Change Sign Detection with Two-Stage MDL Change Statistics

二段階MDL変化統計量による変化予兆検知

48-196233 結城 凌
指導教員 山西 健司教授
数理第6研究室

1/27(水) 修士論文審査

Agenda

1. 研究の背景
2. 2段階MDL変化統計量
3. 人工データと実データによる実験
4. まとめ

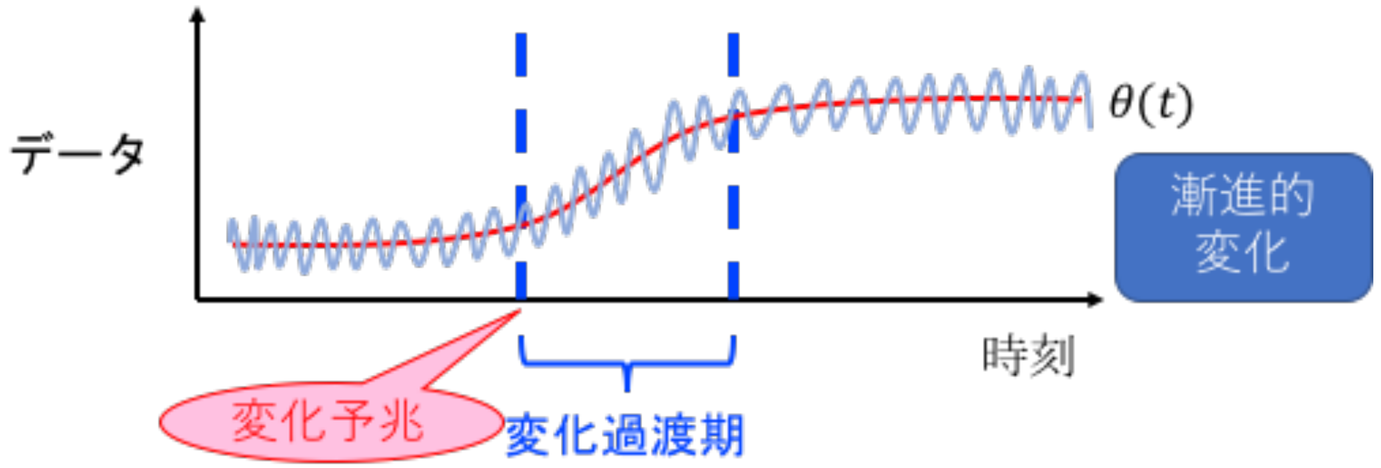
Agenda

1. 研究の背景

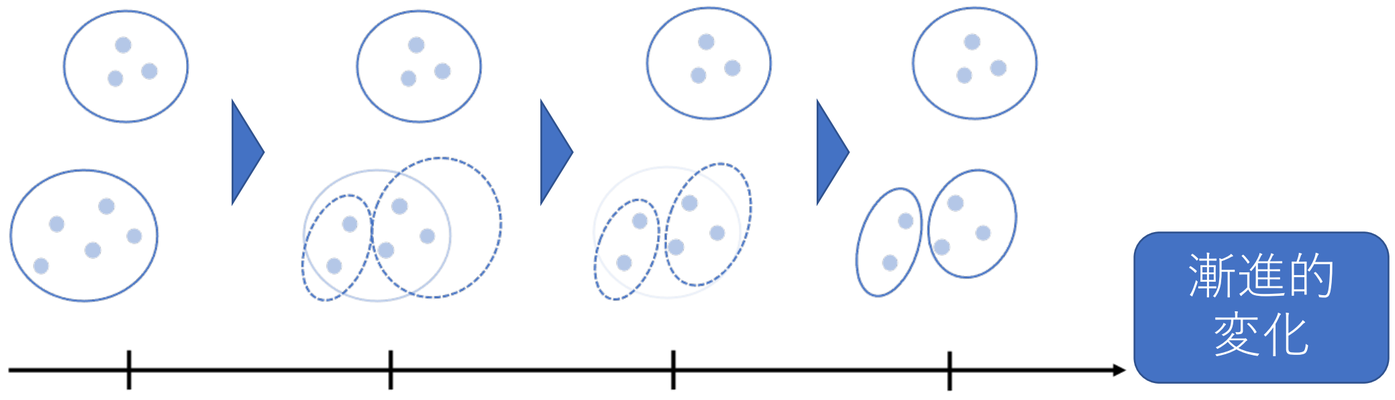
本研究の概要

パラメータ変化と構造的変化の漸進的変化開始を、
2段階minimum description length(MDL)変化統計量で検知

パラメータ漸進的変化



構造的漸進的変化

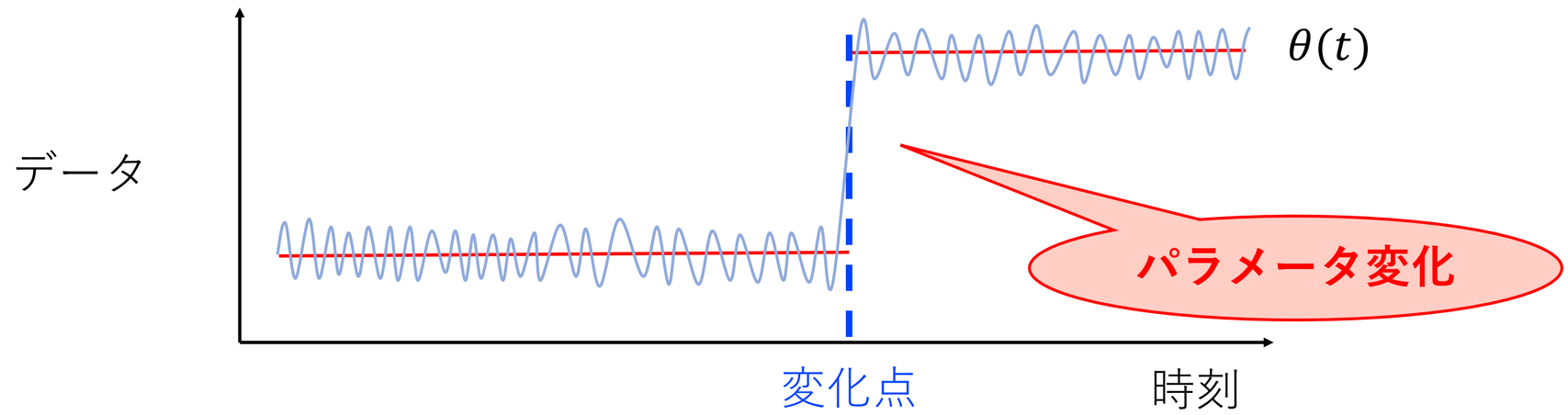


→漸進的変化検知に対し、2段階変化検知というアプローチを提案

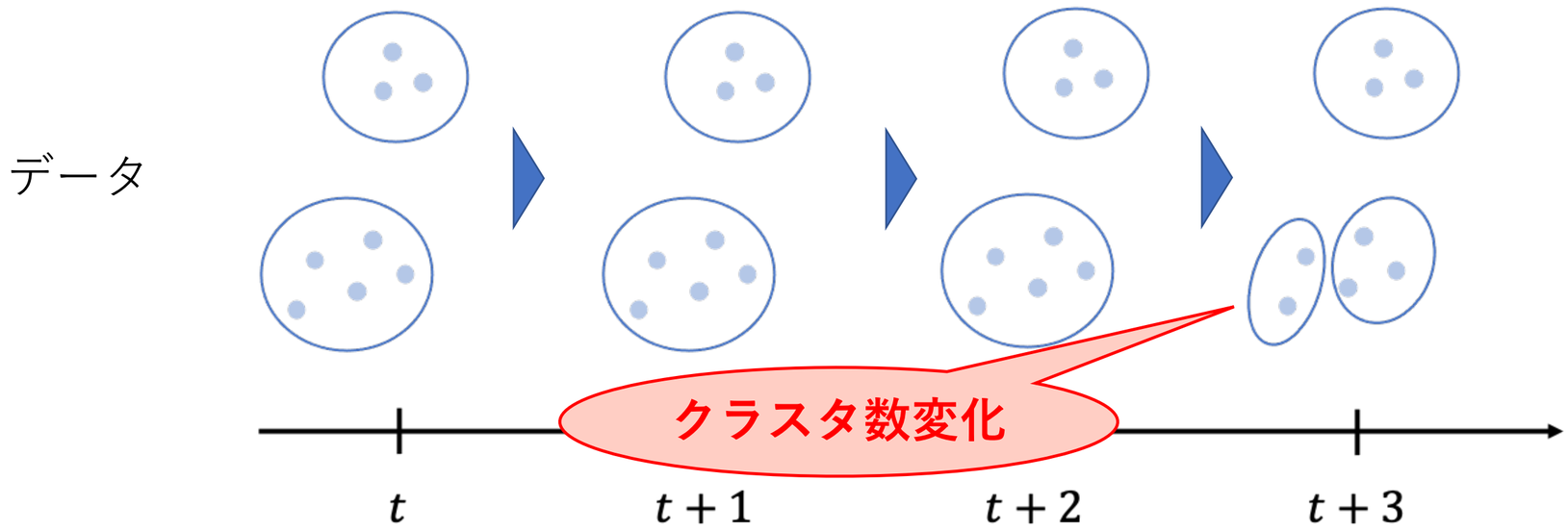
背景

変化検知とは…
データ発生確率分布が変化する時刻を検知する問題

パラメータ変化



構造的変化



変化とイベント

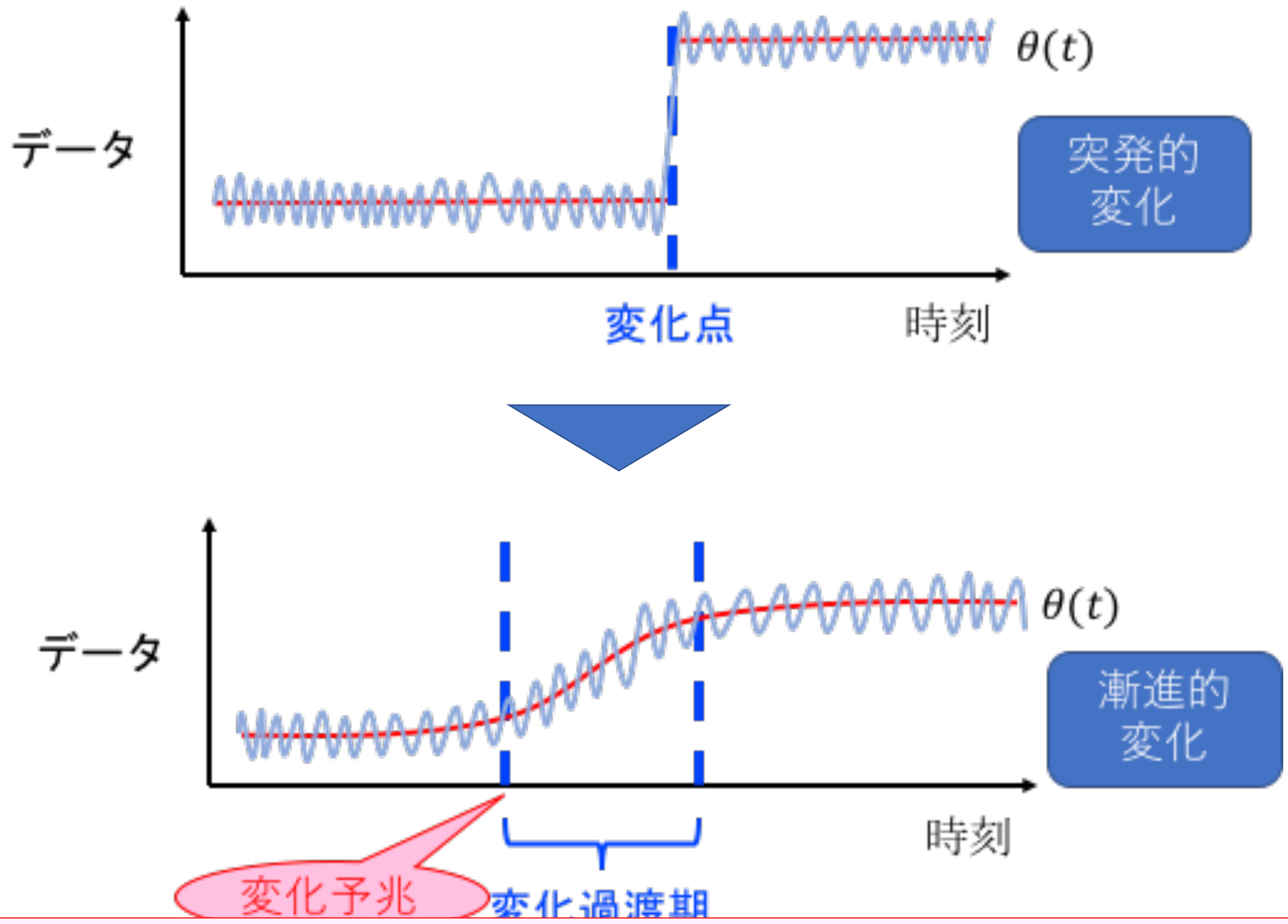
変化と重要イベントは頻繁に結びつく
→イベントの気づくきっかけ・原因解析

時系列データと対応イベント(山西編著 2019)

時系列データ	対応イベント
Webアクセスログ	SQLインジェクション(Yamanishi and Miyaguchi 2016)
工場センサーデータ	事故発生 (Yamanishi and Miyaguchi 2016)
マーケティングデータ	市場トレンドの変化 (Hirai and Yamanishi 2012, 2018, 2019)
SNSデータ	新規話題の出現 (Takahashi et al. 2012)
平均株価	政治的イベント(Adams and MacKay 2007)

変化予兆検知への研究展開

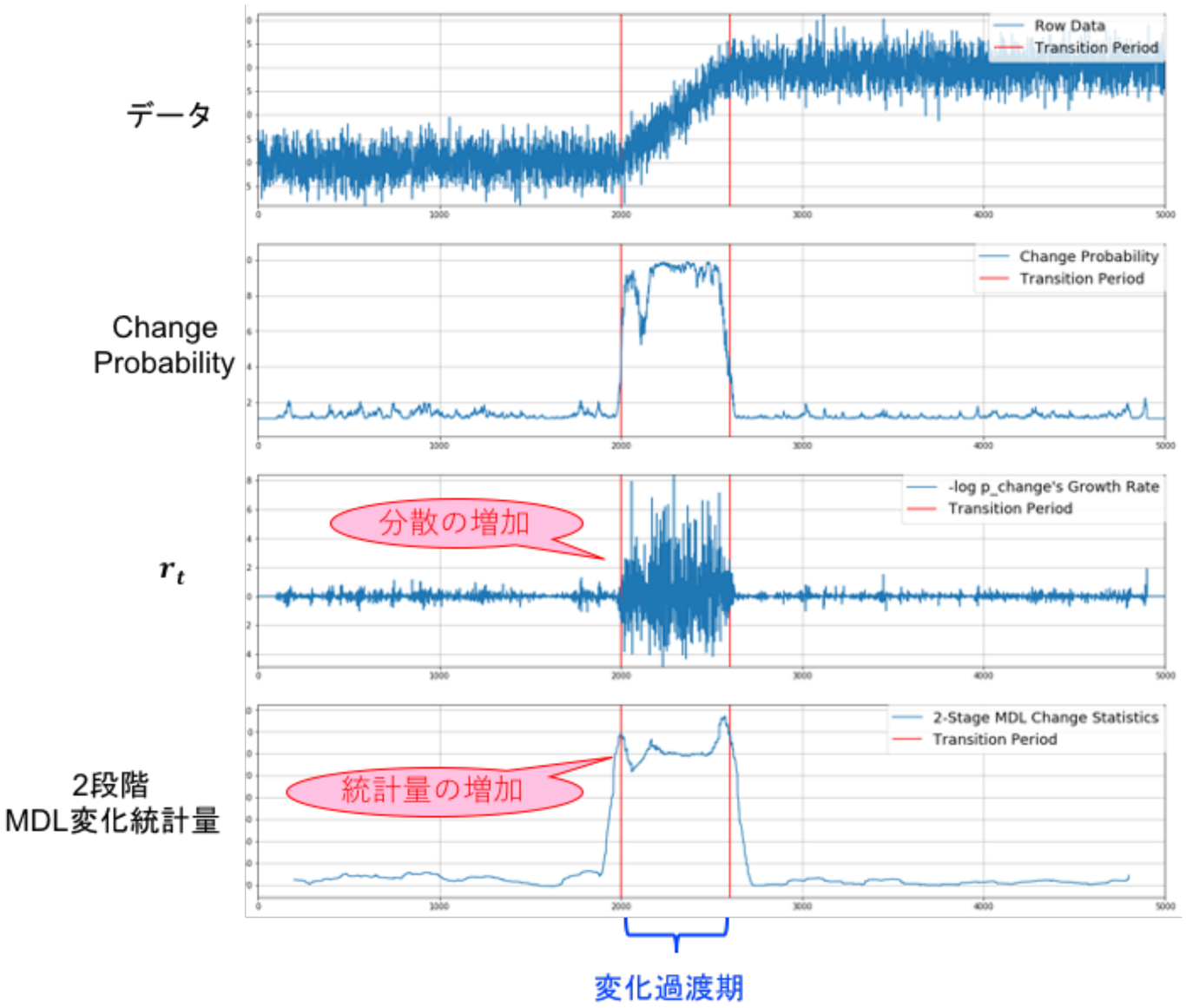
突発的変化検知から漸進的変化検知・変化予兆検知へ



→より一般に、変化予兆検知研究への展開

提案手法の概要

変化過渡期に性質が変化する統計量を変化スコアから作成
新たに作成した統計量に対し2段階目の変化検知



$x_1^n = x_1, \ldots, x_n$

1段階目 $p_{change}^t(x_1^n)$

$-\log p_{change}^t$ の増加率 r_t

2段階目 $\Phi_{2-stage}^t$

主要な関連研究

突発的変化検知に関する研究

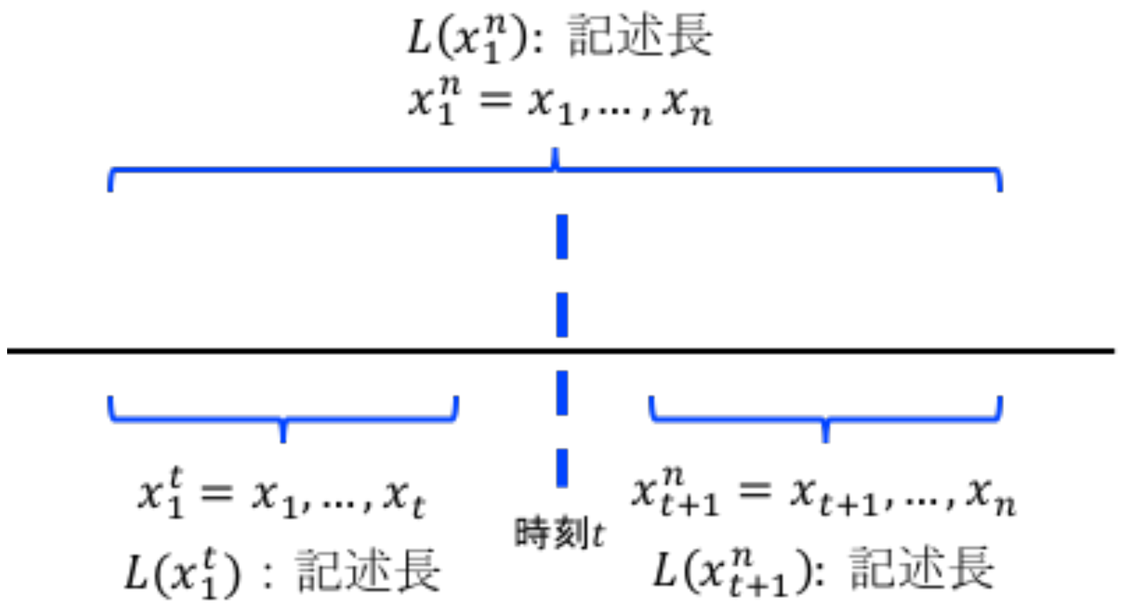
- 統計的検定に基づく手法(Bassevile et al. 1993)
- Baysian online change point detection (Adams and MacKay 2007)
- ChangeFinder(Takeuchi and Yamanishi 2006)
- など
- **漸進的変化では変化度が小さくなりやすく**、誤検知、検知遅れ・見逃しに繋がる。

漸進的変化検知・変化予兆検知に関する研究

- MDL変化統計量(Yamanishi and Miyaguchi 2016)
- 微分的MDL変化統計量(Yamanishi et al. 2020)
- Continuous Change Detection (Miyaguchi and Yamanishi 2017)
- Volatility Shift (Huang, David Tse Jung, et al. 2014)
- Graph-based Entropy (Ohsawa 2018)
- など
- **二段階の変化検知で変化予兆を捉える研究**は存在せず。

MDL変化統計量(先行研究)

変化度をMDL原理(Rissanen 1978)から定量化する手法
「変化なし」「変化あり」どちらのモデルが適しているか?



MDL変化統計量の概念図

Yamanishi and Miyaguchi 2016
Yamanishi and Fukushima 2018

$\rightarrow L(x_1^n) - \{L(x_1^t) + L(x_{t+1}^n)\} > n\epsilon$
であるか?($\epsilon > 0$)

Yes: 時刻tは変化点と推定。
No: 変化点ではないと推定。

正規化最尤符号長(Shtarkov 1987)

$$L(x_1^n) \stackrel{\text{def}}{=} -\log \frac{\max_\theta p(x_1^n;\theta)}{\sum_{y_1^n} \max_\theta p(y_1^n;\theta)},$$

$$= -\log \max_\theta p(x_1^n;\theta) + \log \sum_{y^n} \max_\theta p(y_1^n;\theta).$$

→パラメータ・構造的変化検知の両方が可能

MDL変化統計量(先行研究)

MDL変化統計量を用いた検定は
第1種・2種誤り確率に関する性能保証を持つ

仮説検定

$$H_0 : x_1^n \sim p(x_1^n, \theta_0),$$
$$H_1 : x_1^t \sim p(x_1^t, \theta_1), x_{t+1}^n \sim p(x_{t+1}^n, \theta_2)(\theta_1 \neq \theta_2).$$
$$where\ \theta_0, \theta_1, and\ \theta_2 : unknown$$

誤り確率の上界(Yamanishi and Miyaguchi 2016)

$$\text{Type I error prob.} \leq \exp(-n(\epsilon - \frac{\log C_n}{n})),$$
$$\text{Type II error prob.} \leq \exp(-\frac{n}{2}(D - \frac{\log C_t C_{n-t}}{n} - \epsilon)).$$

D: 分布の乖離度, $\log C_n$: Parametric Complexity

*構造的変化についても、類似した結果が得られている(Yamanishi and Fukushima 2018)。

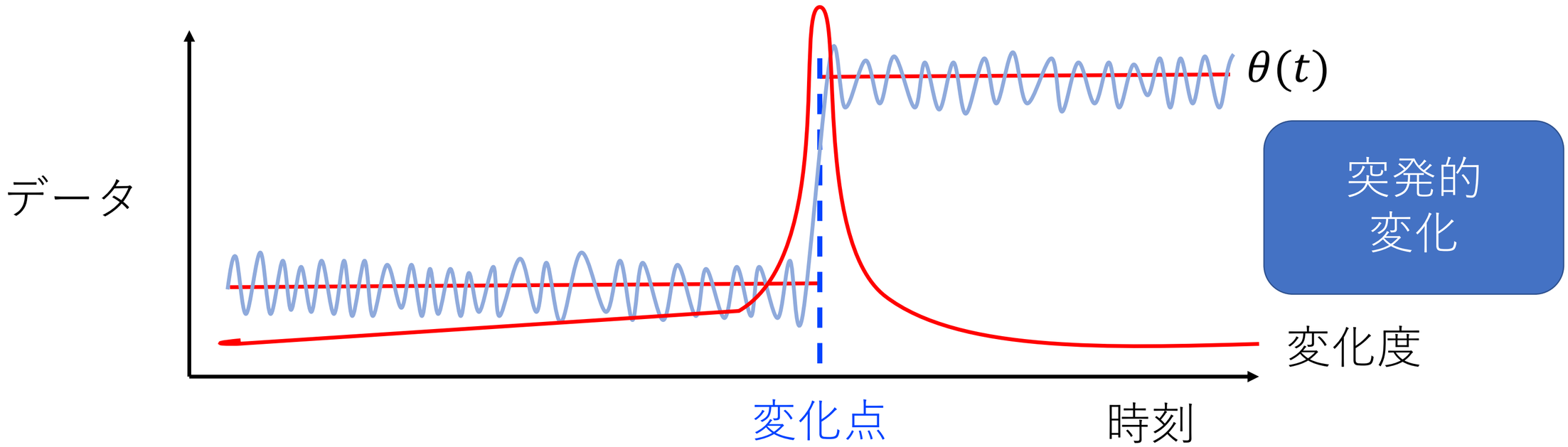
Agenda

2. 2段階MDL変化統計量

突発的変化の変化度の定義

2段階
MDL変化統計量

KLダイバージェンスが変化度の定義

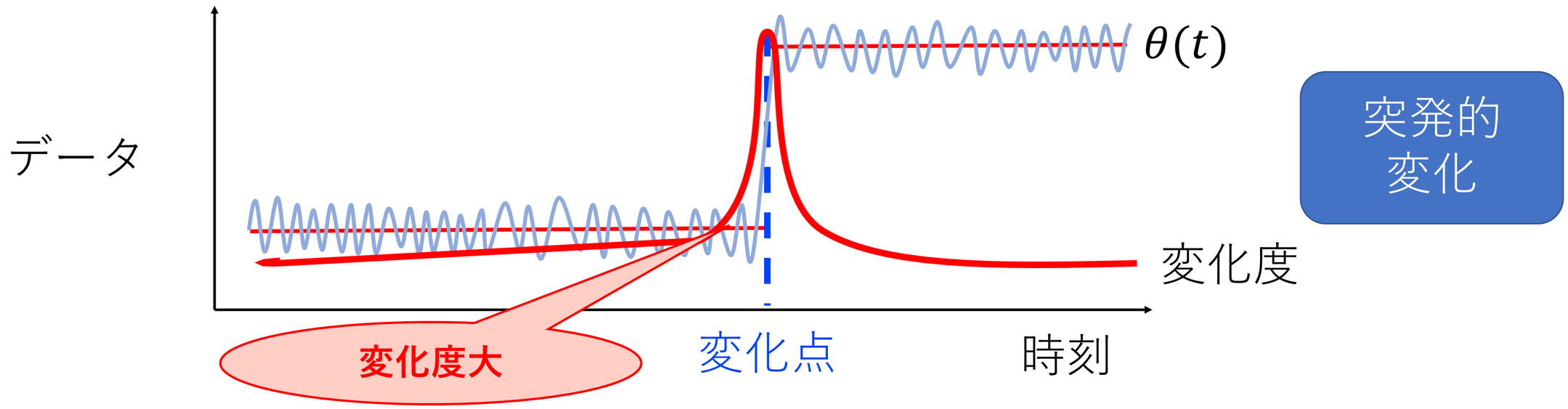


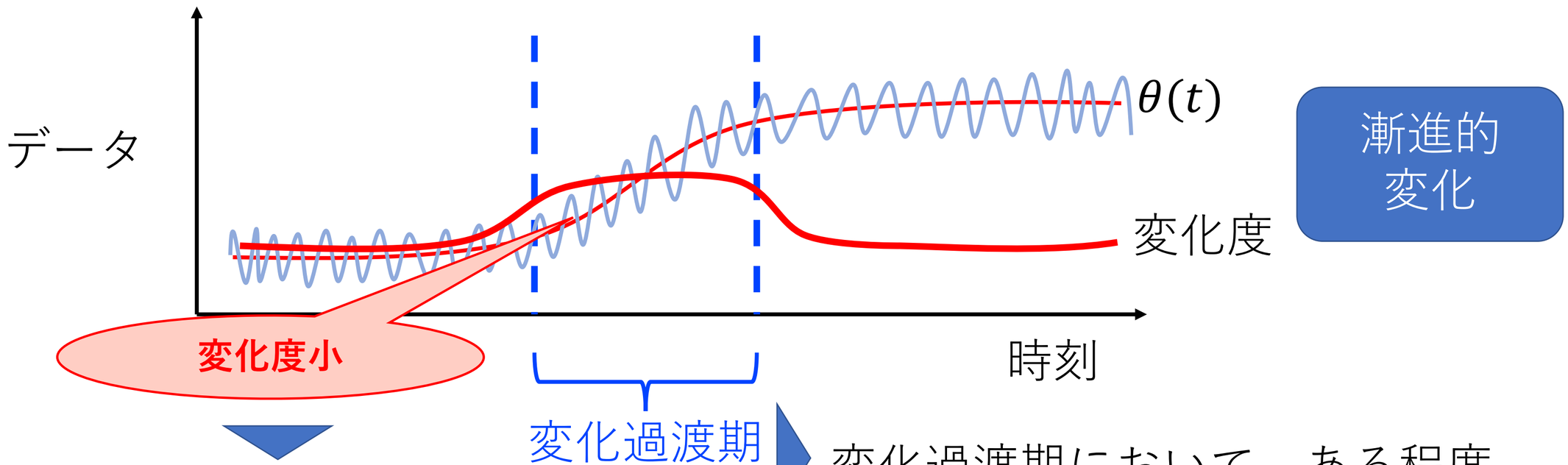
変化度の定義=$D(p_{\theta_{t+1}}||p_{\theta_t})$,

$$D(p_2||p_1) = \sum_x p_2(x) \log \frac{p_2(x)}{p_1(x)}$$ Kullback-Leibler (KL) divergence

漸進的変化における変化度

2段階
MDL変化統計量





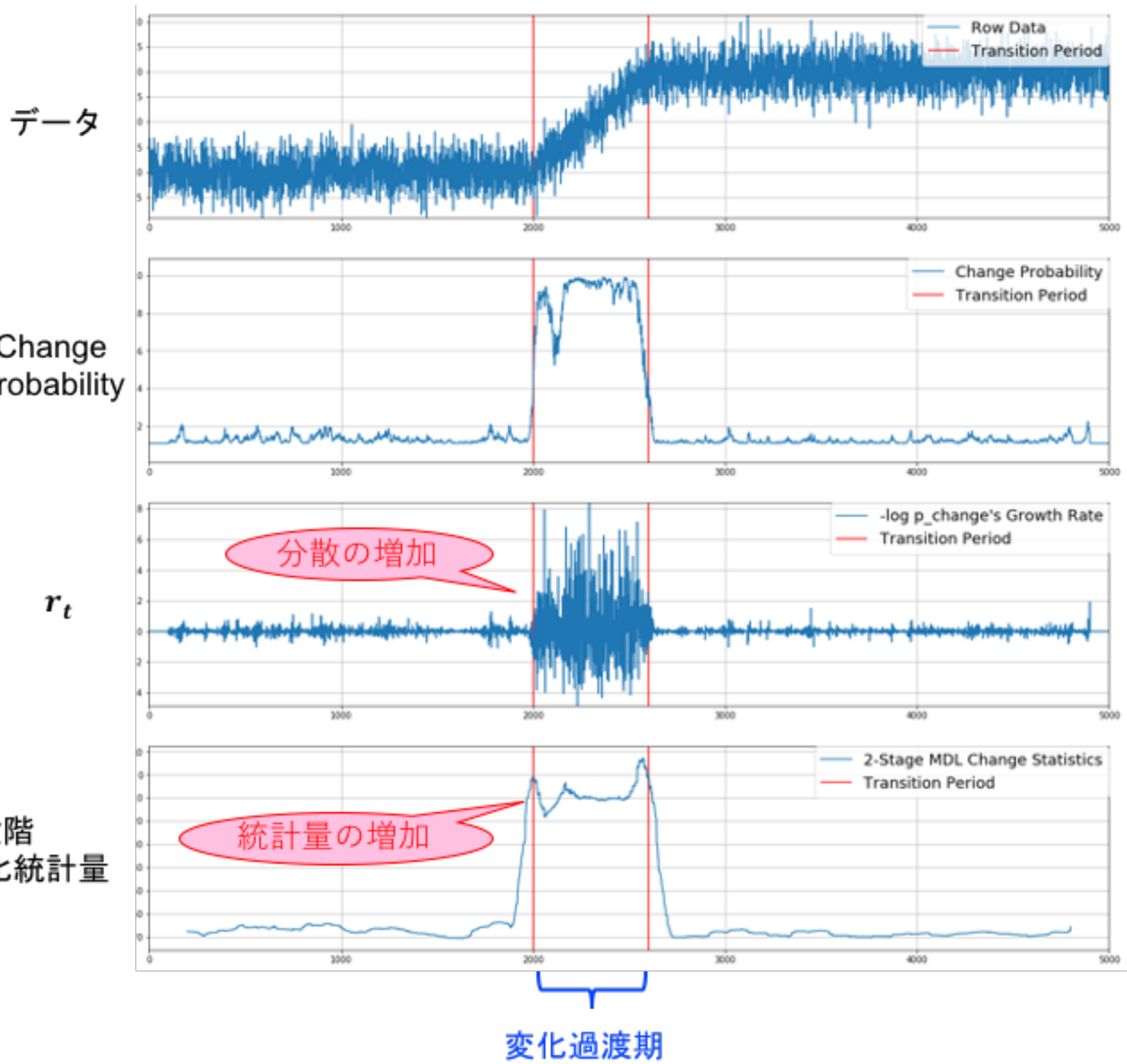
誤検知、検知遅れや見逃しに繋がる

変化過渡期において、ある程度
高い値を継続的にとる

本研究の概要(再掲)

2段階
MDL変化統計量

2段階の変化検知で、漸進的変化の開始点を検知



提案手法の概要

$x_1^n = x_1, \ldots, x_n$

1段階目 $p_{change}^t(x_1^n) \in (0,1)$

$-\log p_{change}^t$ の増加率 $r_t = \dfrac{-\log p_{change}^t + \log p_{change}^{t-1}}{-\log p_{change}^{t-1}}$

2段階目 $\Phi_{2-stage}^t$

p_{change}^tの増加につれて分散増加に寄与

Change Probability

各時刻が変化点である確率をモデリングしたもの

考える仮説検定

$$H_0 : x_1^n \sim p(x_1^n; \theta_0),$$
$$H_1 : x_1^t \sim p(x_1^t; \theta_1), x_{t+1}^n \sim p(x_{t+1}^n; \theta_2),$$

Change Probability (但し$\beta > 0$)

$$p_{change}^t(x_1^n) = \frac{\exp\left(-\beta\{L(x_1^t) + L(x_{t+1}^n)\}\right)}{\exp\left(-\beta L(x_1^n)\right) + \exp\left(-\beta\{L(x_1^t) + L(x_{t+1}^n)\}\right)}.$$

→有意水準δ $(0 < \delta < 0.5)$に対し、$p_{change}^t(x_1^n) > 1 - \delta$か？
Yes: 時刻tは変化点と推定
No: 変化点ではないと推定

アルゴリズム

2段階
MDL変化統計量

Sequential MDL (S-MDL) アルゴリズム(Yamanishi and Miyaguchi 2016)

- データを一定の長さの窓で切り出し逐次的にスコア計算。

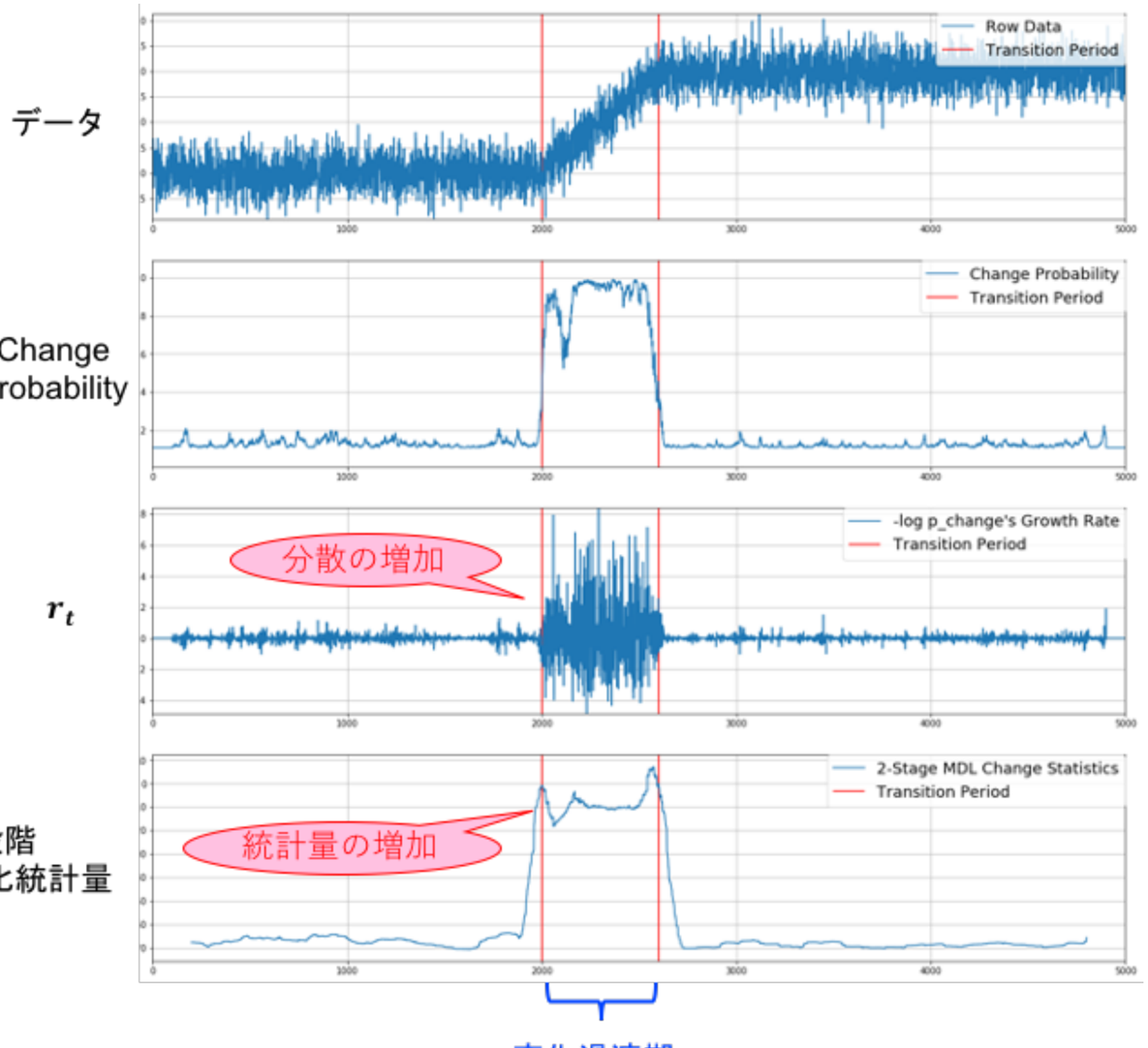
$x_1^n = x_1, \ldots, x_n$

$score_1$

$score_2$

…

→オンラインにスコア計算



データ　$x_1^n = x_1, \ldots, x_n$

▼ S-MDL的に計算

1段階目　$p_{change}^t(x_1^n)$

▼

$-\log p_{change}^t$ の増加率　r_t

▼ S-MDLを適用し計算

2段階目　$\Phi_{2-stage}^t$

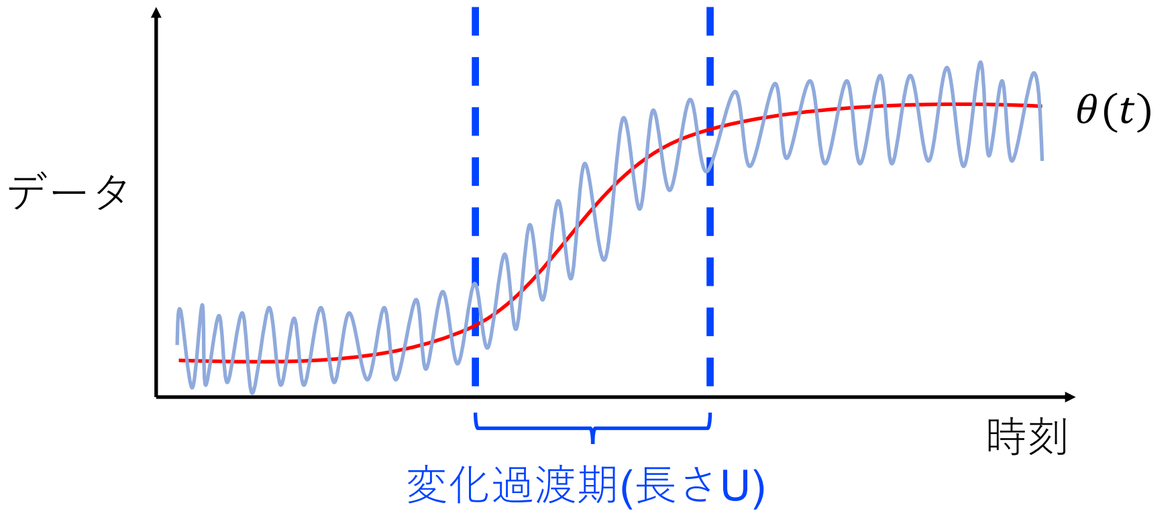
Agenda

3. 人工データと実データによる実験

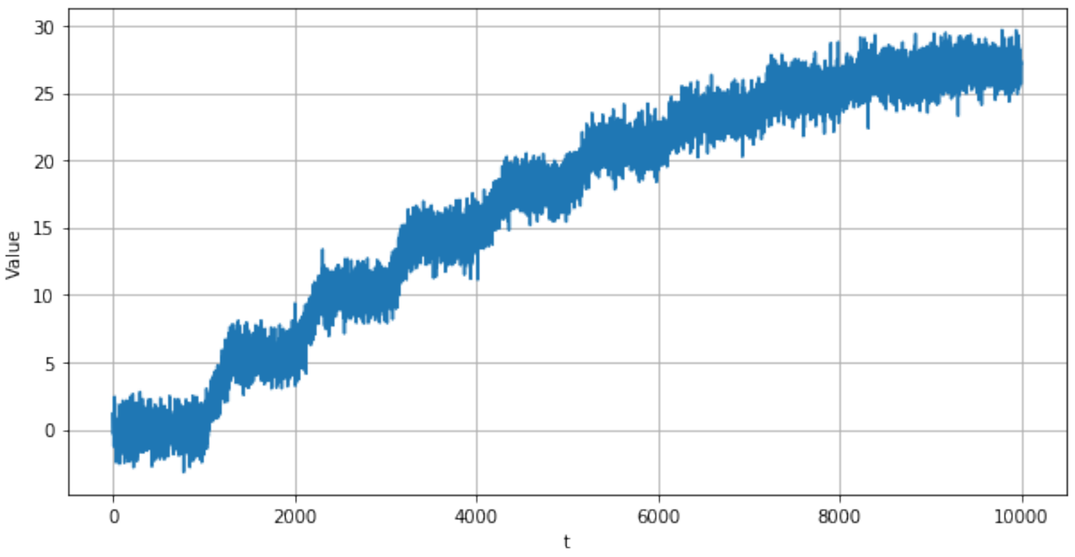
発生させた人工データ

人工データ

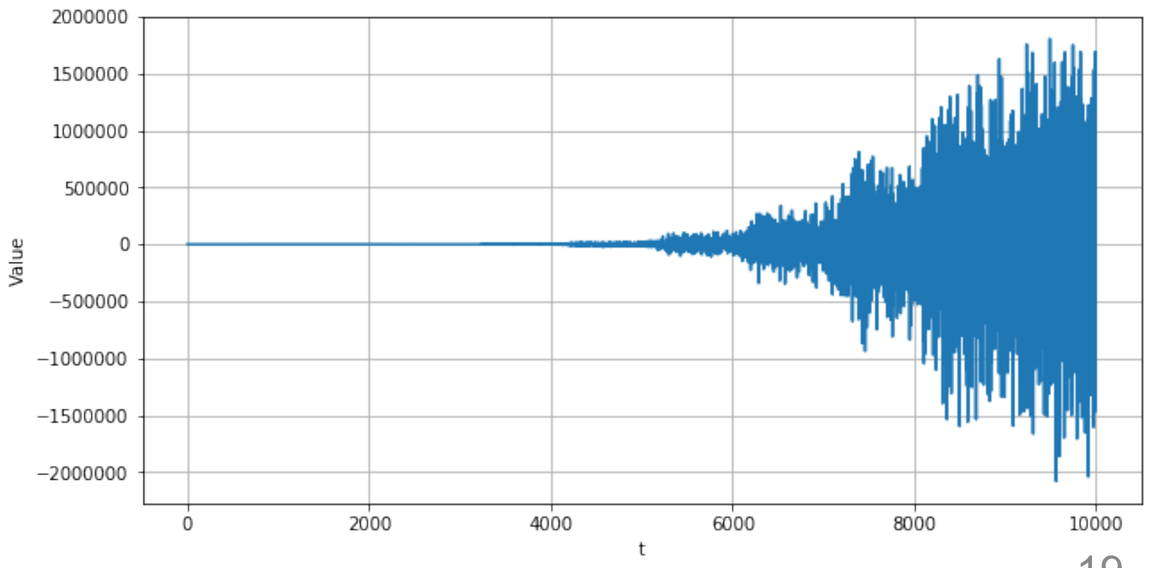
- 平均・分散の各漸進的変化データを**各20本準備**。
- **1本**でパラメータチューニング、残り**19本**でスコア計算。



・U=0, 100, 200, 300, 400と変化させ、検知性能を比較。
・Uが大きいほど検知が難しくなる。



平均が徐々に変化するデータ

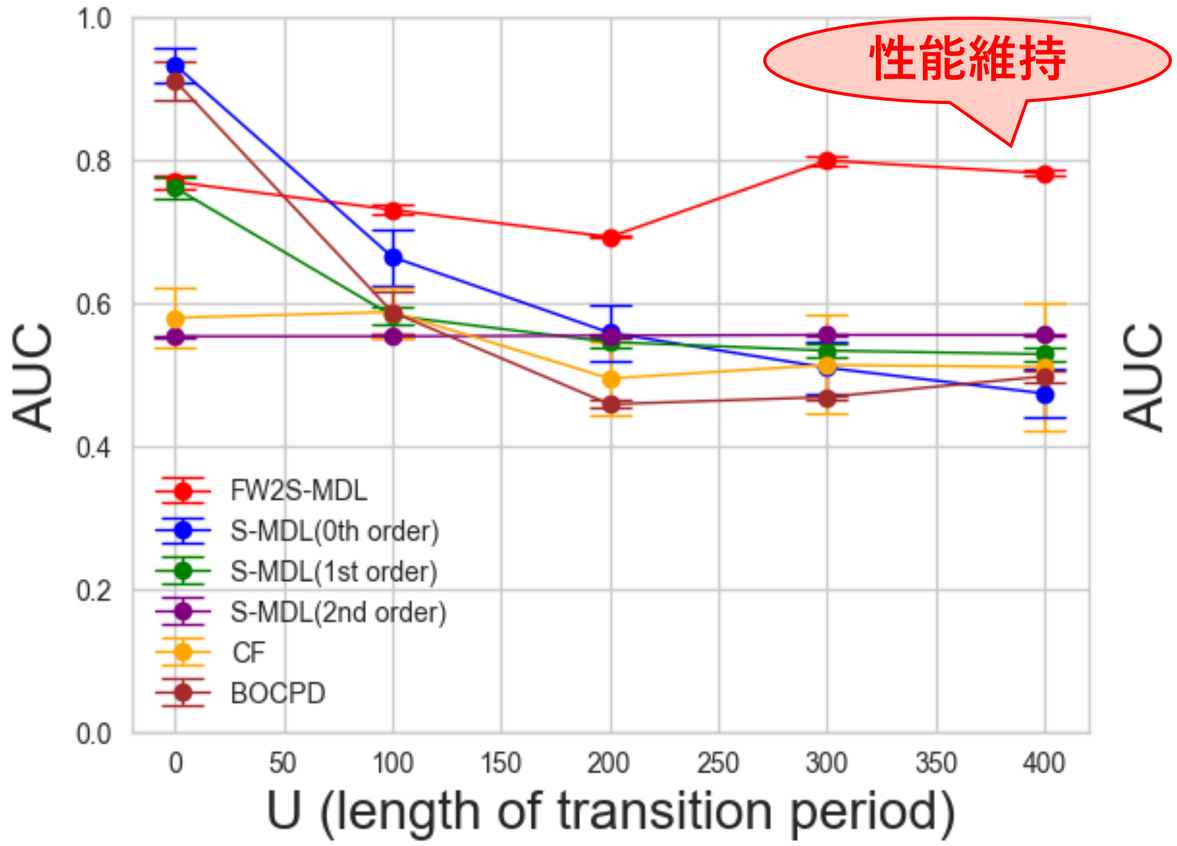


分散が徐々に変化するデータ

人工データ実験結果

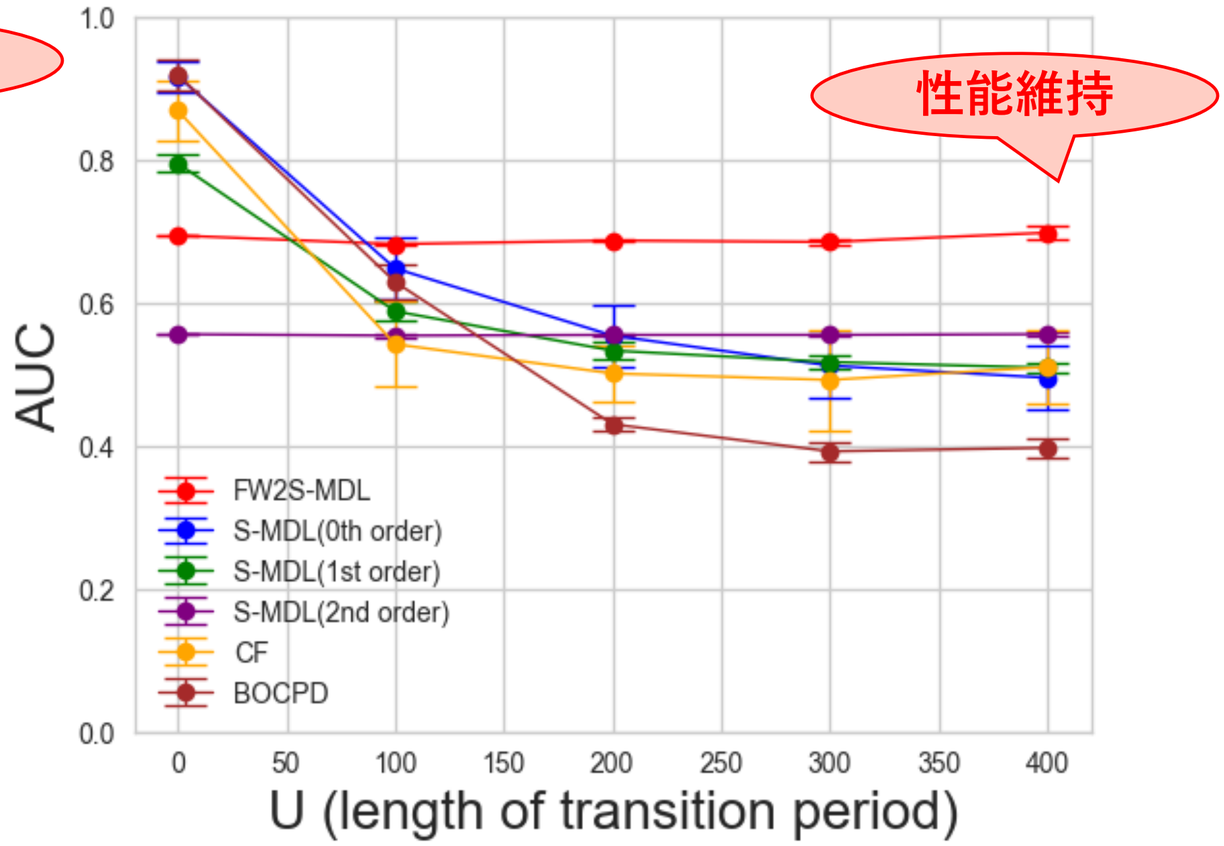
提案手法は変化過渡期が長くなっても性能を維持
→対抗手法は性能低下

検知難易度の上昇 →



平均が変化するデータセットでの結果

検知難易度の上昇 →



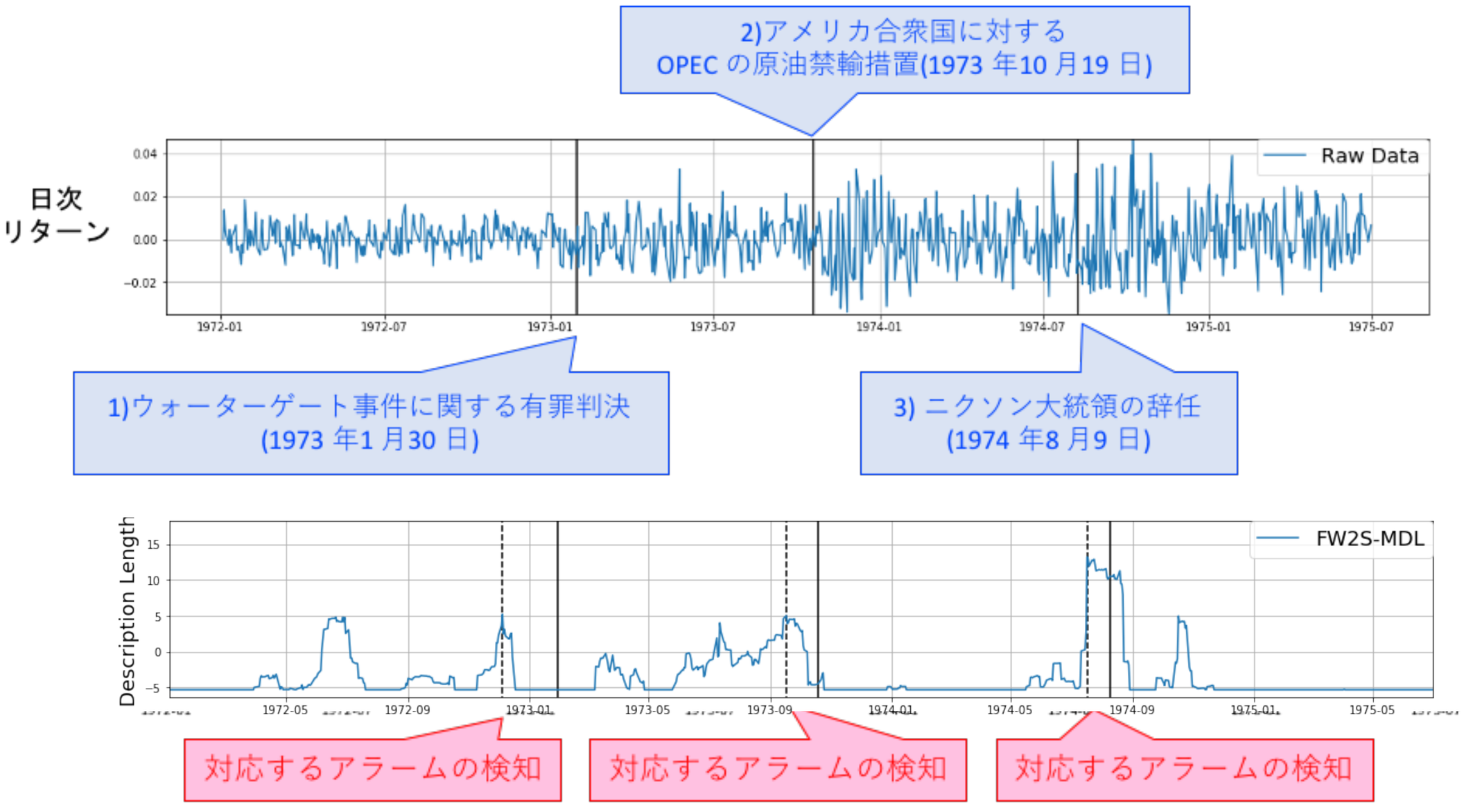
分散が変化するデータセットでの結果

→Change Probabilityの-logの増加率を考える操作が
漸進的変化を突発的変化検知に変換している可能性を示唆

*構造的変化についても、漸進的変化をもつ人工データを生成し、提案手法が対抗手法より良い性能を示すことを確認。

実データによる実験結果

現実でのイベントに対応するアラームを検知



Dow Jones 日次リターンに対する提案手法の結果

Agenda

4. まとめ

まとめ

導入

- 変化検知は、**時系列データの発生分布が変化**する点を検知する問題。
- **漸進的変化検知・変化予兆検知**への研究の展開がある。

2段階MDL変化統計量

- **2段階の変化検知**により漸進的変化の変化予兆検知を実現。
- パラメータ変化および構造的変化においても定義が可能。

実験

- **人工データ**では、定量的に良い性能。
- **実データ**では、現実でのイベントに対応したアラームを検知。

今後の展望

- **予兆検知の意味での性能保証**に関する理論解析。
  - MDL変化統計量が持っていた性能保証を利用。

参考文献I

- 山西健司編著. データサイエンスの数理数理で読み解くデータの価値 数理科学2019年06 月号. サイエンス社, 2019 年6 月.
- Yamanishi, Kenji, and Kohei Miyaguchi. "Detecting gradual changes from data stream using MDL-change statistics." *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016.
- Hirai, So, and Kenji Yamanishi. "Detecting changes of clustering structures using normalized maximum likelihood coding." *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. 2012.
- Hirai, So, and Kenji Yamanishi. "Detecting latent structure uncertainty with structural entropy." *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.
- Hirai, So, and Kenji Yamanishi. "Detecting Model Changes and their Early Warning Signals Using MDL Change Statistics." *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.
- Takahashi, Toshimitsu, Ryota Tomioka, and Kenji Yamanishi. "Discovering emerging topics in social streams via link-anomaly detection." *IEEE Transactions on Knowledge and Data Engineering* 26.1 (2012): 120-130.
- Adams, Ryan Prescott, and David JC MacKay. "Bayesian online changepoint detection." *arXiv preprint arXiv:0710.3742* (2007).

参考文献II

- Basseville, Michele, and Igor V. Nikiforov. *Detection of abrupt changes: theory and application*. Vol. 104. Englewood Cliffs: Prentice hall, 1993.
- Takeuchi, Jun-ichi, and Kenji Yamanishi. "A unifying framework for detecting outliers and change points from time series." *IEEE transactions on Knowledge and Data Engineering* 18.4 (2006): 482-492.
- Kenji Yamanishi, Linchuan Xu, Ryo Yuki, Shintaro Fukushima, and Chuan hao Lin. Change sign detection with differential mdl change statistics and its applications to covid-19 pandemic analysis, 2020.
- Miyaguchi, Kohei, and Kenji Yamanishi. "Online detection of continuous changes in stochastic processes." *International Journal of Data Science and Analytics* 3.3 (2017): 213-229.
- Huang, David Tse Jung, et al. "Detecting volatility shift in data streams." *2014 IEEE International Conference on Data Mining*. IEEE, 2014.
- Ohsawa, Yukio. "Graph-based entropy for detecting explanatory signs of changes in market." *The Review of Socionetwork Strategies* 12.2 (2018): 183-203.

参考文献III

- Rissanen, Jorma. "Modeling by shortest data description." *Automatica* 14.5 (1978): 465-471.
- Yamanishi, Kenji, and Shintaro Fukushima. "Model change detection with the MDL Principle." *IEEE Transactions on Information Theory* 64.9 (2018): 6115-6126.
- Shtar'kov, Yurii Mikhailovich. "Universal sequential coding of single messages." *Problemy Peredachi Informatsii* 23.3 (1987): 3-17.

βの設定方法

仮説検定

$$H_0 : x_1^n \sim p(x_1^n, \theta_0),$$
$$H_1 : x_1^t \sim p(x_1^t, \theta_1), x_{t+1}^n \sim p(x_{t+1}^n, \theta_2)(\theta_1 \neq \theta_2).$$

MDL変化統計量とChange Probabilityの比較

検定手法	統計量と検定方法	誤り確率(の上界)
MDL変化統計量	$L(x_1^n) - \{L(x_1^t) + L(x_{t+1}^n)\} > n\epsilon$	Type I error prob. $\leq \exp(-n(\epsilon - \frac{\log C_n}{n}))$, Type II error prob. $\leq \exp(-\frac{n}{2}(D - \frac{\log C_t C_{n-t}}{n} - \epsilon))$.
Change Probability	$p_{change}^t(x_1^n) > 1 - \delta$	2. 誤り確率の上界のβ, δによる表示

1. $((\beta, \delta)$とϵを対応させた下での)検定の同一性の証明

3. 誤り確率の上界をδ以下に抑えるβの設定方法の導出

Change ProbabilityとMDL変化統計量の同一性

Lemma 1 (The relationship between the change probability and the MDL change statistics) *The hypothesis testing that is given by the MDL change statistics and that given by the change probability are the same under the condition that* $\beta = (\log((1-\delta)/\delta))/n\epsilon.$ →(β, δ)とϵの対応関係

Proof p^t_{change} can be written with the MDL change statistics in the following form:

$$p^t_{change}(x_1^n) = \frac{1}{1 + \exp\Big(-\beta\big\{L(x_1^n) - \big\{L(x_1^t) + L(x_{t+1}^n)\big\}\big\}\Big)}.$$

Then, $p^t_{change}(x_1^n) > 1 - \delta$ is transformed into the following inequality:

$$\big\{L(x_1^n) - \big\{L(x_1^t) + L(x_{t+1}^n)\big\}\big\} - \frac{1}{\beta}\log\frac{1-\delta}{\delta} > 0,$$

→$n\epsilon$と一致すれば良い

and it is equal to the hypothesis testing given by the MDL change statistics if $\beta = (\log((1-\delta)/\delta))/n\epsilon$.

検定の誤り確率(の上界)とβの設定方法

Lemma 2 (Type I and II probabilities) *The type I and II error probabilities of the hypothesis testing provided by the change probability are upper-bounded as follows:*

$$\text{Type I error prob.} \leq \exp\left(-\frac{1}{\beta}\log\frac{1-\delta}{\delta} + \log C_n\right),$$

$$\text{Type II error prob.} < \exp\left(-\frac{n}{2}d_n(p_{NML}, p_{\theta_1,\theta_2}) + \frac{1}{2}\log C_t C_{n-t} + \frac{1}{2\beta}\log\frac{1-\delta}{\delta}\right).$$

→2. 誤り確率の上界のβ,δによる表示

Theorem 2 *The type I error probability of the hypothesis testing given by the change probability is less than δ if*

$$\beta \leq \frac{\log(1-\delta) - \log\delta}{\log C_n - \log\delta}.$$

→3. 誤り確率の上界をδ以下に抑えるβの設定方法の導出

This condition is obtained by transforming the following inequality:

$$\exp\left(-\frac{1}{\beta}\log\frac{1-\delta}{\delta} + \log C_n\right) \leq \delta.$$

With this theorem, we set $\beta = \frac{\log(1-\delta)-\log\delta}{\log C_n - \log\delta}$ because the hypothesis testing with a low β is too conservative.

モデル変化への拡張の概要

モデル変化におけるMDL変化統計量の利用

- モデル変化におけるMDL変化統計量は既に提案されている(Yamanishi and Fukushima 2018)。

$$\Phi_t(x_1^n) \stackrel{\text{def}}{=} L_0(x_1^n) - L_1(x_1^n) - n\epsilon,$$

$$L_0(x_1^n) \stackrel{\text{def}}{=} \min_{M_0}\{L_{NML}(x_1^n; M_0) + L(M_0)\},$$

$$L_1(x_1^n) \stackrel{\text{def}}{=} \min_{M_1, M_2}\{L_{NML}(x_1^t; M_1) + L_{NML}(x_{t+1}^n; M_2) + L(M_1, M_2)\}.$$

- モデル変化におけるChange Probabilityも同様に定義できる。

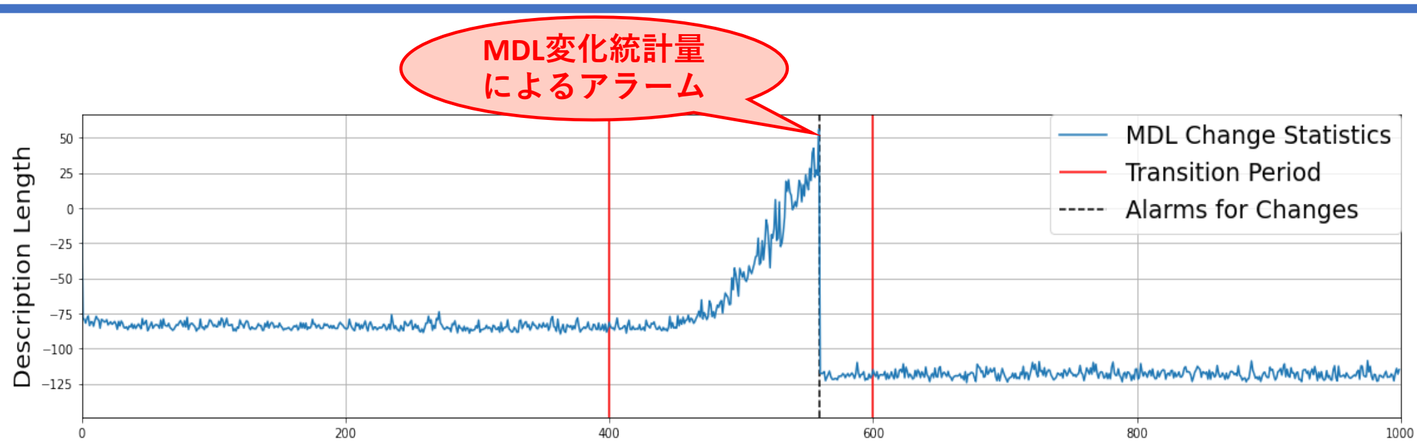
$$p_{change}^t(x_1^n) \stackrel{\text{def}}{=} \frac{\exp(-\beta L_1(x_1^n))}{\exp(-\beta L_0(x_1^n)) + \exp(-\beta L_1(x_1^n))}.$$

- Type-I error probablityにおける議論を同様に行え、βを設定できる。

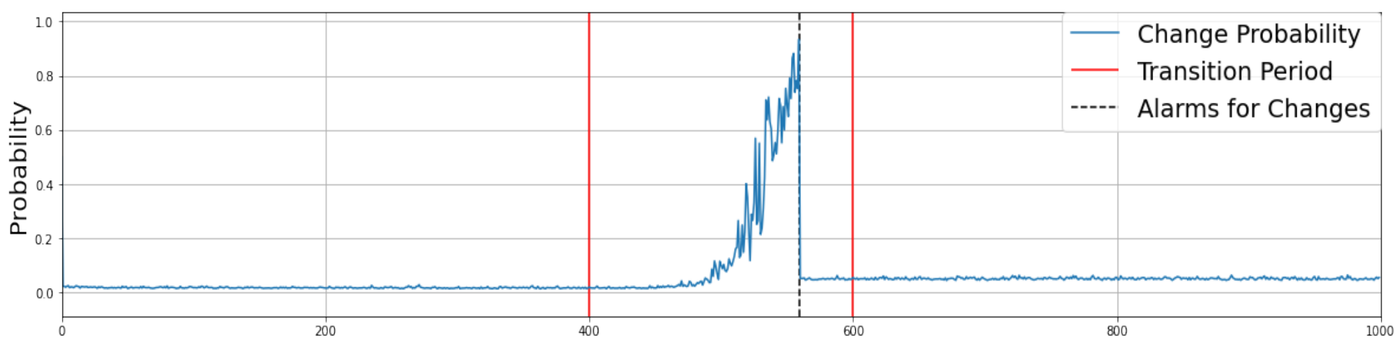
$$\beta \leq \frac{\log(1-\delta) - \log\delta}{\log C_n(M_0^*) + L(M_0^*) - \log\delta}.$$

モデル変化での2段階MDL変化統計量の例

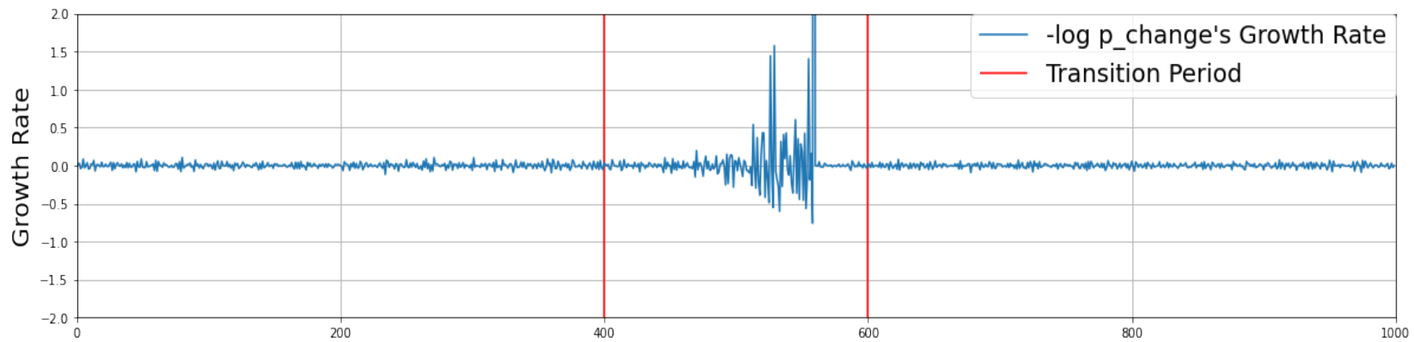
MDL
変化統計量



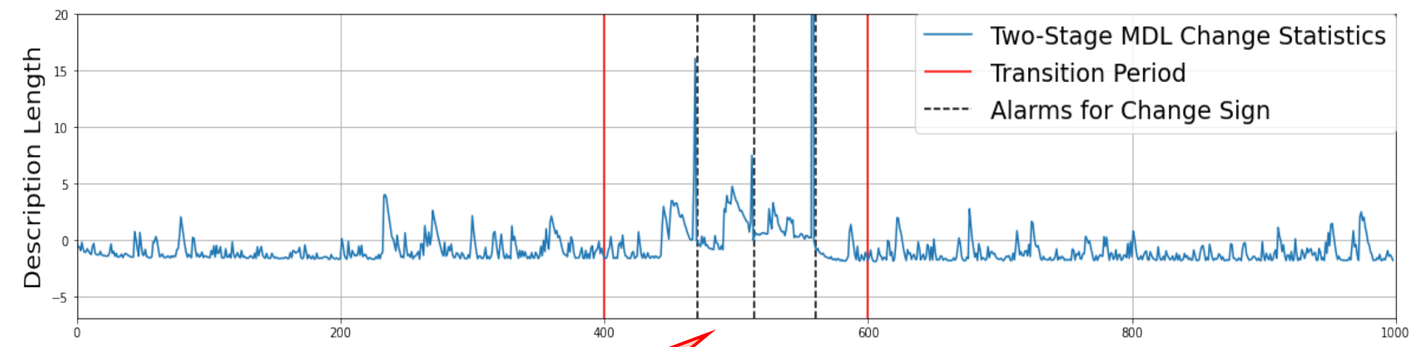
Change
Probability



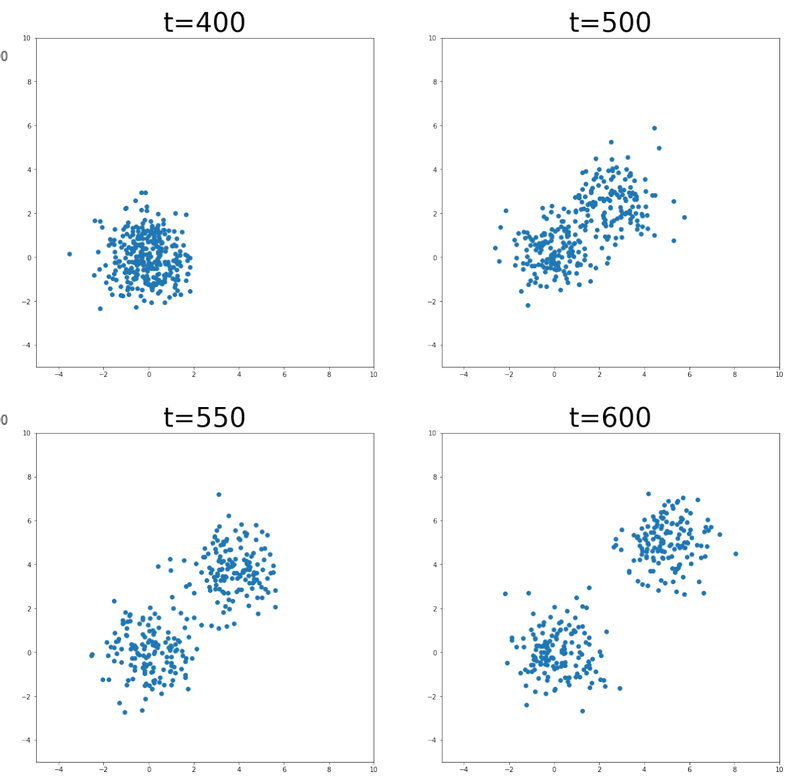
-log p_change
の増加率



2段階MDL
変化統計量



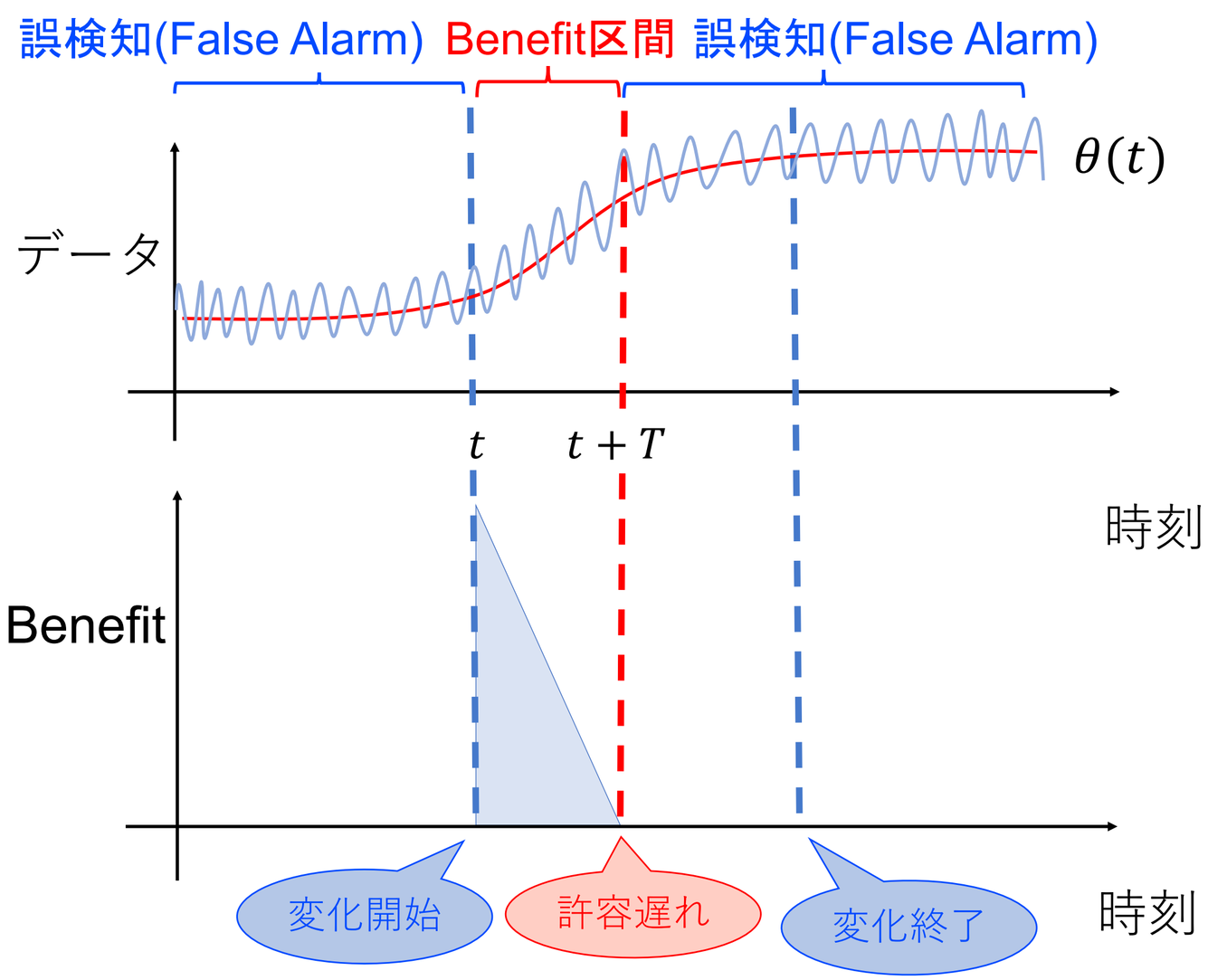
提案手法による
アラーム



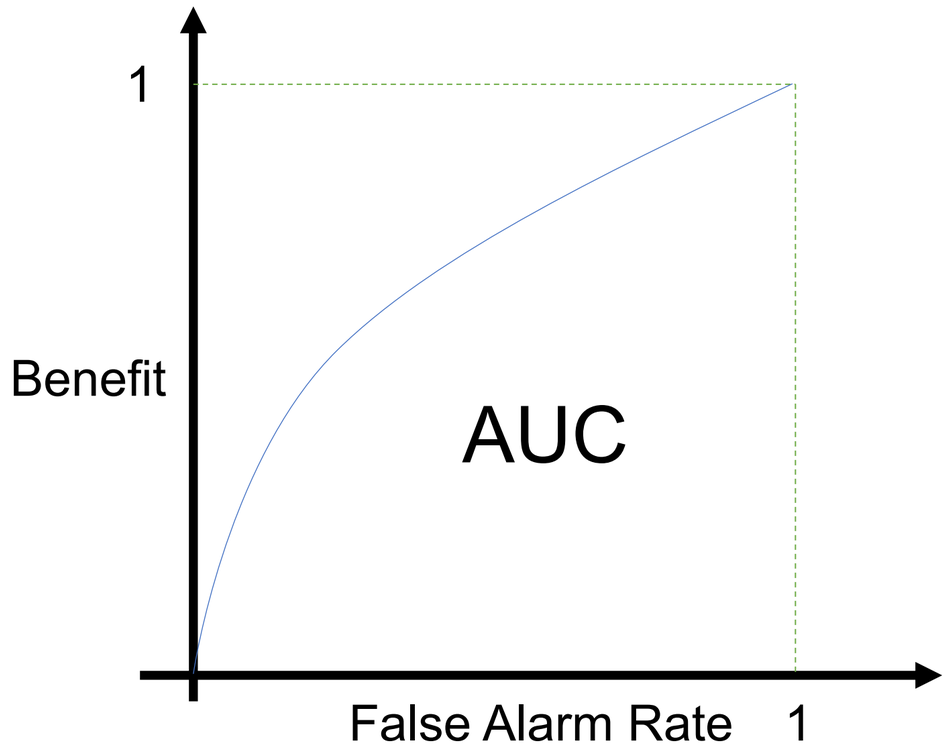
データの時系列(一部)

Area Under Curve (AUC)

閾値を変えながら
誤検知率 vs. 予兆検知の早さのグラフのAUCを計算



→アラームが上がれば対応するBenefitを与える



$$Benefit = \frac{|t-t^*|}{T},$$

$where$

$t: estimated\ changepoint$

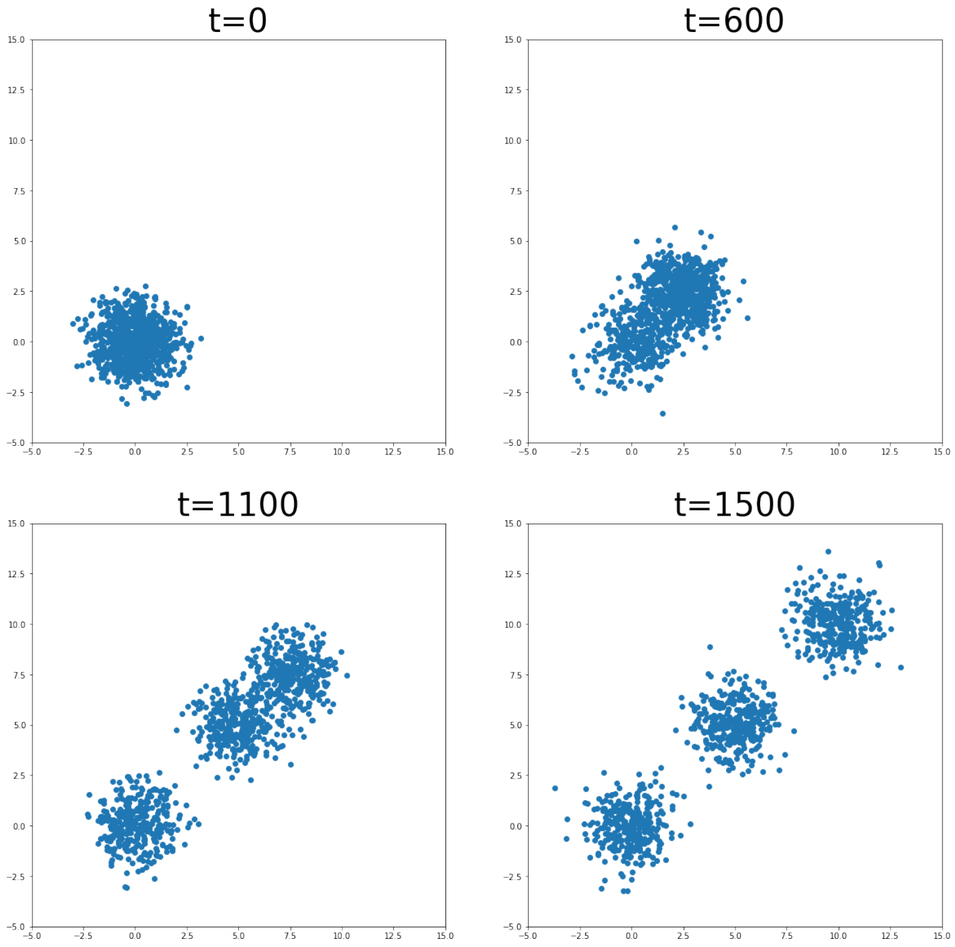
$t^*: true\ changepoint$

$T: maximum\ tolerant\ delay$

生成したデータ

Gaussian Mixture Models (GMMs) クラスタ数増加

- クラスタ数が1→2→3個に変化するデータセットを生成(変化点は二つ)
- クラスタ中心が動くことでクラスタ数が変化する。移動スピードはパラメータαによってコントロールされる。



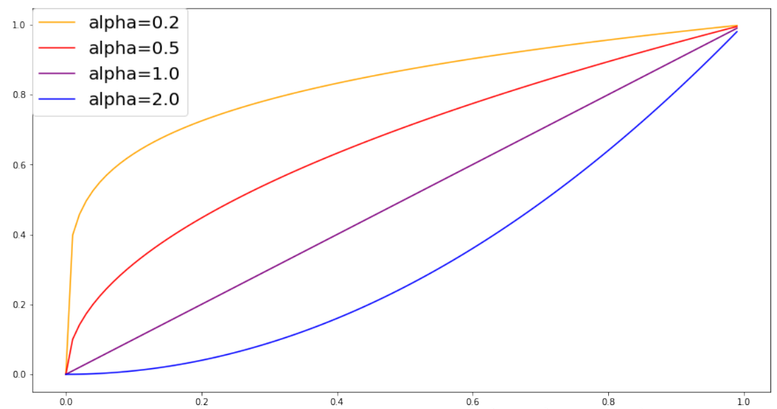
データの時系列(一部)

$$\begin{cases} K^* = 1, \mu = (\mu_1) & (1 \le t \le t_1), \\ K^* = 2, \mu = (\mu_1, u_1(t)) & (t_1 + 1 \le t \le t_1 + U), \\ K^* = 2, \mu = (\mu_1, \mu_2) & (t_1 + U + 1 \le t \le t_2), \\ K^* = 3, \mu = (\mu_1, \mu_2, u_2(t)) & (t_2 + 1 \le t \le t_2 + U), \\ K^* = 3, \mu = (\mu_1, \mu_2, \mu_3) & (t_2 + U + 1 \le t \le T), \end{cases}$$

where

$$u_i(t) = (1 - \Delta t_i^\alpha) \cdot \mu_i + \Delta t_i^\alpha \cdot \mu_{i+1},$$

$$\Delta t_i = \frac{t - \tau_i}{U}.$$



クラスタ中心の移動の速さ
αが大きいほど、早期に検知しやすくなる。

Result

結果

- データを10本生成。1本でパラメータチューニングし、残り9本で性能評価。
- FARが0.001を下回る条件下で、Benefitを最大化するようパラメータをチューニング
- 提案手法は、αが大きくなってもBenefitの減衰が少ないことを確認。

$\alpha = 0.0$	Benefit	FAR
DMS	$\mathbf{1.0000 \pm 0.0000}$	0.0000 ± 0.0000
SMCS	$\mathbf{1.0000 \pm 0.0000}$	0.0000 ± 0.0000
Prop.	$\mathbf{1.0000 \pm 0.0000}$	0.0000 ± 0.0000

提案手法

$\alpha = 0.2$	Benefit	FAR
DMS	0.9861 ± 0.0171	0.0000 ± 0.0000
SMCS	$\mathbf{0.9950 \pm 0.0000}$	0.0000 ± 0.0000
Prop.	$\mathbf{0.9950 \pm 0.0000}$	0.0001 ± 0.0003

提案手法

$\alpha = 0.5$	Benefit	FAR
DMS	0.8536 ± 0.0102	0.0000 ± 0.0000
SMCS	0.8611 ± 0.0072	0.0008 ± 0.0008
Prop.	$\mathbf{0.9608 \pm 0.0258}$	0.0004 ± 0.0006

提案手法

$\alpha = 1.0$	Benefit	FAR
DMS	0.6192 ± 0.0123	0.0000 ± 0.0000
SMCS	0.6058 ± 0.0342	0.0003 ± 0.0004
Prop.	$\mathbf{0.7431 \pm 0.0557}$	0.0007 ± 0.0007

提案手法

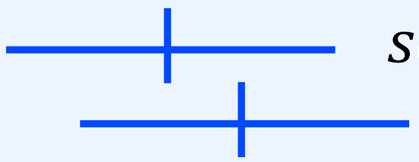
$\alpha = 2.0$	Benefit	FAR
DMS	0.3806 ± 0.0087	0.0000 ± 0.0000
SMCS	$\mathbf{0.4975 \pm 0.1041}$	0.0009 ± 0.0007
Prop.	$\mathbf{0.5114 \pm 0.0983}$	0.0003 ± 0.0004

提案手法

アルゴリズムの計算量

Sequential MDL (S-MDL) アルゴリズム(Yamanishi and Miyaguchi 2016)

- データを一定の長さの窓で切り出し逐次的にスコア計算。



$score_1$

$score_2$

…

→オンラインにスコア計算

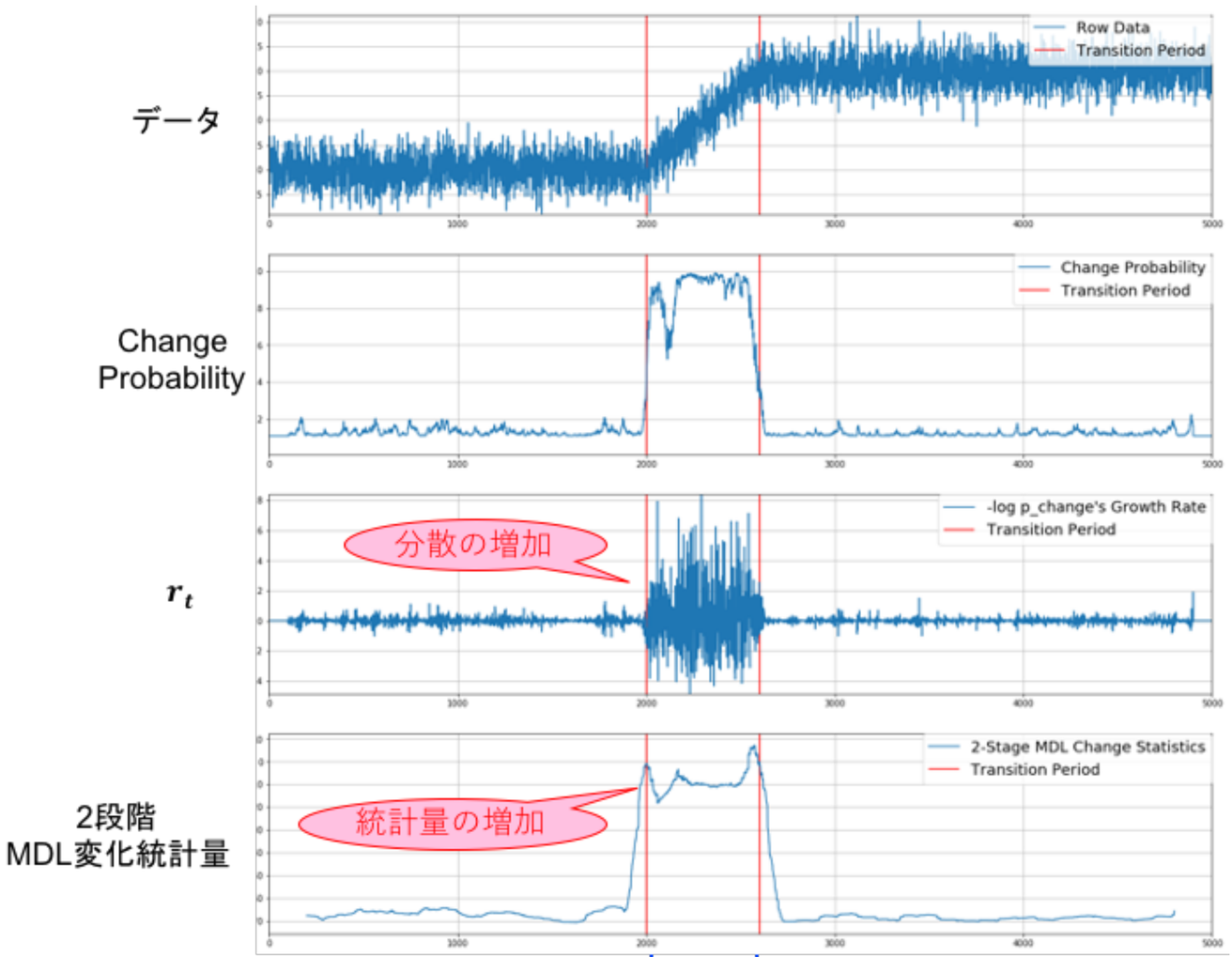
$x_1^n = x_1, \ldots, x_n$

計算量: $\mathbf{O(nf(h))}$

n:データ数

f: 最尤推定量の計算コスト

h: ウィンドウサイズ



データ　$x_1^n = x_1, \ldots, x_n$

▼ S-MDL的に計算

1段階目　$p_{change}^t(x_1^n)$

▼

$-\log p_{change}^t$ の増加率　r_t

▼ S-MDLを適用し計算

2段階目　$\Phi_{2-stage}^t$

→ 総計算量$\mathbf{O(nf(h_1) + nh_2)}$

分散増加に関する考察

変化が十分gradualに変化することを仮定。

$$p_{change}^{t+1} = p_{change}^{t} + \Delta p. \quad \Delta p / p_{change}^{t} << 1.$$

r_tに関するTaylor展開

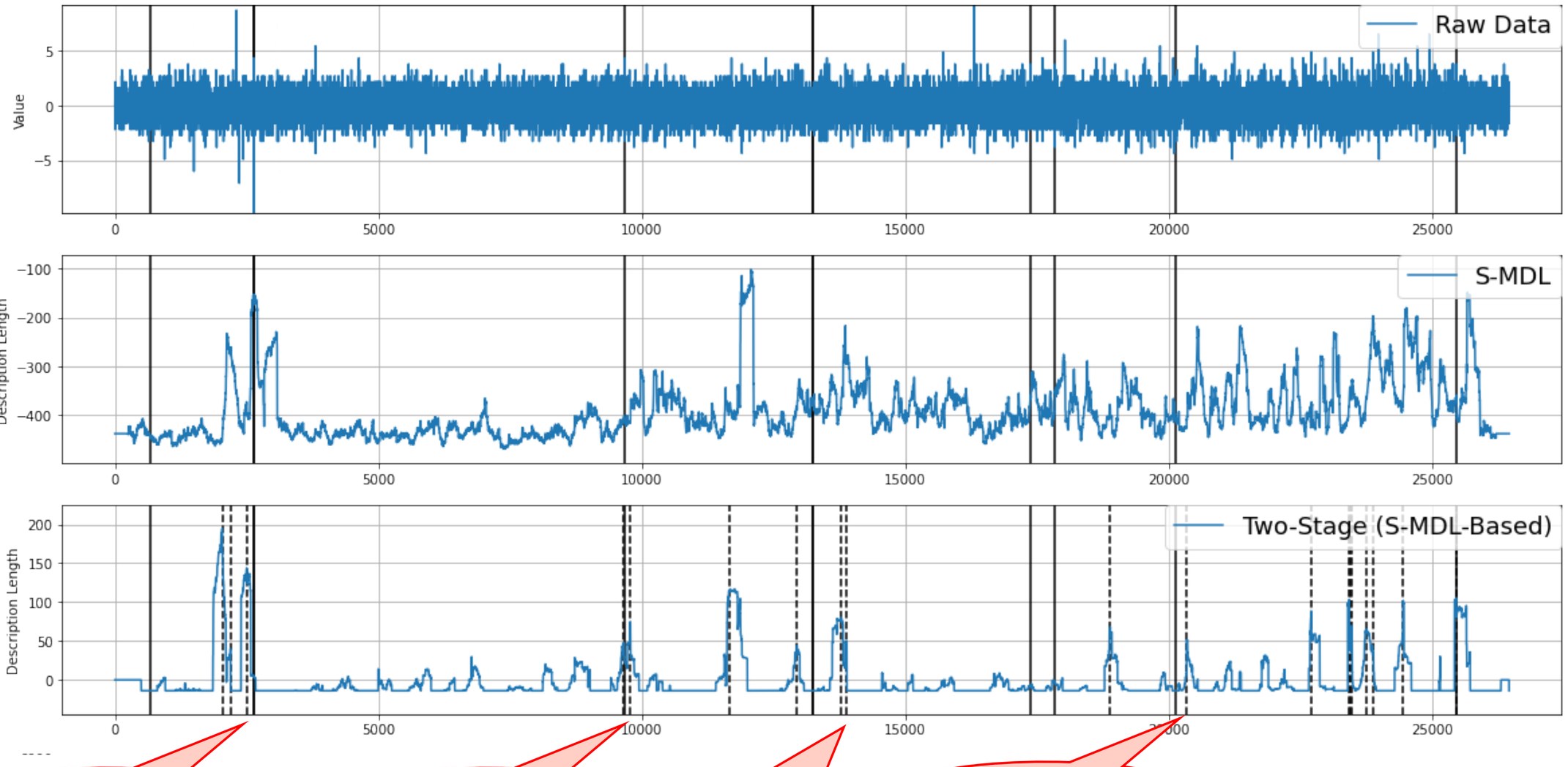
$$\begin{aligned}
r_{t+1} &= \frac{-\log p_{change}^{t+1}}{-\log p_{change}^{t}} - 1, \\
&= \frac{-\log(p_{change}^{t} + \Delta p)}{-\log p_{change}^{t}}, \\
&= \frac{-\log p_{change}^{t}(1 + \frac{\Delta p}{p_{change}^{t}})}{-\log p_{change}^{t}} - 1, \\
&= \frac{-\log(1 + \frac{\Delta p}{p_{change}^{t}})}{-\log p_{change}^{t}}, \\
&= \frac{-1}{-\log p_{change}^{t}} \left(\frac{\Delta p}{p_{change}^{t}} + o(\Delta p) \right), \\
&= \frac{-\Delta p}{-p_{change}^{t} \log p_{change}^{t}} + o(\Delta p).
\end{aligned}$$

→p_{change}^{t}の増加につれ分散の増加に寄与

Production Condition Dataset

現実でのイベントに対応するアラームを検知

データ列
(8次元)



S-MDL

Two-Stage

検知　検知　検知　検知