

潜在空間を高度活用した ディープナレッジの発見

山西健司

東京大学大学院情報理工学系研究科

主たる共同研究者

朝岡亮

聖隷クリストファー大、静岡大学電子工学研究所、光産業創成大学院大学 光産業創成研究科

2021年12月1日

内容

- 0. 目標設定
- 1. 潜在空間埋め込みの性能解析・構造最適化
- 2. 潜在変数モデル選択とその応用
- 3. 連続パラメータの変化予兆検知
- 4. 離散構造の変化予兆検知
- 5. 計算論的眼科学
- 6. 今後の展望

目標設定

CREST

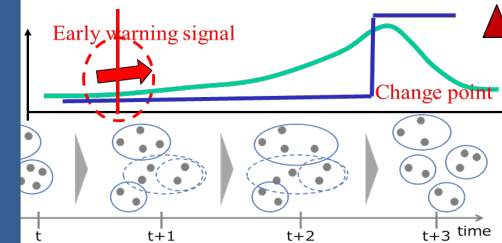
複雑データからの
ディープナレッジ
の発見と価値化

潜在空間の新しい活用の方法論を創出し、より深い知識を発掘する

理論研究

- ・潜在空間埋め込み
- ・潜在構造最適化

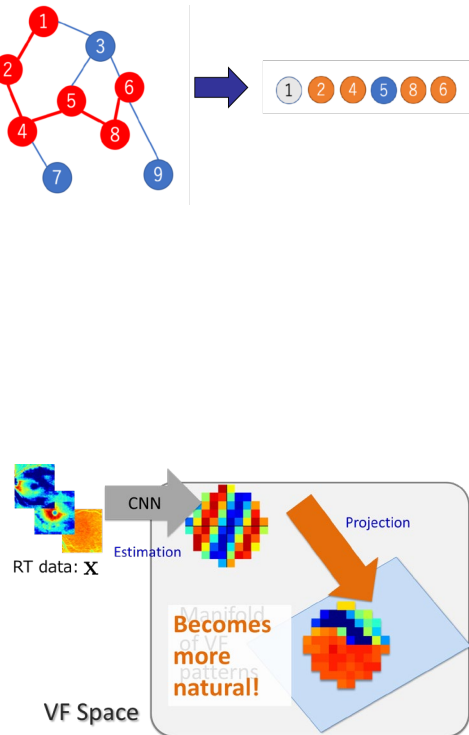
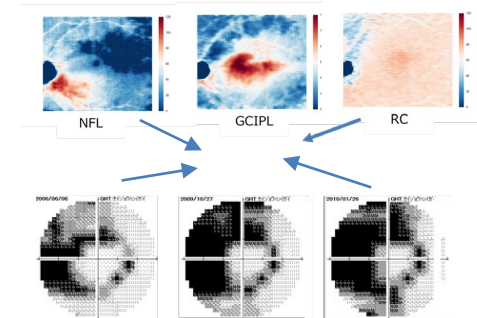
変化予兆検知



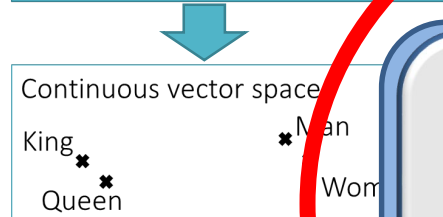
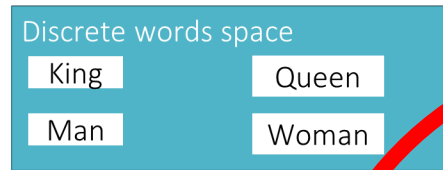
計算論的眼科学

大規模データを
用いた医学的再検証

ヘテロ情報を用いた
緑内障診断・予測



潜在空間埋め込みの研究



埋め込み次元推定

**Dimension
Selection of
Embedding**

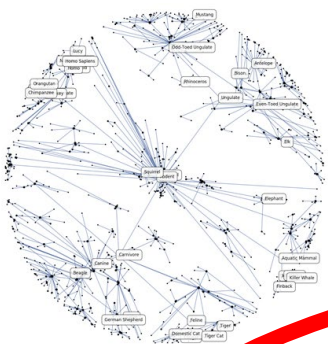
[Hung, Yamanishi
Entropy 21]

潜在空間
埋め込み

ヘテロNW埋め込み

**Heterogeneous
NW Embedding**

[Xu, Wang, Cao, He, Yu,
Yamanishi TKDE 21]



双曲空間埋め込みの理論的解明

**Hyperbolic Ordinal
Embedding**

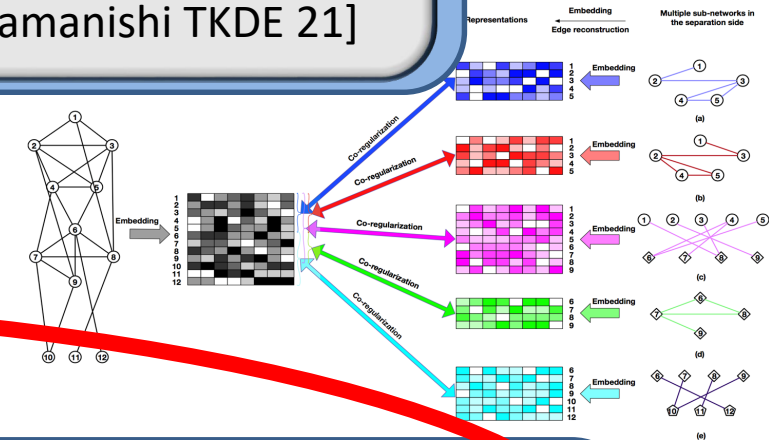
[Suzuki, Wang, Tian, Nitanda,
Yamanishi ACM 19]

**Gen.Error for Hyperbolic
Ordinal Embedding**

[Suzuki, Nitanda, Wang, Xu,
Yamanishi, Cavazza ICML19]

**Gen.Error for Hyperbolic
Graph Embedding**

[Suzuki, Nitanda, Wang, Xu,
Yamanishi, Cavazza NeuLIPS19]



1. 潜在空間埋め込み性能解析、表現最適化

[Suzuki, Wang, Tian, Nitanda, Yamanishi ACML2019]

[Suzuki, Nitanda, Wang, Xu, Yamanishi, Cavazza ICML2021]

[Suzuki, Nitanda, Wang, Xu, Yamanishi, Cavazza NeurIPS2021]

[Hung and Yamanishi Entropy 2021]

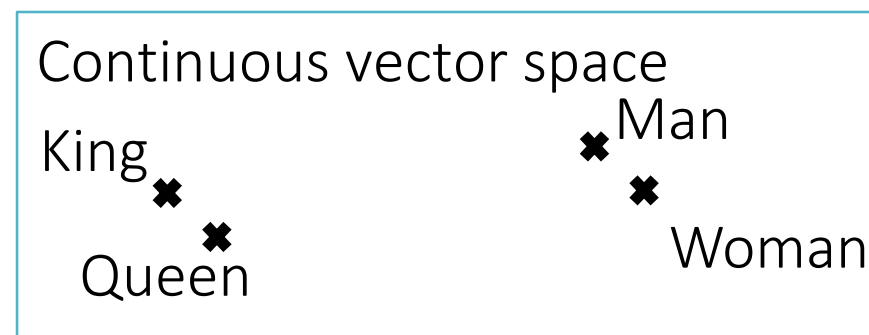
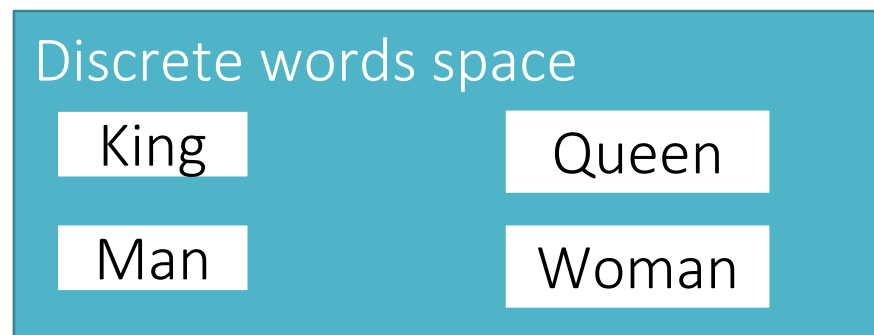
[Xu, Wang, Yamanishi IEEE TKDE 2021]

単語埋め込みの次元選択

単語空間(離散変数)を低次元実数値空間に埋め込み、類似単語を予測

[Mikolov et al. 2013]

[Pennington et al. 2014][Levy and Goldberg 2014]



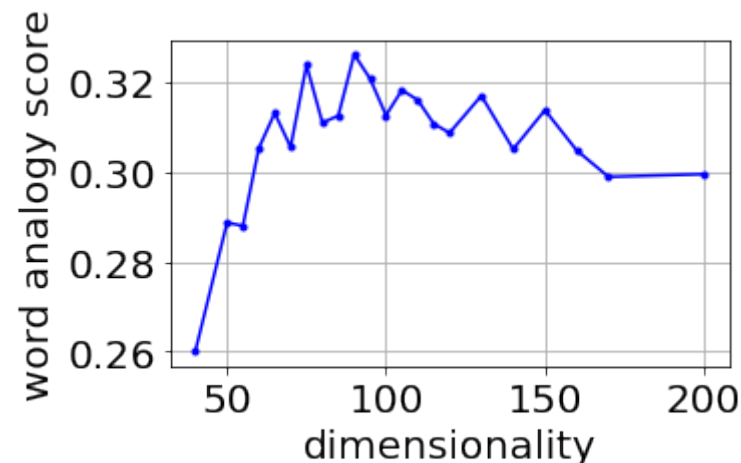
Model performance

$$\overrightarrow{King} - \overrightarrow{Queen} = \overrightarrow{Man} - \overrightarrow{Woman}$$

Problem

埋め込み空間の次元は高いほど
単語推定(識別)精度が高いとは限らない

⇒ 最適な次元をいかに選ぶか？



単語埋め込みの確率モデル

埋め込み次元 = 指数部分の非負値行列因子分解のランク

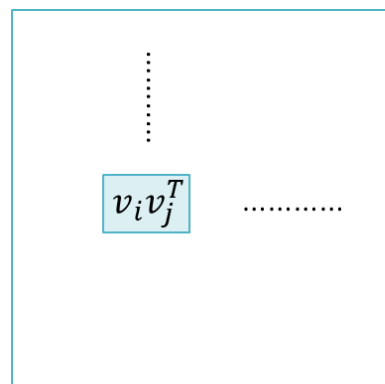
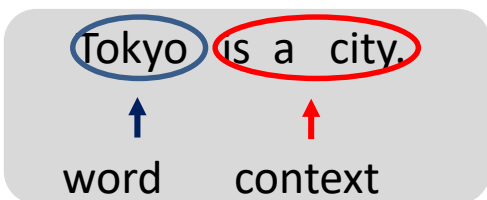
- SkipGram:

$$P_{SG}(\mathbf{c}|\mathbf{w}; E, F) = \prod_{i=1}^n P(c_i|w_i; E, F) = \prod_{i=1}^n \frac{\exp(w_i^T E F c_i)}{\sum_{c' \in V_C} \exp(w_i^T E F c')}$$

context word

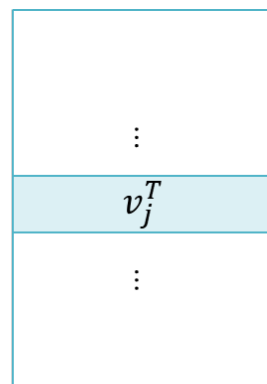
d: 埋め込み次元

[Yin and Shen 2018]



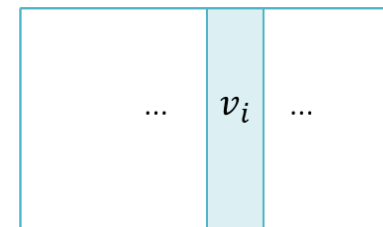
signal matrix

=



Word embedding
Matrix (E)

×



Word context
Matrix (F)

SNML符号長に基づく埋め込み次元選択

記述長最小(MDL)原理に基づきSNML符号長最小化規準で埋め込み次元を選択

■ 逐次的正規化最尤符号長規準 (Sequentially Normalized Maximum Likelihood)

.....逐次的に計算するために1回あたりの計算量が軽い

正規化最尤符号長
(NML符号長)

$$\mathcal{L}_{NML}(\mathbf{c}|\mathbf{w}; \mathcal{M}_d) = -\log P(\mathbf{c}|\mathbf{w}; \hat{\theta}(\mathcal{D})) + \log \left(\sum_{c \in V_c^n} P(\mathbf{c}|\mathbf{w}; \hat{\theta}(\mathcal{D})) \right)$$

最尤推定量

逐次的正規化
最尤符号長

$$\mathcal{L}_{SNML}(\mathbf{c}|\mathbf{w}; \mathcal{M}_d) = \sum_{i=1}^n \mathcal{L}_{SNML}(c_i|w^i, c^{i-1}; \mathcal{M}_d) \Rightarrow \min \text{ w.r.t. } d$$

$$\mathcal{L}_{SNML}(c_i|w^i, c^{i-1}; \mathcal{M}_d) = -\log P(c_i|w^i, c^{i-1}; \hat{\theta}(w^i, c^i)) + \log \underbrace{\sum_{c \in V_c} P(c|w^i, c^{i-1}; \hat{\theta}(w^i, c^{i-1}, c))}_{\text{正規化項}}$$

最尤推定量

where $c^i = c_1 c_2 \dots c_i$ and $w^i = w_1 w_2 \dots w_i$

■ さらなる計算量削減のための施策

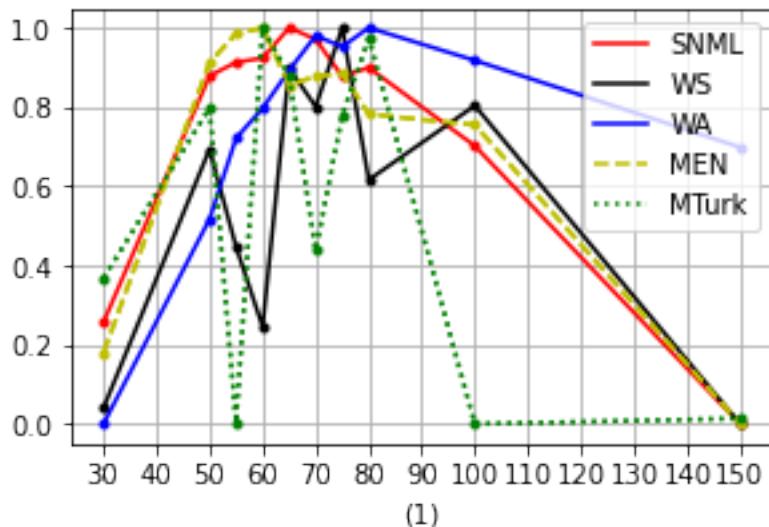
[Hung and Yamanishi *Entropy* 2021]

- ・正規化項をImportance Samplingで近似
- ・部分列のみから次元選択

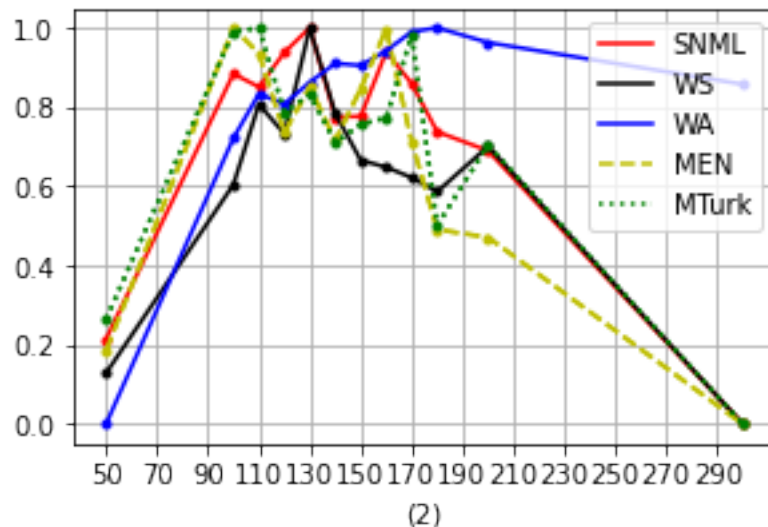
実験結果

SNMLが他規準に比べてNLPタスクに最も近い次元を推定している

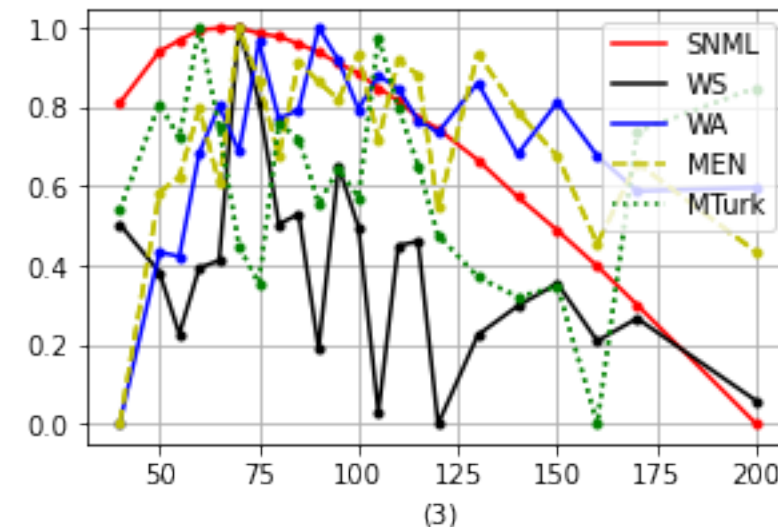
[Hung and Yamanishi *Entropy* 2021]



oSG with Text8



oSG with Wikipedia



SGNS with Text8

提案手法

	SNML	WS	WA	MEN	MTurk	CV	AIC	BIC	PIP
Text8 (oSG)	65	75	80	60	60	65	u	u	120
Wikipedia (oSG)	130	130	180	100	110	130	200	u	u
Text8 (SGNS)	70	70	95	70	60	60	70	u	105

NLP Task

PIP.... minimizing Pairwise Inner Product loss

[Zin and Shen 2018]

双曲順序付き埋め込み

Hyperbolic Ordinary Embedding

双曲空間は階層データの表現に適している

[Lamping Rao 1994] [Nickel Kiela 2017]

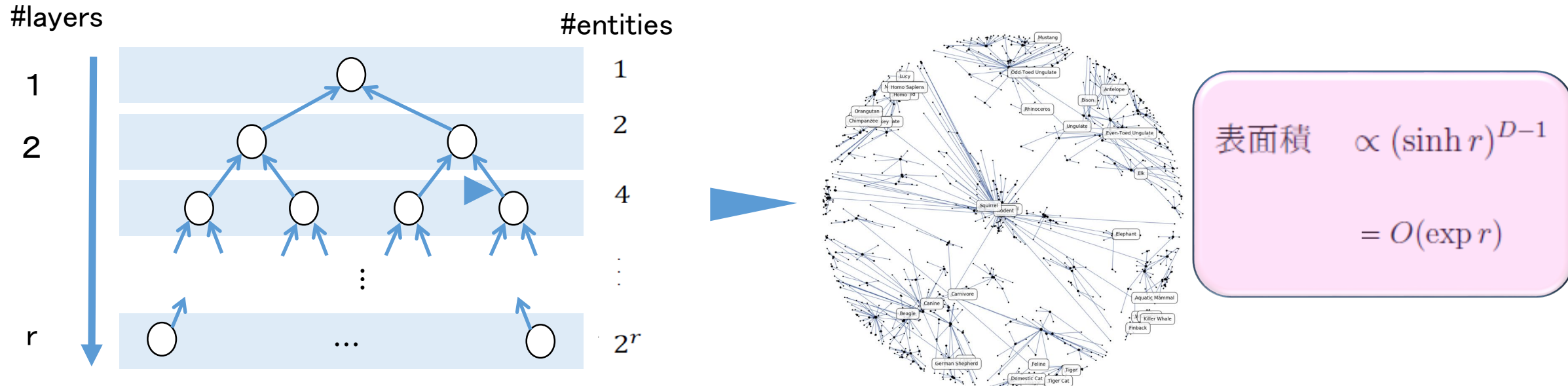


Figure 2 (b) in [Nickel & Kiela, NIPS, 2017]

- Hyperbolic Ordinary Embedding

$i \leftrightarrow j$ is more similar than $i \leftrightarrow k \Rightarrow d_{\mathbb{H}^D}(x_i, x_j) \leq d_{\mathbb{H}^D}(x_i, x_k)$

- 距離関数 $d_{\mathbb{H}^D}(x, y) = \text{arcosh} \left(1 + \frac{2(x-y)^\top(x-y)}{(1-x^\top x)(1-y^\top y)} \right)$

双曲順序付き埋め込みの確率モデル

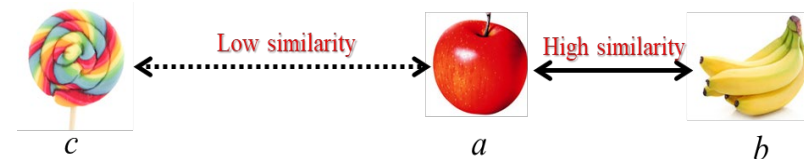
- Conditional Probability

[Suzuki, Wang, Tian, Nitanda, Yamanishi ACML2019]

- $y_t = -1 \Leftrightarrow j_t$ is more similar than k_t to i_t
- $y_t = +1 \Leftrightarrow k_t$ is more similar than j_t to i_t

$$P(y = -1 | (i, j, k); (x_n)_{n \in [N]}) = \frac{f(d_{\mathbb{H}^D}(x_i, x_j))}{f(d_{\mathbb{H}^D}(x_i, x_j)) + f(d_{\mathbb{H}^D}(x_i, x_k))},$$

- $f: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$: kernel function (decreasing function)



- Objective: Minimizing Log Likelihood

$$\text{Minimize } \sum_{t \in [T]} [-\log P(y_t | (i_t, j_t, k_t); (x_n)_{n \in [N]})], \text{ w. r. t. } (x_n)_{n \in [N]}$$

階層データに対しては双曲空間の方がより少ない損失を達成

Error rates

Cora	$D = 2$	$D = 4$	$D = 8$	$D = 16$
1-EMOE	0.3513 $\pm$ 0.0002	0.3258 $\pm$ 0.0002	0.3131 $\pm$ 0.0002	0.2973 $\pm$ 0.0003
2-EMOE	0.3584 $\pm$ 0.0003	0.3311 $\pm$ 0.0004	0.3091 $\pm$ 0.0002	0.2947 $\pm$ 0.0002
G-EPOE	0.3695 $\pm$ 0.0003	0.3525 $\pm$ 0.0005	0.3348 $\pm$ 0.0004	0.3103 $\pm$ 0.0003
t-EPOE	0.3629 $\pm$ 0.0003	0.3367 $\pm$ 0.0002	0.3156 $\pm$ 0.0002	0.3007 $\pm$ 0.0002
1-HMOE	0.3481 $\pm$ 0.0003	0.3245 $\pm$ 0.0002	0.3074 $\pm$ 0.0004	0.2923 $\pm$ 0.0002
2-HMOE	0.3528 $\pm$ 0.0003	0.3276 $\pm$ 0.0003	0.3051 $\pm$ 0.0002	0.2889 $\pm$ 0.0002
G-HPOE	0.3593 $\pm$ 0.0004	0.3347 $\pm$ 0.0002	0.3124 $\pm$ 0.0003	0.2967 $\pm$ 0.0002
t-HPOE	0.3247 $\pm$ 0.0002	0.2900 $\pm$ 0.0003	0.2789 $\pm$ 0.0003	0.2743 $\pm$ 0.0003

Ordinal Embeddingの汎化誤差

Theorem

[Suzuki, Nitanda, Wang, Xu, Yamanishi, Cavazza *ICML2021*]

- With probability $1 - \delta$, we have

$$\mathcal{R}^* \left(\left(\hat{\mathbf{z}}^{(n)} \right)_{n=1}^N \right) \leq \mathcal{R}^* \left(\left(\mathbf{z}^{*(n)} \right)_{n=1}^N \right) + O \left(\omega_{\mathcal{Z}}(R) \left(\sqrt{\frac{N^2 \ln N}{S}} + \frac{N \ln N}{S} + \sqrt{\frac{\ln \frac{2}{\delta}}{S}} \right) \right)$$

Empirical Risk Minimizer
による期待リスク

最小期待リスク

N: Entityの数, S: 3つ組 順序関係データの数

$$\omega_{\mathcal{Z}}(R) := \begin{cases} (\exp R)^2 & \text{hyperbolic,} \\ (R)^2 & \text{Euclidean.} \end{cases}$$

All representations must exist in a ball with radius R .

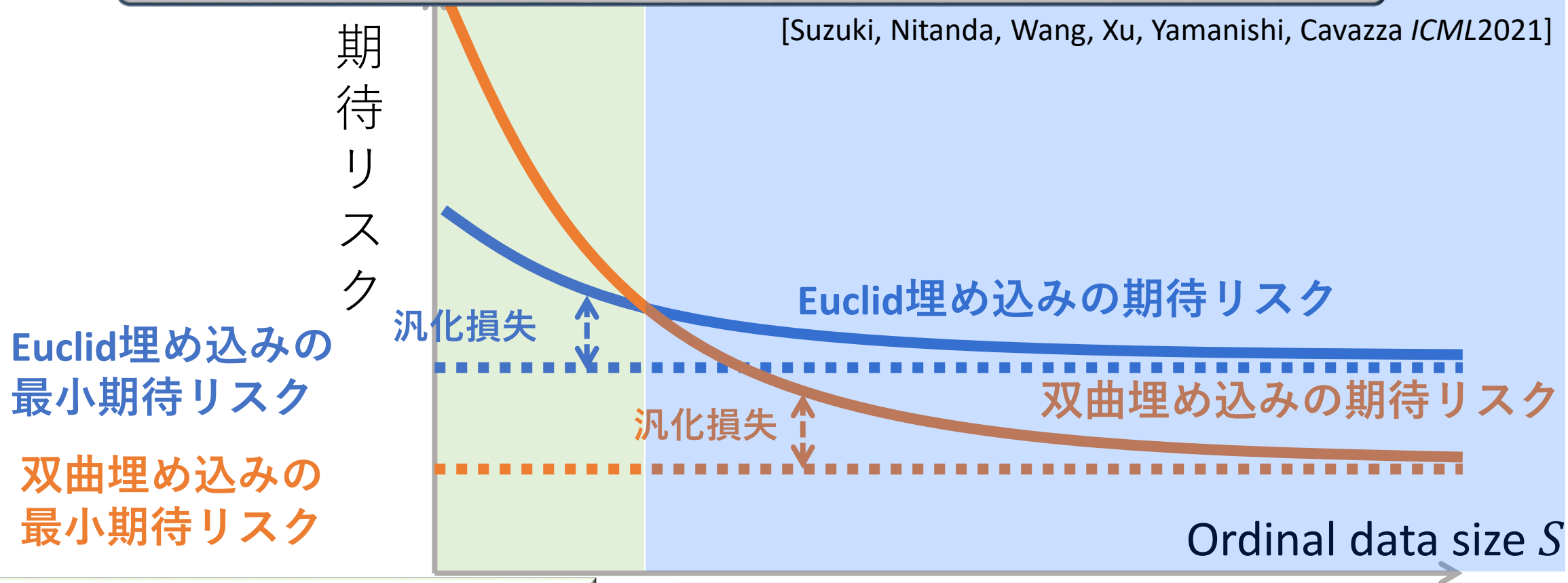
- Proof technique:
 - Convert HOE's loss function to that of the linear model on the matrix space.
 - Rewrite the radius condition to the parameter's spectrum norm's condition.
 - Evaluating the Rademacher complexity.

Euclid埋め込みVS双曲埋め込み

—汎化誤差の観点から—

データ数が増大するほど双曲埋め込みの優位性が際立つ

[Suzuki, Nitanda, Wang, Xu, Yamanishi, Cavazza ICML2021]



S が小 $\rightarrow$ Euclid埋め込みがよい

$\because$ Euclidの汎化誤差が双曲のそれよりも小さい

S が大 $\rightarrow$ 双曲埋め込みがよい

$\because$ 双曲埋め込みのEmpirical Risk minimizerの期待リスクが、Euclidのそれよりも小さい¹³

Graph embeddingの汎化誤差解析

Theorem

[Suzuki, Nitanda, Wang, Xu, Yamanishi, Cavazza *NeurIPS*2021]

With probability $1 - \delta$

$$\mathcal{R}^* \left(\left(\hat{\mathbf{z}}^{(n)} \right)_{n=1}^N \right) - \mathcal{R}^* \left(\left(\mathbf{z}^{*(n)} \right)_{n=1}^N \right)$$

Empirical Risk Minimizer
による期待リスク

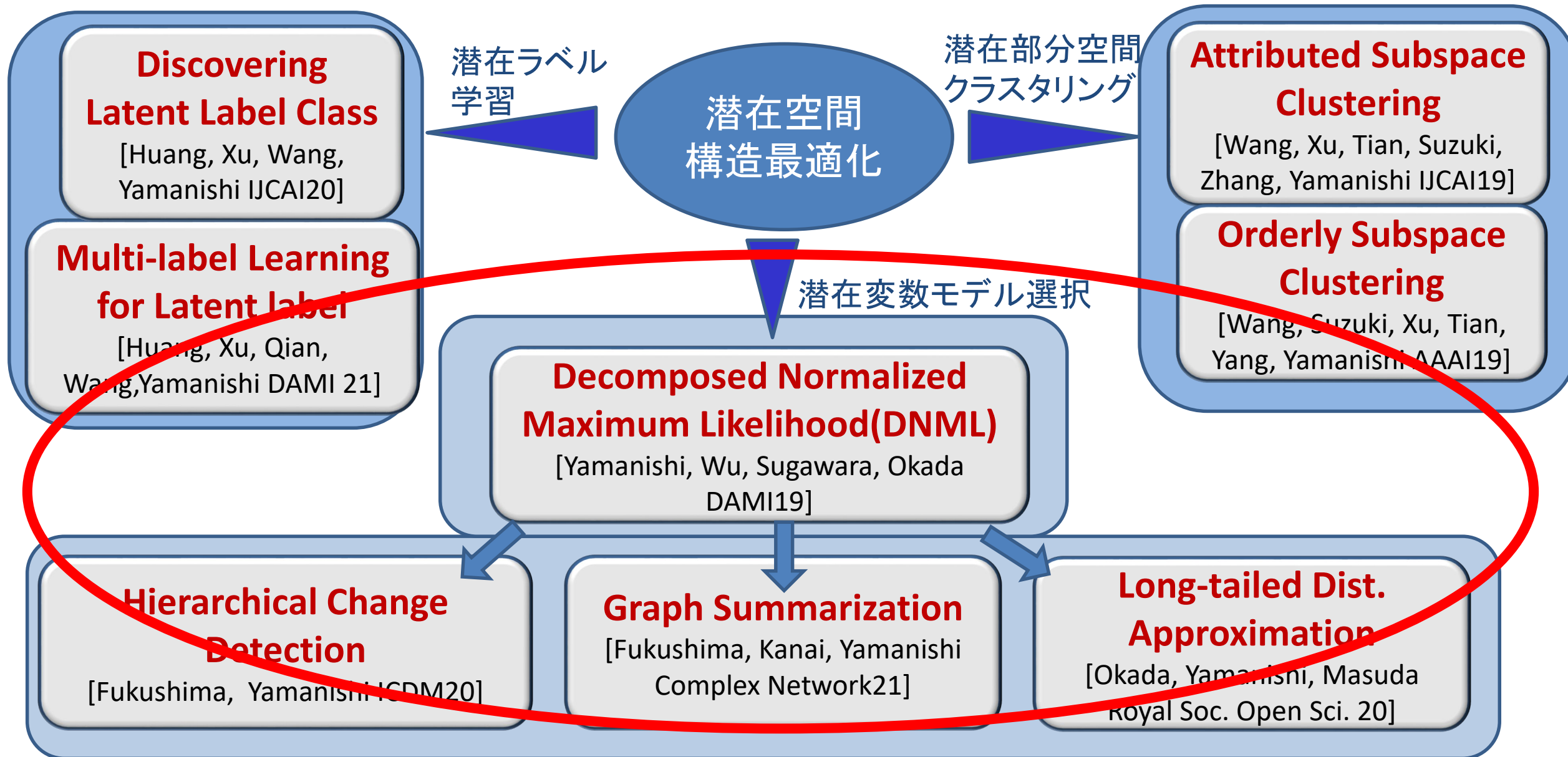
最小期待リスク

$$\leq \frac{2\omega(R)}{M} \sum_{\spadesuit=+,-} K^{\spadesuit} L^{\spadesuit} \left(\sqrt{2M v_{k,z}^{\spadesuit} \log N} + \frac{1}{3} \log N \right) + (b-a) \sqrt{\frac{\log \frac{2}{\delta}}{M}}$$

N: ノードの数, M: 正負を表す対のデータ数

- $\omega_Z(R) := \begin{cases} \cosh^2 R + \sinh^2 R = O((\exp R)^2) & \text{Hyperbolic,} \\ (2R)^2 = O(R^2) & \text{Euclidean.} \end{cases}$
- $O\left(\frac{1}{\sqrt{M}}\right), O(\log N)$
- $v_{k,z}^{\spadesuit}$ depends on the data distribution of $(i_{k,m}^{\spadesuit}, j_{k,m}^{\spadesuit})$ ($\spadesuit = +, -$).
 - Small if the distribution of $(i_{k,m}^{\spadesuit}, j_{k,m}^{\spadesuit})$ is balanced (close to the uniform distribution).

潜在空間構造最適化



2. 潜在変数モデル選択とその応用

[Yamanishi, Wu, Sugawara, Okada Data Mining and Knowledge Discovery 2019]

[Okada, Yamanishi, Masuda Royal Society of Open Science 2019]

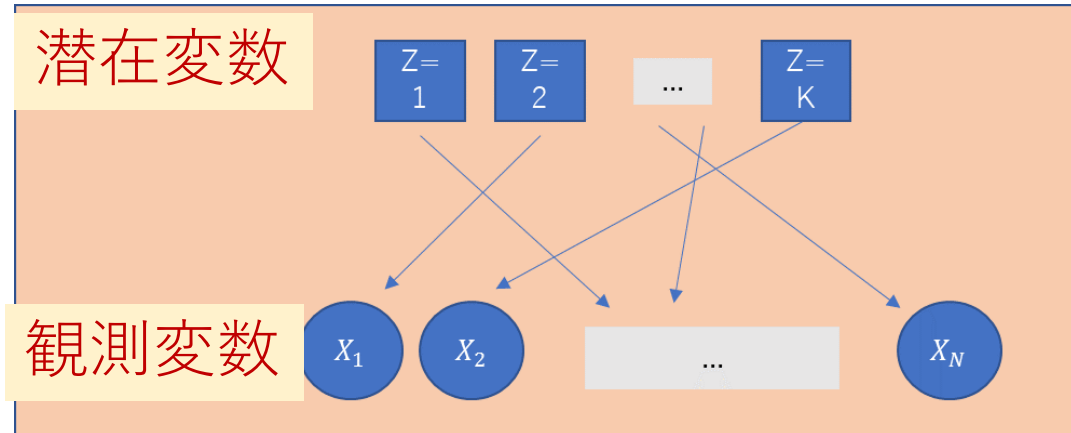
[Fuushima, Yamanishi ICDM2020]

[Fukushima, Kanai, Yamanishi Complex Networks 2021]

[Suzuki, Yamanishi IEEE Transactions on Information Theory 2021]

潜在変数モデル選択

潜在変数の数をいかに推定するか？



[Yamanishi, Wu, Sugawara, Okada
Data Min.Know.Disc. 2019]

M : モデル, x : 観測データ, z : 潜在変数,

分解型正規化

最尤符号長規準

(Decomposed Normalized
Maximum Likelihood: DNML)

$$\mathcal{L}_{\text{DNML}}(x, z; M) \stackrel{\text{def}}{=} \mathcal{L}_{\text{NML}}(x|z; M) + \mathcal{L}_{\text{NML}}(z; M) \implies \min \text{ w.r.t. } M$$

正規化最尤符号長

理論的保証 : 潜在変数のMinimax Estimation Optimality
計算効率性 : $O(n+M)$ で計算可能

正規化最尤符号長

正規化最尤符号長
(NML符号長)

パラメトリック
コンプレキシティ

$$\mathcal{L}_{\text{NML}}(\boldsymbol{x}) = -\log \max_{\theta} p(\boldsymbol{z}; \theta) + \log C_n$$

$$\log C_n = \log \sum_{\boldsymbol{x}} \max_{\theta} p(\boldsymbol{x}; \theta)$$

$$\approx \frac{k}{2} \log \frac{n}{2\pi} + \log \int \sqrt{|I(\theta)|} d\theta$$

Estimation Optimality of DNML Estimator

DNML estimator is the only one that attains the minimax risk.

Theorem (Estimation Optimality of DNML Estimator)

[Yamanishi, Wu, Sugawra, Okada DAMI 2019]

$\bar{k}(x, z)$: general estimator of model

k^* : true latent variable model

$$p_{\bar{k}}(x, z) = \frac{p(x|z, \bar{k}(x, z))p(z; \bar{k}(x, z))}{\bar{C}}$$

$$\min_{\bar{k}} \max_{k^*} D(p_{k^*} || p_{\bar{k}})$$

is achieved by **DNML estimator**

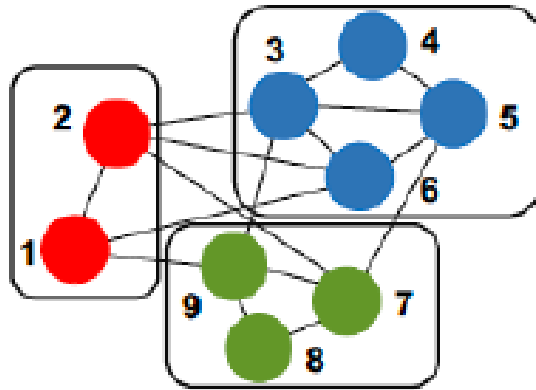
$$\hat{k}(x, z) = \operatorname{argmin}_k \{L_{\text{NML}}(x|z; k) + L_{\text{NML}}(z|k)\}$$

where $D(*||*)$ is the Kullback-Leibler divergence.

DNMLに基づくGraph Summarization

[Fukushima, Kanai, Yamanishi *Complex Network* 2021]

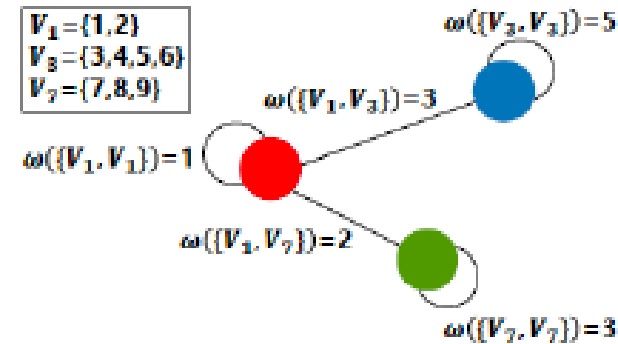
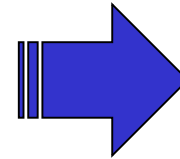
グラフを潜在変数モデルとして捉え、DNML規準に基づき要約



グラフ

	1	2	3	4	5	6	7	8	9
1	0	1	0	0	0	1	0	0	1
2	1	0	1	0	0	1	1	0	0
3	0	1	0	1	1	1	0	0	1
4	0	0	1	0	1	0	0	0	0
5	0	0	1	1	0	1	1	0	0
6	1	1	1	0	1	0	0	0	0
7	0	1	0	0	1	0	0	1	1
8	0	0	0	0	0	0	1	0	1
9	1	0	1	0	0	0	1	1	0

隣接行列



要約

	1	2	3	4	5	6	7	8	9
1	0	1	3/8	3/8	3/8	3/8	1/3	1/3	1/3
2	1	0	3/8	3/8	3/8	3/8	1/3	1/3	1/3
3	3/8	3/8	0	5/6	5/6	5/6	0	0	0
4	3/8	3/8	5/6	0	5/6	5/6	0	0	0
5	3/8	3/8	5/6	5/6	0	5/6	0	0	0
6	3/8	3/8	5/6	5/6	5/6	0	0	0	0
7	1/3	1/3	0	0	0	0	0	1	1
8	1/3	1/3	0	0	0	0	1	0	1
9	1/3	1/3	0	0	0	0	1	1	0

復元隣接行列

$$z_i \sim \text{Multinomial}(z_i; \xi, k),$$

$$x_{ij} \sim \text{Bernoulli}(x_{ij}; \eta, k) \quad (i, j = 1, \dots, N),$$

G : グラフ $\bar{G}$: 要約

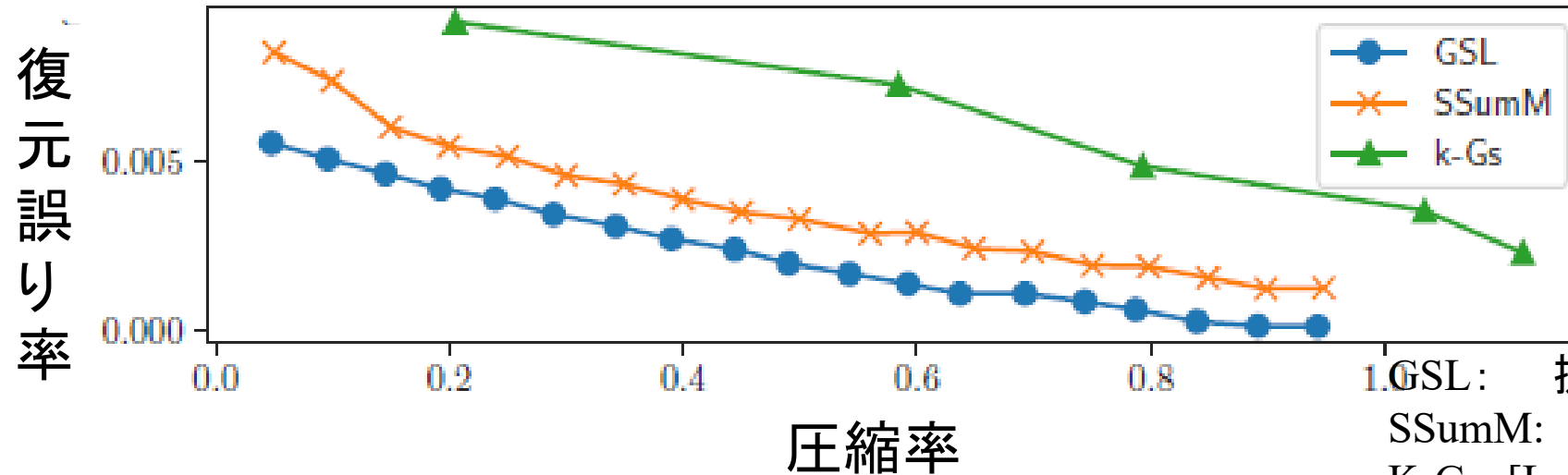
$$\min_{\bar{G}} \mathcal{L}(G; \bar{G}) = \min_{\bar{G}} \{ \mathcal{L}(G | \bar{G}) + \mathcal{L}(\bar{G}) \}$$

$$\text{DNML符号長} = \min_M \{ \mathcal{L}_{\text{NML}}(x | z; M) + \mathcal{L}_{\text{NML}}(z; M) \}$$

実験結果：ベンチマークデータ

[Fukushima, Kanai, Yamanishi *Complex Network* 2021]

圧縮率と復元誤り率のトレードオフにおいて他手法よりも高圧縮



GSL: 提案手法

SSumM: [Lee et al. KDD2020]

K-Gs: [Lefevre et al. ICDM2010]

Fig. 2. Evaluation results for GSL, SSumM, and k-GS for the ego-Facebook dataset.

ego-Facebook dataset

<http://snap.stanford.edu/data/ego-Facebook.html>.

The number of nodes is 4039 and the number of edges is 88234.

潜在変数モデルの階層的変化検知

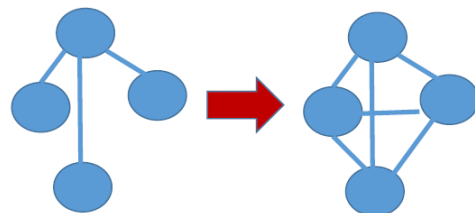
[Fukushima, Yamanishi *ICDM* 2020]

変化はどこから来るのか？ $\Rightarrow$ 階層的変化検知で説明

Level1.
パラメータの変化

θ_1

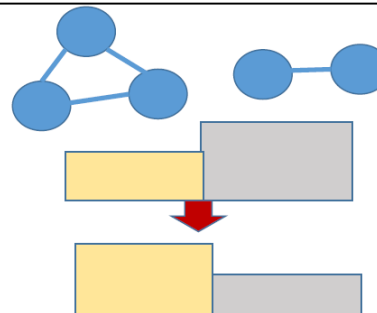
Level 1. Data distribution change



Level2.
潜在変数の変化

Z

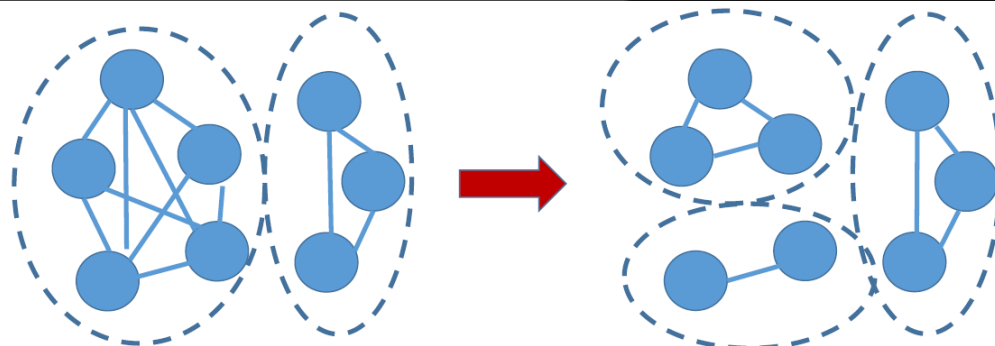
Level 2. Latent variable change



Level3.
モデルの変化

M

Level 3. Model change



分解型
MDL変化統計量

$$\Delta_t \mathcal{L}_{\text{DNML}}(x, z; M) = \Delta_t \mathcal{L}_{\text{NML}}(x|z; M) + \Delta_t \mathcal{L}_{\text{NML}}(z; M) + \Delta_t \mathcal{L}(M)$$

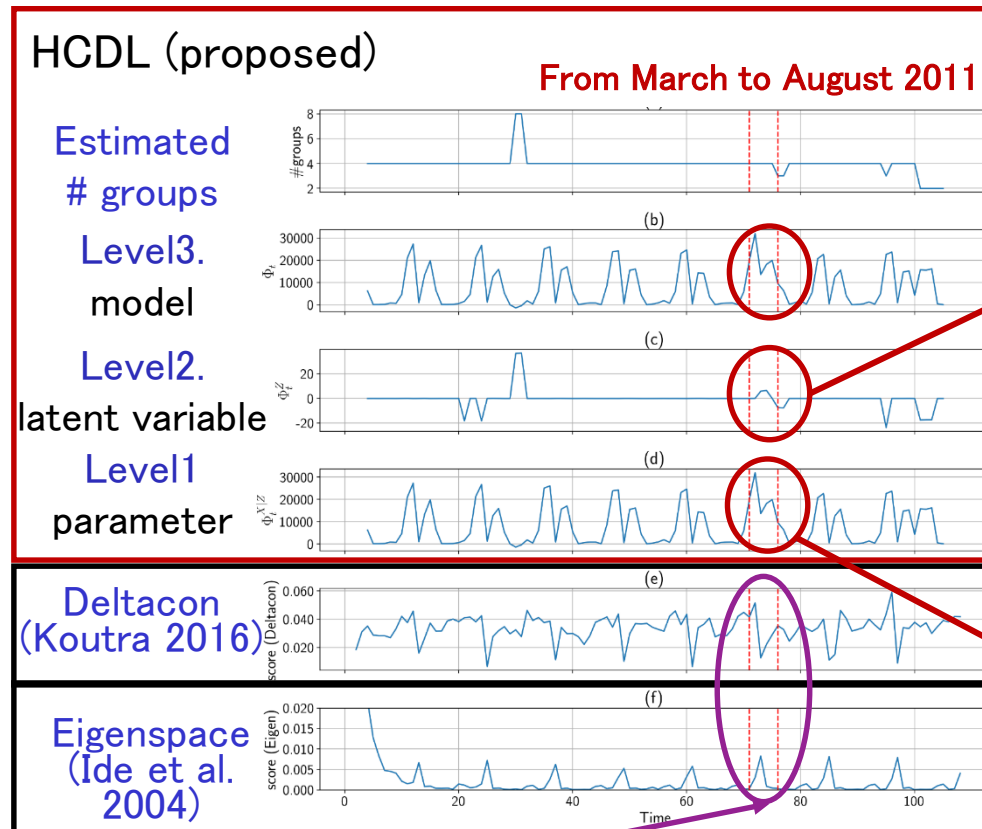
実験結果:人口移動データ

大震災に伴う人口移動の異なるレベル(県間単位、ブロック単位)の変化を検知

月間単位の県間人口移動時系列データ

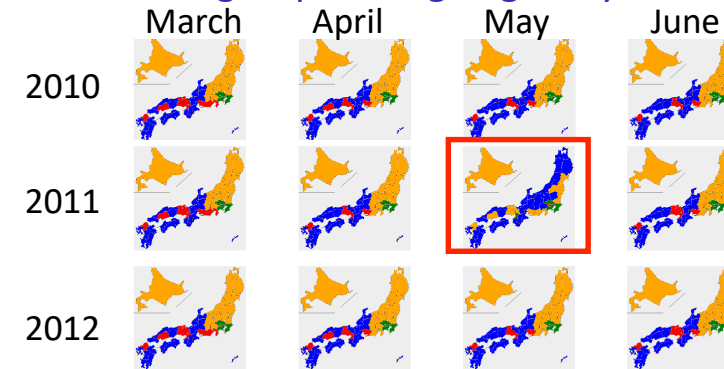
<https://www.e-stat.go.jp/en/stat-search/database> (April 2005–March 2014)

[Fukushima, Yamanishi ICDM 2020]

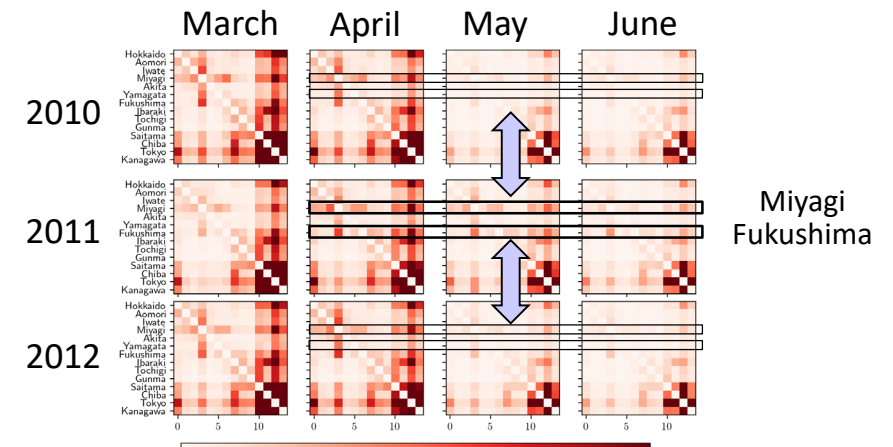


Sensitivity analysis for increase of changes is not possible

Level2
Estimated groups changed greatly in May 2011

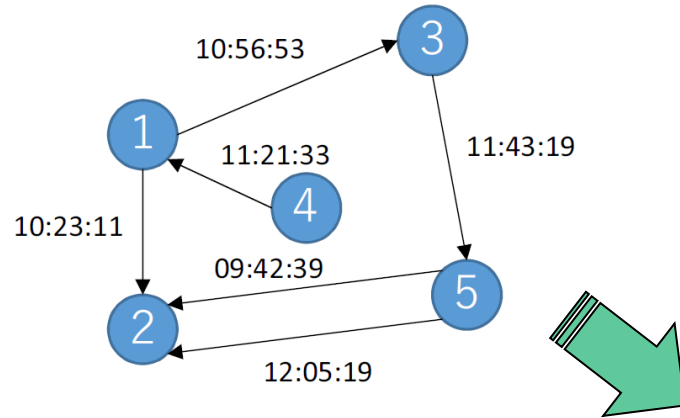


Level1
More people moved from Miyagi and Fukushima in Apr. –June 2011



混合指数分布によるべき分布近似

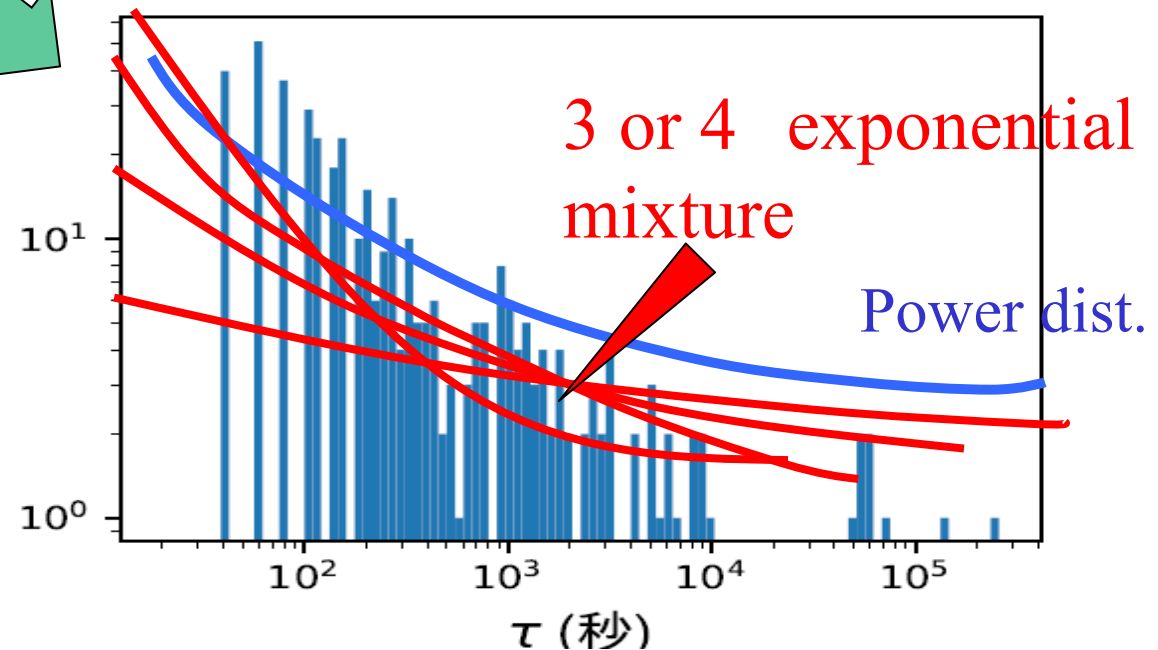
テンポラルネットワーク上のべき分布は混合数3~4の混合指数分布で近似できる



[Okada, Masuda, Yamanishi *Royal Soc. Open Sci.* 2020]

ネットワーク上のイベントの生成区間分布
= べき分布

DNMLの結果
ほとんどのネットワークの
イベント生成区間分布は
3-4個の混合指数分布で
表される

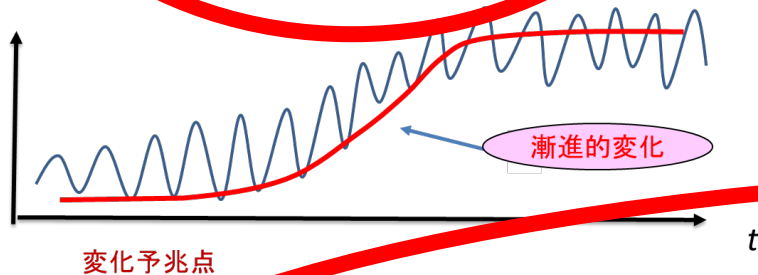


変化予兆検知

連続パラメータの変化予兆検知

Differential MDL Change Statistics

[Yamanishi, Xu, Yuki, Fukushima, Lin Scientific Reports 21]



階層的变化検知

潜在構造変化検知

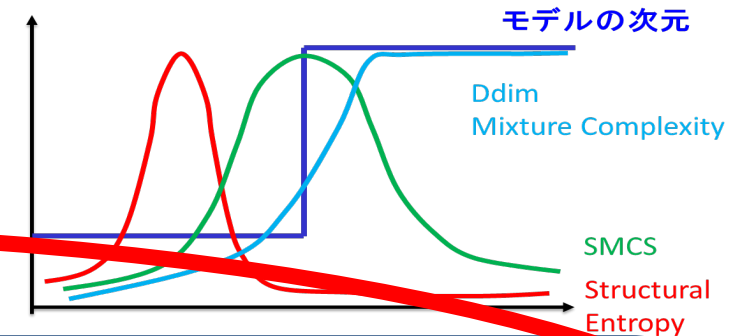
Meta- Change Detection

[Fukushima, Yamanishi Entropy 19]

Hierarchical Change Detection

[Fukushima, Yamanishi ICDM20]

離散モデルの変化予兆検知



変化予兆指標

Structural Entropy

[Hirai, Yamanishi BigData18]

Sequential MDL Change Statistics

[Hirai, Yamanishi BigData19]

非整数潜在次元

Descriptive Dimensionality

[Yamanishi Hirai IEEE TKDE 21]

Kernel Complexity

[Hirai, Yamanishi IEEE BigData 21]

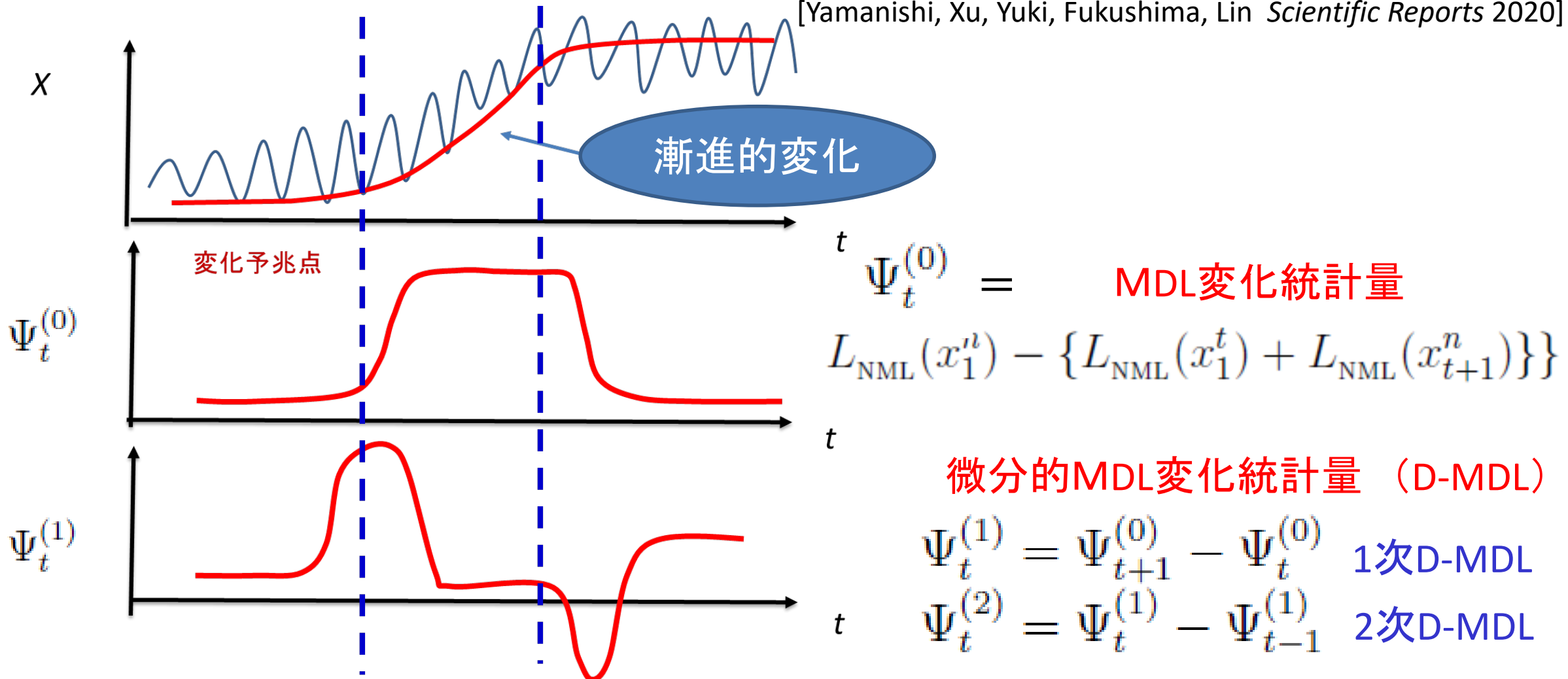
3. 連続パラメータの変化予兆検知

[Yamanishi, Xu, Yuki, Fukushima, Lin Scientific Reports 2021]

連続パラメータ変化予兆検知

漸進的変化の開始点(変化予兆点)をMDL変化統計量の微分で捉える

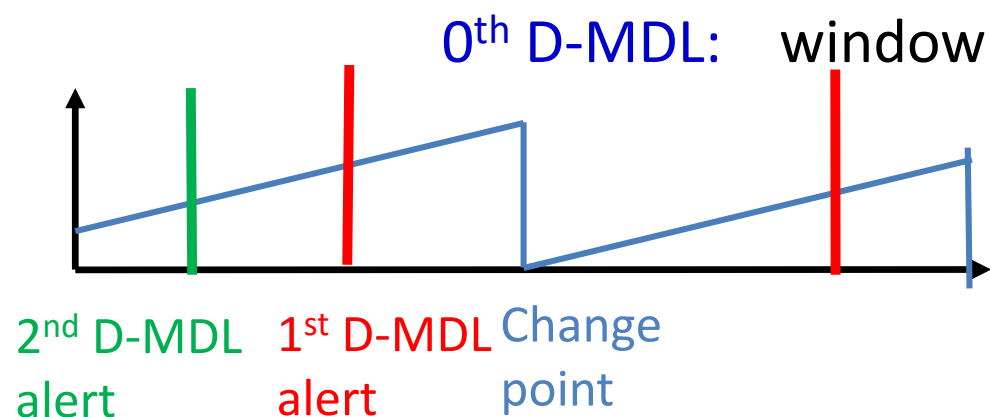
[Yamanishi, Xu, Yuki, Fukushima, Lin *Scientific Reports* 2020]



階層的変化/変化予兆検知

理論的閾値を基に、変化点と予兆点を階層的に検知

[Yamanishi, Xu, Yuki, Fukushima, Lin *Scientific Reports* 2020]



$$\epsilon = (2 + d/2 + \delta) \log w + \log(1/\delta)$$

d : parameter dim, δ : confidence, w : window size

s.t. Total #Type I error $< \infty$

1st D-MDL: raise 1st D-MDL sign alert when $w \Psi^{(1)} > \epsilon_1$

$$\epsilon_1 = d \log(w/2) + \log(1/\delta_1) \quad \text{s.t. Type I error prob.} < \delta_1$$

2nd D-MDL: raise 2nd D-MDL sign alert when $w \Psi^{(2)} > \epsilon_2$

$$\epsilon_2 = 2(d \log(w/2) + \log(1/\delta_2)) \quad \text{s.t. Type I error prob.} < \delta_2$$

COVID-19 Pandemic解析への応用

■ データ

European Centre for Disease Prevention and Control (ECDC)

<https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>

■ 目的

1. COVID-19 の感染爆発(outbreak)の検知 $\Rightarrow$ 0次のMDL変化統計量
2. 感染爆発の予兆(sign)の検知 $\Rightarrow$ 1次、2次のD-MDL
3. 基本再生産数 R_0 の変化とその予兆検知

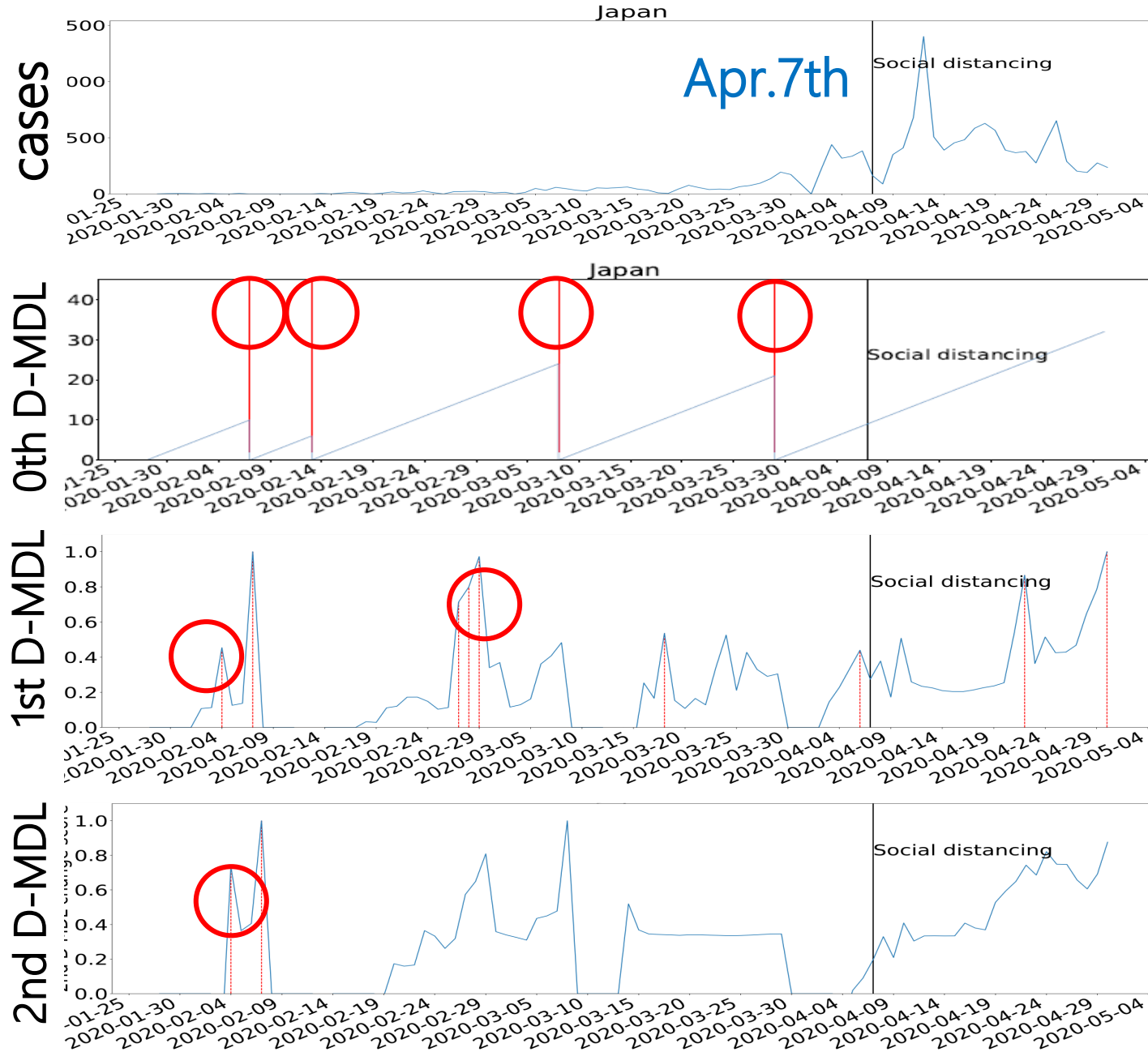
■ Model

- **Gaussian Model** (of daily new cases)
parameter changes $\Rightarrow R_0 > 1$ (increase of cases) or $R_0 < 1$ (decrease)
- **Exponential Growth Model** (of cumulative cases) [Malthus et al. 1992]

$$\log C(t) = rt + \log C(0) + \varepsilon \quad (\text{Malthusian growth model})$$

parameter changes $\Rightarrow$ changes of R_0 ($R_0 - 1 \propto r$)

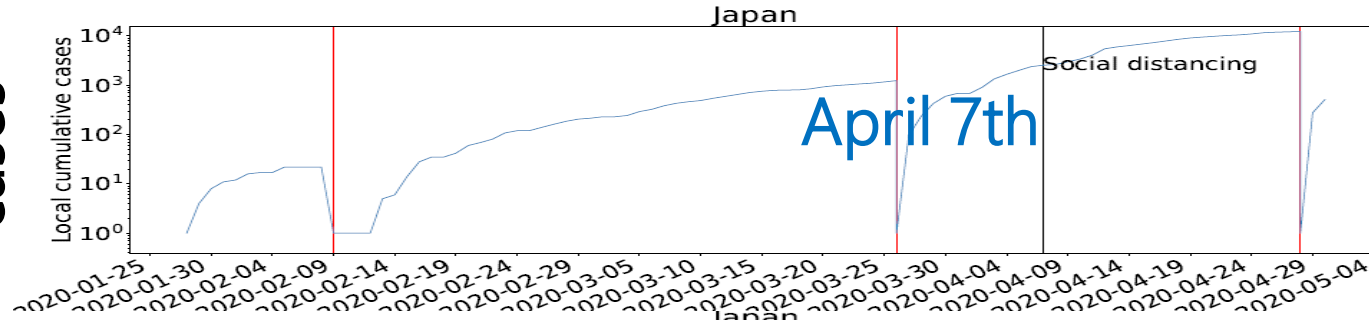
ガウスモデルによる新感染者数時系列の変化解析



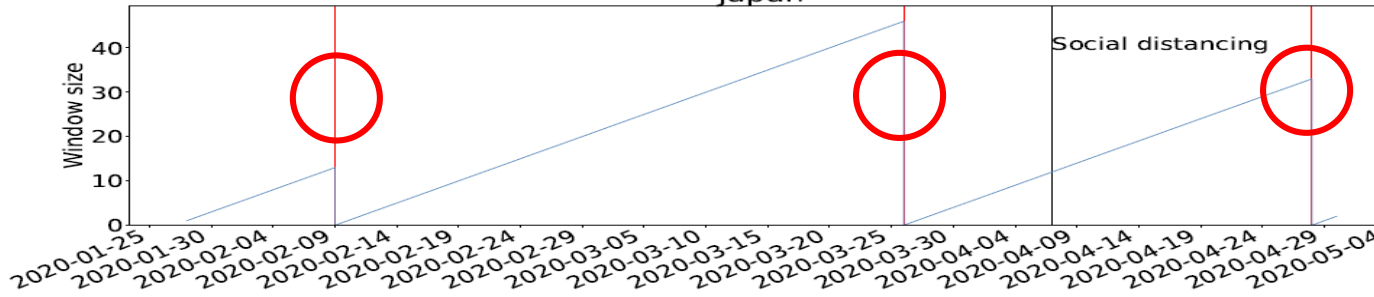
- **感染爆発**
緊急事態宣言(4/7)前に発現 ($R_0 > 1$)
- MDL変化統計量アラート
感染爆発時点に対応 ($R_0 > 1$)
- 1次のD-MDLアラート
変化の速度
感染爆発前に上がっている
- 2次のD-MDLアラート
変化の加速度
感染爆発前に上がっている

指数成長モデルによる累積感染者数の変化解析

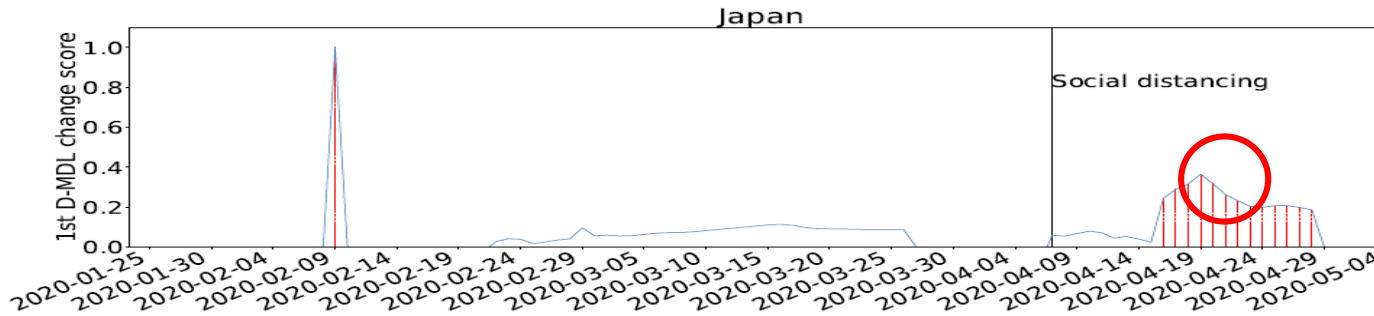
Cumulative cases



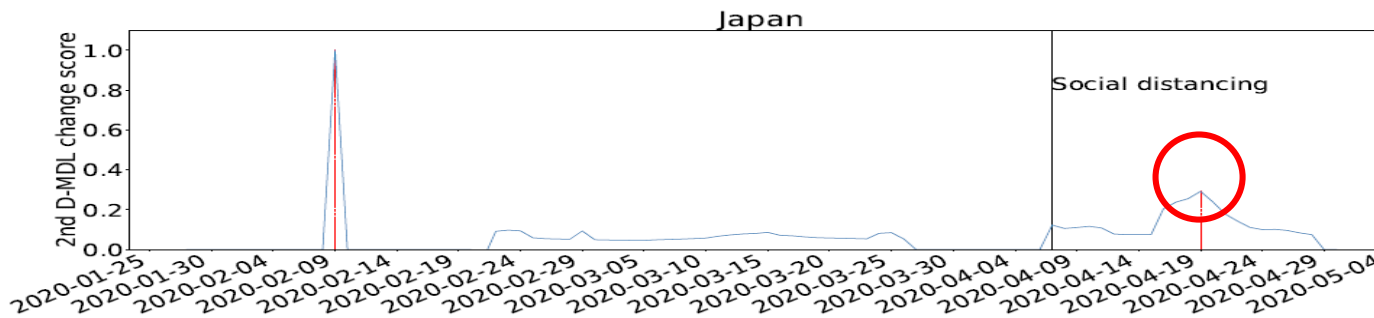
0th D-MDL



1st D-MDL



2nd D-MDL



- 感染爆発(4/7) これを機に $R_0 > 1$ から $R_0 < 1$ に転じている

- MDL変化量アラート
感染爆発(4/7)前に2個、後に1個のアラートを上げている。

- 1次のD-MDLアラート
変化の速度
 $R < 1$ に応じてアラートが上がっている

- 2次のD-MDLアラート
変化の加速度
 $R < 1$ に応じてアラートが上がっている

Findings

[Yamanishi, Xu, Yuki, Fukushima, Lin *Scientific Reports* 2020]

- 2021年4月30日までに1万以上の累積感染者数が出た37か国に対して**106 件の変化点が感染爆発 (outbreak)として検知された。**
- 感染爆発の64%については、**平均約6日以上前にその予兆を検知できていた。**
- 緊急事態宣言(social distancing)前に高々 **5 個以内の変化点/予兆アラートにとどまった国はそうでない国に比べて感染者数減が優位に早かった。**
- 早期に緊急事態宣言を発令していた国は**減少の変化点までの期間を約30日まで抑えられていた** (その他の国は37日以上)。
- SIRモデルのような予測シミュレーションモデルとは異なる角度から感染爆発を分析。.

4. 離散構造の変化予兆検知

[Hirai, Yamanishi, IEEE Bigdata 2019]

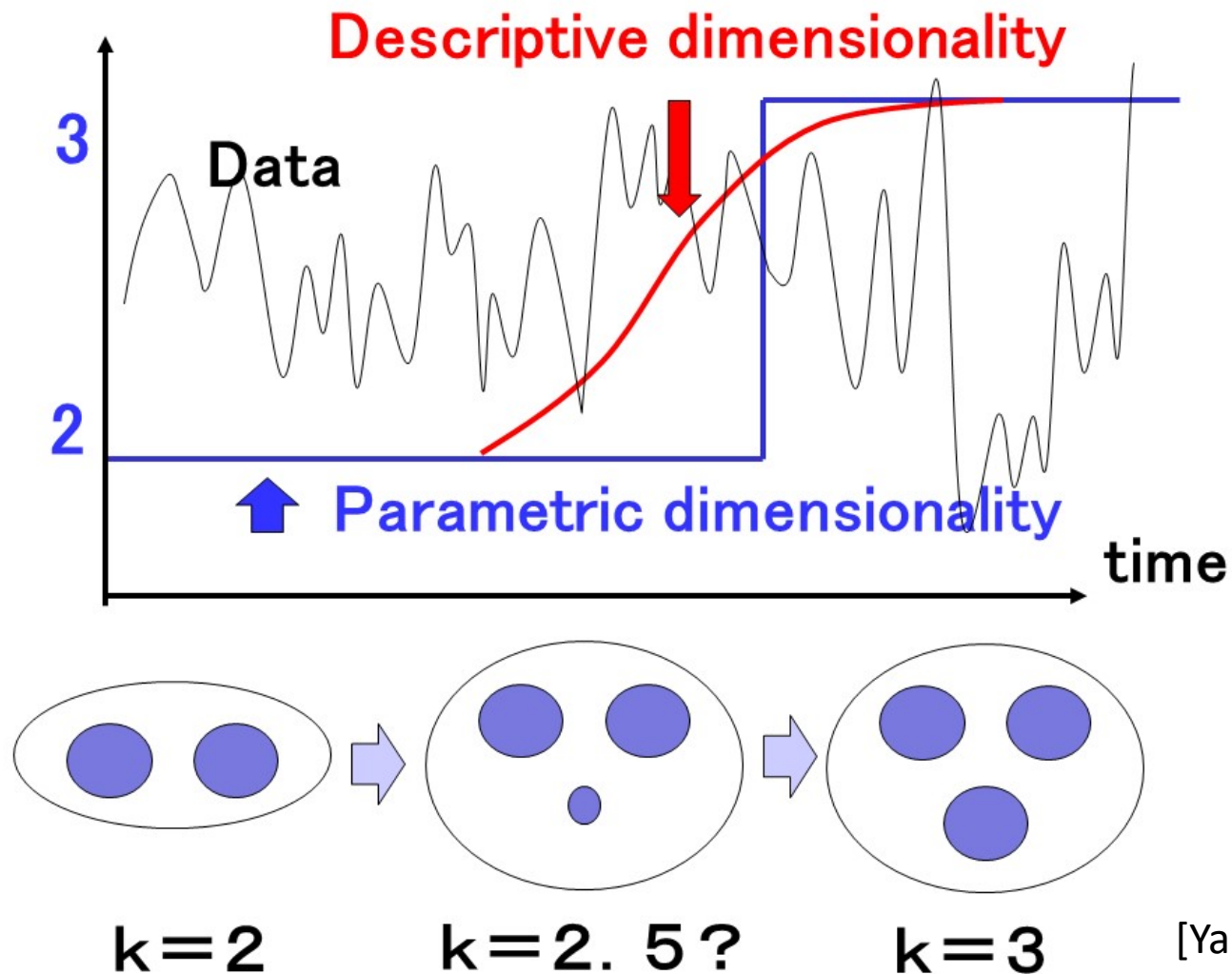
[Hirai, Yamanishi, IEEE Bigdata 2020]

[Yamanishi, Hirai 2021]

[Yamanishi Arxiv 2020]

潜在構造変化予兆検知

モデル変化の過渡期の次元を非整数値次元で測る



連続的
モデル選択

[Yamanishi Hirai 2021]

記述次元: Descriptive Dimensionality

パラメトリックコンプレキシティのフラクタル次元を測る

$\mathcal{P} = \{p(X^n)\}$: probability model class

[Yamanishi Arxiv 2020]

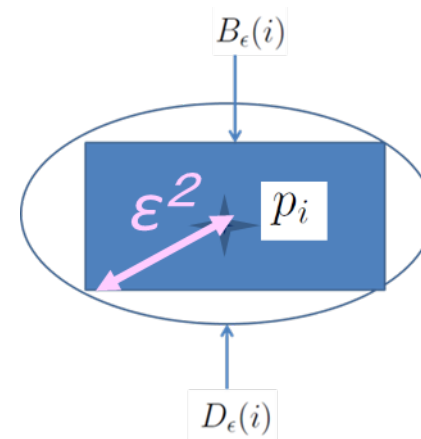
Parametric complexity of $\mathcal{P}$ for data size n :

$$\log \mathcal{C}_n(\mathcal{P}) \stackrel{\text{def}}{=} \log \sum_{\mathbf{x}} \max_{p \in \mathcal{P}} p(\mathbf{x})$$

$$C_n(\mathcal{P}) \approx \overline{C}_n(\overline{\mathcal{P}}) \stackrel{\text{def}}{=} \sum_{i=1}^{m(\epsilon; \mathcal{P})} Q_\epsilon(i)$$

$$Q_\epsilon(i) \stackrel{\text{def}}{=} \int_{\hat{p} \in D_\epsilon(i)} g(\hat{p}, \hat{p}) d\hat{p}$$

$$g(\hat{p}, p) \stackrel{\text{def}}{=} \sum_{\mathbf{x}: \max_p p(\mathbf{x}) = \hat{p}} p(\mathbf{x})$$



Descriptive dimension(記述次元)

$$\text{Ddim}(\mathcal{P}) \stackrel{\text{def}}{=} \lim_{\epsilon \rightarrow 0} \frac{\log m(\epsilon : \mathcal{P})}{\log(1/\epsilon)}$$

Theorem(Single parametric model)

$\mathcal{P}_k = \{p(x; \theta) : \theta \in \mathbb{R}^k\}$: k -dim. parametric class

$$\text{Ddim}(\mathcal{P}_k) = k$$

Theorem(Model fusion)

$\mathcal{F}^\odot = \mathcal{P}_1 \odot \cdots \odot \mathcal{P}_s$, $P_i \sim p(\mathcal{P}_i)$ ($i = 1, \dots, s$)

$$\text{Ddim}(\mathcal{F}^\odot) \geq \sum_i p(\mathcal{P}_i) \text{Ddim}(\mathcal{P}_i) = \underline{\text{Ddim}}(\mathcal{F}^\odot)$$

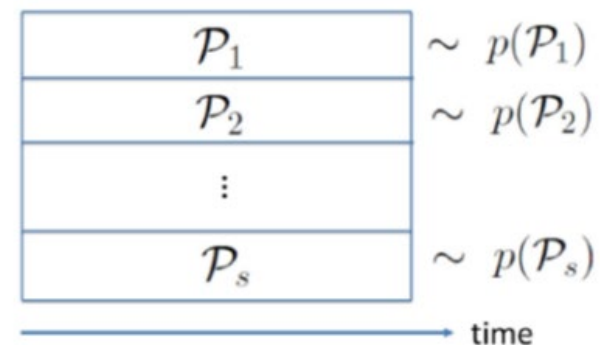
Pseudo Ddim

Theorem(Generalization error)

d : Bhattacharya distance, $\hat{p}$: MDL estimator

$$E[d(\hat{p}, p_{\text{true}})] \leq O\left(n^{\underline{\text{Ddim}}(\mathcal{F}^\odot)} e^{-Cn}\right).$$

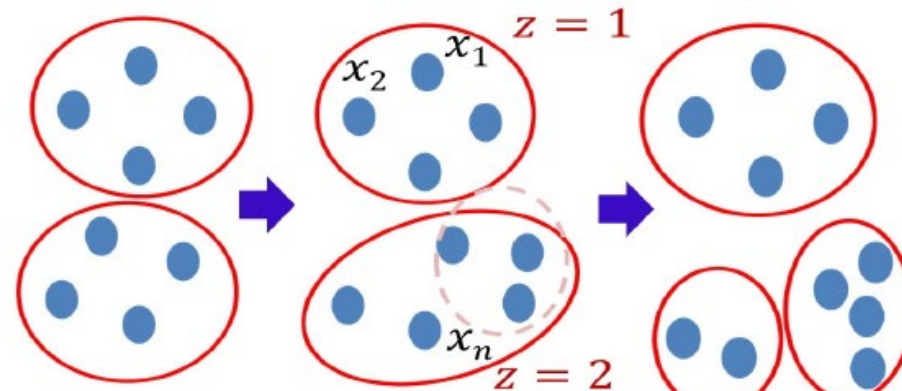
$$\mathcal{P}_1 \odot \mathcal{P}_2 \cdots \odot \mathcal{P}_s$$



(a) Model fusion

時変ガウス混合モデルの記述次元

モデル変化の過渡期におけるモデルフュージョンの度合いを記述次元で捉える



Model fusion

$y_1, y_2, \dots, y_t, \dots$
時系列データ

$$\mathbf{y}_t = (x_1, \dots, x_n, z_1, \dots, z_n)$$

$$\text{Ddim} = 2 \quad 2 < \text{Ddim} < 3 \quad \text{Ddim} = 3$$

$$\overline{\text{Ddim}}(\mathcal{F}^\odot)_t \approx \sum_k p(k|\mathbf{y}_t)k.$$

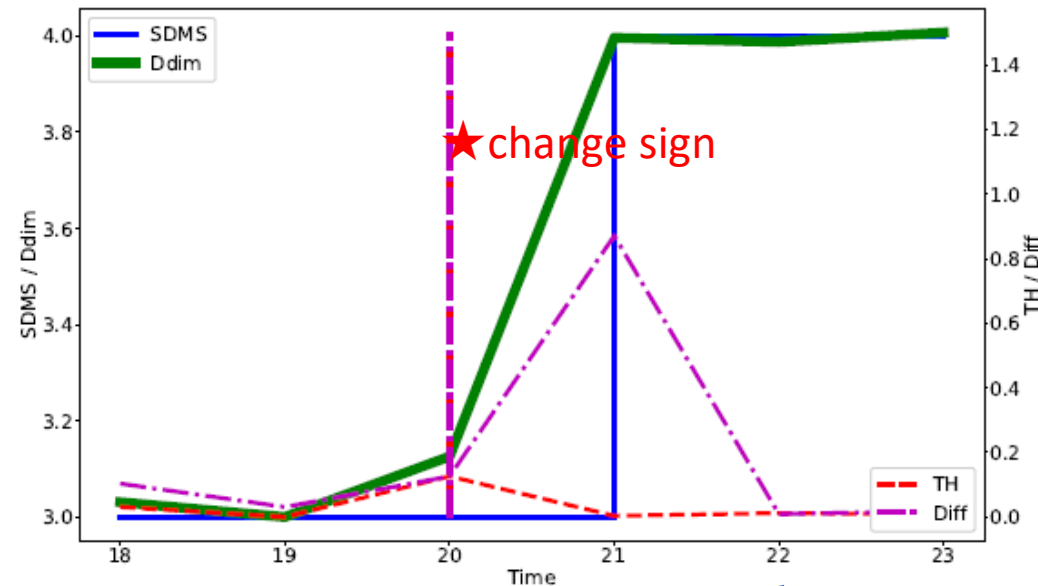
Descriptive Dimensionality

$$p(k|\mathbf{y}_t) = \frac{\exp(-\beta(L_{\text{NML}}(\mathbf{y}_t; k) - \log p(k|k^{t-1})))}{\sum_{k'} \exp(-\beta(L_{\text{NML}}(\mathbf{y}_t; k') - \log \max_p p(k'|k^{t-1})))}$$

$$L_{\text{NML}}(\mathbf{y}_t; k) = -\log \max_p p(\mathbf{y}; k) + \log \sum_{\mathbf{y}'} \max p(\mathbf{y}'; k).$$

実験:電力消費データ

■ 336 houses, every other minute from Dec. 17, 2006 to Dec. 10, 2010,
3-dim data ((1) in kitchen and laundry rooms, 2) by electric water heaters and 3) by
air-conditioners.



UC machine learning
depository, 2017

Cluster collapse detected
Emergence of a cluster of
heavy m1 users

Consumption
category

The week from May 14th

cluster

c	m1	m2	m3	#
0	0.01	2.73	0.01	163
1	0.00	2.80	6.16	92
2	5.18	3.05	4.74	57

The week from May 21st

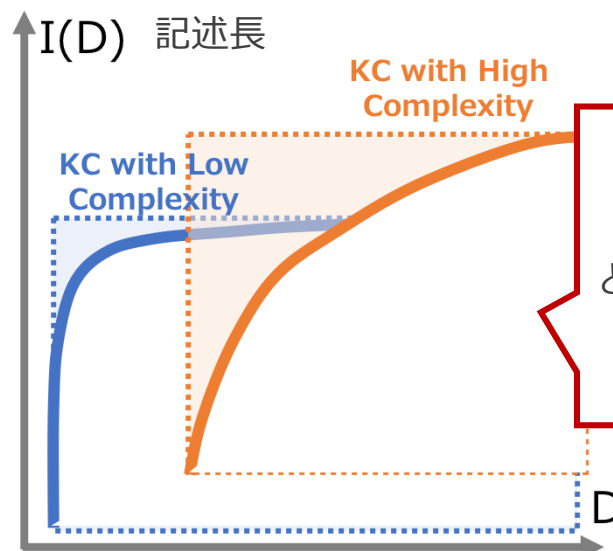
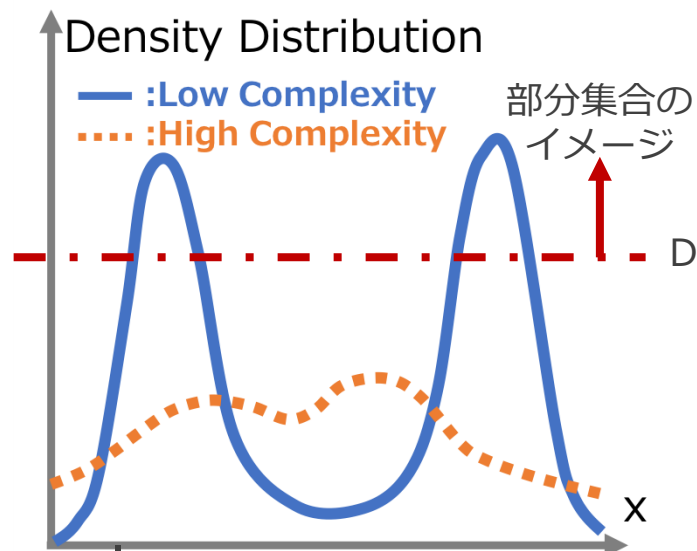
c	m1	m2	m3	#
0	0.01	2.58	0.01	168
1	0.00	2.49	6.22	99
2	5.65	2.64	4.34	45

The week from May 28th

c	m1	m2	m3	#
0	0.01	2.70	0.01	150
1	0.00	3.01	6.38	99
2	1.73	2.72	5.87	8
3	6.20	3.11	4.91	55

Kernel Complexity

ノンパラメトリックな分布のComplexityを定量化する



[Hirai Yamanishi IEEE BigData 2021]

となることを期待する。
(記述長の変化に
Gini係数を適用)

- ・ 情報の偏りが大きい (青線) → 変化のばらつきが大きい
- ・ 情報の偏りが小さい (オレンジ線) → 変化のばらつきが小さい

$$KC = 1 - \frac{\int_D \text{正規化最尤符号長 } I(D) dD - \frac{1}{2} \Delta L \cdot \Delta D - L_0 \cdot \Delta D}{\frac{1}{2} \Delta L \cdot \Delta D}$$

ΔL : $I(D)$ の最大最小の差分
 L_0 : $I(D)$ の最小値
 ΔD : D の最大最小の差分

Kernel分布のNML符号長

[Hirai Yamanishi *IEEE BigData* 2021]

$$p(\mathbf{x}; h) = \frac{1}{n} \sum_{j=1}^n K(\mathbf{x} - \mathbf{x}_j; h),$$

$$K(\mathbf{x}; h) = \frac{1}{(2\pi h^2)^{m/2}} \exp \left\{ -\frac{\|\mathbf{x}\|^2}{2h^2} \right\}.$$

正規化最尤符号長(NML)

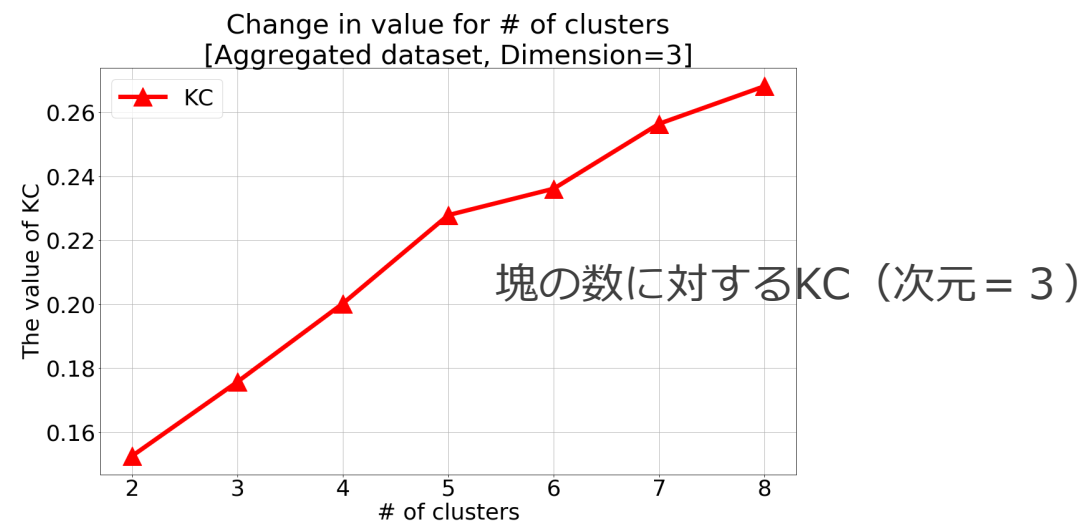
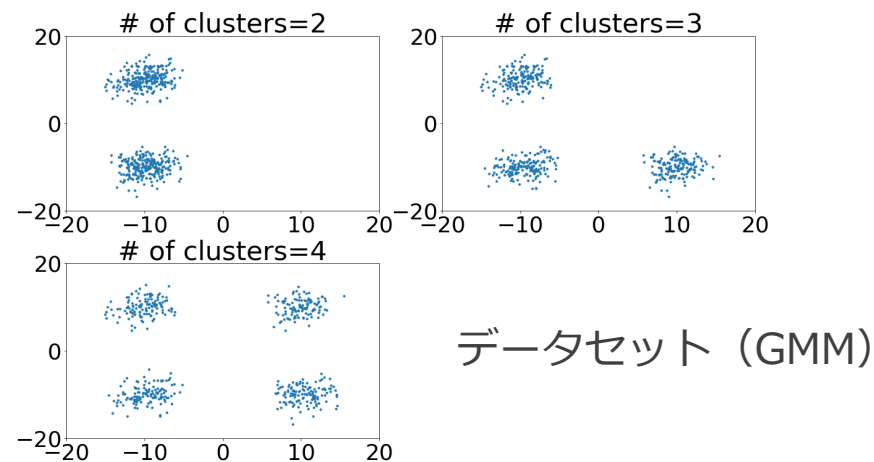
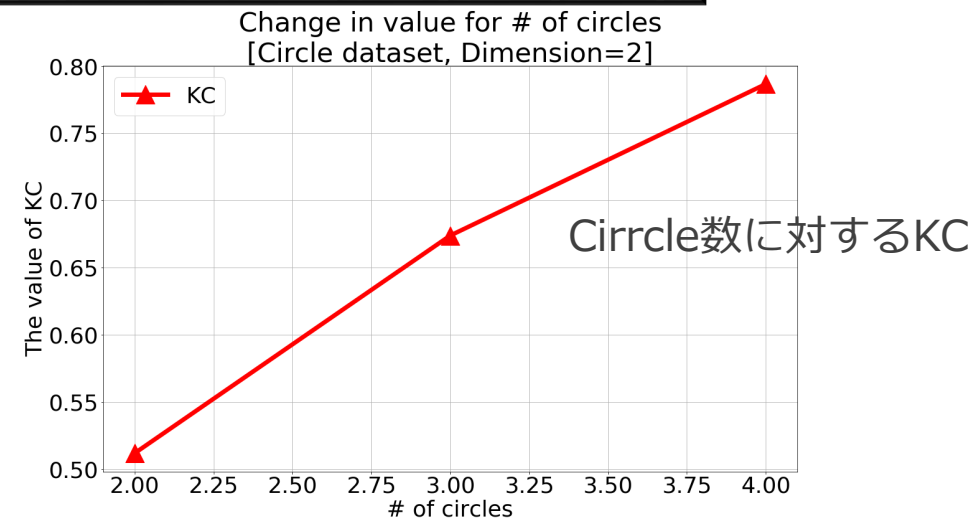
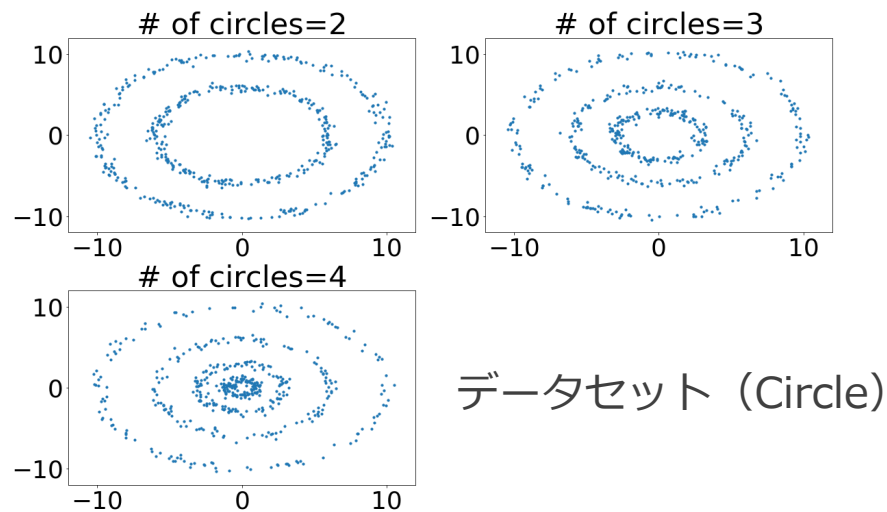
$$\begin{aligned} & \frac{nm}{2} \log \left\{ \sum_{i \in B} \frac{1}{N(A_i)} \sum_{j \in A_i} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right\} - \sum_{i \in B} \left(\frac{m}{m+4} \log n - m \log \epsilon \right) \\ & + nm \log \left(\frac{n^{\frac{1}{m+4}}}{\epsilon} \right) + \log \log \left(\frac{2\pi D \cdot n^{\frac{2}{m+4}}}{m\epsilon^2} \right) + \frac{nm}{2} \log(\pi) - \log \Gamma \left(\frac{nm}{2} \right). \end{aligned}$$

$$A_i = \left\{ j \mid \|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq (1 + \gamma)D \right\} \quad (i = 1, \dots, n),$$

$$B = \left\{ i \mid \frac{1}{N(A_i)} \sum_{j \in A_i} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq D \right\}.$$

Kernel Complexityの性質

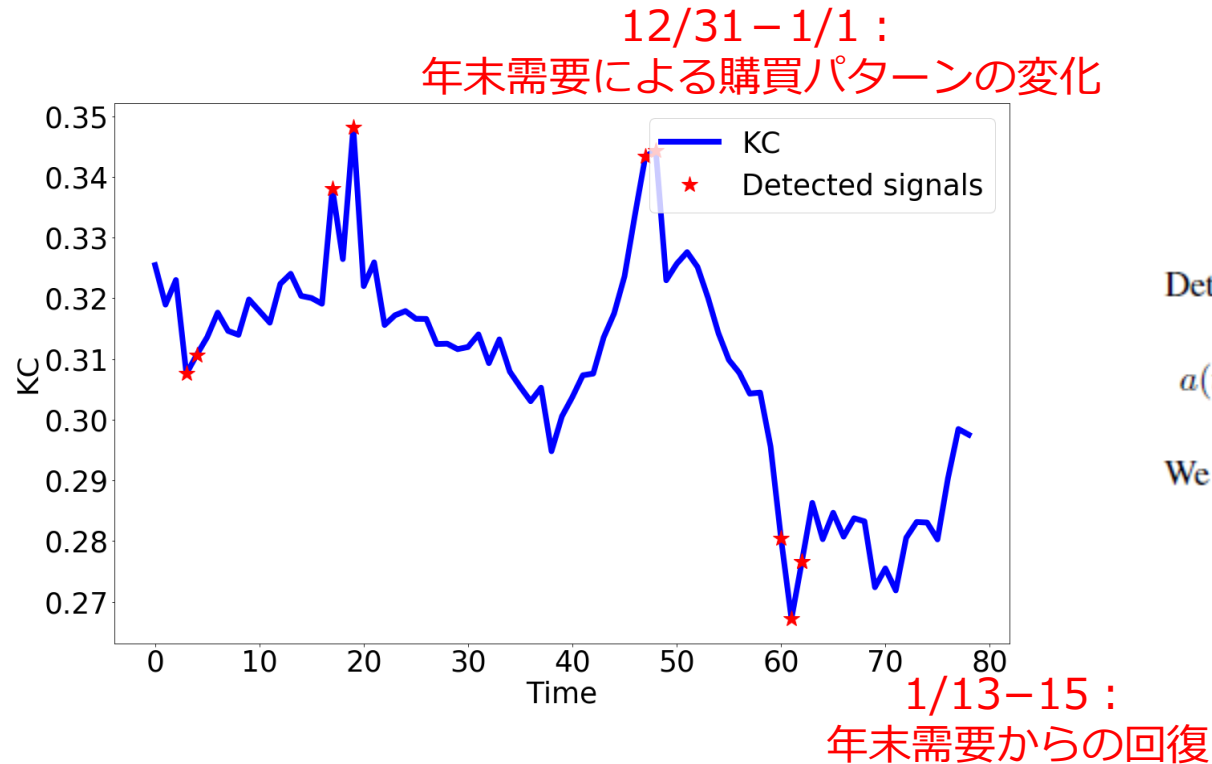
塊 (cluster) の数とKCとはほぼ比例関係、KCはその間を埋める



実験：マーケティングデータ

年末需要によるパターンの変化と年末需要の落ち着きに対応した変化を検知

[Hirai Yamanishi *IEEE BigData* 2021]



Detect changes with following conditions:

$$a(t) = \begin{cases} 1 & \text{if } |KC_t - E[KC^{t-1}]| > \eta * Std[KC^{t-1}], \\ 0 & \text{otherwise.} \end{cases}$$

We can detect changes in the case where $a(t) = 1$.

Data: Hakuhodo, Inc. M-CUBE 提供

5. Progress on Computational Ophthalmology 計算論的眼科学

[Hashimoto, Asaoka, Kiwaki, Sugiura, Asano, Murata, Matsuura, Miki, Mori, Ikeda, Kanamoto, Yamagami, Inoue, Tanito, Yamanishi: *British J. Ophthalmology*2020]

[Xu, Asaoka, Kiwaki, Murata, Fujino, Matsuura, Hashimoto, Asano, Miki, Mori, Ikeda, Kanamoto, Yamagami, Inoue, Tanito, Yamanishi: *American J. Ophthalmology*2020]

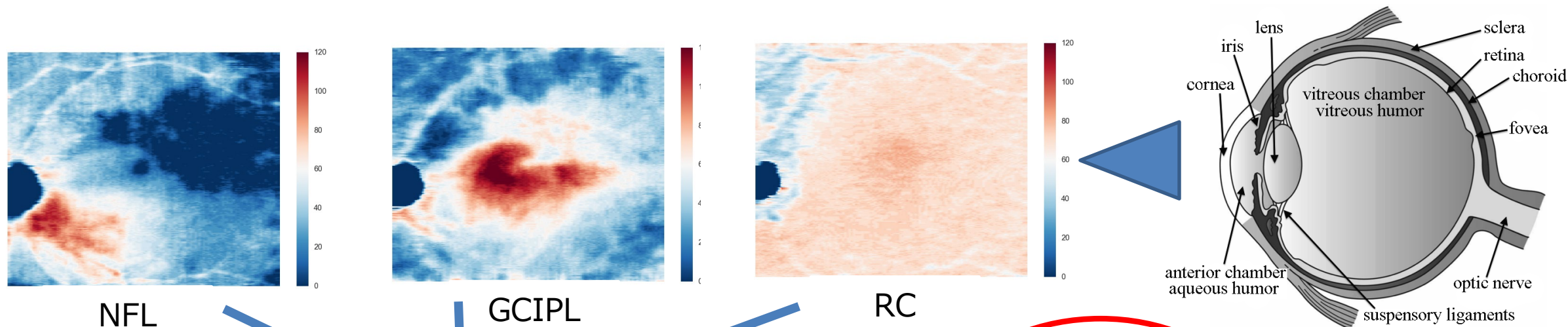
[Xu, Asaoka, Murata, Kiwaki, Zheng, Matsuura, Fujino, Tanito, Mori, Ikeda, Kanamoto, Yamanishi: *Ophthalmology Glaucoma* 2020]

[Xu, Asaoka, Kiwaki, Murata, Fujino, Yamanishi *KDD*2021]

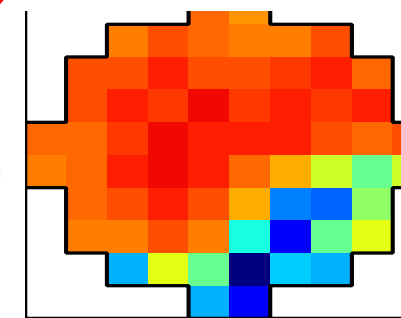
Hashimoto, Kiwaki, Sugiura, Asano, Murata, Fujino, Matsuura, Miki, Mori, Ikeda, Kanamoto, Yamagami, Inoue, Tanito, Yamanishi, Asaoka,: *Trans. Vision Sci. and Tech.* 2021.

[Asaoka, Xu, Murata, Kiwaki, Matsuura, Fujino, Tanito, Mori, Ikeda, Kanamoto, Inoue, Yamagami, Yamanishi *Ophthalmology Science*2021]

網膜層厚 (RT)と 視野感度(VF)



網膜層厚データ
(Retinal Thickness: RT)
Easy, time saving



視野感度データ
(Visual Field Sensitivity: VF)
Time consuming, noise sensitive



網膜層厚(RT)と視野感度(VF)を用いた緑内障研究

RTからVFの推定
(Unbiased データ)

CNN-based
pointwise regression
[Uesaka et al. KDD2017]

医学的再検証

Tensor regression
[Xu et al. AJO2020]

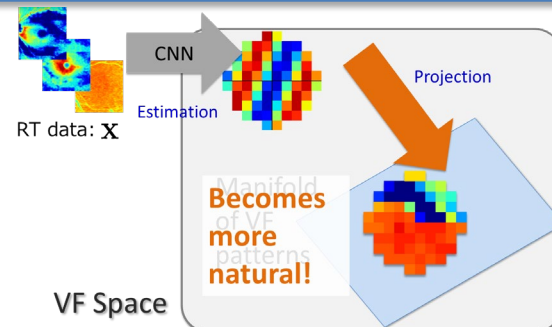
RTからVFの推定
(Biasedデータ)

Pattern-based
regularization
[Sugiura et al. KDD2018]

医学的再検証

Data augmented PBR
[Hashimoto et al. BJO2020]

Modified PBR
[Hashimoto et al. TVST2021]

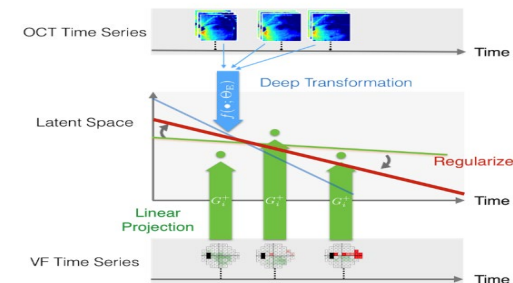


RTとVF双方を用いた
緑内障進行予測

Deep Latent Linear
Regressor
[Zheng et al. KDD2019]

医学的再検証

Re-examination
[Xu et al.
Ophthalmology
Glaucoma 2020]

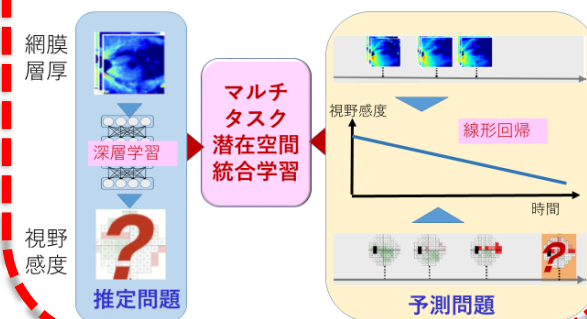


VFの推定と緑内障
進行予測の統合

Joint Estimation
and Prediction
[Xu et al. KDD2021]

医学的再検証

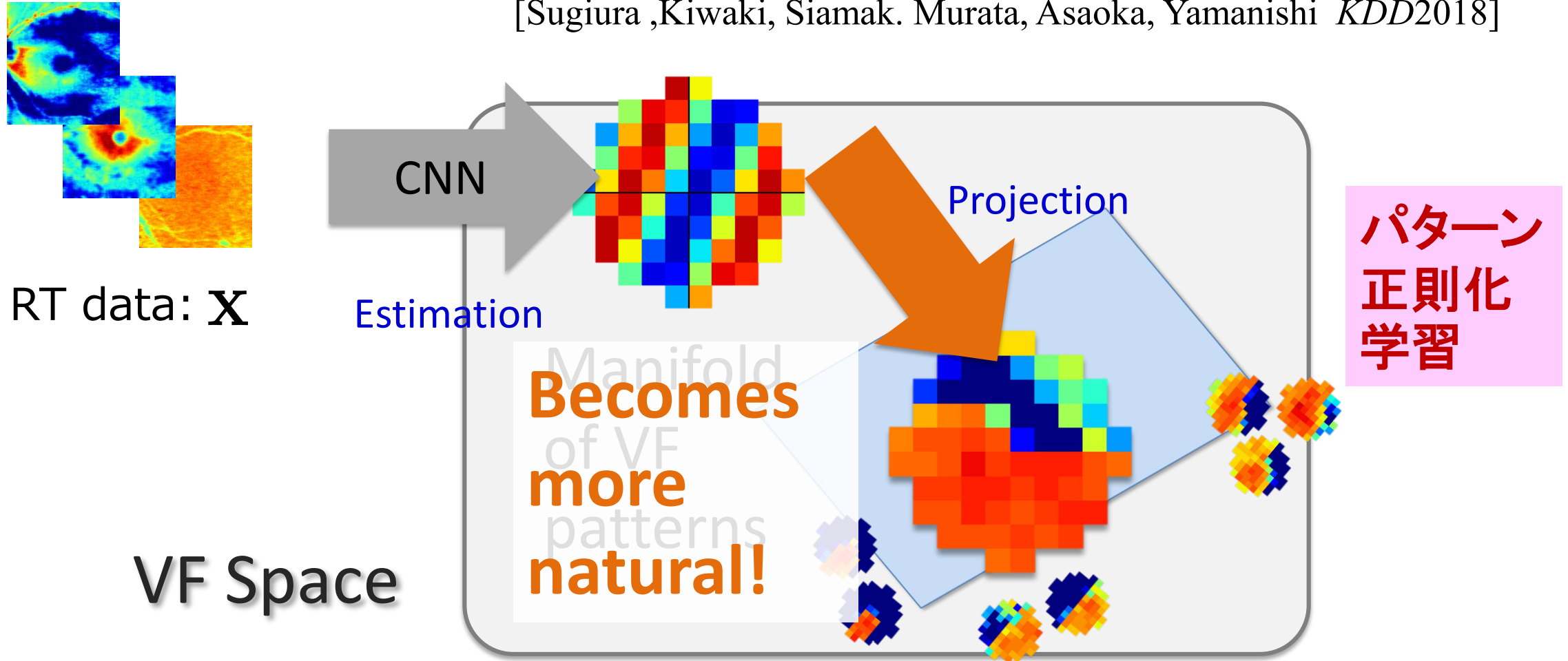
Re-examination
[Asaoka et al.
Ophthalmology.Science
2021]



網膜層厚(RT)からの視野感度(VF)の推定 ーパターン正則化技術ー

Biased data に対して、出力側(VF側)のパターンを教師なし学習、パターン空間へ射影

[Sugiura ,Kiwaki, Siamak. Murata, Asaoka, Yamanishi *KDD*2018]



[Hashimoto et al. *Br. J. Ophthalmology*2020]にて再検証

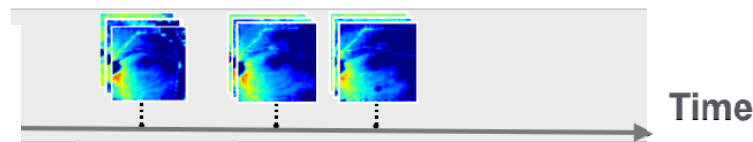
網膜層厚と視野感度の統合予測

—DLLR (Deeply-regularized Latent Space Linear)—

視野の回帰予測 を 潜在区間にて網膜厚データで正則化、精度改善

[Zheng, Xu, Kiwaki, Wang, Murata, Asaoka, Yamanishi KDD2019]

網膜層厚み時系列



潜在空間

LPPs

$f(\cdot; \Theta_E)$ Deep Transformation

Regularized!

潜在空間
回帰モデル

Linear
Projection

G_0^+

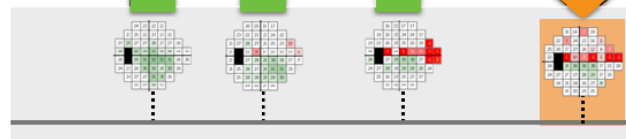
G_0^+

G_0^+

G_0

Time

視野感度時系列



Accurate
Prediction!

Time

$$\mathcal{L}_{\text{DLLR}} = \sum_{i=0}^N \|M_i \odot (F_i - G_i W_i P_i^\top)\|_F^2 + \lambda_0 \sum_{j=1}^N z_{0,j} \|W_0 - W_j\|_F^2 + \lambda_1 \sum_{i=0}^N \|f(C_i; \Theta_E) - W_i\|_F^2,$$

Latent space leaning

Collective learning

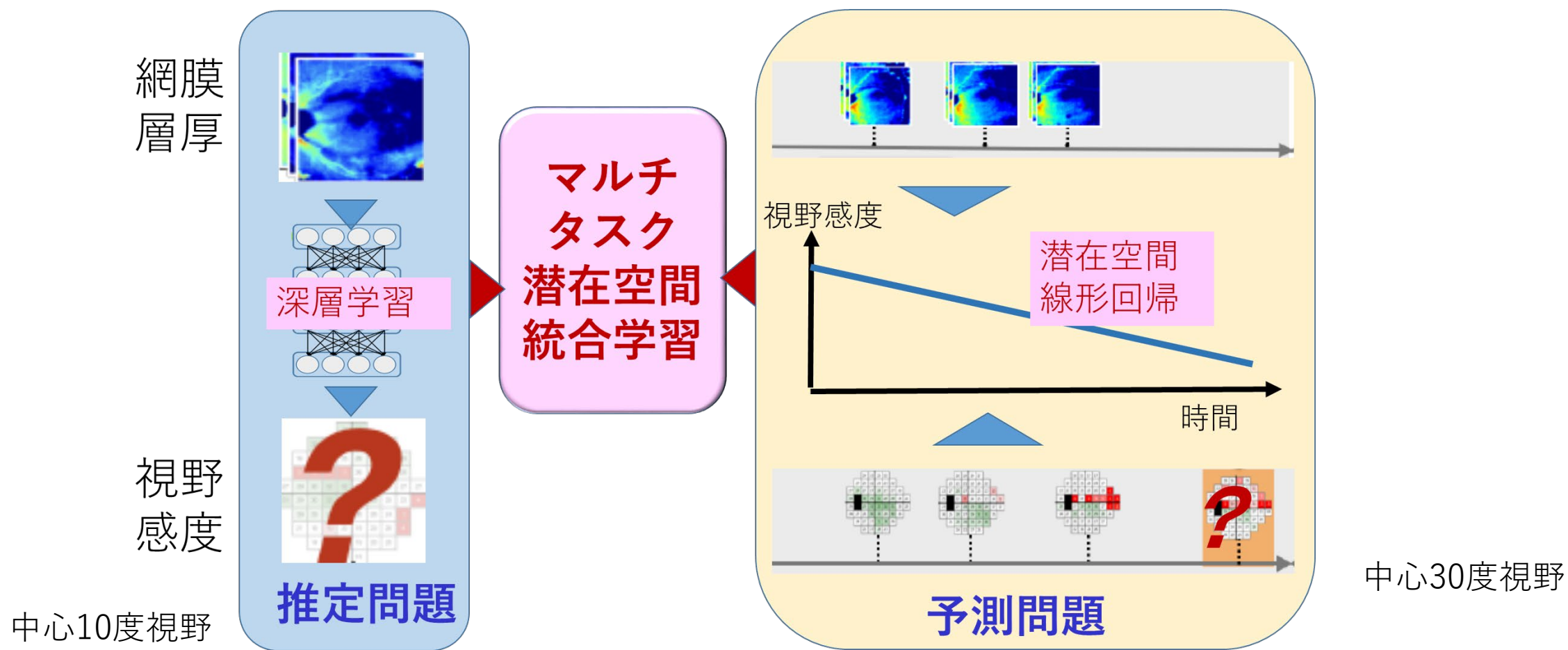
RT to VF transformation with CNN

[Xu, et al. *Ophthalmology Glaucoma* 2020] にて医学的に再検証

網膜層厚からの視野感度の推定 と緑内障進行予測の統合

マルチタスク潜在回帰モデルにより推定と予測の相乗効果で精度増強

[Xu, Asaoka, Kiwaki, Murata, Fujino, Yamanishi *KDD2021*]



実験結果

表 1. 視野感度推定問題における平方根平均二乗誤差

MLR:多変量線形回帰 SVR:サポートベクトル回帰

CNN-PR:畳み込みニューラルネット+パターン正則化

CNN-TR:畳み込みニューラルネット+テンソル回帰

MLR	SVR	CNN-PR	CNN-TR	Our model
8.56 (3.69)	7.18 (3.87)	6.76 (3.86)	6.32 (3.76)	5.92 (3.54)

表 2. 視野感度予測問題における平方根平均二乗誤差

PLR: 視野点毎の線形回帰 DLLR: 深層正則化線形回帰

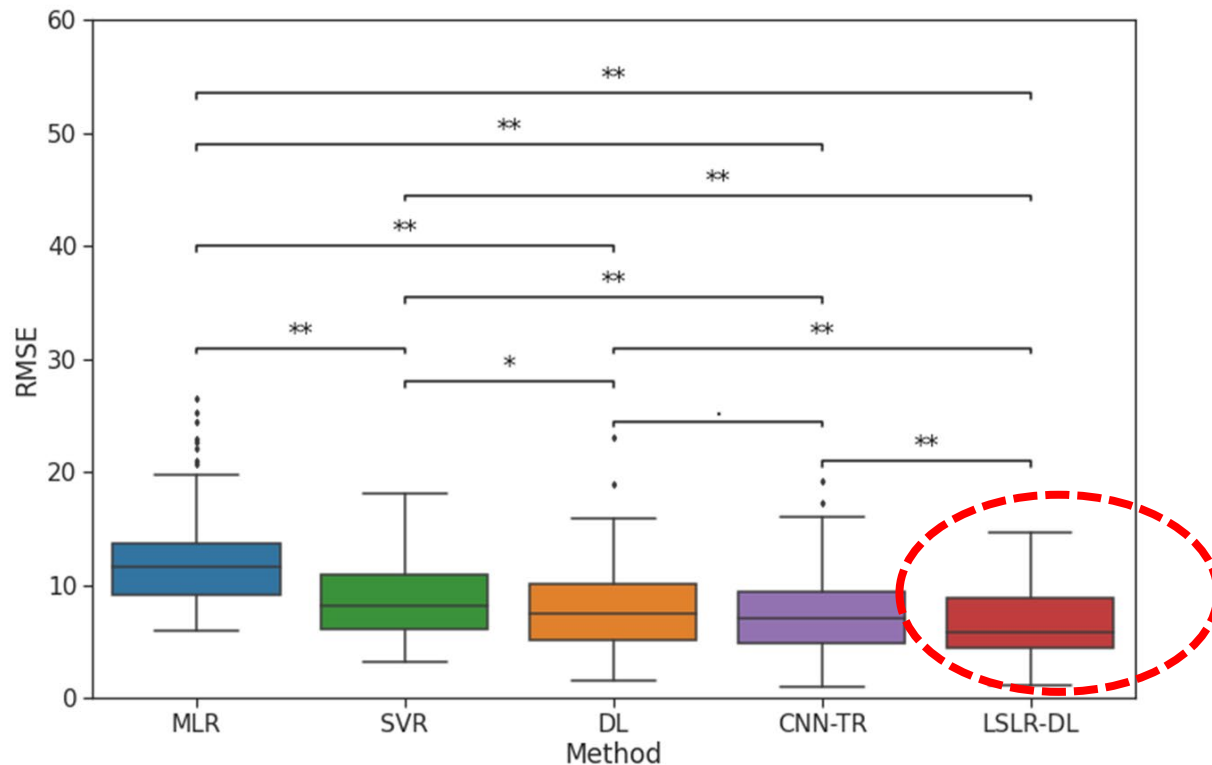
Mean RMSEs		PLR	DLLR	Our model	[Xu, Asaoka, Kiwaki, Murata, Fujino, Yamanishi <i>KDD</i> 2021]
C: 視野検査の回数	C=2	33.11 (33.33)	4.58 (2.70)	4.42 (2.73)	
	C=3	14.52 (8.65)	4.31 (2.57)	4.16 (2.66)	
	C=4	9.06 (5.14)	4.14 (2.57)	4.05 (2.66)	
	C=5	6.40 (3.78)	4.03 (2.45)	3.97 (2.59)	
	C=6	4.91 (2.61)	3.81 (2.18)	3.85 (2.35)	
	C=7	3.93 (2.33)	3.80 (2.16)	3.83 (2.23)	

医学的再検証

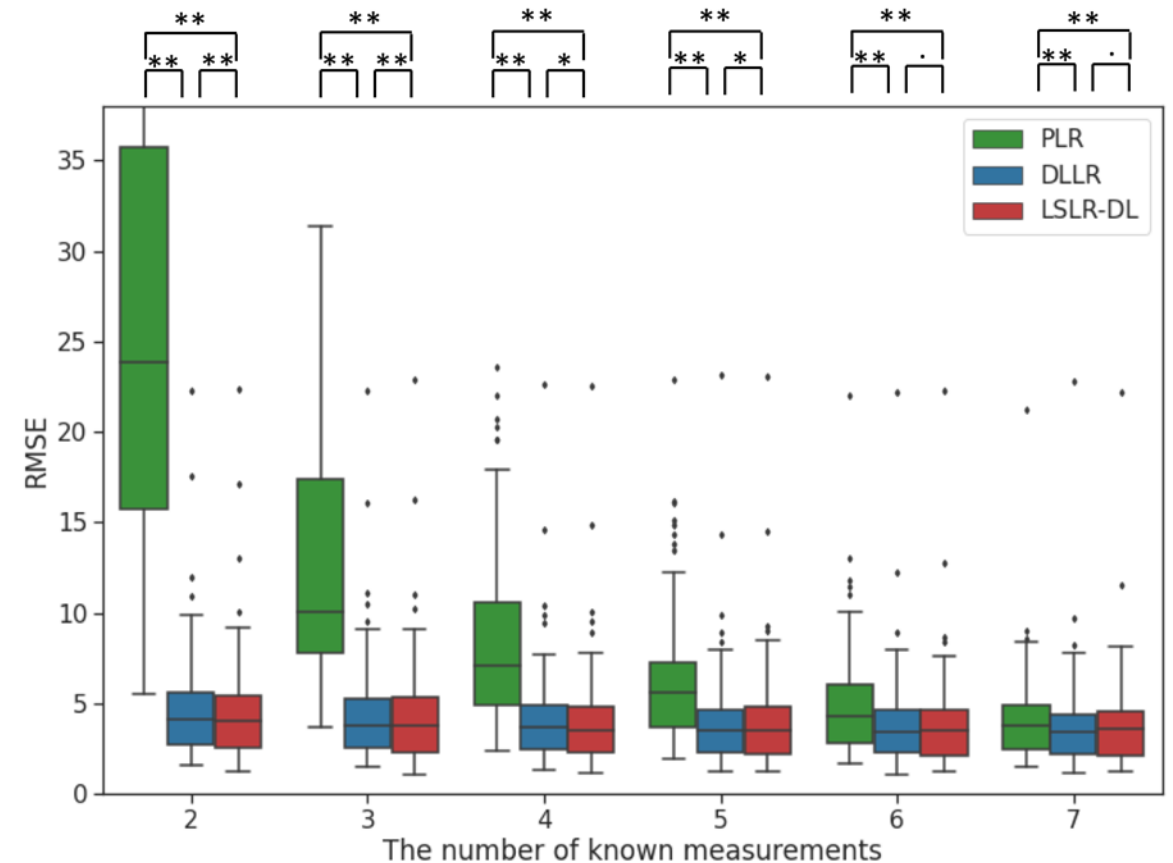
提案手法は推定・予測ともに既存手法を凌駕、最高精度達成

[Asaoka, Xu, Murata, Kiwaki, Matsuura, Fujino, Tanito, Mori, Ikeda, Kanamoto, Inoue, Yamagami, Yamanishi
Ophthalmology Science 2021]

網膜厚からの視野感度推定精度



網膜厚と視野感度からの予測精度



おわりに：今後の展望

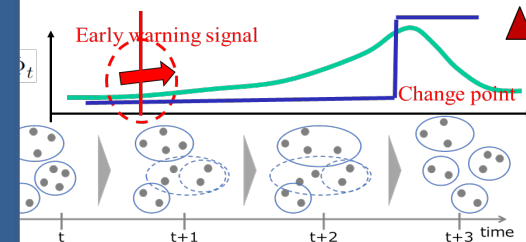
理論構築と医学応用を発展させ、「予兆情報学」に展開する

理論研究

潜在空間埋め込み
潜在構造最適化

変化予兆検知

予兆情報学の発展

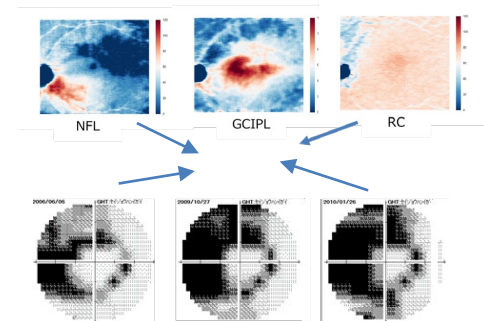
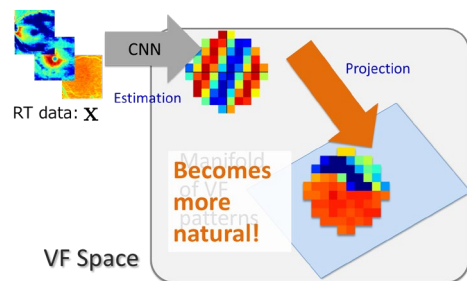


計算論的眼科学

大規模データを用いた
医学的再検証

ヘテロ情報を用いた
緑内障診断・予測

さらなるヘテロ情報の取り込み
精度増強、実用化



公開ソフトウェア

- Word2vec-skipgram 埋め込み次元選択アルゴリズム
<https://github.com/truythu169/snml-skip-gram>
- 分解正規化最尤符号長 (DNML) によるモデル選択アルゴリズム
<https://github.com/tianyi-wu/DNML>
- 混合指数分布の推定及びモデル選択アルゴリズム
https://github.com/naokimas/exp_mixture_model
- 組合せ2元モデル上のパターン学習アルゴリズム
<https://gitlab.com/taitor/GRAB-experiments>
- 階層的変化検知アルゴリズム (HCDL)
<https://github.com/s-fuku/hcdl>
- メタ変化検知アルゴリズム
<https://github.com/s-fuku/metachange>
- 埋め込みによるネットワーク変化検知アルゴリズム
https://bit.ly/RWiLSn_NCD
- 変化予兆検知アルゴリズム <https://github.com/IbarakikenYukishi/differential-mdl-change-statistics>.
及びそのCOVID19感染者数変化検知システム
<https://ibarakikenyukishi.github.io/d-mdl-html/index.html>
- Descriptive Dimensionによる連続的モデル選択アルゴリズム
<https://github.com/conf-post-01/Ddim.git>

生命現象の発生原因など多様で変動性を持つ複雑データの因果関係 解析手法にかかり、直面する問題点・解決策案および将来展望など

■ 計算論的眼科学の研究から

★ヘテロ情報の統合 ⇒ 適切な潜在空間への埋め込みが本質的

潜在表現最適化

埋め込み次元推定、埋め込み空間の決定
潜在空間の説明能力を最大化することが重要

■ マーケティング、人間行動モデリングの研究から

★パターンは変動し、モードを切り替えるのが本質

予兆情報学

変化を見逃さず、その予兆を捉え、説明する
離散的観測から連続的観測へ。