Modern MDL Meets Data Mining Insight, Theory, and Practice

—Part IV—

Dynamic Setting

Kenji Yamanishi

Graduate School of Information Science and Technology, the University of Tokyo

August 4th 2019

KDD Tutorial

Part IV. Dynamic Setting

4.1. Change Detection with MDL Change Statistics
- 4.1.1. Change Detection
- 4.1.2. MDL Change Statistics
- 4.1.3. Sequential Gradual Change Detection
- 4.1.4. Adaptive Windowing

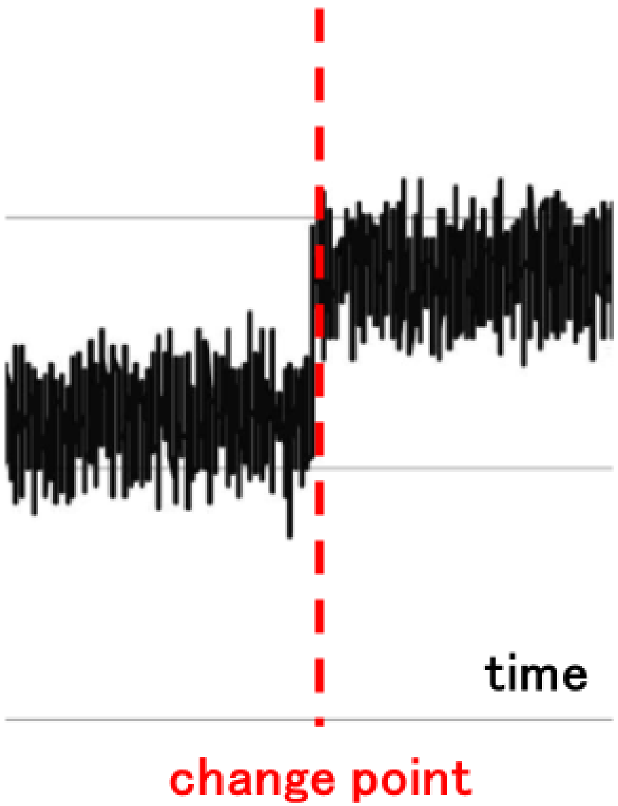
4.2. Model Change Detection with MDL Principle
- 4.2.1. MDL Model Change Statistics
- 4.2.2. Dynamic Model Selection
- 4.2.3. Clustering Change Detection
- 4.2.4. Model Change Sign Detection

4.1. Change Detection with MDL Change Statistics.

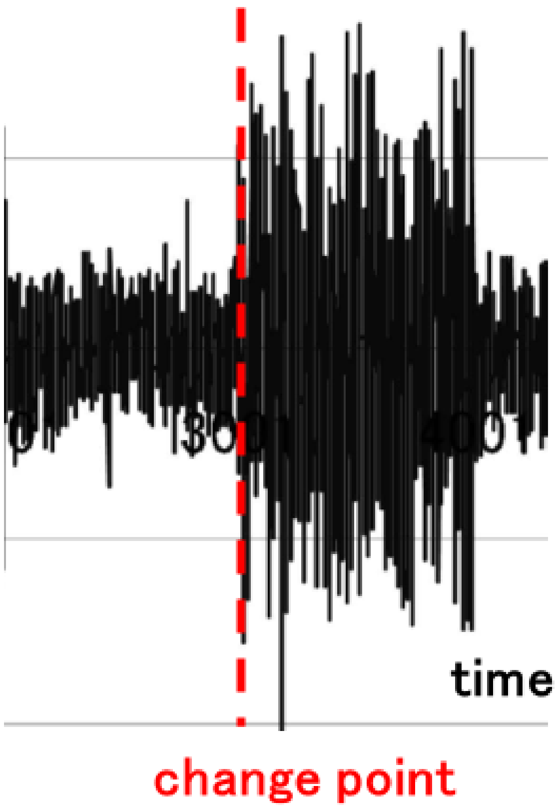
4.1.1 Change Detection

What's Change Detection?

Detecting emergence of bursts of anomalies

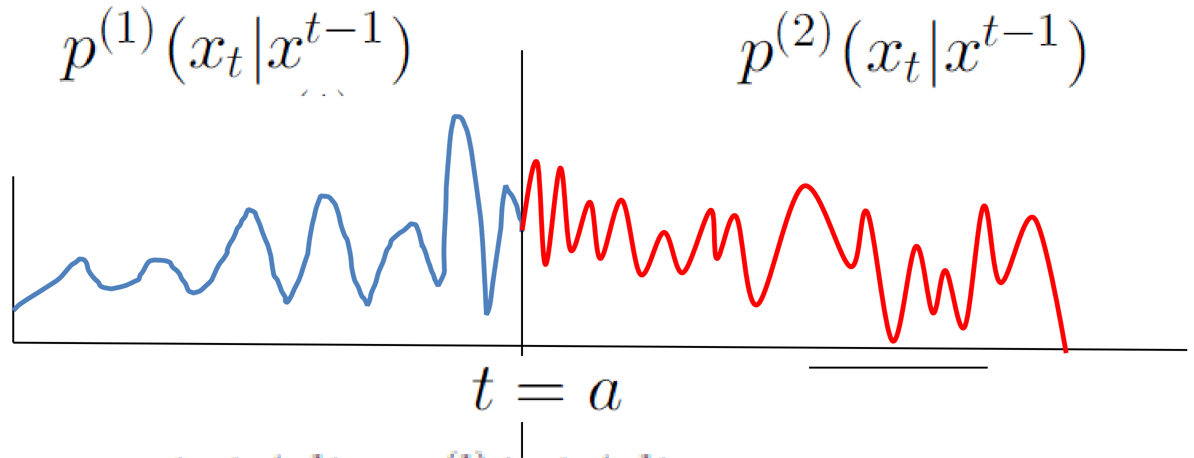


Mean change



Variance change

Definition of Change Point



$$p(x_t|x^{t-1}) = p^{(1)}(x_t|x^{t-1}) \qquad t < a,$$

$$p(x_t|x^{t-1}) = p^{(2)}(x_t|x^{t-1}), \qquad t \geq a.$$

t=a ：
change point

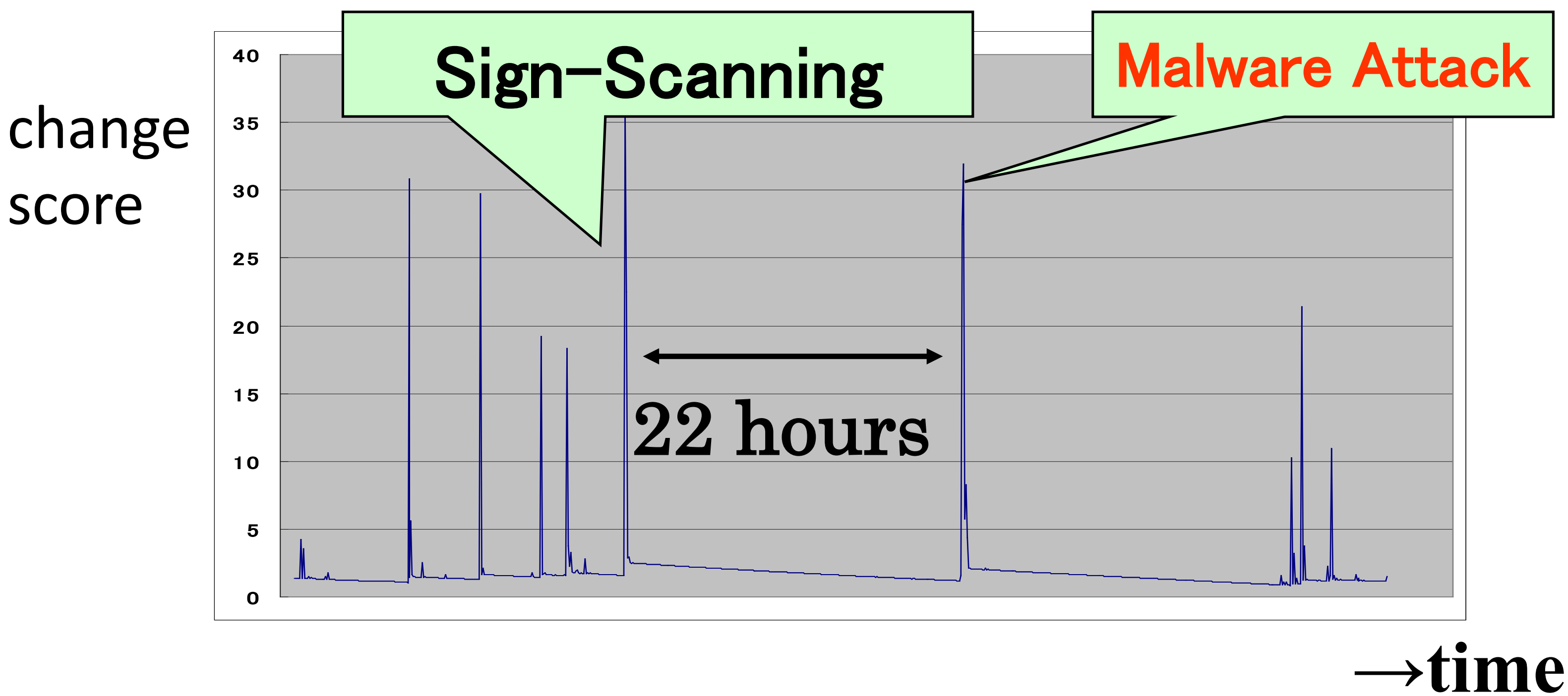
⇔ $D(p^{(2)}||p^{(1)}) \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} E_{p^{(2)}} \log \frac{p^{(2)}(x^n)}{p^{(1)}(x^n)},$

is large

Dissimilarity Measure＝
Kullback–Leibler divergence

Application to Malware Detection

Detecting SQL Injection via change point detection



Why Change Detection?

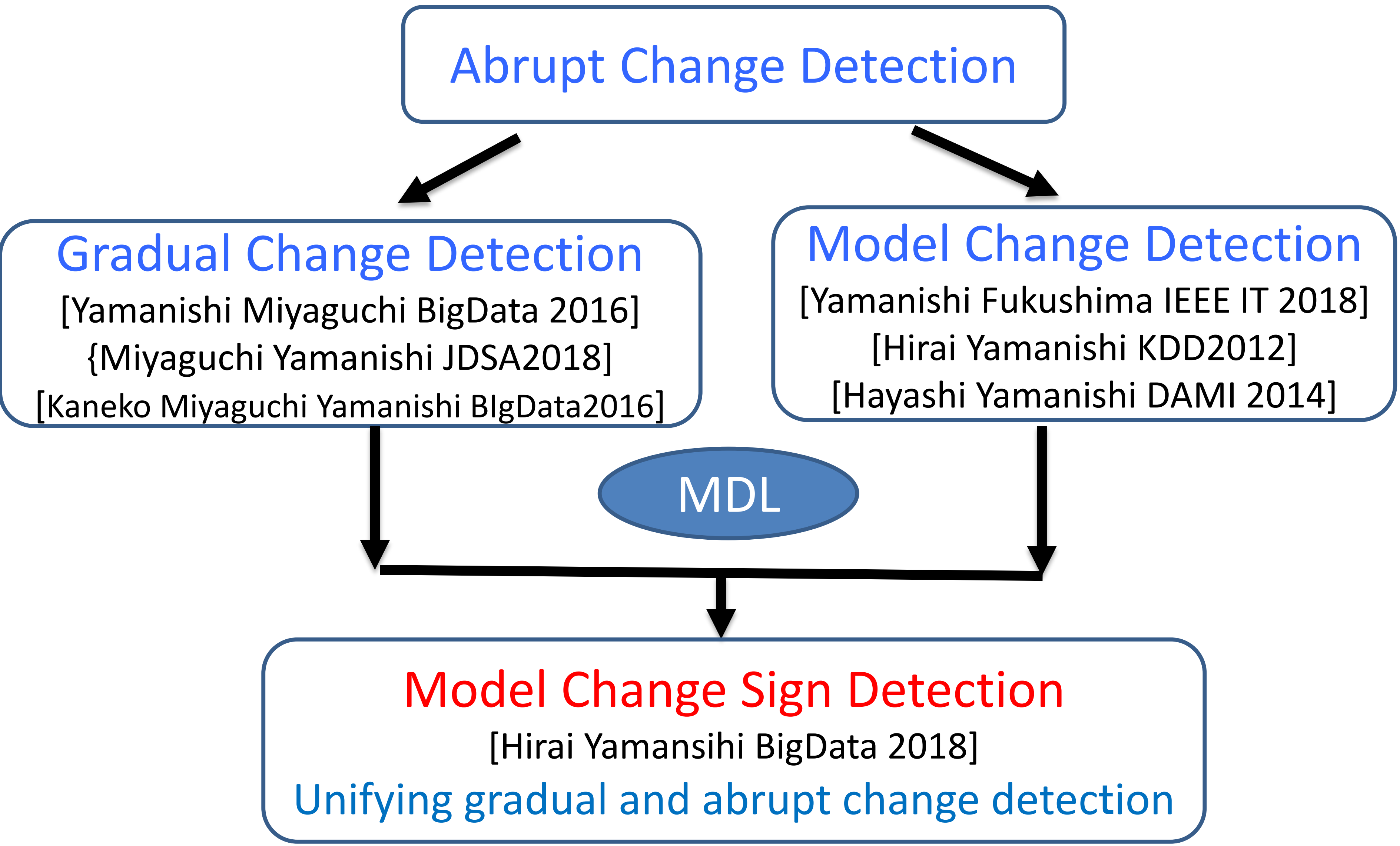
Time Series	Event behind change
Access log	Malware
Computer usage log	Fraud
Syslog	Failure
Sensor data	Accident
Tweet	Topic Emergence
Real estate transaction	Economics crisis
Usage transaction	Market trend
Visual field loss	Glaucoma

Previous Work

- Abrupt Change detection:
  [Hinkley 1970] [Hsu 1977][Basseville, Nikiforov 1993](CUSUM)
  [Guralnik, Srivastava 1998] [Fearnhead, Liu 2007]
- On-line abrupt change detection:
  [Yamanishi,Takeuchi 2002] [Kiefer et al.2004]
  [Takeuchi, Yamanishi 2006] [Adams,MacKay 2007]
- Incremental change detection（Concept drift）
  [Zliobaite 2009] [Gama et al. 2013]
- Continuous change detection
  [Miyaguchi, Yamanishi 2015] [Yamanishi Miyaguchi 2016]

No studies on unifying approaches to detecting gradual changes as well as abrupt ones

New Directions of Change Detection



4.1.2 MDL Change Statistics

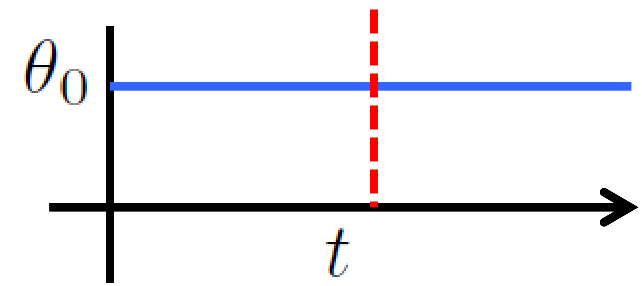
Hypothesis Testing Framework

$$\mathcal{F} = \{p(X;\theta) : \theta \in \Theta\}.$$

parametric class of prob. densities

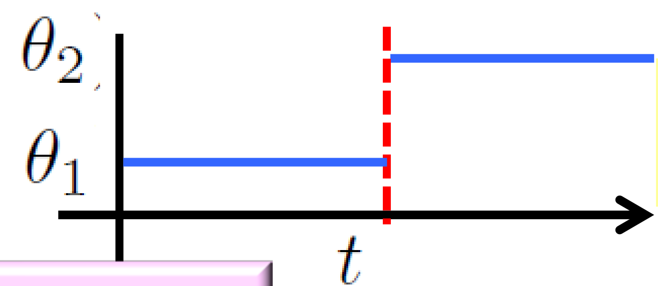
For sample size n, given a time point $t (1 \leq t \leq n)$,

$$H_0 : x_1^n \sim p(x;\theta_0)$$



t is not change pt

$$H_1 : x_1^t \sim p(x_1^t;\theta_1),$$
$$x_{t+1}^n \sim p(x_{t+1}^n;\theta_2)$$



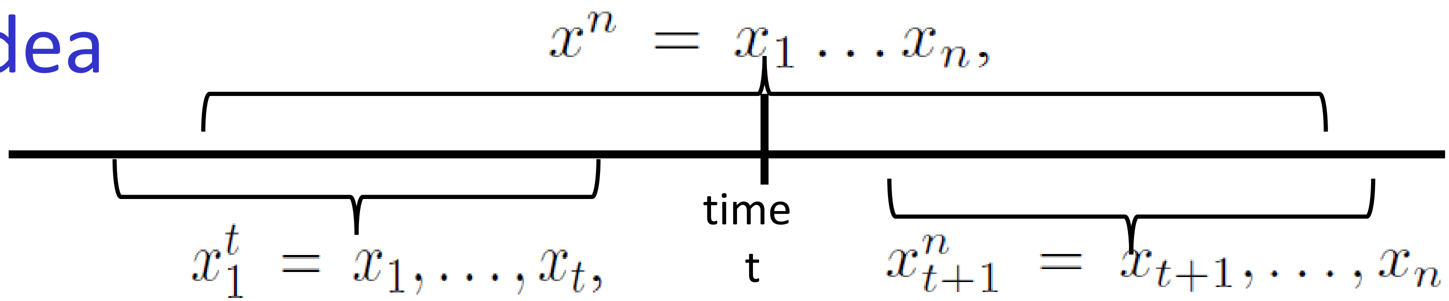
t is change pt

Likelihood test cannot be applied

$\theta_0, \theta_1, \theta_2$ are un… is specified.

MDL Change Statistics

Basic Idea



$$\mathcal{F} = \{p(X;\theta) : \theta \in \Theta\}.$$

$$\text{codelength for } x^n \text{ using } \mathcal{F}$$
$$-\{\text{codelength for } x_1^t \text{ using } \mathcal{F}$$
$$+\,\text{codelength for } x_{t+1}^n \text{ using } \mathcal{F}\}.$$

If the data can be compressed significantly more by changing the distribution at time *t*, then that point may be thought of as a change point.

C.f. [Yamanishi Miyaguchi BigData2016] [Vreeken Leeuwen DAMI2014] [Hooi et al. CIKM2018] [Guralnik and Srivastava KDD1999]

NML Codelength

Parametric model

$$\mathcal{P} = \{p(X^n;\theta):\ \theta \in \Theta\}\ (n = 1, 2, \dots)$$

NML Codelength

（Normalized Maximum Likelihood（NML） Codelength)

$$\begin{aligned}\mathcal{L}(x^n) &= -\log \frac{\max_\theta p(x^n;\theta)}{\sum_{y^n} \max_\theta p(y^n;\theta)} \\ &= -\log \max_\theta p(x^n;\theta) + \log \sum_{y^n} \max p(y^n;\theta) \\ &\approx -\log \max_\theta p(x^n;\theta) + \frac{k}{2}\log\frac{n}{2\pi} + \log \int \sqrt{|I(\theta)|}d\theta\end{aligned}$$

Parametric Complexity

$= C_n$

where 、 $I(\theta) = E_\theta[-\frac{\partial^2 \log p(X;\theta)}{\partial\theta\partial\theta^T}]$ (Fisher Information) k:# parameters

Formal Definition of MDL Change Statistics

MDL-change statistics [Yamanishi Miyaguchi BigData2016]

$x^n = x_1 \dots x_n, \;\; x_1^t = x_1 \dots x_t, \;\; x_{t+1}^n = x_{t+1} \dots x_n$

t: change point candidate, $\epsilon > 0$

NML Code-length for unchange

NML Code-length for change

$$\Phi_t(x^n) \stackrel{\text{def}}{=} \left\{ (-\log \max_{\theta} P(x_1^n;\theta)) + \log C_n \right\} - \left\{ (-\log \max_{\theta} P(x_1^t;\theta)) + \log C_t + (-\log \max_{\theta} P(x_{t+1}^n;\theta)) + \log C_{n-t} \right\} - n\epsilon$$

where $C_n = \sum_{x^n} \max_\theta P(x^n;\theta)$: parametric complexity

$\Phi_t(x^n) > 0 \Rightarrow$ t is a change point

$\Phi_t(x^n) \le 0 \Rightarrow$ t is not a change point

Performance Evaluation Metrics

The performance measure of hypothesis testing

Type I error probability:

=The probability that H_0 is true but H_1 is accepted.

(False alarm rate)

Type II error probability

=The probability that H_1 is true but H_0 is accepted.

(Overlooking rate)

Theoretical Performance of MDL-Test

Theorem 4.1.1（Error probabilities for MDL-test)

[Yamanishi Miyaguchi BigData2016]

$$\text{TypeI error prob.} \leq \exp\left[-n\left(\epsilon - \frac{\log C_n}{n}\right)\right],$$

（False alarm rate)

$$\text{TypeII error prob.} \leq \exp\left[-\frac{n}{2}\left(d_n(p_{\mathrm{NML}}, p_{\theta_1\theta_2}) - \frac{\log (C_t C_{n-t})}{n} - \epsilon\right)\right]$$

（Overlooking rate)

where

$$d_n(p,q) \stackrel{\text{def}}{=} 2\left\{1 - \left(\sum_{x^n}\{p(x^n)\}^{\frac{1}{2}}\{q(x^n)\}^{\frac{1}{2}}\right)^{\frac{1}{n}}\right\}$$

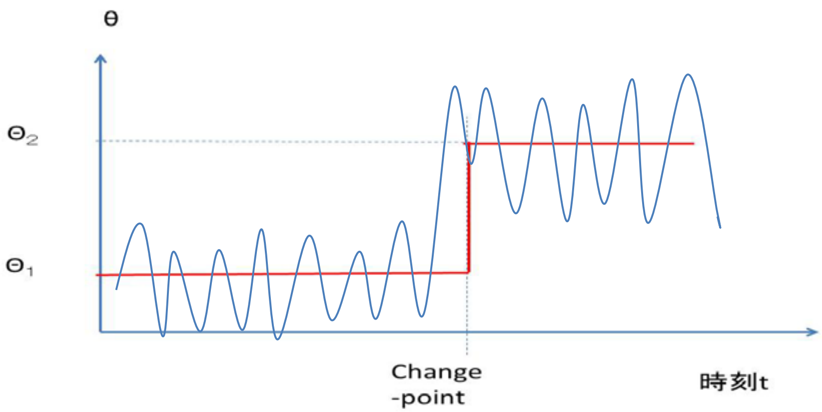
p_{NML} :NML distribution

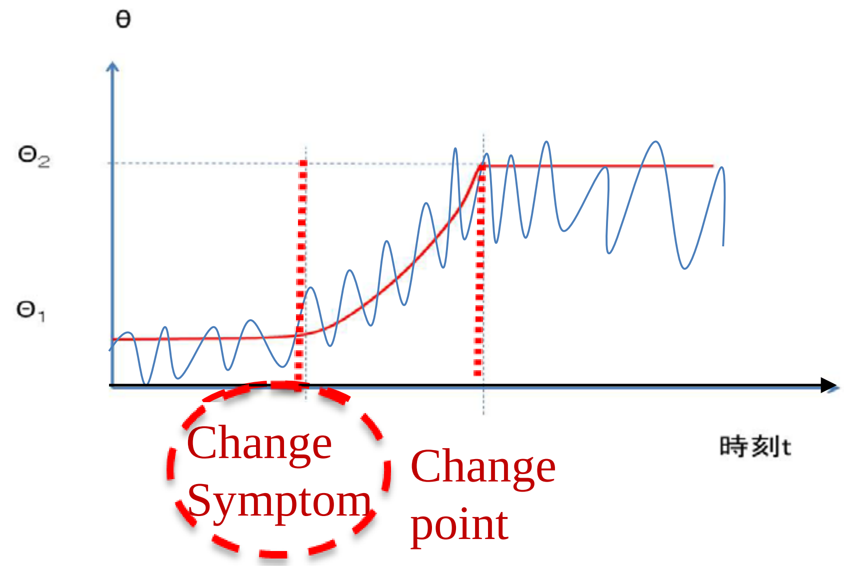
$$p_{\theta_1,\theta_2}(x^n) \stackrel{\text{def}}{=} p(x_1^t;\theta_1)p(x_{t+1}^n;\theta_2),$$

Error probabilities converge to zero exponentially with model complexity-based exponents.

4.1.3.Sequential Gradual Change Detection

Detecting change symptom from data stream





Abrupt change

⇒Conventional target

Gradual change

⇒Our new target

Challenges :

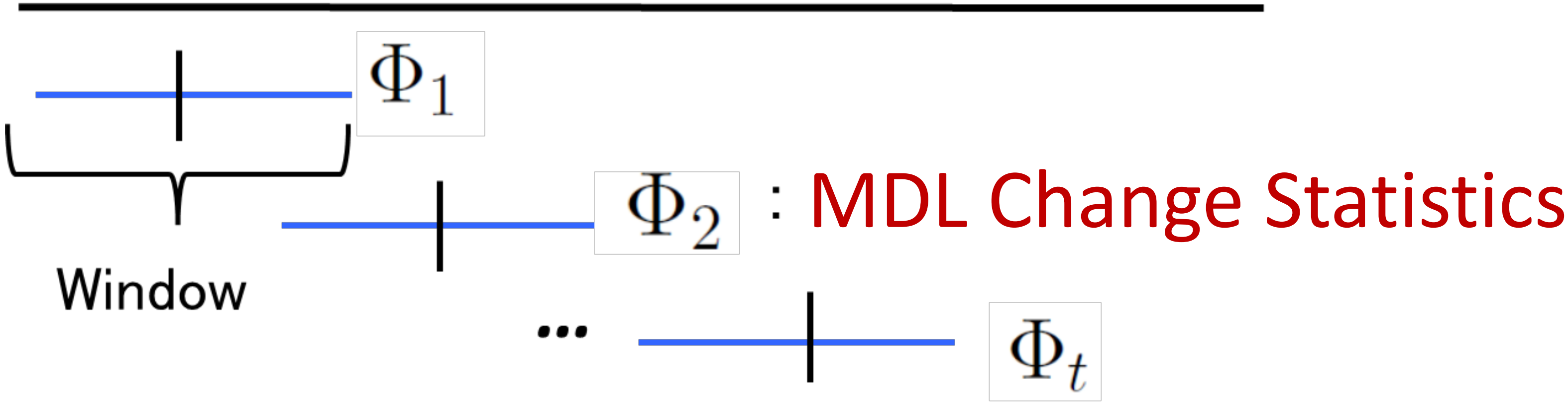
Real-time detection of sign of changes

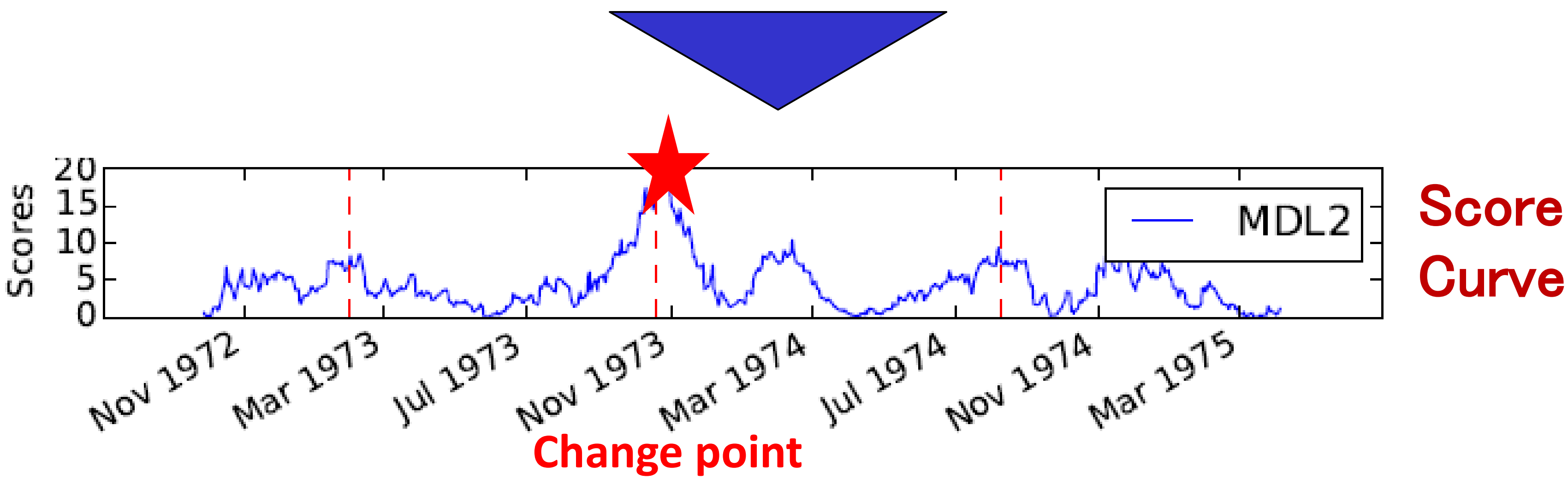
Sequential MDL Change Detection(S-MDL)

Sequentially compute MDL change statistics with fixed window

[Yamanishi, Miyaguchi BigData2016]

$$x^n = x_1 \dots x_n,$$





Sequential MDL Change Detection

Sequential variant

2h: window size

$$\Phi_t \stackrel{\text{def}}{=} \frac{1}{2h}\left\{\min_\theta(-\log P(x_{t-h+1}^{t+h};\theta)) + \log C_{2h}\right\} - \frac{1}{2h}\left\{\min_\theta(-\log P(x_{t-h+1}^{t};\theta)) + \min_\theta(-\log P(x_{t+1}^{t+h};\theta)) + 2\log C_h\right\},$$

Sequential MDL Change Detection Algorithm (S-MDL)

Given: h: window size, T: data length, $\mathcal{F}_M$: model class, β: threshold parameter

for all $t = h+1, \ldots, T-h+1$ **do**

Input $x_{t-h}, \ldots, x_{t+h}$.

Calculate a change score Φ_t at time t.

Make an alarm if and only if $\Phi_t > \beta$.

end for

Runs linearly in window size

Example 4.1.1. (Gaussian distributions)

$$\mathcal{F} = \left\{ P(X;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2} \right), \right.$$
$$\left. \theta = (\mu, \sigma^2) \in (-\mu_{\max}, +\mu_{\max}) \times (\sigma_{\min}, \sigma_{\max}) \right\},$$

where $\mu_{\max} < \infty,\ 0 < \sigma_{\min}, \sigma_{\max} < \infty$

MDL change statistics at time t:

$$\Phi_t = \frac{h}{2} \log \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1 \hat{\sigma}_2} + \log \frac{C_{2h}}{C_h^2},$$

where $\hat{\sigma}_0^2, \hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$: maximum likelihood (ML) estimators

$$\log C_k = \frac{1}{2} \log \frac{16|\mu|_{\max}}{\pi \sigma_{\min}^2} + \frac{k}{2} \log \frac{k}{2\mathrm{e}} - \log \Gamma\left(\frac{k-1}{2} \right)$$

Example 4.1.2. (Poisson distributions)

$$\mathcal{F} = \left\{ P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!},\ \lambda \in (0, \lambda_{\max}) \right\}$$

where $\lambda_{\max} < \infty$. is an upper bound on λ.

MDL change statistics at time t:

$$\Phi_t = -2h \log \left(\hat{\lambda}_h^{\hat{\lambda}_h} / \left(\hat{\lambda}_t^{\hat{\lambda}_t} \hat{\lambda}_{h-t}^{\hat{\lambda}_{n-t}} \right)^{1/2} \right) + \log \frac{C_{2h}}{C_h^2}$$

where $\hat{\lambda}_h$ is the ML estimator from x^n

$$\log C_k = \frac{1}{2} \log \frac{k}{2\pi} + \left(1 + \frac{\lambda_{\max}}{2} \right) \log 2 + \log^* \lambda_{\max}$$

Example 4.2.3. (Linear Regression)

$$X^n = (x_1, \ldots x_n)^\top = W_n^\top \beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

$$W_n = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 2 & 3 & \ldots & n \end{pmatrix}^\top \in \mathbb{R}^{n \times 2}, \ \beta \in \mathbb{R}^2$$

$$\mathcal{F} = \left\{ p(X^n; \theta) = \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{||X - W_n^\top \beta||^2}{2\sigma^2} \right) : \theta = (\beta, \sigma^2) \in \mathbb{R}^3, \ n = 1, 2, \ldots \right\}.$$

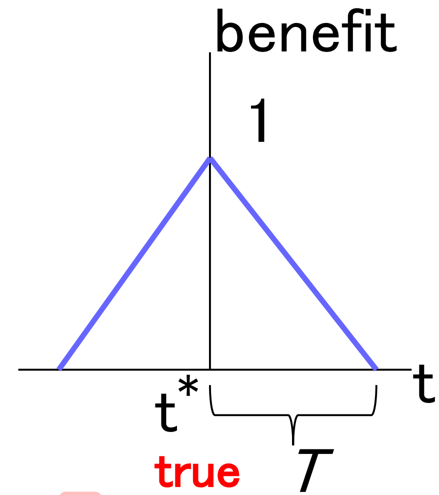
MDL change statistics at time t:

$$\Phi_t = h \log \frac{\hat{\sigma}_h^2}{\hat{\sigma}_t \hat{\sigma}_{h-t}} - \log \frac{R}{\sigma_{\min}^2} - \log \frac{\Gamma(h-1)}{\Gamma(h/2-1)^2} + h \log 2,$$

where $\sigma_{\min}$ and R are hyper-parameters $\hat{\sigma}_i^2 \geq \sigma_{\min}^2$ and $||\hat{\beta}|| \leq nR$.

$\hat{\sigma}_h^2$ $\hat{\sigma}_t^2, \hat{\sigma}_{h-t}^2$ the ML estimator of σ^2

Experiments: Synthetic Data



Evaluation metrics

- **Total Benefit** (How early)

$$b(t;t^*) = \begin{cases} 1 - \frac{|t-t^*|}{T} & (0 \le |t-t^*| < T), \\ 0 & (\text{otherwise}). \end{cases}$$

$$B_\beta(a_0^{n-1}) \stackrel{\text{def}}{=} \sum_{k=0}^{n-1} a_k b(k;t^*).$$

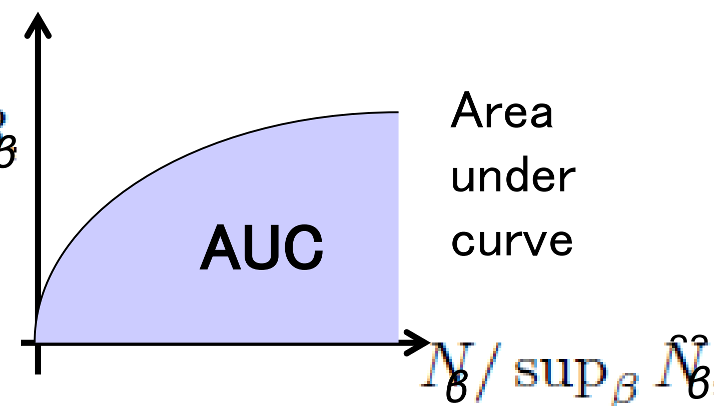
$$a_t \stackrel{\text{def}}{=} \begin{cases} 1 & (s_t > \beta), \\ 0 & (\text{otherwise}). \end{cases}$$

threshold

- **#False Alarms** (How reliably)

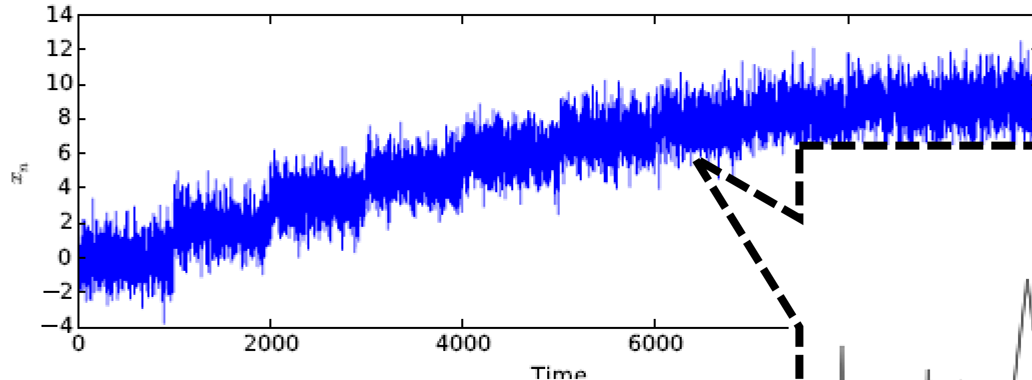
$$N_\beta(a_0^{n-1}) \stackrel{\text{def}}{=} \sum_{k=0}^{n-1} a_k \mathbb{I}(b(k,t^*) = 0),$$

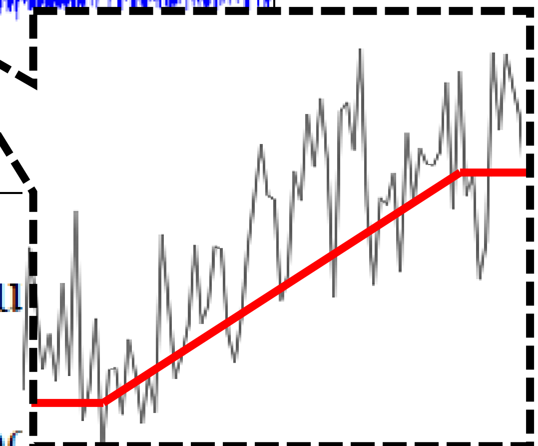
- **Performance Measure** $B_\beta / \sup_\beta B_\beta$



Experiments: Synthetic Data

Jumping means





Abrupt Change

datum was drawn from the Gaussian distribu

$$\mu_n = 0.6 \sum_{i=1}^{9} (10 - i) H(n - 1000i),$$

where $H(x)$ is the Heaviside step function that takes 1 if $x \geq 0$, otherwise 0

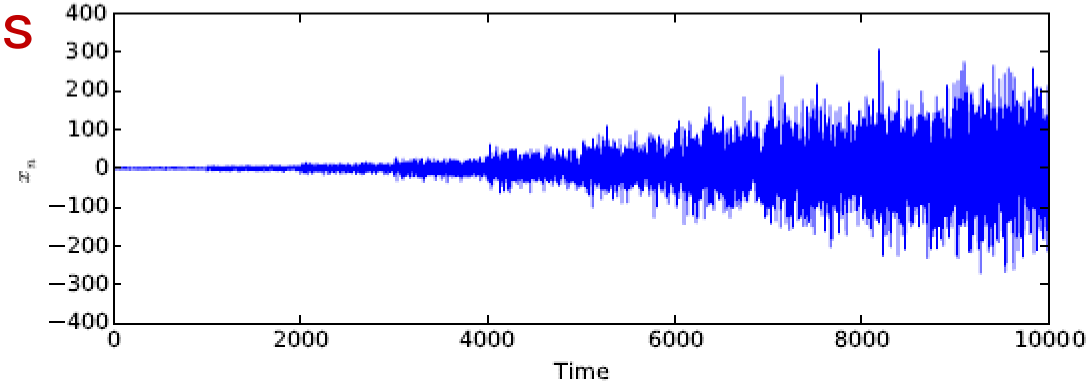
Gradual Change

replacing the step function H(·) with a slope function S(·) s.t.

$$S(x) = \begin{cases} 0 & (x < 0), \\ x/300 & (0 \leq x < 300), \\ 1 & (300 \leq x). \end{cases}$$

Experiments: Synthetic Data

Jumping variances



Abrupt Change

drawn from the Gaussian distribution $\mathcal{N}(0, \sigma_n^2)$ (Fig. 3) s.t.

$$\log \sigma_n = 0.3 \sum_{i=1}^{9} (10 - i) H(n - 1000i).$$

where $H(x)$ is the Heaviside step function that takes 1 if $x \geq 0$, otherwise 0

Gradual Change

replacing the step function H(·) with a slope function S(·) s.t.

$$S(x) = \begin{cases} 0 & (x < 0), \\ x/300 & (0 \leq x < 300), \\ 1 & (300 \leq x). \end{cases}$$

Experiments: Synthetic Data

[Yamanishi, Miyaguchi BigData2016]

Jumping means: AUC

Method	Abrupt	Gradual
IRL	0.467 ± 0.007	0.397 ± 0.014
CF	0.511 ± 0.015	0.495 ± 0.017
MDL1	$\mathbf{0.856 \pm 0.001}$	$\mathbf{0.654 \pm 0.001}$
MDL2	0.796 ± 0.019	0.646 ± 0.001

Jumping variances: AUC

Method	Abrupt	Gradual
IRL	0.514 ± 0.018	0.450 ± 0.021
CF	0.573 ± 0.008	0.493 ± 0.015
MDL1	0.721 ± 0.035	$\mathbf{0.718 \pm 0.035}$
MDL2	$\mathbf{0.731 \pm 0.034}$	0.711 ± 0.025

IRL: Inverse Run Length [Adams and MacKay 2007]

CF: ChangeFinder [Takeuchi and Yamanishi 2006]

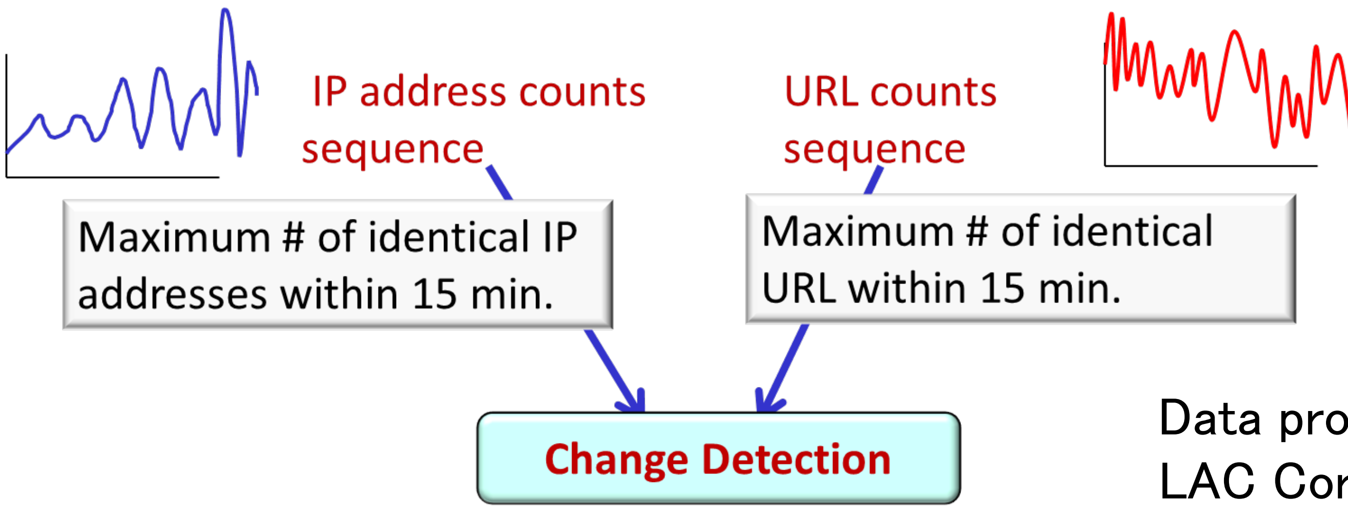
MDL1: Proposed method with independent Gaussian

MDL2: Proposed method with linear regression

Experiments: Real Data(Security)

[Yamanishi, Miyaguchi BigData2016]

SQL injection symptom detection



Data provided by LAC Corporation

■A time series of IP-URL counts, where each datum was the maximum # of total counts of records sent from an identical IP address to an identical URL within 15 minutes.

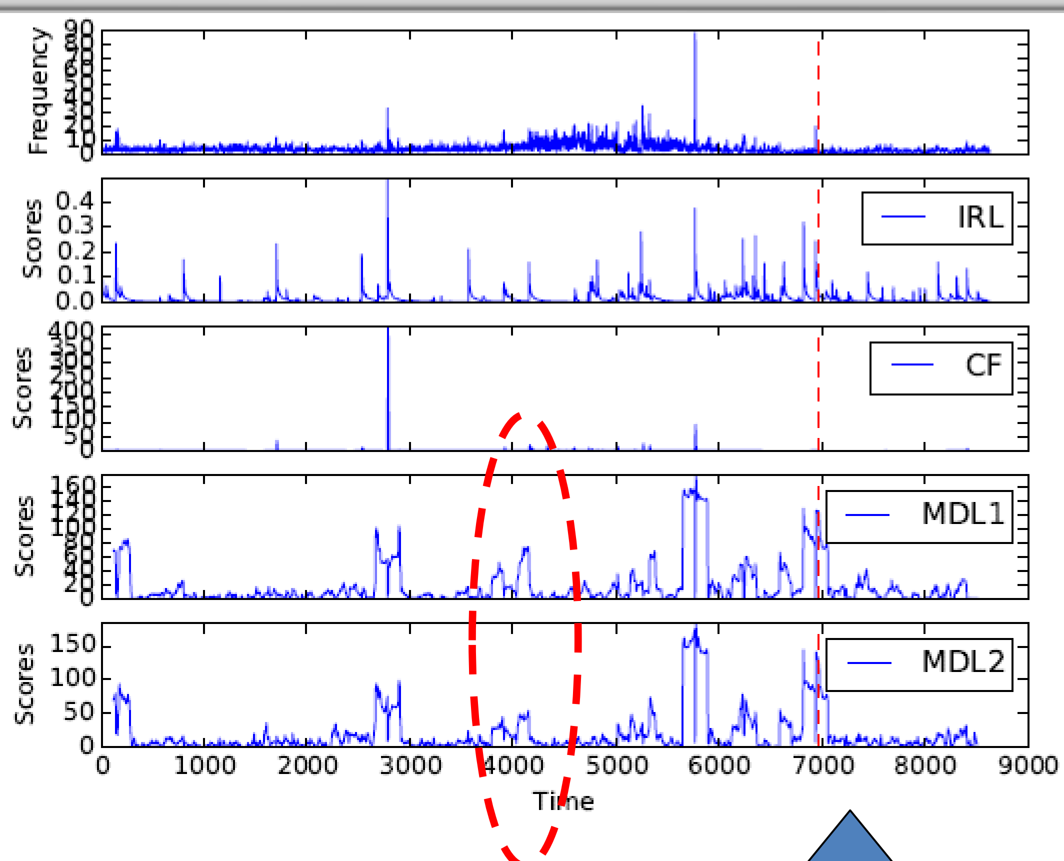
■Total records =8632

■MDL1 and MDL2 employ Poisson distributions

Experiments: Real Data

−SQL injection symptom detection−

Detected symptom caused by gradual increase of IP−URL accounts



Real symptom
security analysts confirmed

SQL injection
Attack

How do you choose window size?

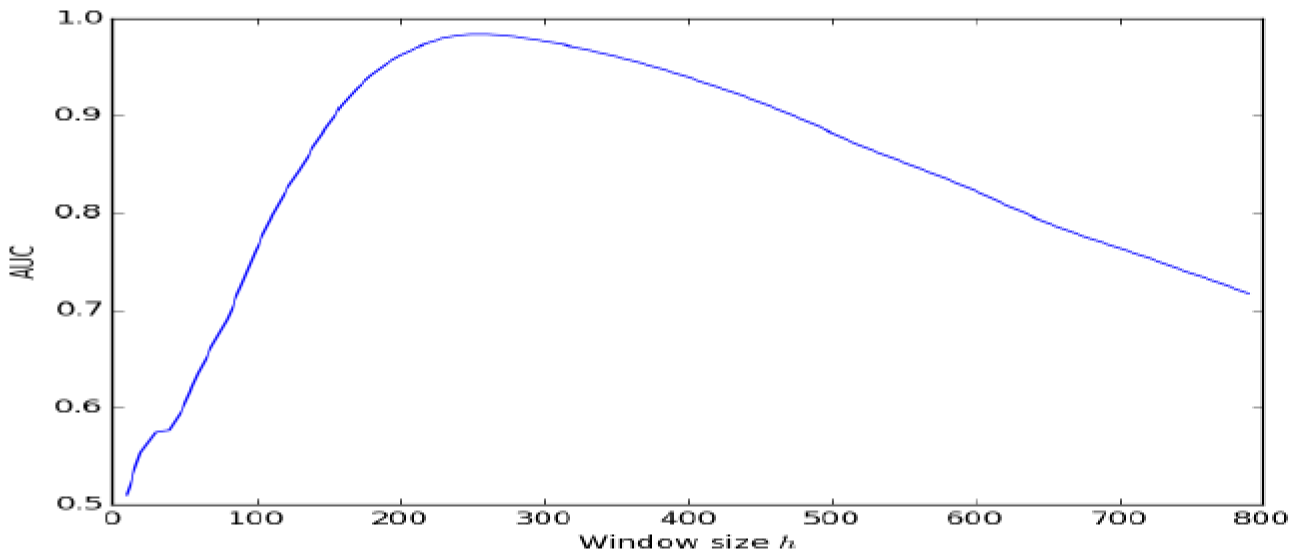


Figure 1. AUC vs window size

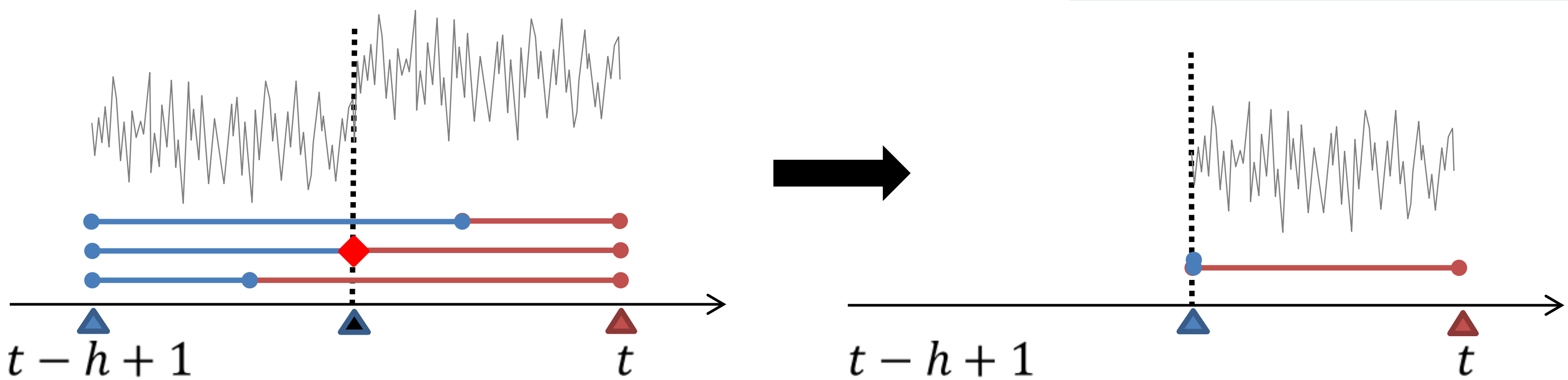
4.1.4. Adaptive Windowing

SCAW: Sequentially compute MDL change statistics with Adaptive Windowing (ADWIN) [Bifet & Gavaldà SDM07]

[Kaneko, Miyaguchi, Yamanishi BigData2017]

Compute statistics for all division points in the window

Determine window size



- If a statistics value exceeds threshold, it shrinks its window
  → no need to choose window size h heuristically
- Cost-saving version (ADWIN2)
  - Narrowing down the number of division points from $O(|W|)$ to $O(\log|W|)$

Asymptotic Reliability

Definition (Asymptotic reliability) *Algorithm $\mathcal{A}$ is asymptotically reliable if and only if, for all $\theta_0 \in \Theta$,*

$$X_1^\infty \sim p(x_1^\infty; \theta_0) \Rightarrow \lim_{n\to\infty} |\mathcal{T}_\mathcal{A}(X_1^n)| < \infty,$$

where $|\mathcal{T}_\mathcal{A}(X^n)|$ denotes the number of change points estimated by $\mathcal{A}$.

- Asymptotic reliability assures:
  "the number of false-alarms stays finite as the data size grows when the target process does *not* contain any changes."

Theorem 4.1.2 [Kaneko, Miyaguchi, Yamanishi BigData2017]

Let d be the dimension of a data sequence. Then, SCAW is asymptotically reliable if there exists a hyper parameter $\delta > 0$ that satisfies

$$\epsilon_h \geq \log\frac{1}{\delta} + (1+\delta+\frac{d}{2})\log h + const.$$

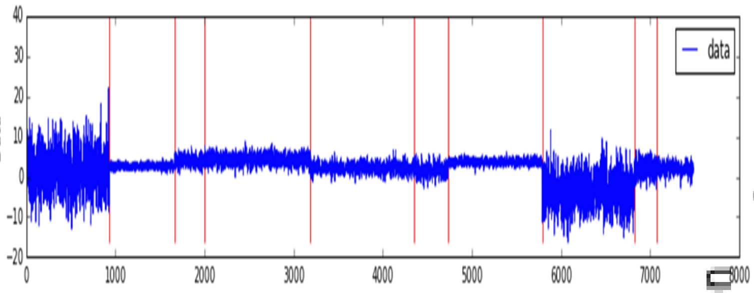
Threshold

Hyperparameter

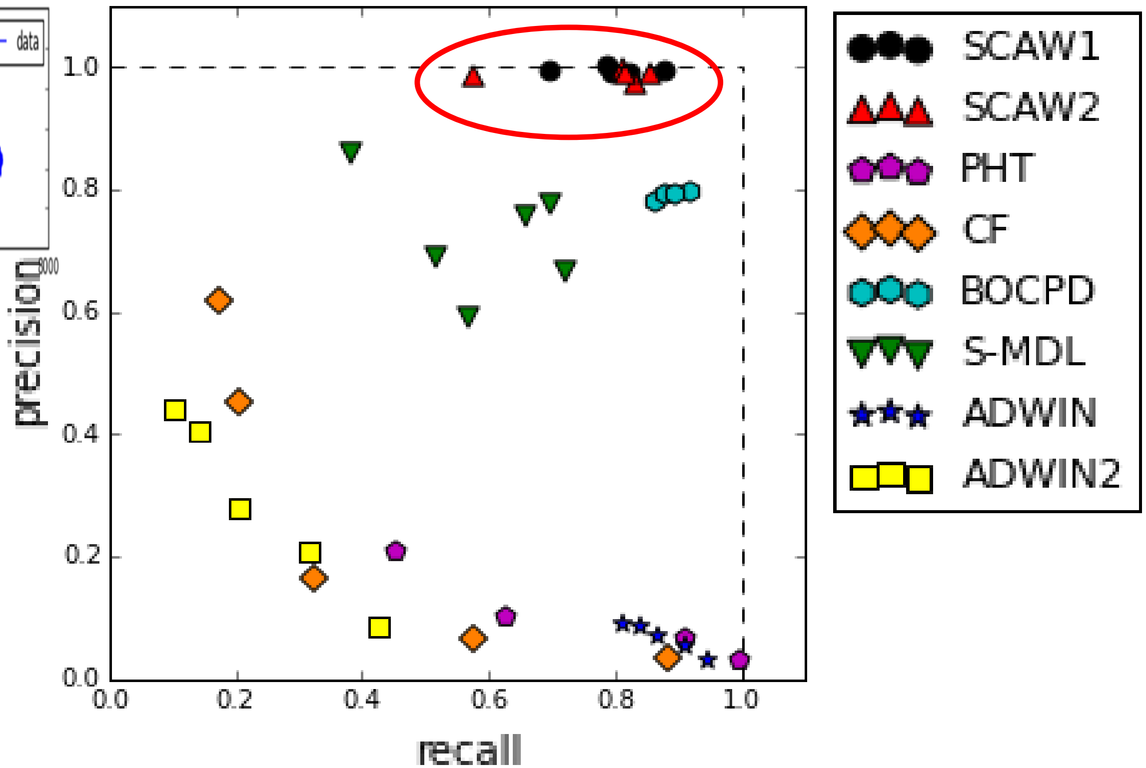
Experimental Result: Synthetic Data

SCAW achieves highest performance

[Kaneko, Miyaguchi. Yamanishi BigData2017]



- Gaussian distribution with different means and variances



- Precision-recall plots

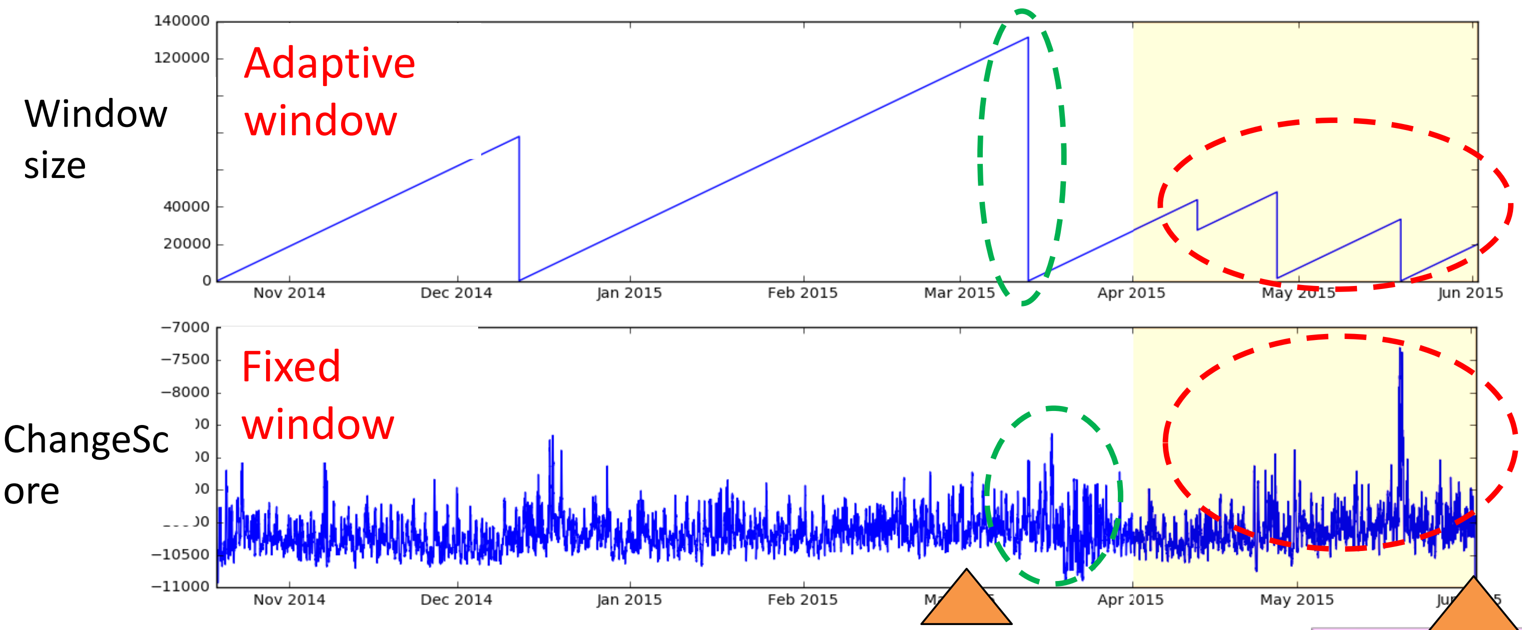
PHT: Page-Hinkley Test [Hinkle 70] ADWIN [Bifet & Gavaldà 07]
CF: ChangeFinder [Takeuchi & Yamanishi 06]
BOCPD: Bayesian online chnagepoint detection [Adams & MacKay 07]

Experimental Results: Real Data
—Failure Sign Detection—

Detected signs of real failures in an industrial boiler system

[Kaneko, Miyaguchi. Yamanishi BigData2017]

- Increase in the amount of an ingredient from early Apr. in 2015
- A temporary stop of the boiler system on Mar. 15^{th} in 2015



- time series data 217×325,440
- Data provided and evaluated by Toray Corp.

Signs of failures

Real Failure

SCAW is the better choice as a stream change detection

4.2. Model Change Detection with MDL Principle

Related Work

- Tracking Piecewise Stationary Sources

  [Shamir Merhav IEEE IT1999]

  [Killick, Fearnhead, Eckley JASA2012] [Davis, Yau EJS2013]

- Switching Distribution

  [Erven, Grunwald, Rooij JRoyalStat 2013]

- Tracking Best Experts / Derandomization

  [Herbster, Warmuth JML 1998] [Vovk ML99]

- Dynamic Model Selection

  [Yamanishi, Maruyama KDD2005, IEEE IT2007]

  [Davis, Lee, Rodriguez JASA 2006]

  [Hirai Yamanishi KDD2012] [Yamanishi Fukushima IEEE IT2019]

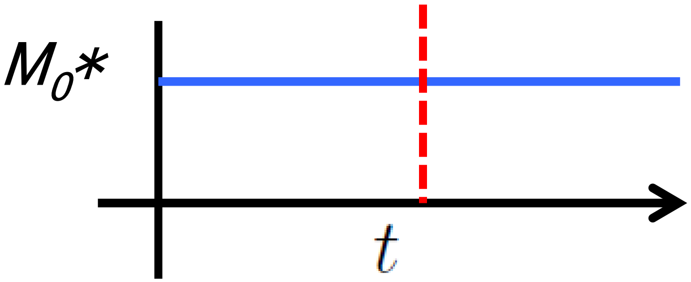
- Concept Drift

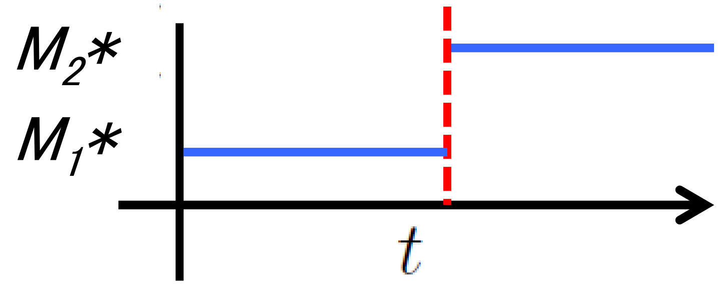
  [J. Gama, I. Zlibait, A. Bifet, M. Pechenizkiy, Bouchachia, ACM Survey 2013]

4.2.1. MDL Model Change Statistics

$$\mathcal{P} = \{p(X;\theta, M) : \theta \in \Theta_M,\ M \in \mathcal{M}\}$$

parameter model





$$H_0 : x_1^n \sim p(X_1^n; \theta_0^*, M_0^*),$$

$$H_1 : x_1^t \sim p(X_1^t; \theta_1^*, M_1^*),$$
$$x_{t+1}^n \sim p(X_{t+1}^n; \theta_2^*, M_2^*)$$

MDL-Change Statistics [Yamanishi Fukushima IEEE Inform Theory 2018]

$x^n = x_1 \ldots x_n$ $\quad t$: change point candidate

$$\Phi_t(x^n) \stackrel{\text{def}}{=} \min_M \{\mathcal{L}_{\text{NML}}(x^n; M) + \mathcal{L}(M)\}$$

NML codelength for unchange

$$- \min_{M_1, M_2} \left\{\mathcal{L}_{\text{NML}}(x_1^t; M_1) + \mathcal{L}_{\text{NML}}(x_{t+1}^n; M_2) + \mathcal{L}(M_1, M_2)\right\} - n\epsilon,$$

NML codelength for change

$$\mathcal{L}_{\text{NML}}(\boldsymbol{x}^n : M) = -\log \max_\theta p(\boldsymbol{x}^n : \theta, M) + \log C_n(M),$$

$$C_n(M) = \sum_{\boldsymbol{y}^n} \max_\theta p(\boldsymbol{y}^n; \theta, M)$$

Parametric Complexity

Theoretical Result on MDL-Test

MDL Test: $\Phi_t(x^n)$: MDL change statistics

$$\Phi_t(x^n) > 0 \Longrightarrow t \text{ is a change point}$$

$$\Phi_t(x^n) \leq 0 \Longrightarrow t \text{ is not a change point}$$

Theorem 4.1.3 [Yamanishi Fukushima IEEE Inform Theory 2018]

$$\text{Type I error prob.} \leq \exp\left[-n\left(\epsilon - \frac{\log C_n(M_0^*) + \mathcal{L}(M_0^*)}{n}\right)\right]$$

(False alarm prob.)

$$\text{Type II error prob.} \leq \exp\left(-nD_n^\alpha(M_1^*, M_2^*, \epsilon)\right)$$

(Overlooking prob.)

$$D_n^\alpha(M_1^*, M_2^*, \epsilon) \stackrel{\text{def}}{=} 2\alpha(1-\alpha)d_n^\alpha(\tilde{p}_{\text{NML}}, p_{M_{1*2}}) - \alpha\frac{\ell_n(M_1^*, M_2^*, \epsilon)}{n}$$

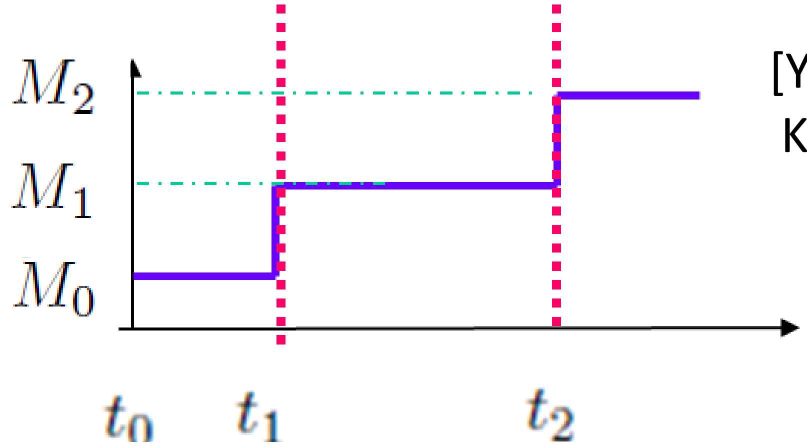
$$\ell_n(M_1^*, M_2^*, \epsilon) \stackrel{\text{def}}{=} \log C_t(M_1^*) + \log C_{n-t}(M_2^*) + \mathcal{L}(M_1^*, M_2^*) + \log \tilde{C}_n + n\epsilon.$$

Type I and II error probabilities converge exponentially to zero where exponents depend on parametric complexities

4.2.3. Dynamic Model Selection (DMS)

-Multiple model change detection-

Find a model sequence that minimizes total description length



[Yamanishi and Maruyama KDD2005, IEEE IT 2007]

$\mathcal{P} = \{P(X;\theta, M) : \theta \in \Theta_M, M \in \mathcal{M}\}$: Model class

DMS（Dynamic Model Selection）criterion

$$\sum_{t=1}^{T}(-\log P(x_t|x^{t-1};M_t)) + \sum_{t=1}^{T}(-\log P(M_t|M^{t-1})) \Longrightarrow \text{Min} \quad \text{w.r,t,} \quad M_1,\ldots,M_T$$

Predictive Codelength for data sequence

PredictiveCodelength for model sequence

Computable via Dynamic Programming

Probabilistic Setting of DMS

■ Predictive distribution for data sequence

$$P(x_t|x^{t-1};M_t) = P(x_t;\hat{\theta}(x^{t-1}),M_t) :$$ Maximum Likelihood Prediction

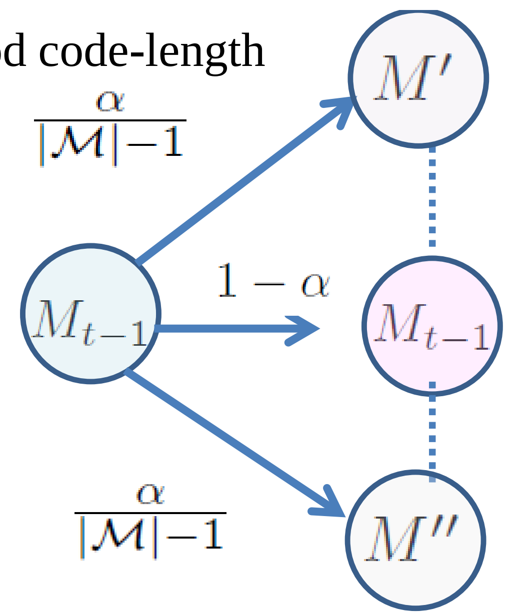
$$P(x_t|x^{t-1};M_t) = \int P(x_t;\theta,M_t)p(\theta|x^{t-1};M_t)d\theta :$$ Bayes Prediction

$$P(x_t|x^{t-1};M_t) = \frac{P(x_t|x^{t-1};\hat{\theta}(x_t\cdot x^{t-1}),M_t)}{\int P(X|x^{t-1};\hat{\theta}(X\cdot x^{t-1}),M_t)dX} :$$ SNML Prediction

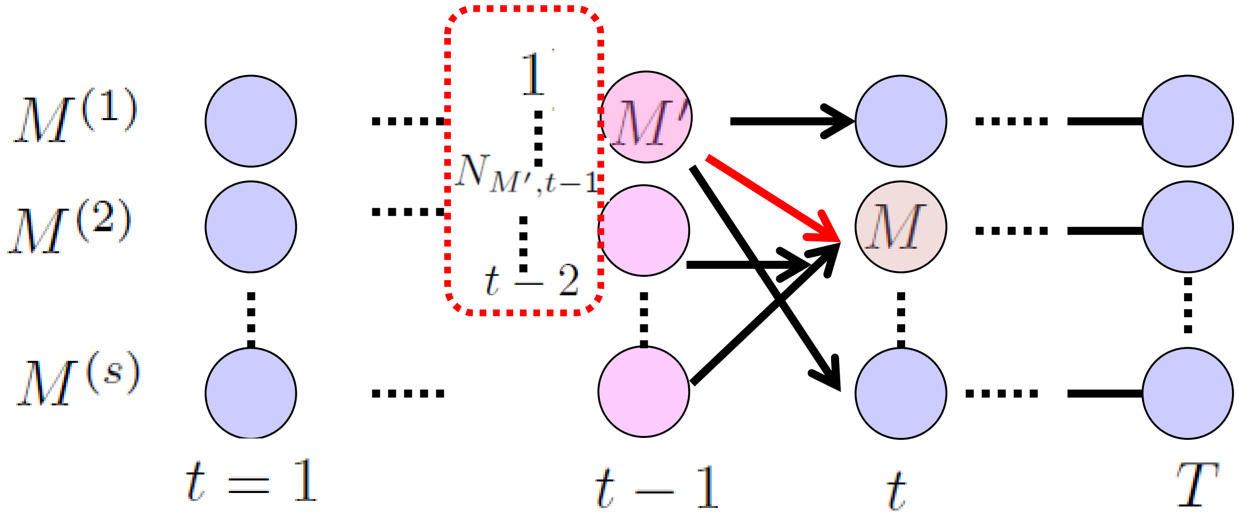
Sequentially normalized maximum likelihood code-length

■ Model transition probability

$$P(M_t|M^{t-1};\alpha) = \begin{cases} 1-\alpha & (M_t = M_{t-1}), \\ \frac{\alpha}{|\mathcal{M}|-1} & (M_t \neq M_{t-1}). \end{cases}$$



DMS Algorithm



1） Model sequence selection using dynamic programming

$$S(M, N_{M,t}, t) = \min_{M', N_{M',t-1}} \{S(M', N_{M',t-1}, t-1) - \log P(x_t|x^{t-1}, M_{t-1}) - \log P(M|M', \alpha(N_{M',t-1}))\}$$

$N_{M,t}$: # change points needed to be *M* at time *t*

2)Estimating model transition prob. via Krischevsky−Trofimov estimator

$$\alpha(N_{M,t}) = \frac{N_{M,t} + 1/2}{t}$$

Application to Failure Detection from Syslog

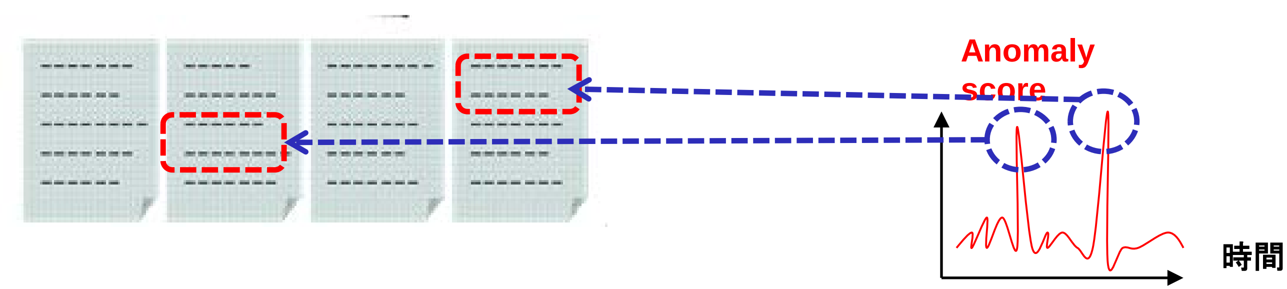
[Yamanishi, Maruyama KDD2005]

■What's Syslog?

- Event sequences collected with BSD syslog protocol
- Warning messages about devices

ID	Time stamp	Event Severity	Att1	Att2	Message
##	Nov 13 00:06:23:	ERR	bridge:	!brdgursrv:	queue is full. discarding a message.
##	Nov 13 10:15:00:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/1 Lock-Up!!
##	Nov 13 10:15:10:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/2 Lock-Up!!
##	Nov 13 10:15:20:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/3 Lock-Up!!

Detect failures early and identify their patterns



Syslog Modeling with HMM Mixtures

Syslog sessions are modeled with HMM mixtures

j-th session of syslog : $\mathbf{y}_j = (y_{j1}, \ldots, y_{jT_j})$

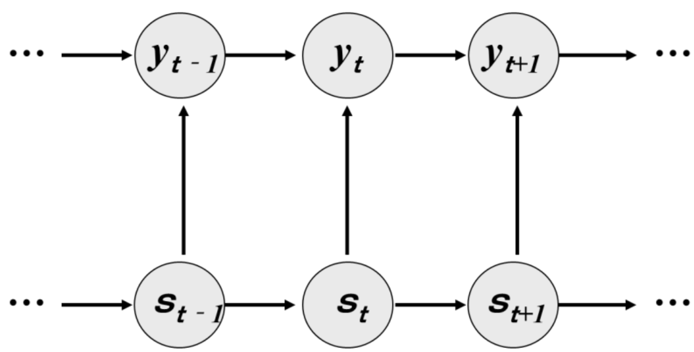
$$P(\boldsymbol{y}_j \mid \theta) = \sum_{k=1}^{K} \pi_k P_k(\boldsymbol{y}_j \mid \theta_k)$$

K: #syslog behavior patterns

where $$P_k(\boldsymbol{y}_j \mid \theta_k) = \sum_{(x_1, \ldots, x_{T_j})} \gamma_k(x_1) \prod_{t=1}^{T_j - 1} a_k(x_{t+1} \mid x_t) \prod_{t=1}^{T_j} b_k(y_t \mid x_t)$$

Syslog seq.

State seq..



T_j :session length

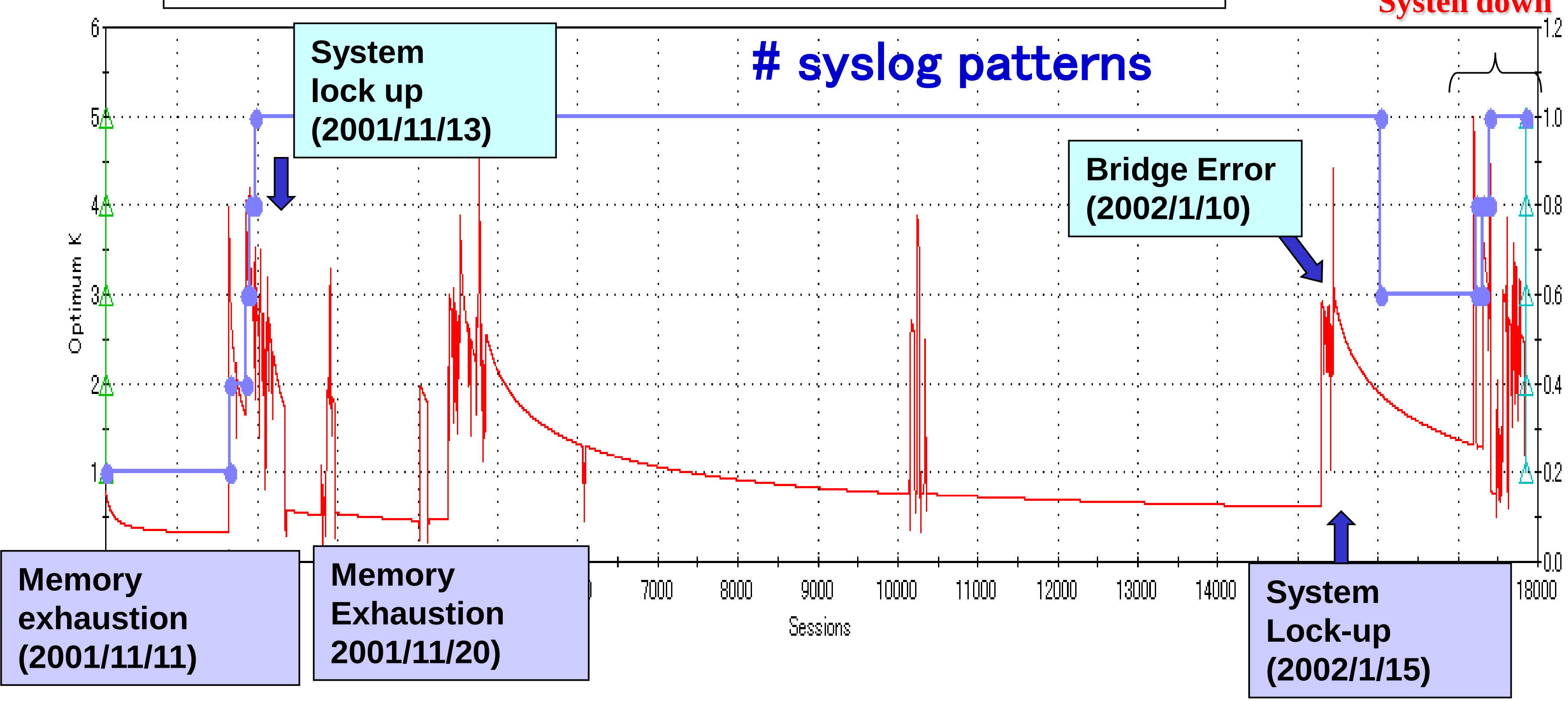
$(x_1, \ldots, x_{T_j})$:latent variables

Experiments: Failure Detection

#syslog patterns changed two days before system down.

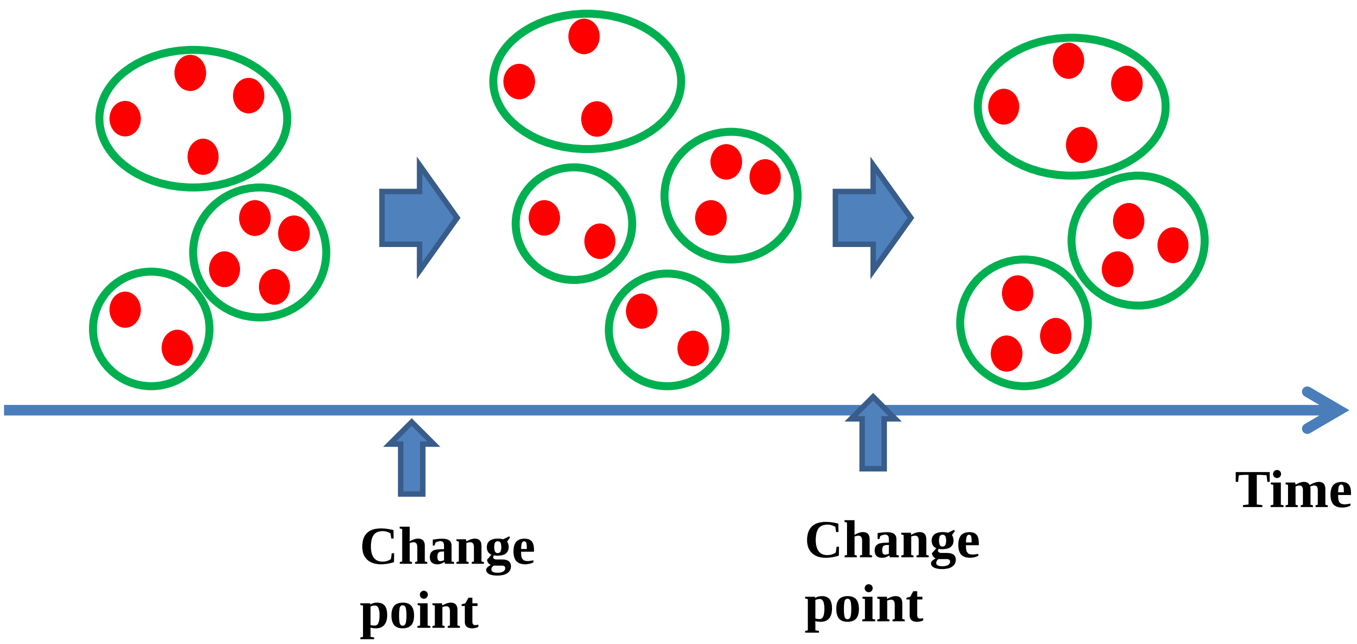
http://fbi-award.jp/sentan/jusyou/2005/nec.pdf

33025:Jan 15 15:03:59 WARN:swsig:sw_SigGetMem: alloc failed(256)
33026:Jan 15 15:03:59 WARN:swsig:sw_SigGetMem: alloc failed(256)
⋮
42253:Jan 19 22:26:33 ERR :bridge:!brdgursrv: queue is full. discarding a message.



4.2.3. Clustering Change Detection

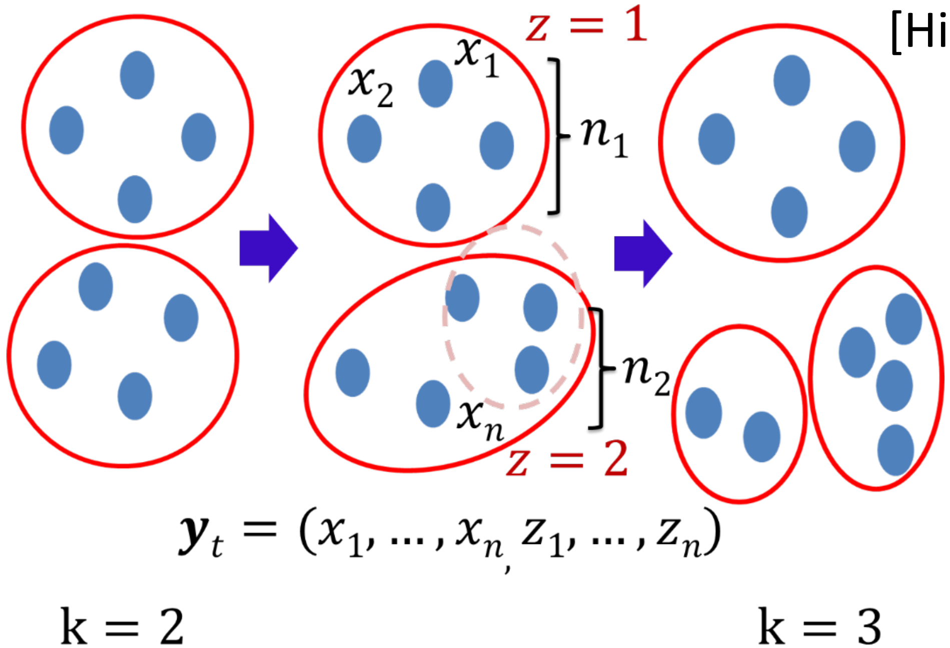
Detecting changes of number of clusters and clustering assignments



DMS for Complete Variable Model

Incrementally Application of DMS to complete variable model

[Hirai Yamanishi KDD2012]



Z: latent variable
…Cluster index of X

At each time $t(=1, \ldots, T)$, observe

$X_t = \boldsymbol{x}_t^n = \boldsymbol{x}_{t1}, \ldots, \boldsymbol{x}_{tn}$: observed data of n objects

$Z_t = \boldsymbol{z}_t^n = z_{t1}, \ldots, z_{tn}$: latent variable seqience

$\boldsymbol{x}_{ti} = (x_{ti1}, \ldots, x_{tim})^\top \in \mathbb{R}^m$: m-dimenstional data for each object

$\mathcal{P} = \{p(X, Z; \theta, M)\}$: Complete variable model

Incremental DMS Criterion

Slice total codelength time-wisely, then select # clusters and cluster assignment at each time

$X^T = X_1, \ldots, X_T$: data sequence
$Z^T = Z_1, \ldots, Z_T$: latent variable sequence
$M^T = M_1, \ldots, M_T$: model sequence

[Hirai Yamanishi KDD2012]
See also [Sun et al. KDD2007]
[Satoh Yamanishi ICDM2013]

$$\mathcal{L}(X^T, Z^T) + \mathcal{L}(M^T)$$
$$= \sum_{t=1}^{T} \{\mathcal{L}_{\mathrm{NML}}(X_t, Z_t | X^{t-1}, Z^{t-1}; M_t \cdot M^{t-1}) + \mathcal{L}(M_t | M^{t-1})\} \Longrightarrow \min \ \text{w.r.t.} \ M^T$$

Slice time-wisely

$$\forall t, \quad \mathcal{L}_{\mathrm{NML}}(X_t, Z_t | X^{t-1}, Z^{t-1}; M_t \cdot M^{t-1}) + \mathcal{L}(M_t | M^{t-1}) \Longrightarrow \min \text{w.r.t.} \ M_t$$

NML codelength for Clustered data sequence

Codelength for cluster change

Application to Gaussian Mixture Model

Complete variable model of Gaussian mixture model

$$f(\mathbf{x}^n, z^n; \mu, \Sigma) = \prod_{k=1}^{K} \pi_k^{h_k} \times \prod_{x_i \in z_k} \frac{1}{(2\pi)^{\frac{mh_k}{2}} \cdot |\Sigma_k|^{\frac{h_k}{2}}} \times \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^\top \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\right\}.$$

Upper bound on NML codelength for GMM

$$L_{\mathrm{uNML}}(\mathbf{x}^n, z^n ; \mathcal{M}(K)) = -\log f(\mathbf{x}^n, z^n; \mathcal{M}(K), \hat{\theta}(\mathbf{x}^n, z^n)) + \log \mathcal{C}_{\mathrm{u}}(\mathcal{M}(K), n),$$

$$\mathcal{C}_{\mathrm{u}}(\mathcal{M}(K), n) = \sum_{h_1, \cdots, h_K} \frac{N!}{h_1! \cdot \cdots \cdot h_K!} \prod_{k=1}^{K} \left(\frac{h_k}{N}\right)^{h_k} \times B(m, R, \epsilon) \cdot \left(\frac{h_k}{2\mathrm{e}}\right)^{\frac{mh_k}{2}} \frac{1}{\Gamma_m\left(\frac{h_k - 1}{2}\right)}$$

$$B(m, R, \epsilon) \stackrel{\mathrm{def}}{=} \frac{2^{m+1} R^{\frac{m}{2}} \prod_{j=1}^{m} \epsilon_{1j}{}^{-\frac{m}{2}}}{m^{m+1} \cdot \Gamma\left(\frac{m}{2}\right)}.$$

[Hirai and Yamanishi IEEE IT 2019]

Experimental Results: Real Data

-Market Structure Change Detection-

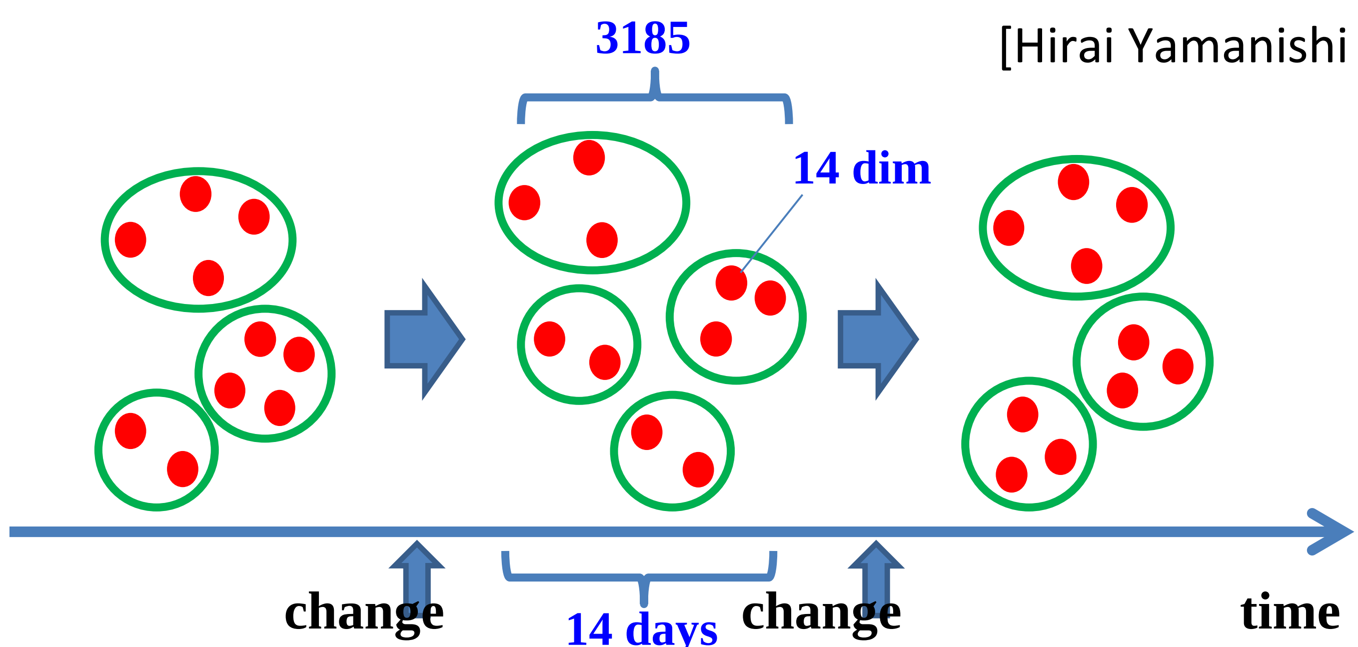
Tracking changes of customer structures from beer transaction behavior data (QPR)

Data provided by M-Cube

Period: Nov.2011-Jan. 2012 #customers: 3185

Data for each customer at t = consumption volume of 14 brands beer during 14 days until time t

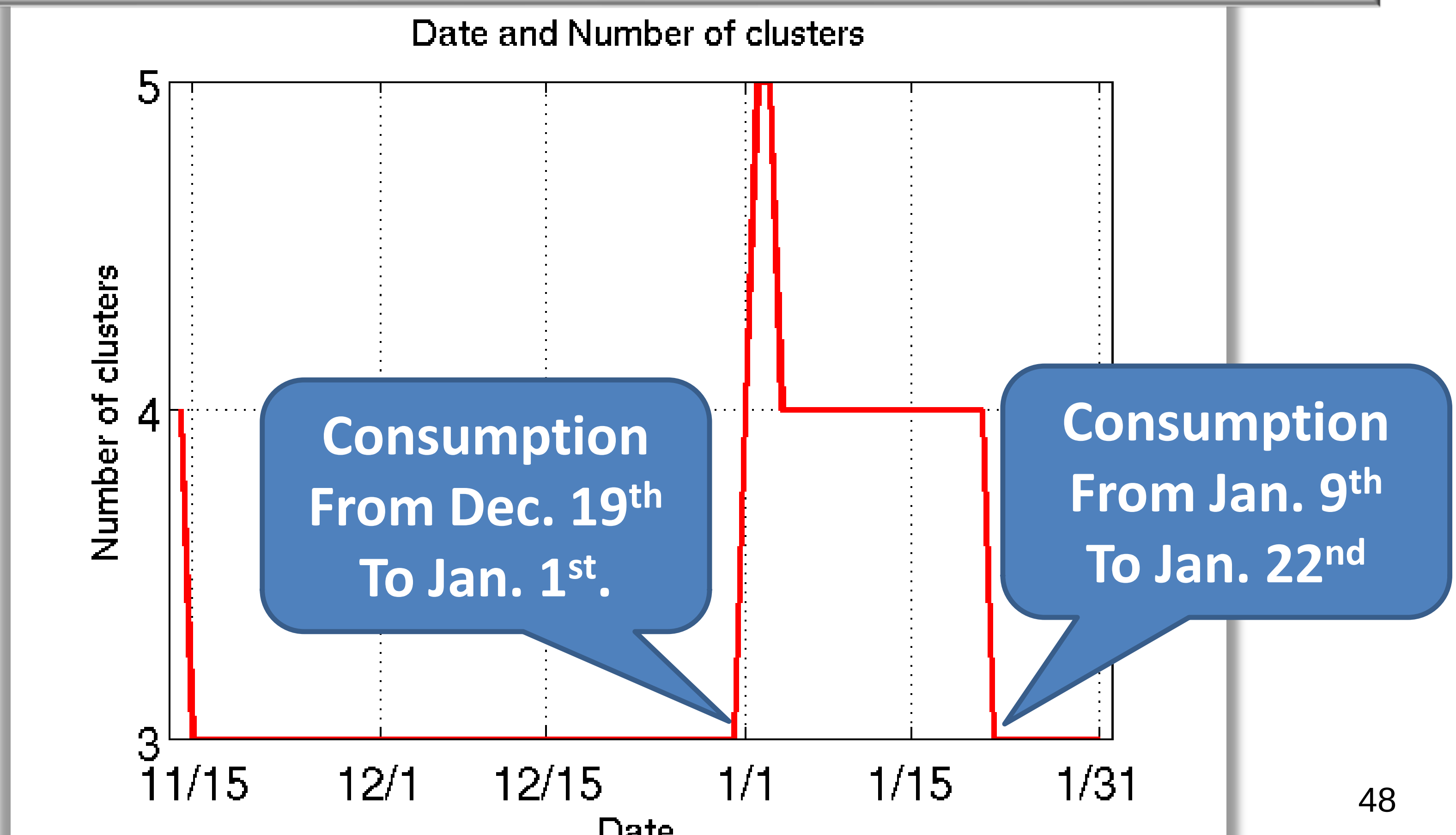
[Hirai Yamanishi KDD2012]



Experimental Results: Real Data

-Market Structure Change Detection-

Change of #clusters was detected at time when year-end demand increased vastly.



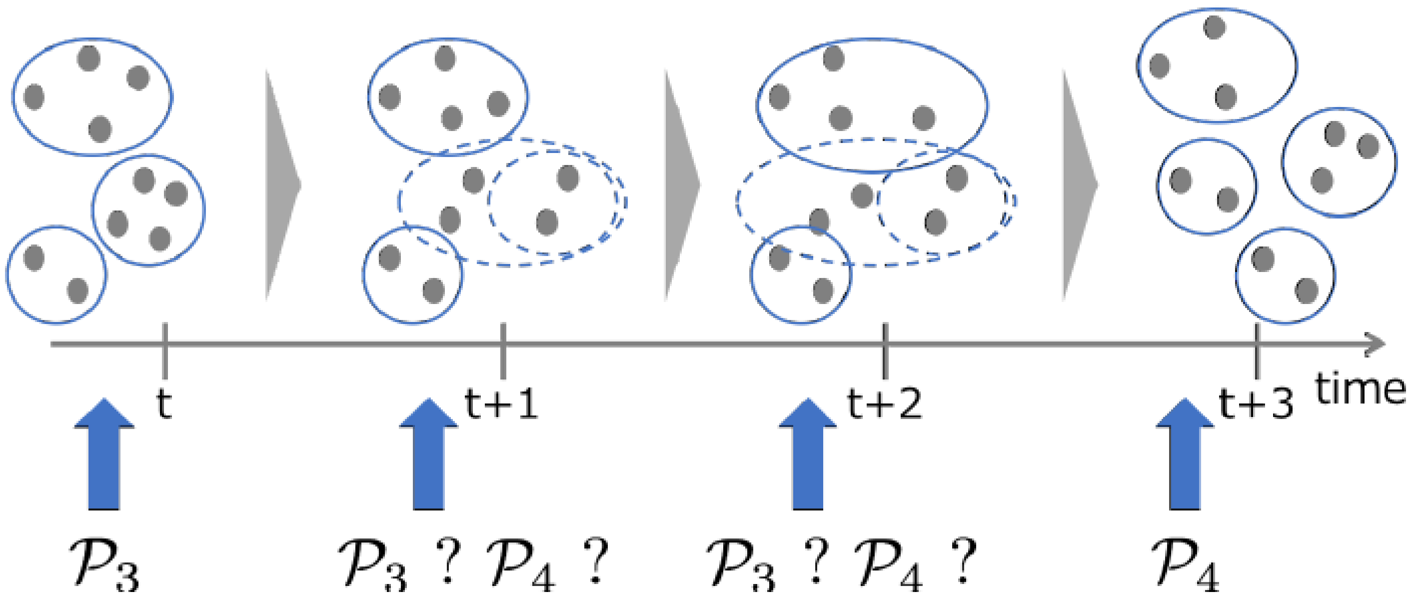
Clustering Structure Change

平均消費量(ml)	cluster 1	cluster 2	cluster 3
ビールA	184	0	117
ビールB	91	0	95
プレミアムA	108	0	80
プレミアムB	[illegible]	[illegible]	[illegible]
ビールC	0	[illegible]	[illegible]
ビールD	0	[illegible]	[illegible]
第三のビールA	[illegible]	[illegible]	[illegible]
第三のビールB	0	[illegible]	[illegible]
第三のビールC	0	[illegible]	[illegible]
第三のビールD	0	[illegible]	[illegible]
発泡酒A	0	[illegible]	[illegible]
オフA	0	0	157
オフB	0	114	34
オフC	0	0	83
総購入量	589	852	1373
人数(人)	598	376	311

cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
84	0	131	50	229
123	0	248	0	0
153	0	174	73	0
[illegible]	[illegible]	[illegible]	0	0
[illegible]	[illegible]	[illegible]	122	0
[illegible]	[illegible]	[illegible]	192	0
[illegible]	[illegible]	[illegible]	0	0
[illegible]	[illegible]	[illegible]	0	131
[illegible]	[illegible]	[illegible]	46	236
[illegible]	[illegible]	[illegible]	0	0
[illegible]	[illegible]	[illegible]	0	0
0	0	169	138	0
0	215	74	0	0
0	0	61	83	0
637	796	2348	705	596
397	190	123	162	363

- Year-end demands of Beer A and 3rd world Beer C rapidly increased, they led to form new additional clusters

4.2.4. Model Change Sign Detection

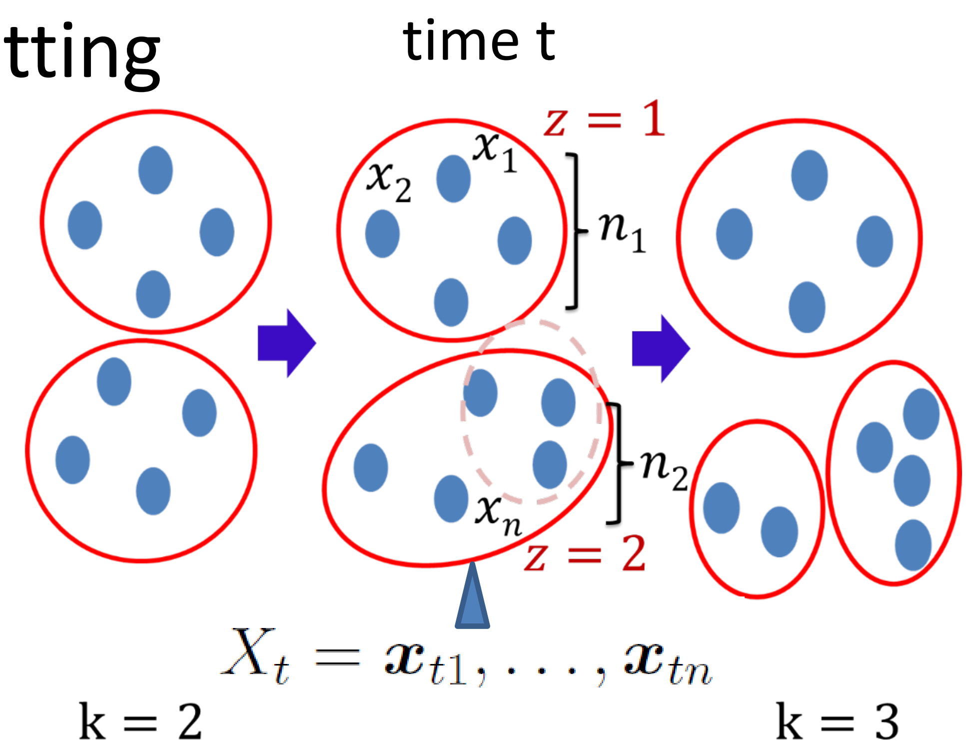


k=3 k=? k=4

Model uncertainty increases

Problem Setting

Problem setting



At each time $t(= 1, \ldots, T)$, observe

$X_t = \boldsymbol{x}_t^n = \boldsymbol{x}_{t1}, \ldots, \boldsymbol{x}_{tn}$: observed data of n objects

$Z_t = \boldsymbol{z}_t^n = z_{t1}, \ldots, z_{tn}$: latent variable seqience

$\boldsymbol{x}_{ti} = (x_{ti1}, \ldots, x_{tim})^\top \in \mathbb{R}^m$: m-dimenstional data for each object

$\mathcal{P} = \{p(\boldsymbol{x}; \theta, k) : \theta \in \Theta_k\}$: model class

Structural Entropy

Structural Entropy [Hirai Yamanishi BigData 2018]

… measuring uncertainty of model selection

$$\mathrm{SE}_t = \sum_k (-p(k|X_t)\log p(k|X_t))$$

where

$$p(k|X_t) = \frac{\exp(-\beta L_t(k))}{\sum_{k'} \exp(-\beta L_t(k'))}$$

$0 < \beta \leq 1$: temparature parameter

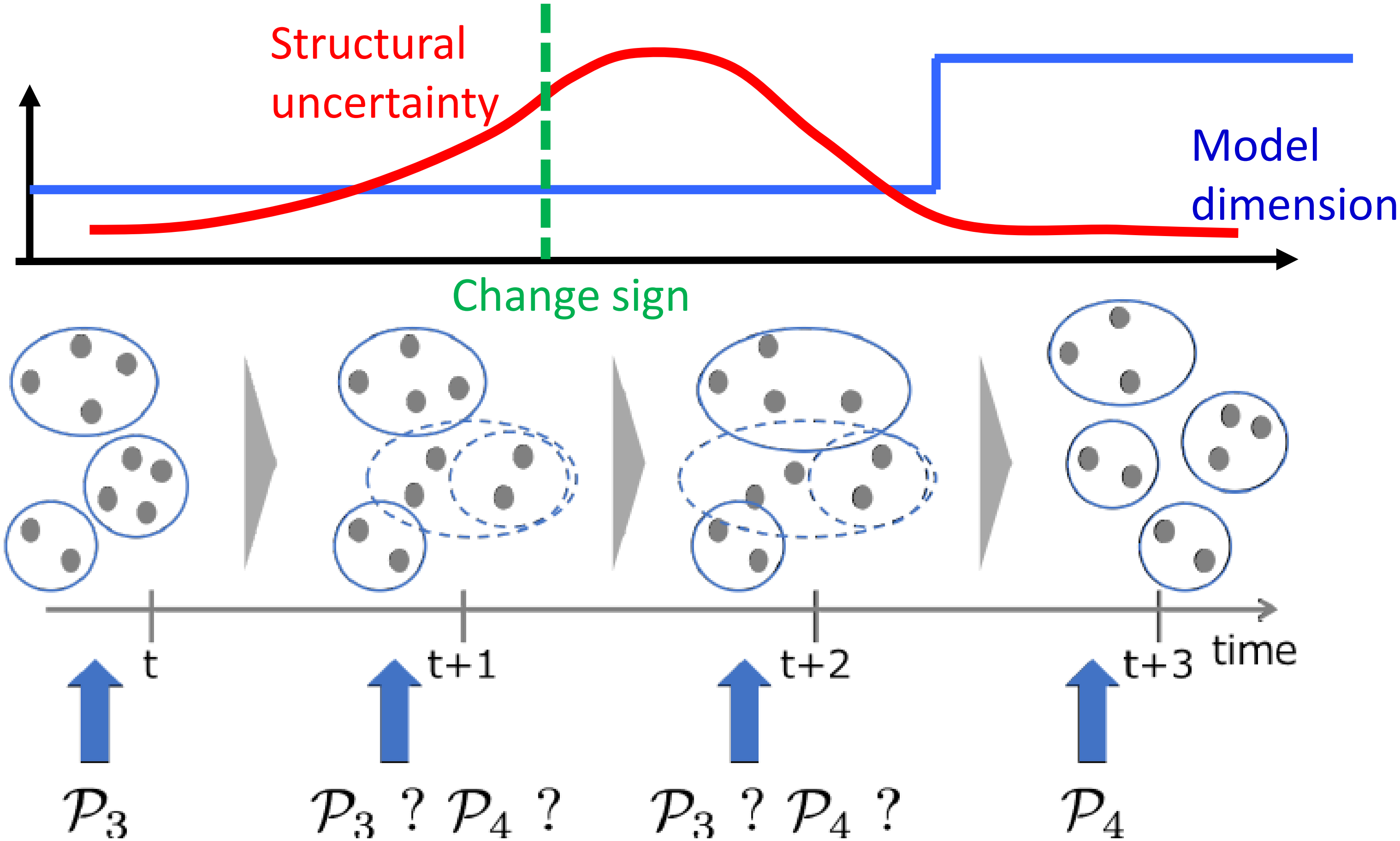
$$\begin{aligned} L_t(k) &= \mathcal{L}_{\mathrm{NML}}(X_t|X^{t-1};k) \\ &= -\log \max_\theta p(X_t|X^{t-1};\theta,k) + \log \sum_Y \max_\theta p(Y|X^{t-1};\theta,k) \end{aligned}$$

Or for complete variable model

$$\begin{aligned} L_t(k) &= \mathcal{L}_{\mathrm{NML}}(X_t, Z_t|X^{t-1}, Z_{t-1}k) \\ &= -\log \max_\theta p(X_t, Z_t|X^{t-1}, Z^{t-1};\theta) + \log \sum_{Y,W} \max_\theta p(Y, W|X^{t-1}, Z^{t-1};\theta) \end{aligned}$$

Model Change Sign Detection via Structural Entropy

[Hirai Yamanishi BigData 2018]
See also [Ohsawa RevSNS 2018]



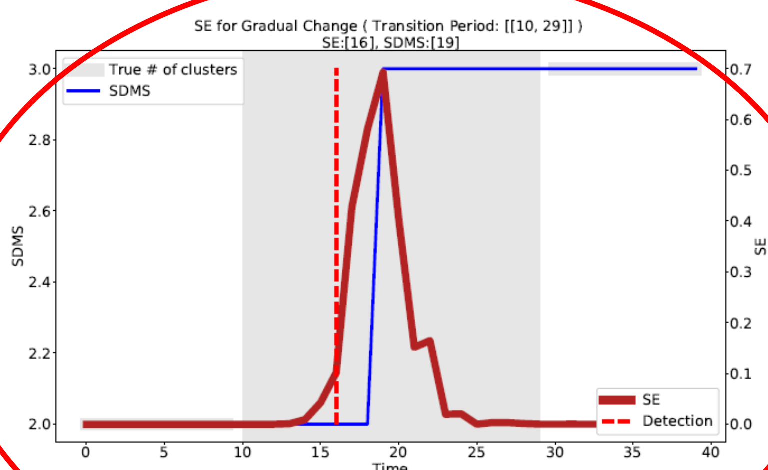
Experimental Results: Synthetic Data

Change sign can be detected by looking at rise up of structural entropy

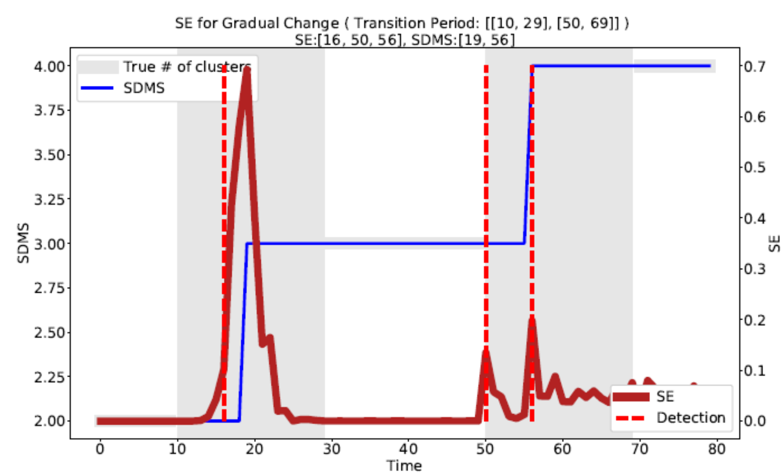
$$
\begin{cases}
K^* = 2,\ \ \mu = (\mu_1, \mu_2) & \text{if } 1 \le t \le \tau_1, \\
K^* = 3,\ \ \mu = (\mu_1, \mu_2, u(t)) & \text{if } \tau_1 + 1 \le t \le \tau_2, \\
K^* = 3,\ \ \mu = (\mu_1, \mu_2, \mu_3) & \text{if } \tau_2 + 1 \le t \le T,
\end{cases}
$$

$$
\text{where } u(t) = \frac{(\tau_2 - t)\mu_2 + (t - \tau_1)\mu_3}{\tau_2 - \tau_1}.
$$

[Hirai Yamanishi BigData2016]



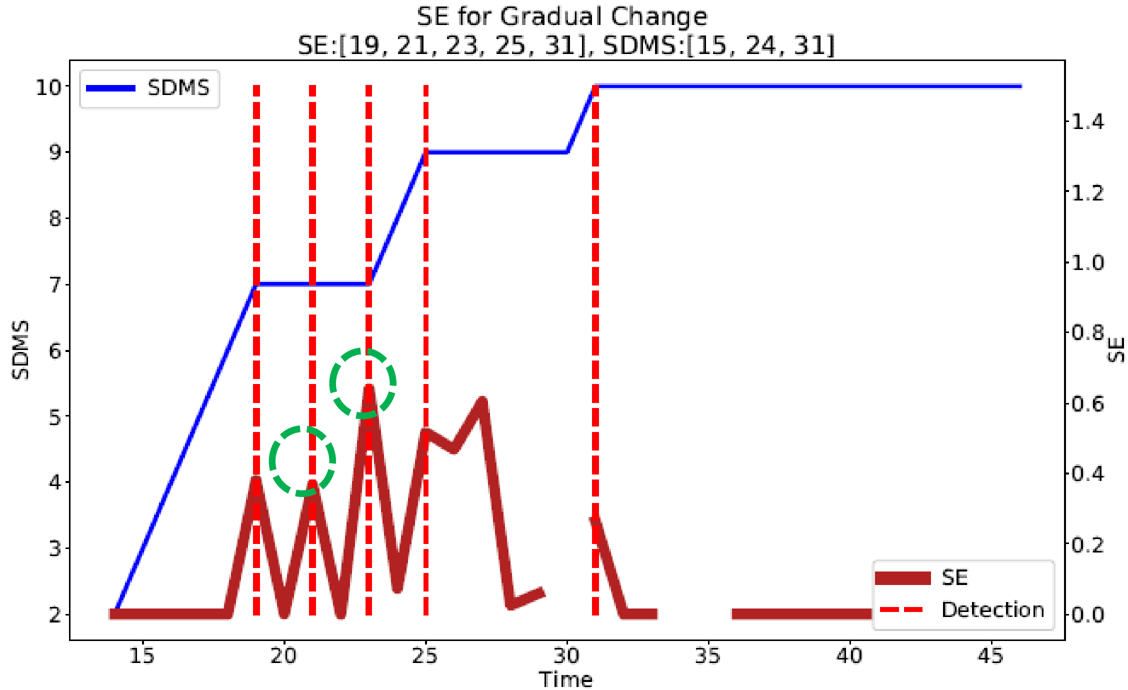
(a) Single Change



(b) Multi Change

Experimental Results: Real Data

Signs of customer clustering structure changes can be detected by looking at rise up of structural entropy



Time 22

brand	clu-1	clu-2	clu-3	clu-4	clu-5	clu-6	clu-7
A	3397	0	16	14	22	6	21
B	12	126	19	7	49	13	36
C	0	0	2328	0	15	10	1815
D	0	0	0	3079	5	7	1551
E	0	0	0	0	559	0	0
F	0	0	0	0	0	2371	0
num	307	368	259	269	15	159	132

Time 24

brand	clu-1	clu-2	clu-3	clu-4	clu-5	clu-6	clu-7	clu-8
A	3782	10	18	9	30	5	23	0
B	0	3118	14	0	26	10	136	0
C	0	0	2492	0	18	6	111	0
D	0	0	0	3296	0	5	1818	0
E	0	0	0	0	638	0	0	0
F	0	0	0	0	0	2466	0	0
num	206	319	248	197	12	156	202	169

Summary

- The MDL change statistics is a theoretically justified methodology for measuring the change score either for parameter changes or model changes.
- For gradual change detection, apply sequential MDL statistics with adaptive/non-adaptive windowing to conduct real-time event detection.
- For multiple model change detection, conduct Dynamic Model Selection(DMS) to obtain optimal model sequences.
- For clustering structure change detection, apply DMS to latent variable models sequentially to catch up latent structure changes.
- Signs of model changes may be detected by looking at structural entropy measuring model uncertainty.

References

■ 4.1. MDL change statistics

- J. Vreeken, M. van Leeuwen, A. Siebes: “Krimp: mining itemsets that compress,” *Data Mining and Knowledge Discovery*, Vol. 23, 1, pp 169-214, 2011.
- K.Yamanishi and K.Miyaguchi: “Detecting gradual changes from data stream using MDL- change statistics,“ *Proceedings of 2016 IEEE International Conference on BigData* (IEEE BigData2016), pp:156-163, 2016.
- R. Kaneko, K.Miyaguchi, and K.Yamanishi : “Detecting Changes in Streaming Data with Information-Theoretic Windowing," *Proceedings of 2017 IEEE International Conference on Big Data* (BigData2017), pp: 646-655, 2017.
- B.Hooi, L.Akoglu,D.Eswaran,A.Pandey, A.Jereminov,L.Pileggi, C.Faloutsos: “ChangeDAR: Online localized change detection for sensor data on a graph,” in Proceedings of the 27th *ACM International Conference on Information and Knowledge Management*, pp:507-516, 2018.

  C.f. Adaptive window algorithm
- A.Bifet and R.Gavaldà: “Learning from time-changing data with adaptive windowing,” in *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

  C.f. Predictive change statistics
- V.Guralnik and J.Srivastava: “Event detection from time series data,” in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp:33–42, 1999.

References

■ 4.2. Dynamic Model Selection

- K.Yamanishi and Y.Maruyama: “Dynamic syslog mining for network failure monitoring,” *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD2005), pp：499—508, 2005.
- K.Yamanishi and Y.Maruyama: “Dynamic model selection with its applications to novelty detection,” *IEEE Transactions on Information Theory*, Vol. 53, NO. 6, pp:2180-2189, 2007.
- K.Yamanishi and S. Fukushima: " Model change detection with the MDL principle", *IEEE Transactions on Information Theory*, 64(9), pp: 6115-6126, 2018.

■ 4.2. Topics Related to Dynamic Model Selection

- M.Herbster and M.Warmuth: “Tracking the best experts,” *Machine Learning*, 32, pp:151–178,1998.
- V. Vovk: “Derandomizing stochastic prediction strategies," *Machine Learning,* vol. 35, no. 3, pp. 247-282, 1999.
- J.Kleinberg: “Bursty and hierarchical structure in stream,” *Data Mining and Knowledge Discovery*, 7, pp:373—397, 2003.

References

■ 4.2. Topics Related to Dynamic Model Selection(Cont.)

- R.A.Davis, T.C.M.Lee, G.A.Rofriguez-Yam: "Structural break estimation for nonstationary time series models," *Journal of American Statistical Associations*, 101, pp:223-239, 2006.
- X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series," *Proceedings of the 24th International Conference on Machine Learning, (ICML2007*), pp.1055--1062, 2007.
- T.Erven, P.Grunwald, and S.Rooij: "Catching up by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma," *Jr.Royal Stat.Soc.Ser.B*, vol. 74, no. Issue 3, pp. 361–417, 2012.
- R.Killick, P.Fearnhead, and I.A.Eckley: "Optimal detection of changepoints with a linear computational cost," *Journal of American Statistical Associations*, 107:500, pp:1590-1598, 2012.
- J. Gama, I. Zlibait, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Survey*, 2013.
- Y.Hayashi and K.Yamanishi: "Sequential network change detection with its applications to ad impact relation analysis," *Data Mining and Knowledge Discovery*: Vol. 29, Issue 1 ,pp: 137-167, 2015.

References

■ 4.2.3. Clustering Change Detection

- M. Song and H.Wang: "Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering," *Intelligent Computing*, 2005.
- J. Sun, C. Faloutsos, S.Papadimitriou,P. S. Yu: "GraphScope: parameter-free mining of large time-evolving graphs," *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2007)*, pp: 687-696, 2007.
- S. Hirai and K.Yamanishi: "Detecting changes of clustering structures using normalized maximum likelihood coding." *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2012),* pp:343-351, 2012
- S.Sato and K.Yamanishi: "Graph partitioning change detection using tree-based clustering," *Proceedings of IEEE International Conference on Data Mining (ICDM2013)*, pp:1169-1174, 2013.

■ 4.2.4. Model Change Sign Detection

- S. Hirai and K.Yamanishi: "Detecting Latent Structure Uncertainty with Structural Entropy", Proceedings of IEEE International Conference on BigData (BigData2018), Dec. 2018.
- Y. Ohsawa: "Graph-based entropy for detecting explanatory signs of changes in market," The Review of Social Network Strategies, Vol 12, 2, pp:183-203, 2018.