limited resources by Thompson sampling
- Effective Online Data Monitoring and Saving Strategy: intelligently select and record the most informative extreme values in the simulation data
- A Spatial Adaptive Sampling Procedure: leverage the spatial information and adaptively and intelligently integrate two seemingly contradictory ideas (Wide and deep searches)
- A Rank-based Sampling Algorithm by Data Augmentation: automatically augment information for unobservable variables based on the online observations
- Online Nonparametric Monitoring and Sampling for High-Dimensional Heterogeneous Processes: Seamlessly integrate the Thompson sampling (TS) algorithm with a quantile-based nonparametric cumulative sum (CUSUM) procedure

Publication

- **Liu, K.** and Shi, J. (2013), "Objective-Oriented Optimal Sensor Allocation Strategy for Process Monitoring and Diagnosis by Multivariate Analysis in a Bayesian Network", *IIE Transactions*, 45, 630–643.
- Jin, R. and **Liu, K.** (2013), "Multistage Multimode Process Monitoring Based on a Piecewise Linear Regression Tree Considering Modeling Uncertainty", *IIE Transactions*, 45, 617-629.
- **Liu, K.**, Zhang, X., and Shi, J. (2014), "Adaptive Sensor Allocation Strategy for Process Monitoring and Diagnosis in a Bayesian Network", *IEEE Transactions on Automation Science and Engineering*, 11, 2, 452-462.
- **Liu, K.**, Mei, Y., and Shi, J. (2015), "An adaptive sampling strategy for online high-dimensional process monitoring", *Technometrics*, 57, 3, 305-319.
- Zhou, C., **Liu, K.**, Zhang, X., Zhang, W., and Shi, J. (2016), "An automatic process monitoring method using recurrence plot in progressive stamping processes", *IEEE Transactions on Automation Science and Engineering*, 13, 2, 1102 - 1111.
- Song, C., **Liu, K.**, Zhang, X., Chen, L., and Xian, X. (2016), "An Obstructive Sleep Apnea Detection Approach Using a Discriminative Hidden Markov Model from ECG Signals", *IEEE Transactions on Biomedical Engineering*, 63, 7, 1532 – 1542
- Li, J., **Liu, K.**, and Xian, X. (2017), "Causation-based Process Monitoring and Diagnosis for Multivariate Categorical Processes", *IISE Transactions*, 49, 3, 332-343.
- Xian, X., Wang, A., and **Liu, K.** (2018), "A Nonparametric Adaptive Sampling Strategy for Online Monitoring of Big Data Streams", *Technometrics*, 60, 1, 14-25.
- Xian, X., Archibald, R., Mayer, B., **Liu, K.**, and Li, J. (2019), "An Effective Online Data Monitoring and Saving Strategy for Large-Scale Climate Simulations", *Quality Technology & Quantitative Management*, 16, 3, 330-346.
- Wang, A., Xian, X., Tsung, F., and **Liu, K.** (2018), "A Spatial Adaptive Sampling Procedure for Online Monitoring of Big Data Streams", *Journal of Quality Technology*, 50, 4, 329-343.
- Xian, X., Li, J., and **Liu, K.** (2019), "Causation-based Monitoring and Diagnosis for Multivariate Categorical Processes with Ordinal Information", *IEEE Transactions on Automation Science and Engineering*, 16, 2, 886-897.
- Feng, T., Qian, X., **Liu, K.**, Huang, S. (2019), "Dynamic Inspection of Latent Variables in State-Space Systems", *IEEE Transactions on Automation Science and Engineering*, 16, 3, 1232-1243.
- Xian, X., Zhang, C., Bonk, S., and **Liu, K.** (2021), "Online Monitoring of Big Data Streams: A Rank-based Sampling Algorithm by Data Augmentation", *Journal of Quality Technology*, 53, 2, 135-153.
- Kim, M., Ou, E., Loh, P., Allen, T., Agasie, R., and **Liu, K.** (2020), "RNN-Based Online Anomaly Detection in Nuclear Reactors for Highly Imbalanced Datasets with Uncertainty", *Nuclear Engineering and Design*, 364, 110699.
- Ye, H., and **Liu, K.** (2021), "A generic online nonparametric monitoring and sampling strategy for high-dimensional heterogeneous processes", *IEEE Transactions on Automation Science and Engineering*, accepted.
- Ye, H., Xian, X., Cheng, J. C., Hable, B., Shannon, R. W., Elyaderani, M. K. and **Liu, K.** (2022), "Online Nonparametric Monitoring of Heterogeneous Data Streams with Partial Observations based on Thompson Sampling", *IISE Transactions*, accepted. (This paper received the Best Student Paper Finalist award in the QCRE Section of Industrial and Systems Engineering Research Conference (ISERC), 2020).

Thank you!

Questions?

Phone: (608) 890-3546
E-mail: kliu8@wisc.edu
Homepage: https://kaibo.ie.wisc.edu/index.html