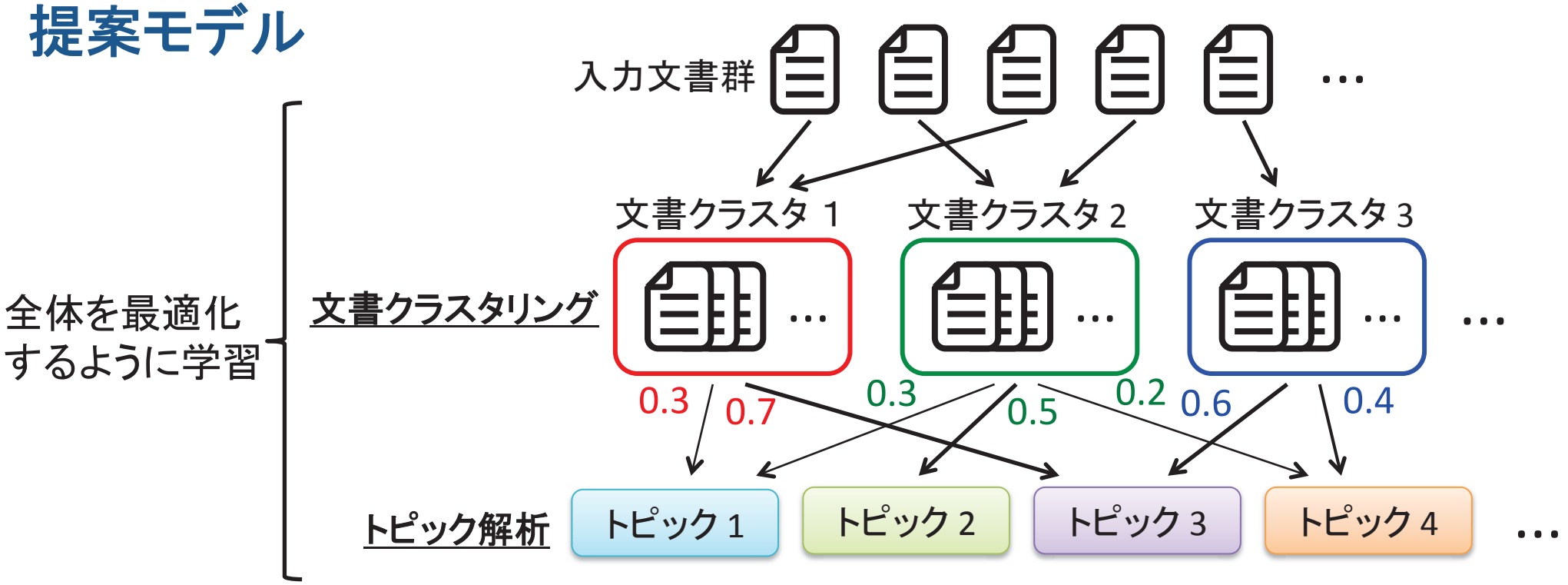
T-23

入れ子階層ディリクレ過程による文書–トピック同時クラスタリング

富永将至，下坂正倫，福井類，佐藤知正（東大）

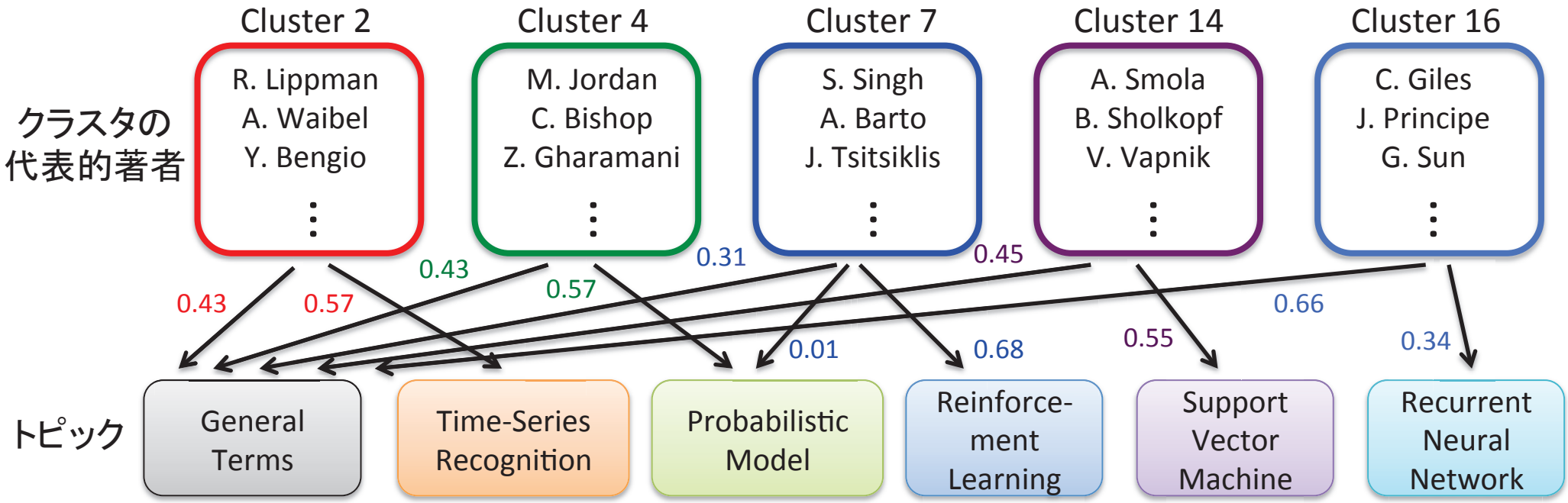
提案モデル



- 文書クラスタ数とトピック数はデータから自動で推定する
- 上記実現のためのノンパラメトリック混合モデルの提案
  - nested-Hierarchical Dirichlet Process (nHDP) Mixtures
  - 閉形式での変分推論が可能

$$G_0^* \sim \mathrm{DP}(\gamma, H), \quad Q \sim \mathrm{DP}(\alpha, \mathrm{DP}(\beta, G_0^*)), \quad \{G_d\}_{1,\ldots,D} \overset{\text{i.i.d}}{\sim} Q$$

NIPS※1 '87～'99 (論文数1740) クラスタリング例 (抜粋)



予測実験

Per-word log-likelihood to test dataset

nHDP-LDA	hLDA [1]	HDP-LDA [2]	nDP-LDA※2
-7.80±0.25	-7.84±0.26	-8.12±0.20	-8.02±0.24

- 既存のトピックモデルと比較し，新規データの予測性能が向上

※1http://books.nips.cc/ ※2比較用にLDAの事前分布をnested DP[3]として実装

[1] Blei. J. ACM. 2010 [2] Wang. AISTATS2011 [3]Rodriguez. JASA. 2008

T-33

Classification and Numbering on Posterior Dental Radiography using Histogram Intersection

Agus Zainal Arifin[1], Ahmad Mustofa Hadi[1], Anny Yuniarti[1], Wijayanti Nurul Khotimah[1], Arya Yudhi Wijaya[1], and Eha Renwi Astuti[2]

1Department of Informatics, Institut Teknologi Sepuluh Nopember (ITS) Kampus ITS, Surabaya, 60111, Indonesia,
2 Department of Dental Radiography Airlangga University

ABSTRAK

Classification and numbering on posterior dental radiography are useful for forensic and biomedical application. This paper proposed a newmethod that simple but robust for classification and numbering on posterior dental radiography using histogram intersection. In this method, we utilized the different characteristic between molar and premolar. Molar has wide shape and two roots while premolar has slander shape and one root. We computed the distance between centroid and boundary in each tooth. 36 points in the b oundary were extracted. A histogram was created from the distance between centroid and those points. Then similarity among the histograms was computed. From experiment, every tooth has been assigned according to universal dental numbering and classified as their sequence order. Our system achieved average c lassification precision of 90 %.

Forensic application usually use physical characteristic such as face, fingerprint, palm print, eyes, and DNA. However, many of those characteristics are only suitable for ante mortem (AM) identification when a person to be identified is still alive not for post mortem (PM) especially in the case of decay or severe body damage caused by fire or collision. In the other side dents have a special characteristic that can be used to identify victims. Although identifying victims using dents is secure-less but it is required in special cases, where the victim's body has serious defect and analyzing another part of body is difficult. That is because dents are strong part that usually remains when another part of body difficult to be identified.

BACKGROUND

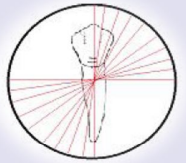
METHOD

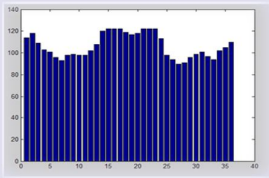
Preprocessing

1. Contrast enhancement and morphological operation.
2. Binarization
3. Image separation (horizontal integral projection and vertical projection).

Process

1. Feature extraction (distance between centroid and boundary of each points)





2. Computed histogram intersection between histogram from the testing data and histogram of molar or premolar from training.

From this process, we determined wether the teeth is molar or premolar.

3. We computed similarity matrix between two default pattern (MMMPPP-PPPMMM) to define the position of the teeth, in left or in right. The result of this process is the dental number.

RESULT

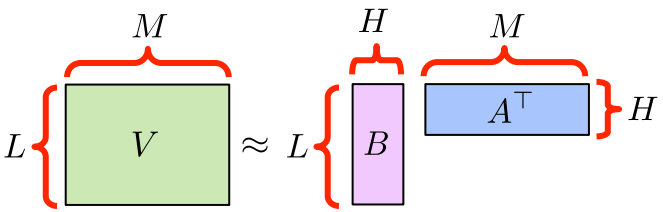
Kumamoto University

T-34: On Dimensionality Recovery Guarantee of Variational Bayesian PCA

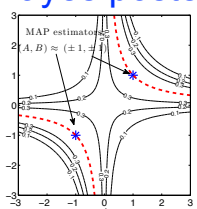
Shinichi Nakajima (Nikon), Ryota Tomioka (Tokyo Univ.),
Masashi Sugiyama (Tokyo Tech.), S. D. Babacan (Illinois Univ.)

VB-PCA (matrix factorization)

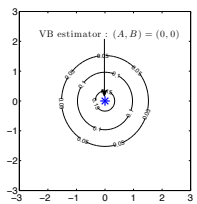
$$V = BA^\top + \mathcal{E}$$

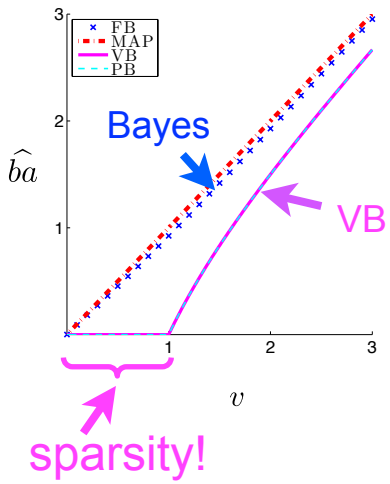


Bayes posterior



VB posterior

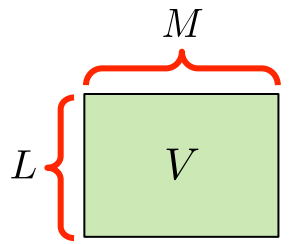




Bayes and VB behave quite differently!

Like to directly prove good property of VB!

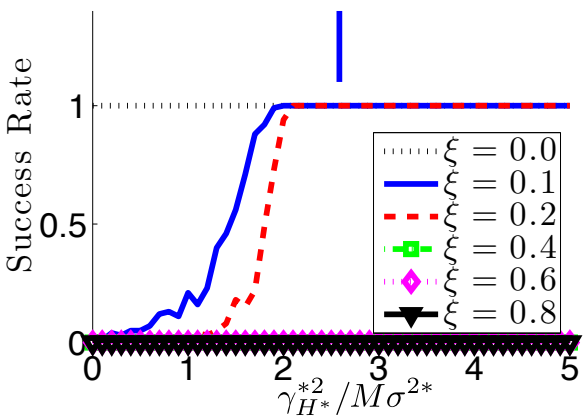
Perfect dimensionality recovery



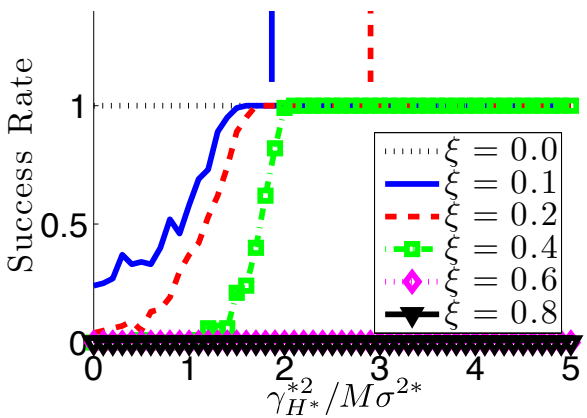
Theorem: In the large scale limit when $L, M \to \infty$, the true PCA-dimensionality is recovered if

$$\xi < \frac{1}{x} \quad \text{and} \quad \gamma_{H^*}^{*2} > \left(\frac{x-1}{1-x\xi} - \alpha \right) \cdot M\sigma^{2*}$$

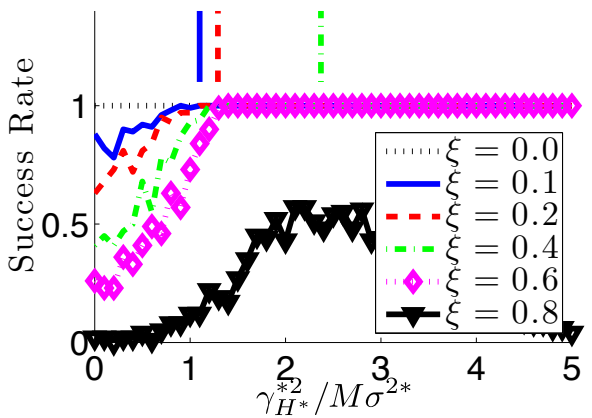
$$\alpha = \frac{L}{M}, \qquad \xi = \frac{H^*}{L}.$$



(a) $\alpha = 1$



(b) $\alpha = 0.5$



(c) $\alpha = 0.1$

We derived a sufficient condition for VB-PCA to perfectly recover the true dimensionality.

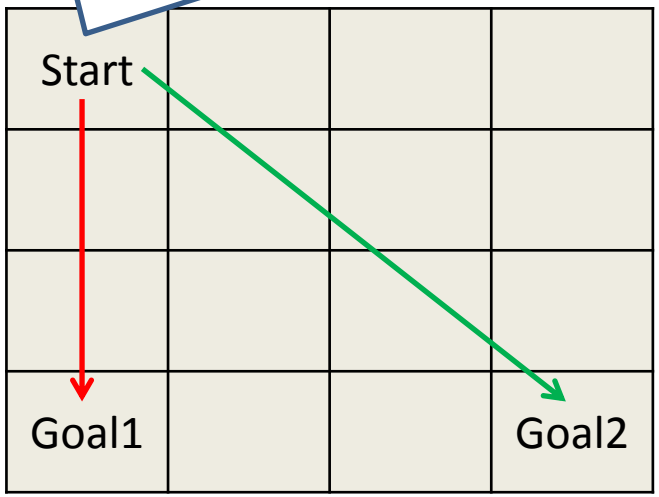
T-35　Concurrent Q LearningとSarsa、Q学習の動的環境への適応能力

村上和謙　尾関智子　（東海大学）

Concurrent Q Learning

あらゆるゴールの可能性を同時に解決する強化学習手法

$Q^{Goal1}(s,a)$ と、$Q^{Goal2}(s,a)$ の両方の価値を同時に学習する

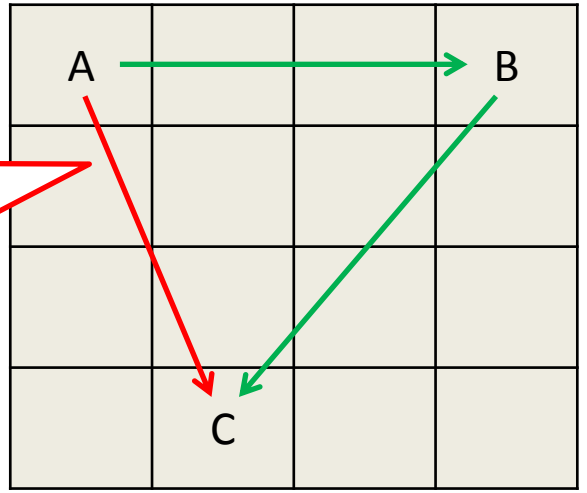


Relaxation

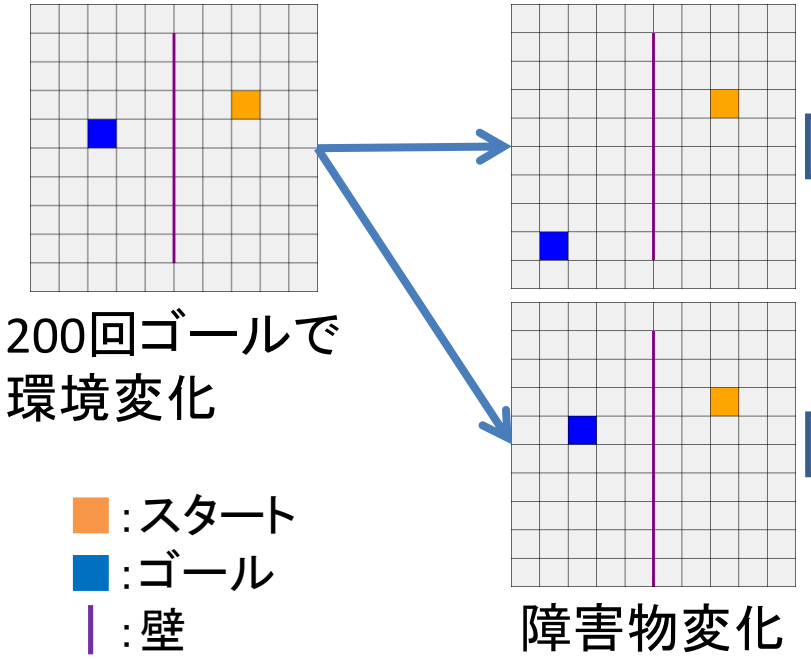
$\overline{AC} \le \overline{AB} + \overline{BC}$ なら

$Q^C(A,a) \ge Q^B(A,a) * \max_a Q^C(B,a)$ である

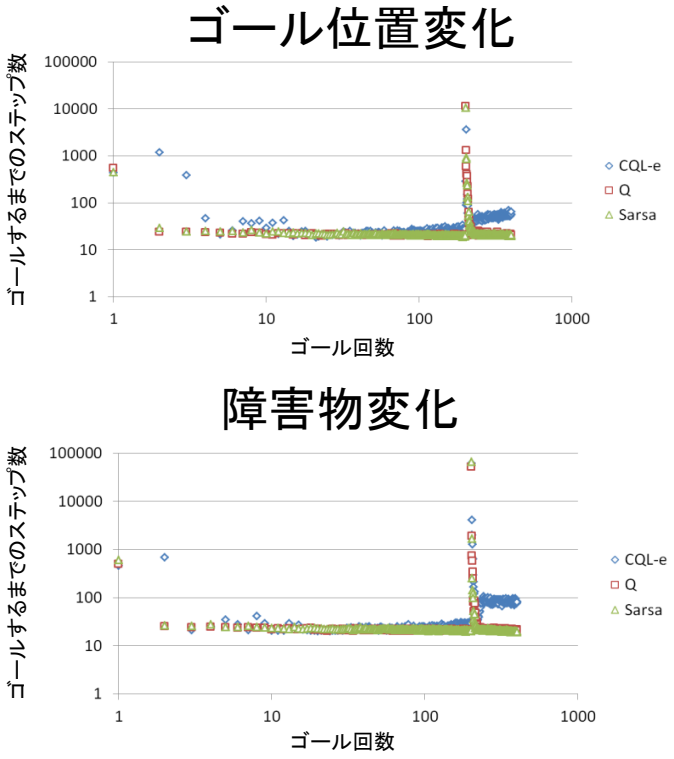
上式を満たすように修正する



実験



実験結果



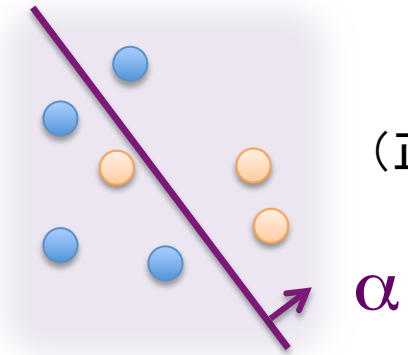
まとめ

- CQL-eはゴール位置の変化に対し、ほかの手法より対応が速い
- しかし、障害物の変化への対応について、優位性は見られない

T-36 Efficient AUC Maximization by Approximation Reduction of Ranking SVMs

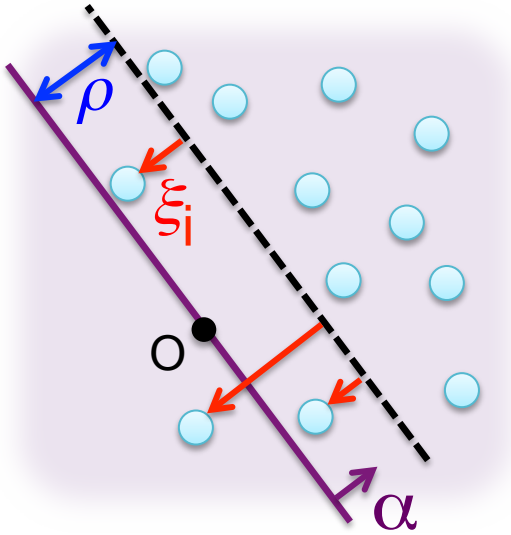
末廣大貴・畑埜晃平・瀧本英二（九州大学）

2部ランキング問題



帰着
（正例と負例の各ペアに対し，
1つの正例を作成）

Ranking SVM



入力：p 個の正例と n 個の負例
出力：線形関数 α

目標：$\mathrm{AUC} = \Pr_{\substack{x^+ \sim \text{正例} \\ x^- \sim \text{負例}}}(\alpha \cdot x^+ > \alpha \cdot x^-) \rightarrow$ 大

入力：*pn* 個の正例
出力：線形関数 α

目標：$\rho - (1/\nu)\sum_i \xi_i \rightarrow$ 大

問題点

- 問題のサイズが *p* + *n* から *pn* に増大
- 既存の効率化手法はマージンに関する理論保証なし

提案手法

・最適化問題の $p \times n$ 制約条件を $p + n$ 制約条件に近似（計算量改善）
・非凸な制約条件を線形制約に近似（マージンに理論保証有り）

実験結果

AUC比較

Datasets		Ranking SVM	Our Method
Noise	K		
5%	10	0.9891	**0.9909**
	30	0.9611	**0.9648**
	50	0.8465	**0.8687**
10%	10	0.9836	**0.9873**
	30	**0.9109**	0.9055
	50	0.8759	**0.8852**

計算時間比較 (sec.)

Sample size	Ranking SVM	Our Method
25	0.206	0.104
50	0.3014	0.162
100	49.328	0.238
200	809.506	0.530
400	N/A	1.814

※共にノイズ5%の人工データ

・AUC はナイーブな手法に匹敵
・計算時間はナイーブより超高速

T-37 線形代数的アプローチによる ノード置換不変な行列カーネルの構築

広瀬俊亮 (SAS Institute Japan)

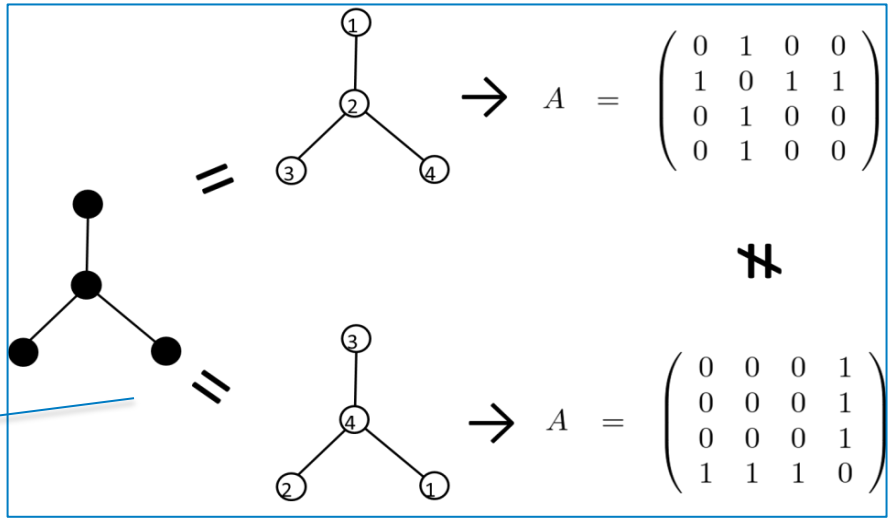
行列を引数とするカーネルを構築したい

- 構造データの比較は、データマイニングの重要な応用の一つ。
  - » NWの隣接行列や変数間の相関行列など、行列型のデータは多い。

要請：ノード置換不変な行列カーネル

1. 追加の情報なしで、単に行列同士を比較できること。
2. 次元の異なる行列を入力とできること。
3. ノード置換不変性
   - » ノード置換をしてもカーネル関数の値が不変：

$$K(A, A') = K(UAU^{\dagger}, U'A'U'^{\dagger})$$



ノード置換
行列成分のラベルを入れ換え。
構造の実体は同じでも表現行列は変化。

T-37 線形代数的アプローチによる ノード置換不変な行列カーネルの構築

提案手法: 線形代数的なアプローチで行列カーネルを構成

- 射影演算子: 入力行列を形式的に同じ次元に揃える。
  - » 異なる次元の行列を比較することが可能に。

$$K(A,A') = \sum_{i,j=1}^{s} \left(\mathcal{S}(s\leftarrow n)A\mathcal{S}(s\leftarrow n)^{\dagger}\right)_{ij} \left(\mathcal{S}(s\leftarrow n')A'\mathcal{S}(s\leftarrow n')^{\dagger}\right)_{ij}$$

n次元空間からs次元空間への射影演算子

- 固有値分解: 射影演算子の作用を見え易くする。

$$K(A,A') = \sum_{l,m} \lambda_l \xi_m \mathrm{tr}\left[S\psi_l\psi_l^{\dagger}S^{\dagger}S'\phi_m\phi_m^{\dagger}S'^{\dagger}\right]$$

固有値　　固有ベクトル

- 確率分布: 射影後のベクトルとして確率密度関数(無限次元)を採用。

$$S(s\leftarrow n)\psi_l = \sqrt{p(\cdot|\psi_l)}$$

ノード置換不変かつ元のベクトルの情報をある程度保持している。

提案手法(カーネル)の特長	
	1. 入力行列以外の情報を必要としない。
	2. 異なる次元の行列を入力とできる。
	3. ノード置換に対して不変である。

T38 逐次的な重点サンプリングを用いたWAIC計算法

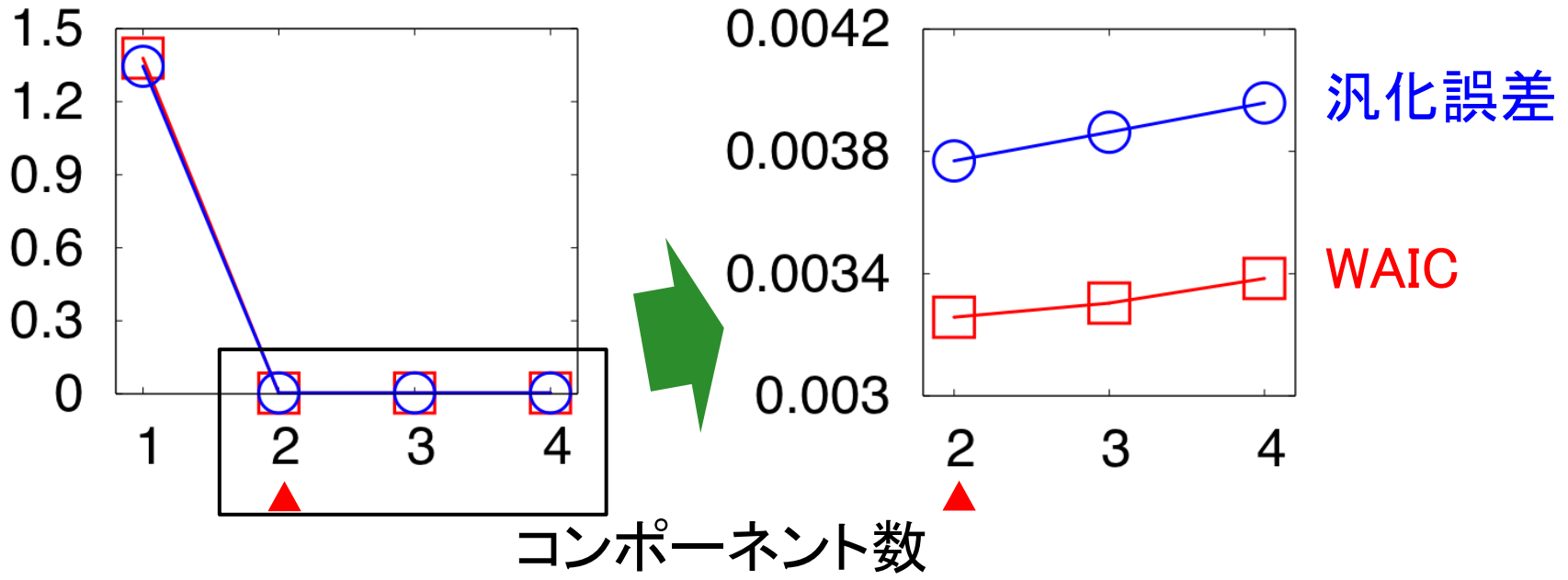
三木・渡辺[東工大]

- WAIC(広く使える情報量規準)
  - 任意の真の分布・学習モデル・事前分布に対して汎化誤差の不偏推定量となる
  - 事後分布での期待値計算を含むので計算に時間がかかる
- 提案
  - 逐次的重点サンプリング法を用いて事後分布を構成し，WAICを計算する
- 結論
  - 十分な精度で効率的に計算できることが分かった

T38 人工データを用いた実験

- 混合正規分布モデル
- 真のコンポーネント数2
- データ数1000
- 120セットの平均

コンポーネント数	1	2	3	4
計算時間[秒/セット]	34	62	96	136



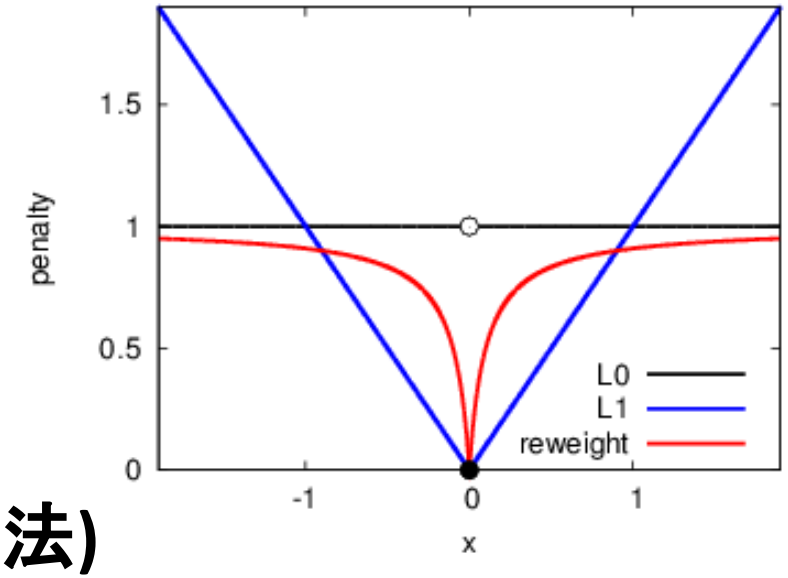
T-39 パス追跡を用いた再重み付け ℓ_1 最小化アルゴリズム

新村 祐紀, 竹内一郎(名工大)

■ 圧縮センシング

■ スパースな信号の再現

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} ||\boldsymbol{x}||_{\ell_1} \text{ s.t. } \boldsymbol{y} = \boldsymbol{\Phi x}$$



■ 再重み付け ℓ_1 最小化 (従来法)

重み付き ℓ_1 最小化

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} ||\mathbf{W}\boldsymbol{x}||_{\ell_1} \text{ s.t. } \boldsymbol{y} = \boldsymbol{\Phi x}$$

⇄

重みの更新

$$\mathbf{W}_{ii} = \frac{1}{|x_i| + \varepsilon}$$

➡ よりスパースな解を得られることが知られている

T-39 パス追跡を用いた再重み付け ℓ_1 最小化アルゴリズム

新村 祐紀, 竹内一郎(名工大)

■ 従来法の問題点

- ■ 繰り返し線形計画問題を解くため，コストがかかる

■ 本研究の目的

効率的な重みの更新を行い，計算コストの削減を図る

■ 提案手法

- ■ パラメトリック計画法 (パス追跡) を利用

T-40

パラメトリック計画法を用いたマルチインスタンスSVM

石原 直樹, 久留美 里織, 竹内 一郎 (名工大)

マルチインスタンス学習

- バッグに対する2クラス分類問題
- バッグ：インスタンスの集合 ⇒ 写真
- インスタンス ⇒ 分割領域

バッグのラベル

- 正：少なくとも一つのインスタンスが正
- 負：全てのインスタンスが負

インスタンスのラベルは未知

T-40

パラメトリック計画法を用いたマルチインスタンスSVM

石原 直樹, 久留美 里織, 竹内 一郎 (名工大)

マルチインスタンスSVM

- SVMの概念をマルチインスタンス学習に適用
- 非凸最適化問題

より良い局所解を求める

提案法

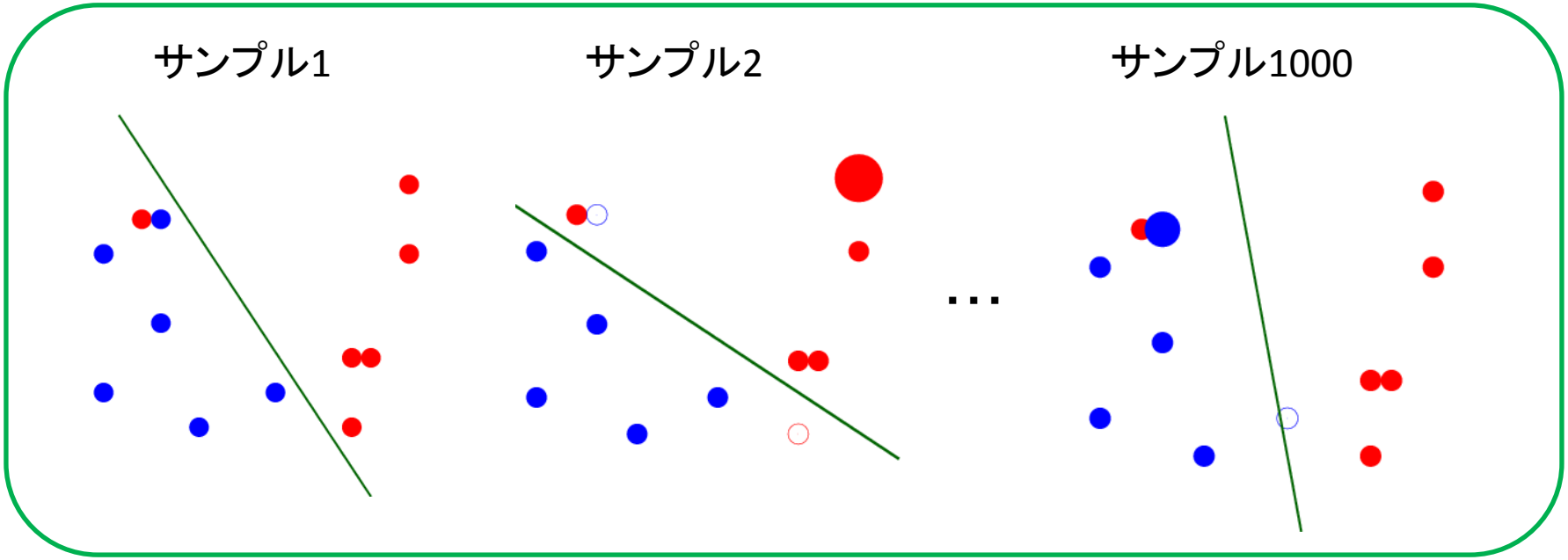
- パラメトリック計画法
- シミュレーテッドアニーリング

T-41

パラメトリック計画法を用いたSVMブートストラップ計算の高速化

鈴木 良規, 小川 晃平, 竹内 一郎(名工大)

- 問題設定
  - ブートストラップ法を用いたSVMの統計的バラツキの推定



多数回の学習に伴う計算コスト

パラメトリック計画法を用いたSVMブートストラップ計算の高速化

鈴木 良規, 小川 晃平, 竹内 一郎(名工大)

- 提案手法
  - パラメトリック計画法による効率的な学習

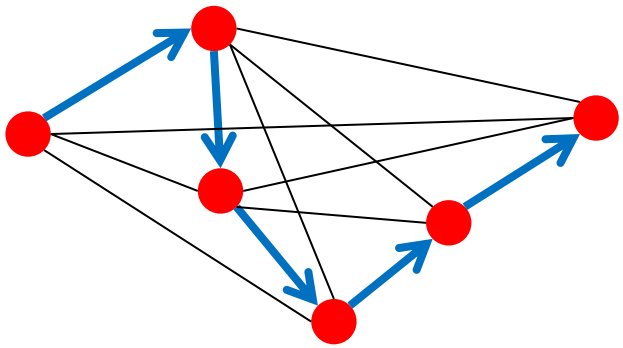
    問題の変化量から解を追跡

    解の変化が小さいときに効率的

  - より効率的な学習のためのスケジューリング

    解の近さ → アクティブセットの違い

    - アクティブセットの変化を予測
    - MSTに基づくスケジューリング



42

圧縮センシングに基づく自由視点画像合成の高効率化

Efficient free-view image synthesis based on the compressed sensing

東京理科大学大学院　工学研究科　電気工学専攻

苫米地　大　　保坂　忠明　　浜本　隆之

自由視点画像合成

複数の実画像を基に，カメラの存在しない仮想視点からの画像を推定する技術

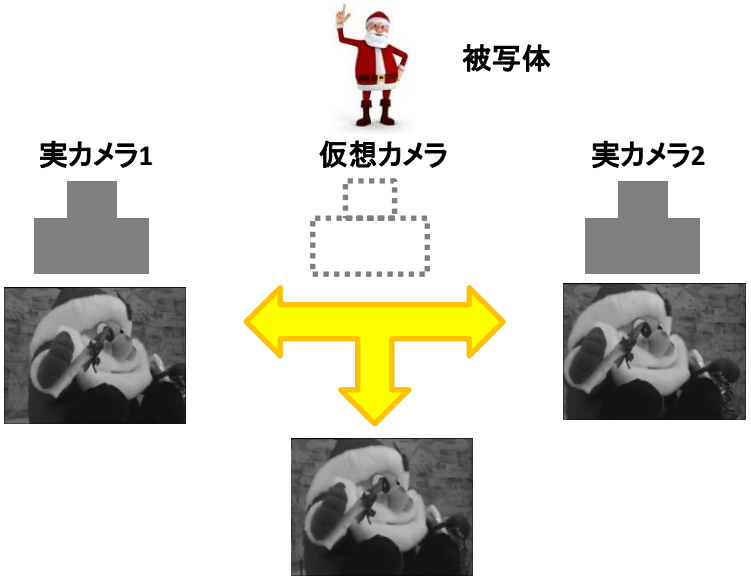
自由視点画像合成の手順

・複数の実視点画像を用いて被写体までの距離を推定

・三次元幾何を基に仮想カメラの画像を推定

問題点

自由視点画像合成におけるデータ量の増大



提案手法

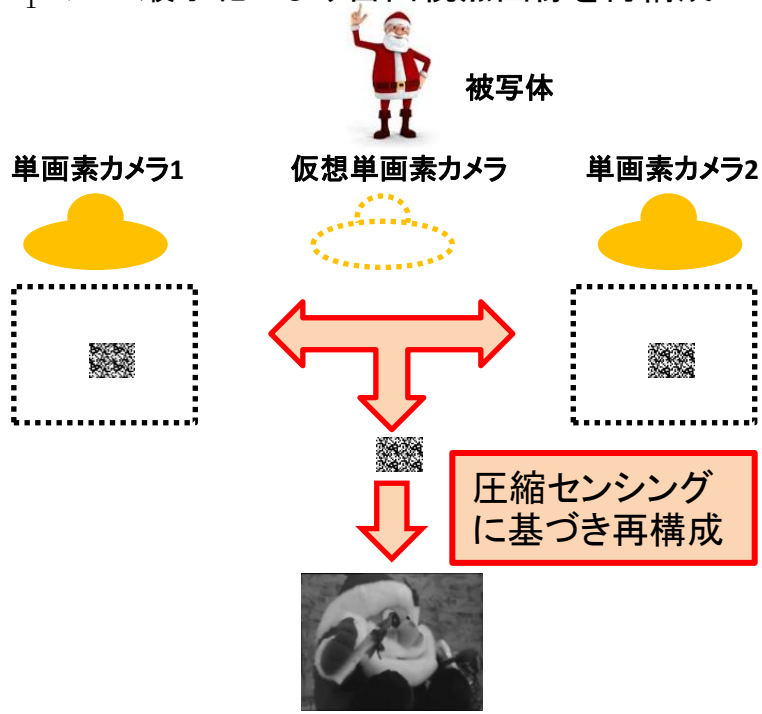
研究目的

圧縮センシングに基づいてデータ量を削減

提案手法の手順

・少数の線形観測値で距離を推定

・l_1ノルム最小化により自由視点画像を再構成



42

圧縮センシングに基づく自由視点画像合成の高効率化

Efficient free-view image synthesis based on the compressed sensing

東京理科大学大学院　工学研究科　電気工学専攻

苫米地　大　　保坂　忠明　　浜本　隆之

評価実験

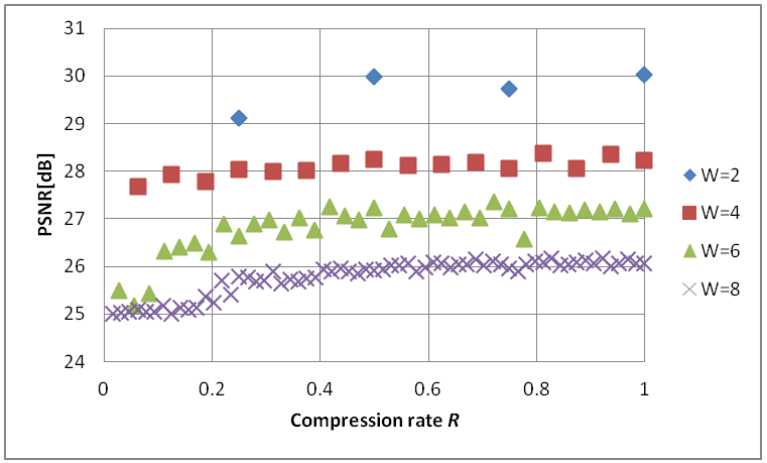
実カメラと単画素カメラから取得されるデータ量の比を圧縮率Rと定義($R \leqq 1$)

Ground Truth

R=1/16
27.69[dB]

R=1/4
28.04[dB]

R=1/2
28.24[dB]



圧縮率Rと画質の関係(Wはパラメータ)

- 圧縮率Rを下げることに伴う画像劣化の程度が小さい
- データ量を1/16に削減しても視覚的に良好な仮想視点画像が得られる

自由視点画像合成における
データ量の削減が可能

T-43: An Efficient Sampling Algorithm for Bayesian Variable Selection

荒木貴光　池田和司

奈良先端科学技術大学院大学

ギブス変数選択法

ベイジアン変数選択におけるサンプリング法

特徴：効率性が擬似事前分布，提案分布のパラメータに強く依存する．

従来法

これらパラメータを**フルモデルに対する予備解析**で決定

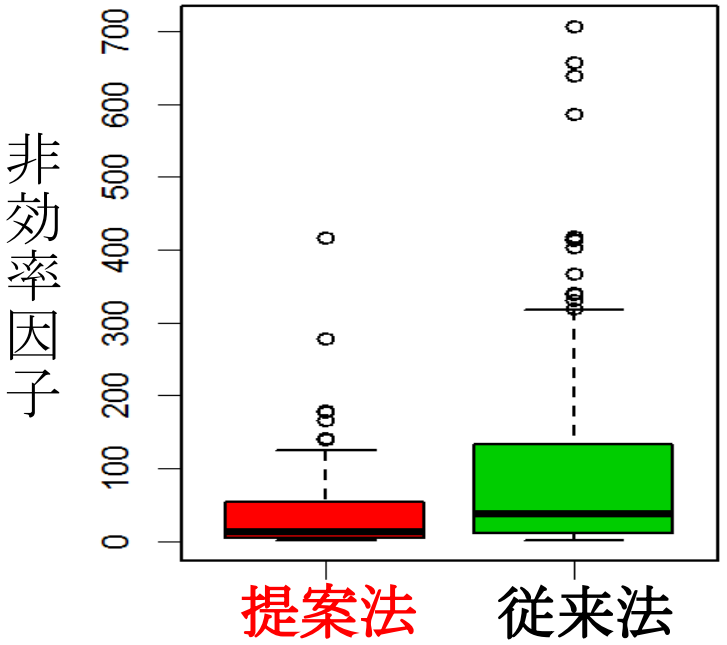
パラメータが不適切な値となりアルゴリズムの効率が下がる．

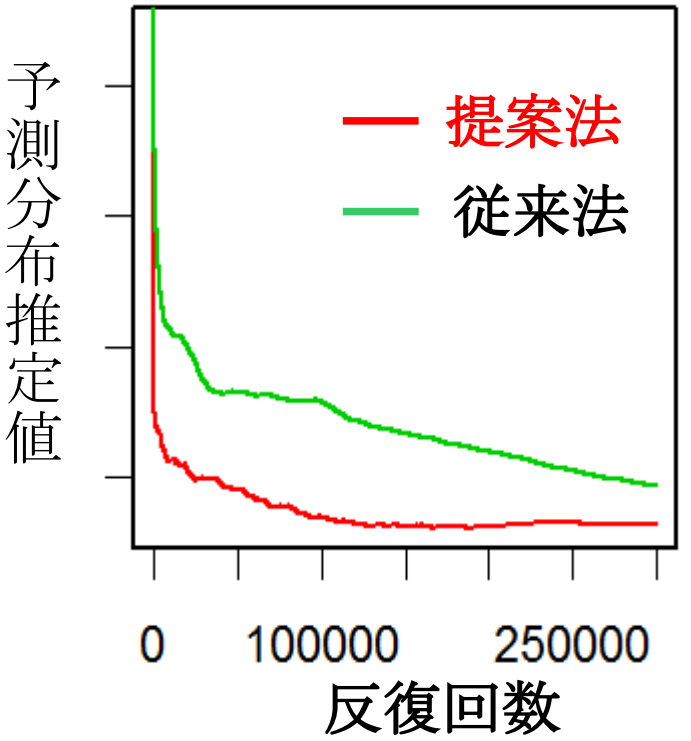
提案法

適応的ギブス変数選択法

… サンプリングしながら擬似事前分布，
提案分布のパラメータを自動調整する．

数値実験





まとめ

適切なパラメータ値を自動で獲得

効率性が向上

T-44

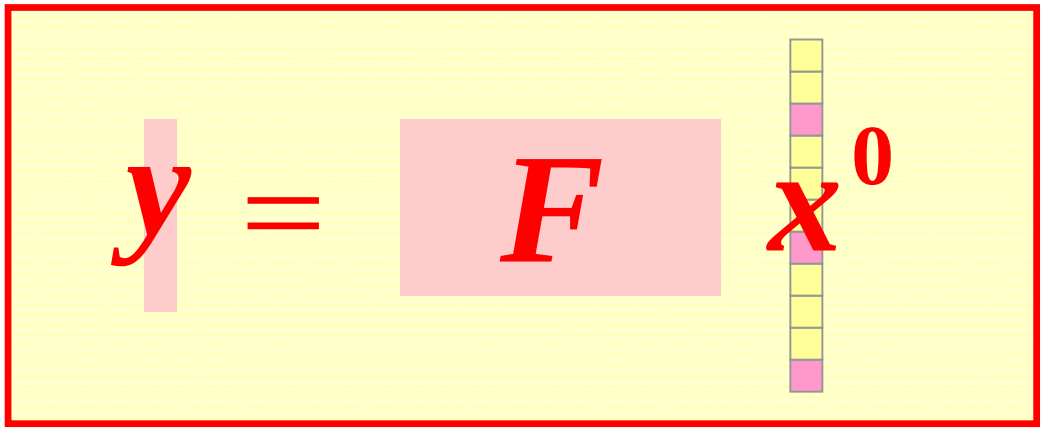
事後確率最大化推定に基づく圧縮センシングのデータ復元アルゴリズム

竹田晃人・樺島祥介（東工大）

★圧縮センシング

低次元のyとFから
x^0が知りたい!

但しx^0は疎



★ℓ_1ノルム最小化解法

ℓ_1ノルム $\sum_i |x_i|$ を最小にする $y = Fx$ の解を求める

線型計画問題… 計算量がデータサイズの3乗

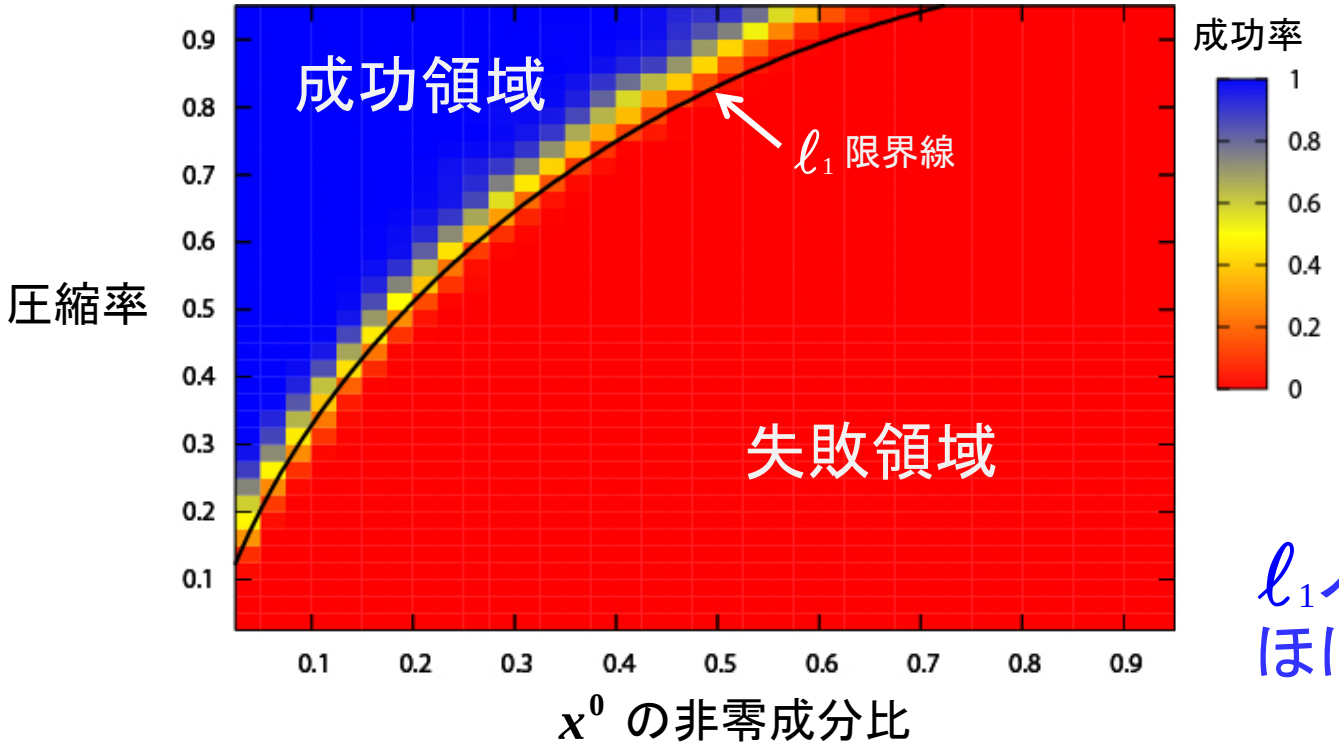
データサイズが大きい場合は計算量がさらに小さい方が良い

★ 事後確率最大化に基づく解法を提案

計算量がデータサイズの2乗

(***F***が疎行列ならデータサイズの1乗)

★ 事後確率最大化解法の性能　(***F***がi.i.d.ランダム行列の場合)



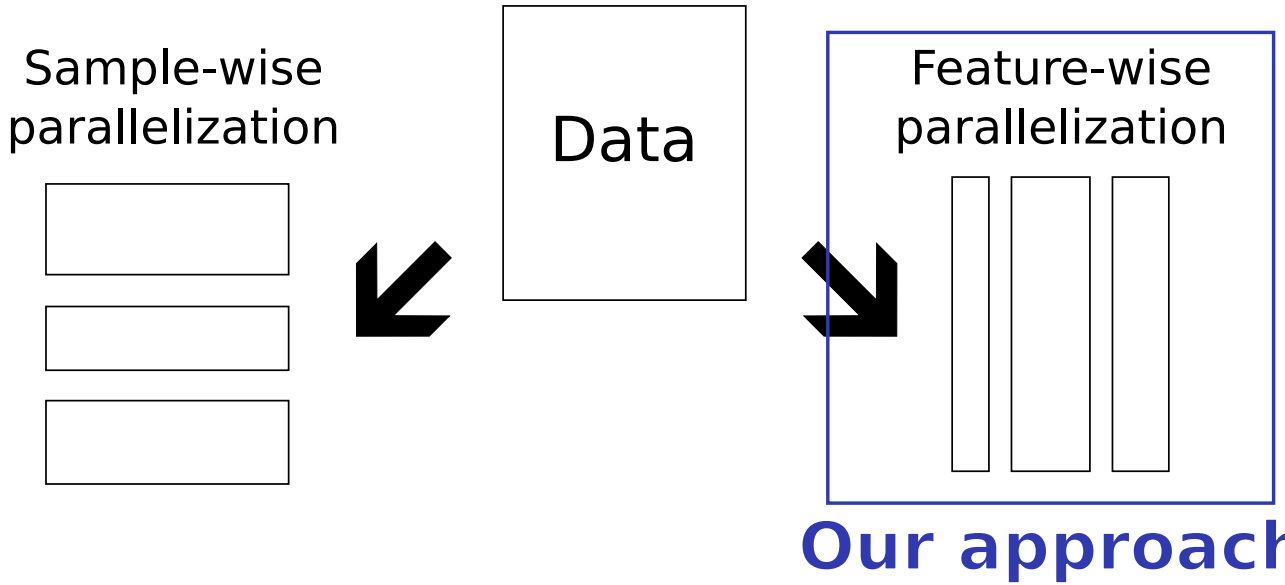
ℓ_1ノルム最小化と
ほぼ同じ性能が出る

★ 既知の確率伝搬法に基づくアルゴリズムとの関係も議論可能

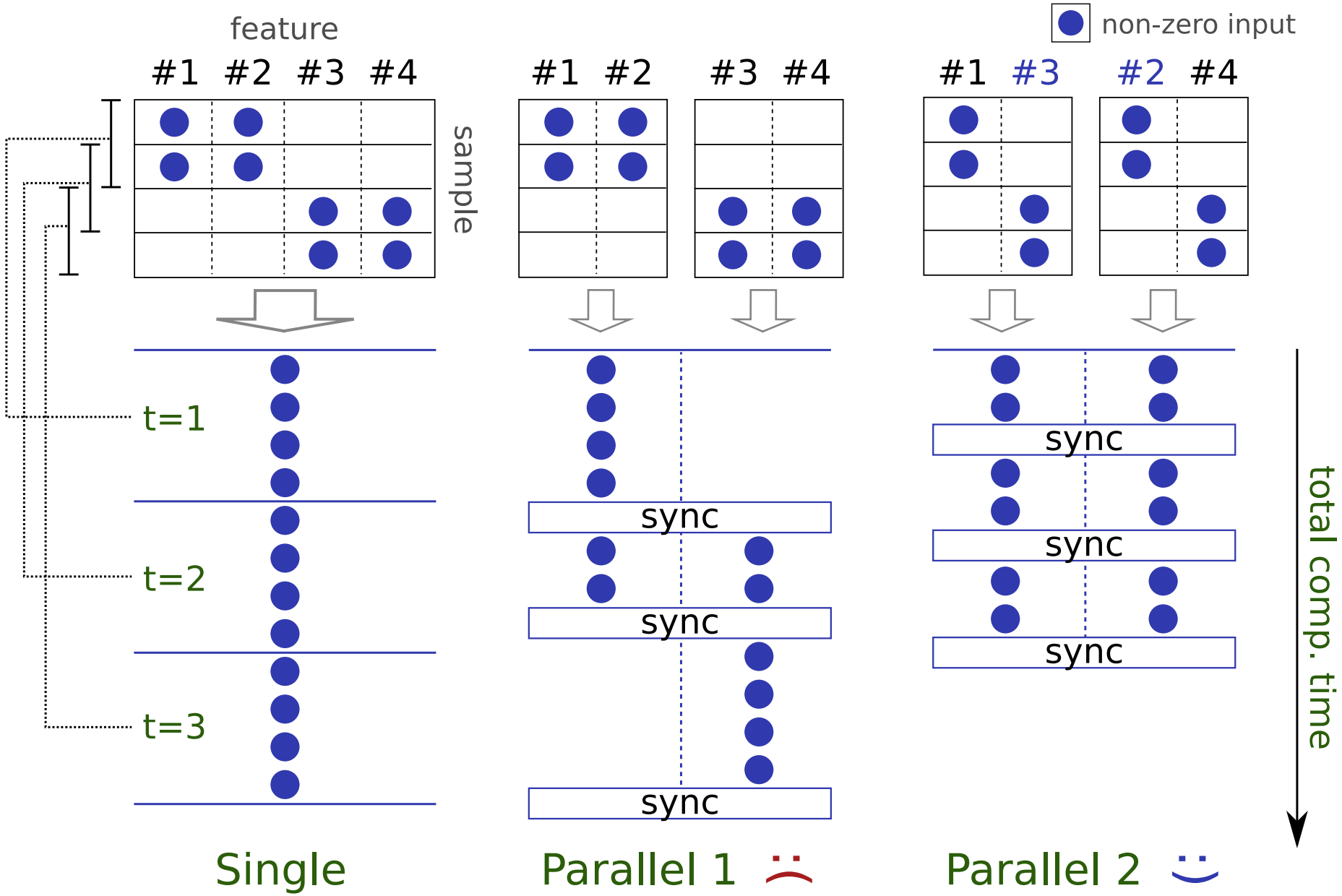
T45 疎データに対する並列SGDのための属性割当最適化

○林浩平 (東大)，藤巻遼平 (NECLA)

目的：SGDを並列化し高速計算



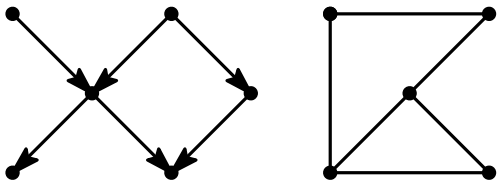
課題：計算時間が属性の割当て方に大きく依存

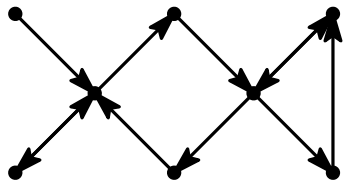


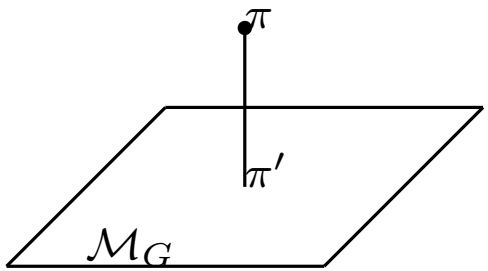
属性割当を最適化問題として定式化

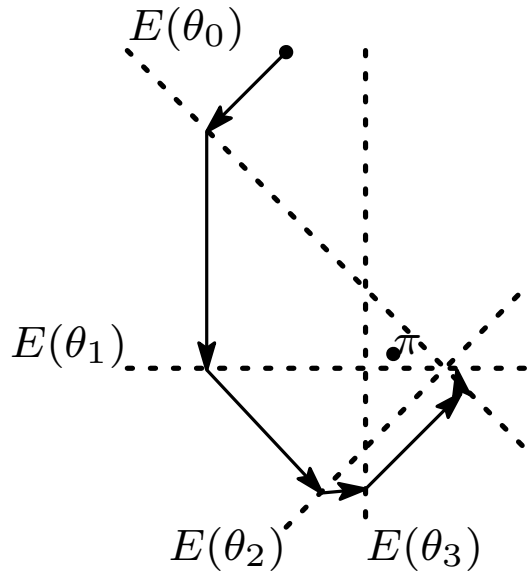
- ２つの近似的解法を理論的・実験的に比較

T-46 新グラフィカルモデル「発火過程ネットワーク」









従来のグラフィカルモデル　系全体で作られる単一のモデル多様体．学習は系全体に関わる重い処理．

提案モデル　ノード毎の多様体．学習はノード毎の軽い処理．

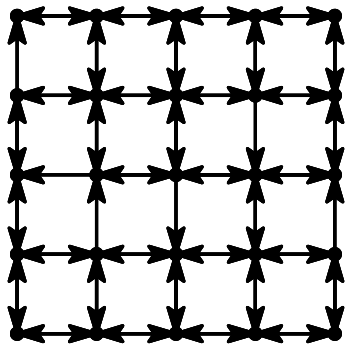
モデル分布の性質) データ分布 π と提案モデルが学習したモデル分布 π' 間の擬距離は π と各ノードの多様体間の KL ダイバージェンスにより抑えられる.

$$FCD(\pi||\pi') \leq \sum_i c(i) KL(\pi||E(\theta_i))$$

またデータ数無限大の極限では π' は π とともにデータが従う真の分布 π_{real} に概収束する.

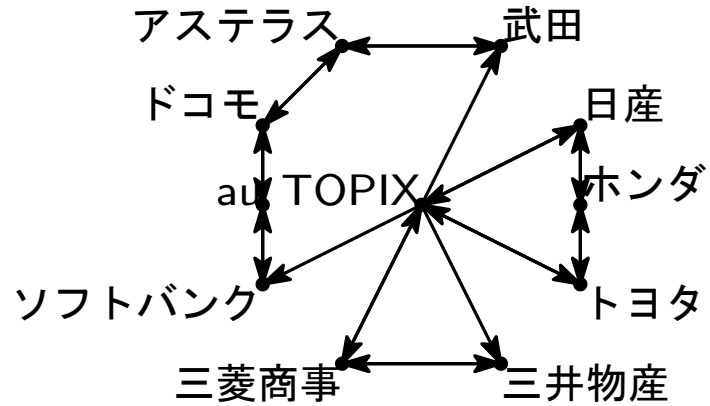
実験結果) ちゃんと動く.

$N = 1000$



5×5 イジングモデル学習結果

$N = 726$



11 銘柄株価陰陽学習結果

T-47　漸近展開による近似精度の予測可能性

野村亮介 (東京大学)，日野英逸 (早稲田大学)，村田　昇 (早稲田大学)，吉田朋広 (東京大学)

1．ヨーロピアン・プット・オプション

確率微分方程式:

$$\begin{cases} dX_t^\epsilon = V_0(X_t^\epsilon, \epsilon)dt + V(X_t^\epsilon, \epsilon)dw_t,\ t \in [0,T] \\ X_0^\epsilon = x_0 \end{cases}$$

となめらかな実数値関数 F に対して，期待値

$$E[\max(K - F(X_T^\epsilon), 0)]$$

をヨーロピアン・プット・オプションと呼ぶ．ここで，K は権利行使価格 (ストライク) と呼ばれる定数である．

漸近展開はこのような期待値を求める手法の一つであるが，その精度は厳密に評価はできない．そこで，漸近展開による近似値をモンテカルロ法による値と比較する．

2．数値実験

2 次漸近展開の値 AE とモンテカルロ法の結果 MC の誤差率を

$$|MC - AE|/|MC|$$

で定義し，この値が 1% を超えるかどうかをランダムフォレストを用いて判別する．

Black–Scholes (BS) モデル:

$$\begin{cases} dX_t^\epsilon = aX_t^\epsilon dt + \epsilon X_t^\epsilon dw_t,\ t \in [0,T] \\ X_0^\epsilon = x_0 \end{cases}$$

と，Constant Elasticity of Variance (CEV) モデル:

$$\begin{cases} dX_t^\epsilon = aX_t^\epsilon dt + \epsilon\sqrt{X_t^\epsilon} dw_t,\ t \in [0,T] \\ X_0^\epsilon = x_0 \end{cases}$$

に対して，ヨーロピアン・プット・オプション $E[\max(K - X_T^\epsilon, 0)]$ を求める問題を考える．

3．実験結果

特徴量として，モデルのパラメータに加えて漸近展開で得られるいくつかの値を用いた．また，それらの二次の交互作用 (各特徴量の積を新たな特徴量とみなす) を考慮した．

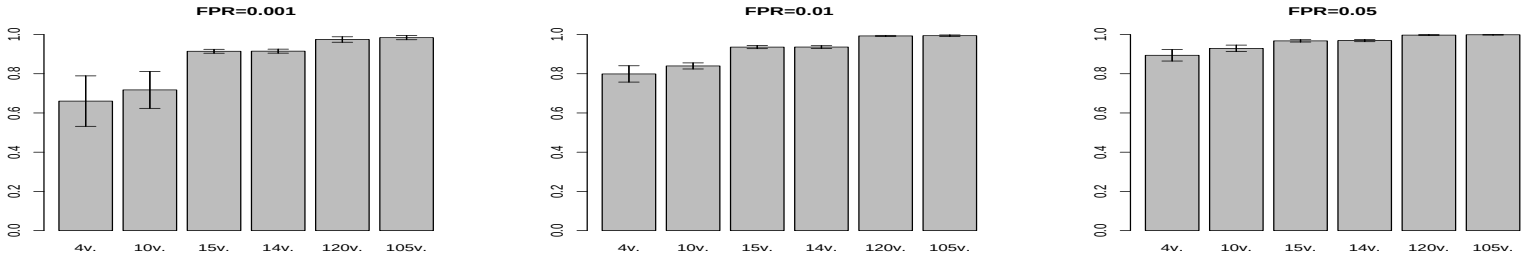


図1　BS モデルにおける各 FPR での TPR の比較

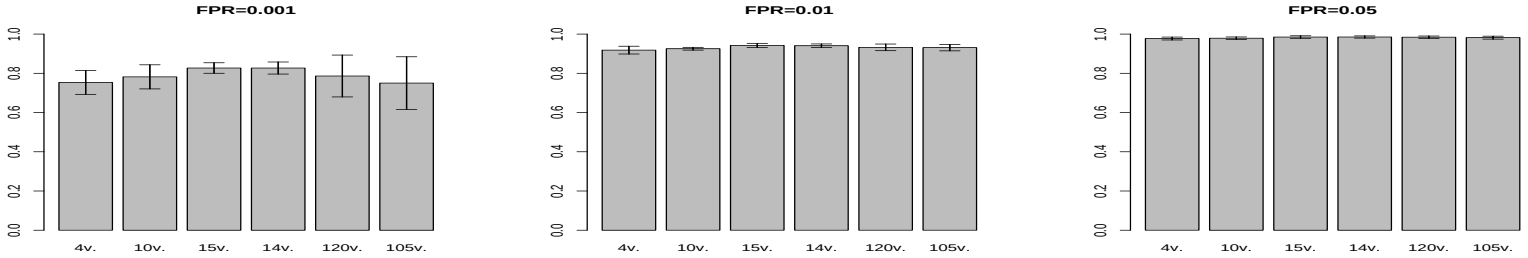


図2　CEV モデルにおける各 FPR での TPR の比較

FPR: 誤差率 1% 以上であるデータ (Negative) を誤って判断する割合
TPR: 誤差率 1% 未満であるデータ (Positive) を正しく判断する割合

収益率が相互に影響する銘柄群と複数の制約条件を持つポートフォリオ最適化問題の情報統計力学

新里隆@秋田県立大学

第15回情報論的学習理論ワークショップ
筑波大学東京キャンパス，発表番号T-48

先行研究（Cilibertiら，新里ら）

各銘柄の**収益率に独立同一分布**を仮定し，**期待収益率も0**と仮定し，**拘束条件が予算制約のみ**の**ポートフォリオ最適化問題**について，最適解の最小リスクと分散投資達成度の典型的な振る舞いを**情報統計力学的手法**を用いて解析してきた．しかしながら，実際の投資市場の特性である，

1.銘柄間に相関関係が含まれている
2.期待収益率が0でない
3.複数の線形不等式制約が課されている

におけるポートフォリオ最適化問題について十分に議論されているとは限らない．

研究目的

各銘柄の**(1)収益率に相関構造**を仮定し，**(2)期待収益率も0ではなく**，**(3)拘束条件が予算制約と複数の線形不等式制約**のある**ポートフォリオ最適化問題**について，最適解の典型的な振る舞いを**情報統計力学的手法**を用いて解析する．

得られた結果

1. 先行研究[1,2]で議論した結果を含むことが示された.
2. 先行研究[3]で提案された「今野・山崎予想」において,

平均分散モデル

$$\chi_u = \frac{\beta}{1+\beta\tilde{\chi}_u}$$

$$q_u = \frac{\beta^2\tilde{q}_u}{(1+\beta\tilde{\chi}_u)^2}$$

$\beta \to \infty$

今野・山崎予想

平均絶対偏差モデル

$$\chi_u \simeq \frac{1}{\tilde{\chi}_u}$$

$$q_u \simeq \frac{\tilde{q}_u}{\tilde{\chi}_u^2}$$

平均分散モデルと平均絶対偏差モデルの最適解が一致する

が得られた.

参考文献

[1]新里隆，安田宗樹：信学技報，Vol. 110, No. 265, pp. 257-264, (2010).

[2]新里隆，若井亮介，嶋崎善章，信学技報，Vol. 111, No. 96, pp. 125-130, (2011).

[3]H. Konno and H. Yamazaki: Man. Sci. Vol. 37, No. 5, pp. 519-531, (1991).

E-LEARNING

JICA Technical Cooperation Project for Research and Education Development on Information and Communication Technology (PREDICT) Phase 2 (2012-2015) INDONESIA - JAPAN

ITS Institut Teknologi Sepuluh Nopember

Kumamoto University

T-49

Usage of Winnowing Algorithm and WordNet for Recognizing the Possibilities of Paraphrasing Sentences in Academic Papers

Diana Purwitasari(1), Umi Laili Yuhana(1), Daniel Oranova Siahaan(1), Achmad Affandi(1), Yoshifumi Chisaki(2), Tsuyoshi Usagawa(2)

(1)Institut Teknologi Sepuluh Nopember, Surabaya-IndonesiA, (2) Kumamoto University, Japan

E-mail: (1)(diana, yuhana, daniel)@if.its.ac.id, affandi@ee.its.ac.id, (2) (chisaki, tuie)@cs.kumamoto-u.ac.jp

Background

With Moodle …
lecturer could share learning materials, student could send assigment reports
…to support ecocampus.
As issue of plagiarism is getting popular, Moodle does not have feature to help lecturer for checking the similarities of submitted reports.

Example: undergraduate students must send referenced papers and final report for their thesis; students might translate referenced texts and submit them as part of the report

Problem Statement

How to recognize the possibilities of cross-plagiarism?
(from English-Indonesian texts)

Note: reference texts are written in English and the student might …
- translate the texts into another language (Indonesian) OR
- make the (Indonesian) translation as paraphrased sentences

(Kent & Salim, 2009) used ALMOST similar procedures for cross-plagiarism between English-Malayu

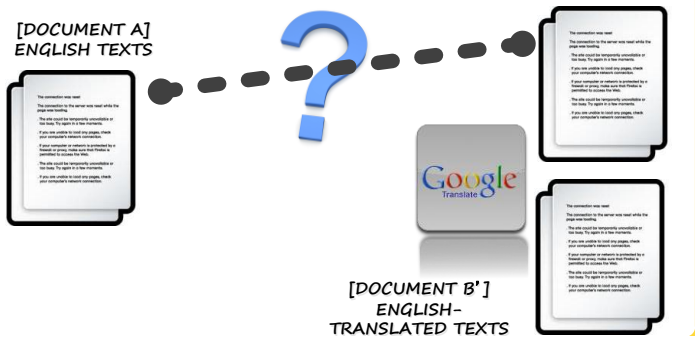
Solution

- recognize paraphrased sentences by considering the usage of words with similar meaning
(ADDITIONAL STEP from procedures of (Kent & Salim, 2009))

- compare fingerprint of documents
NOT
directly compare sequences of text

USED LIBRARY:
GOOGLE TRANSLATOR, WORDNET

Is Plagiarism Happened ?



Procedures

1. translate Indonesian Texts of Document B into English Texts of Document B' using Google Translator

Based on WordNet Library,
2. find pairs of sentences that semantically similar from A and B'
3. replace words inside each pair of sentences from A-B' with their hypernim words
… use semantic similarity value of words from WordNet to identify paraphrased sentences

Based on Winnowing Algorithm,
4. create fingerprints of document A and document B'
5. check similarity between fingerprints
… ASCII values of n-grams from texts become document features

E-LEARNING

JICA Technical Cooperation Project for Research and Education Development on Information and Communication Technology (PREDICT) Phase 2 (2012-2015) INDONESIA - JAPAN

ITS Institut Teknologi Sepuluh Nopember

Kumamoto University

T-49

Usage of Winnowing Algorithm and WordNet for Recognizing the Possibilities of Paraphrasing Sentences in Academic Papers

Diana Purwitasari(1), Umi Laili Yuhana(1), Daniel Oranova Siahaan(1), Achmad Affandi(1), Yoshifumi Chisaki(2), Tsuyoshi Usagawa(2)

(1)Institut Teknologi Sepuluh Nopember, Surabaya-IndonesiA, (2) Kumamoto University, Japan

E-mail: (1)(diana, yuhana, daniel)@if.its.ac.id, affandi@ee.its.ac.id, (2) (chisaki, tuie)@cs.kumamoto-u.ac.jp

Check Paraphased Texts

Princeton University WordNet A lexical database for English

[DOCUMENT A]

··· We propose several heuristics to predict change propagation. We present a framework to measure the performance of our proposed heuristics. ...

[DOCUMENT B]-INDONESIAN

··· Kami mengusulkan beberapa cara heuristik untuk memperkirakan penyebaran perubahan, dan juga menyajikan suatu kerangka kerja untuk mengukur kinerja cara heuristik. ...

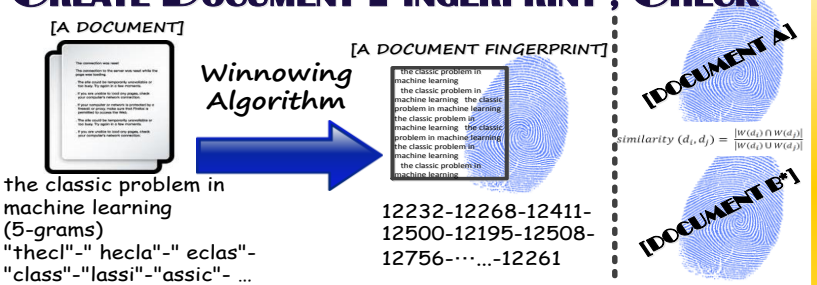
[DOCUMENT B']-ENGLISH

··· We propose a heuristic method to estimate the spread of the change, and also presents a framework for measuring the performance of a heuristic way. ...

[DOCUMENT B*]

··· [some words in Document B' have been changed with their hypernim words] ...

Create Document Fingerprint , Check



Experiment Results

1. Using 2 datasets: Set A (English Texts) and Set B (Indonesian Texts). Each set contains 20 undergraduate thesis abstracts with 200-300 words.
The result is to set parameter for recognizing similarity of fingerprints:
(i) window length = 10-grams; threshold for (ii) semantic similarity of sentences = 0.87 and (iii) fingerprint similarity of documents = 0.89

2. Based on 10 abstracts of computer science journals and 5 abstracts of non-computer science journals as Set A (English ver.), our experts create Set B (Indonesian ver.)
The result is our system could not recognize some abstracts that have paraphrased sentences (fingerprint similarity = 0.27) because there are significant changes on their structures and terms

3. Using the same dataset from previous experiment, we compared results with procedures of (Kent & Salim, 2009) based on Kappa Index (a kind of agreement between two experts: the system and our experts).
Kappa Index for (Kent & Salim, 2009) = 0.55 while Kappa Index for our system = 0.92.
Therefore our proposed procedures still can identify plagiarism with paraphrased sentences to some extent (no significant changes on structures and terms).

References

Myers, E. W. "An O(ND) Difference Algorithm and Its Variations", Algorithmica 1(2): 251-266, 1986.

Schleimer, S., Wilkerson, D.S., & Aiken, A. " Winnowing: Local Algorithms for Document Fingerprinting", Proc. of the 2003 ACM SIGMOD Intl. Conf. on Management of Data, 76-85, 2003

Kent, C., & Salim, N. "Web Based Cross Language Plagiarism Detection", Journal of Computing, 1(1): 39-43, 2009.

Parapar, J. & Barreiro, A. "Evaluation of Text Clustering Algorithms with N-Gram based Document Fingerprints", Proc. of the 31st European Conf. on Information Retrieval Research (ECIR), 645-653, 2009

Barron-Cedeno, A. "On the Mono-and Cross-Language Detection of Text Reuse and Plagiarism", Proc. of the 33rd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 914, 2010.

Pereira, R.C., Moreira, V.P. & Galante, R. " A New Approach for Cross-Language Plagiarism Analysis", Multilingual and Multimodal Information Access Evaluation, Lecture Notes in Computer Science 6360: 15-26, 2010.

Purwitasari, D., Priantara, W.S., Kusmawan, P.Y., Yuhana, U.L., & Siahaan, D.O. "The Use of Hartigan Index for Initializing K-Means++ in Detecting Similar Texts of Clustered Documents as a Plagiarism Indicator", Asian Journal of Information Technology, 10(8-12): 341-347, 2011.

T-50

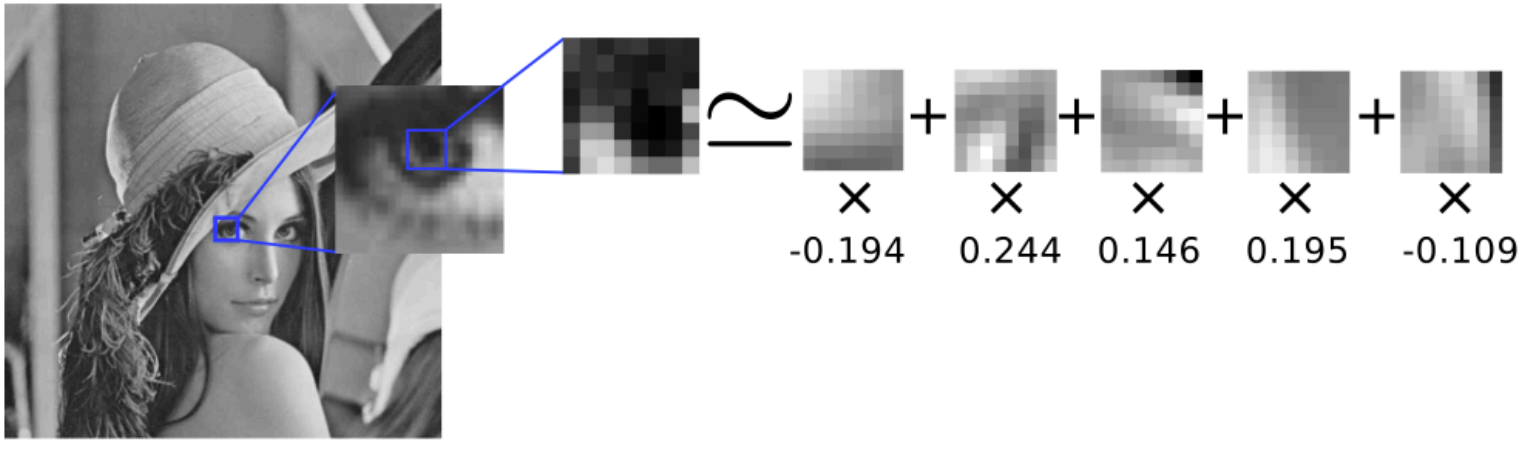
スパースコーディングにおける基底行列生成のための単一母基底の学習

有竹　俊光，日野　英逸，村田　昇（早稲田大学）

スパースコーディングは少数の基底で信号を表現する枠組み

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \text{ subject to } \|\mathbf{x}\|_0 \leq k_0$$

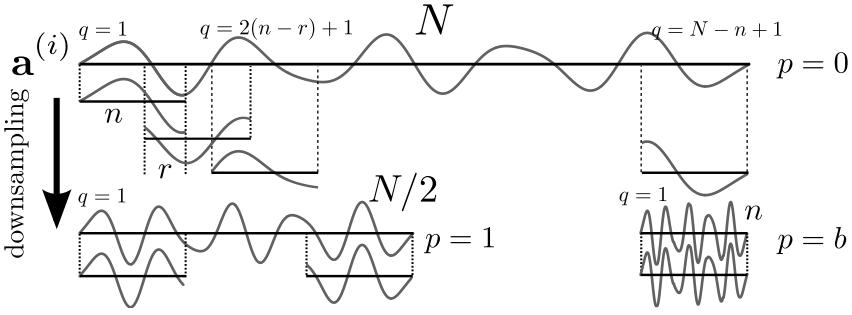
y：観測信号，**D**：基底行列，**x**：係数ベクトル



基底行列を学習により獲得

提案手法

**各基底が母基底と呼ばれるベクトルから窓の移動と
ダウンサンプリングを繰り返して生成されると仮定．(下図)**

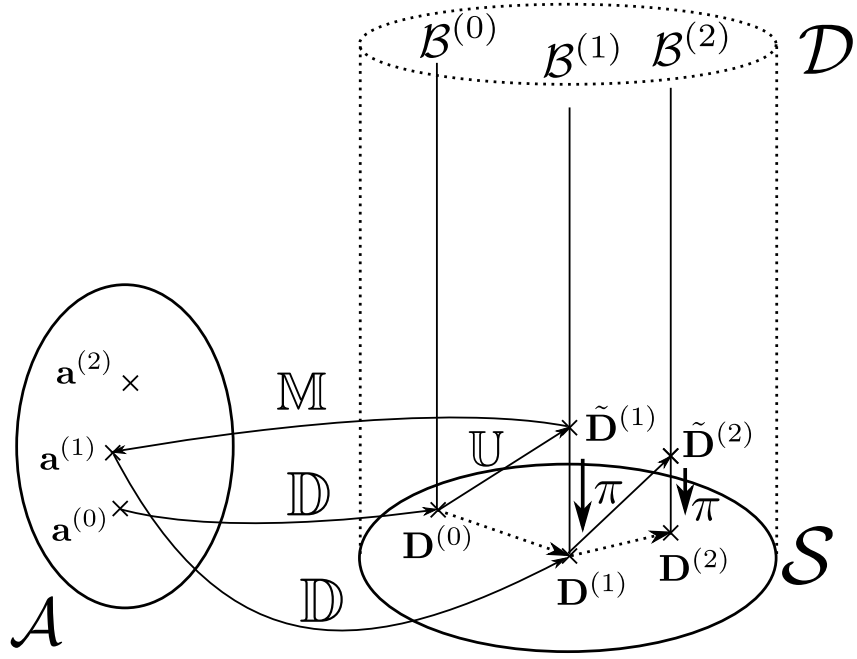


母基底は

- **基底行列の生成，**
- **基底行列の更新，**
- **基底行列の平均化**

**の繰り返しで更新される．
また，直感的な幾何的解釈を
与えることができる．**

アルゴリズムの幾何的解釈



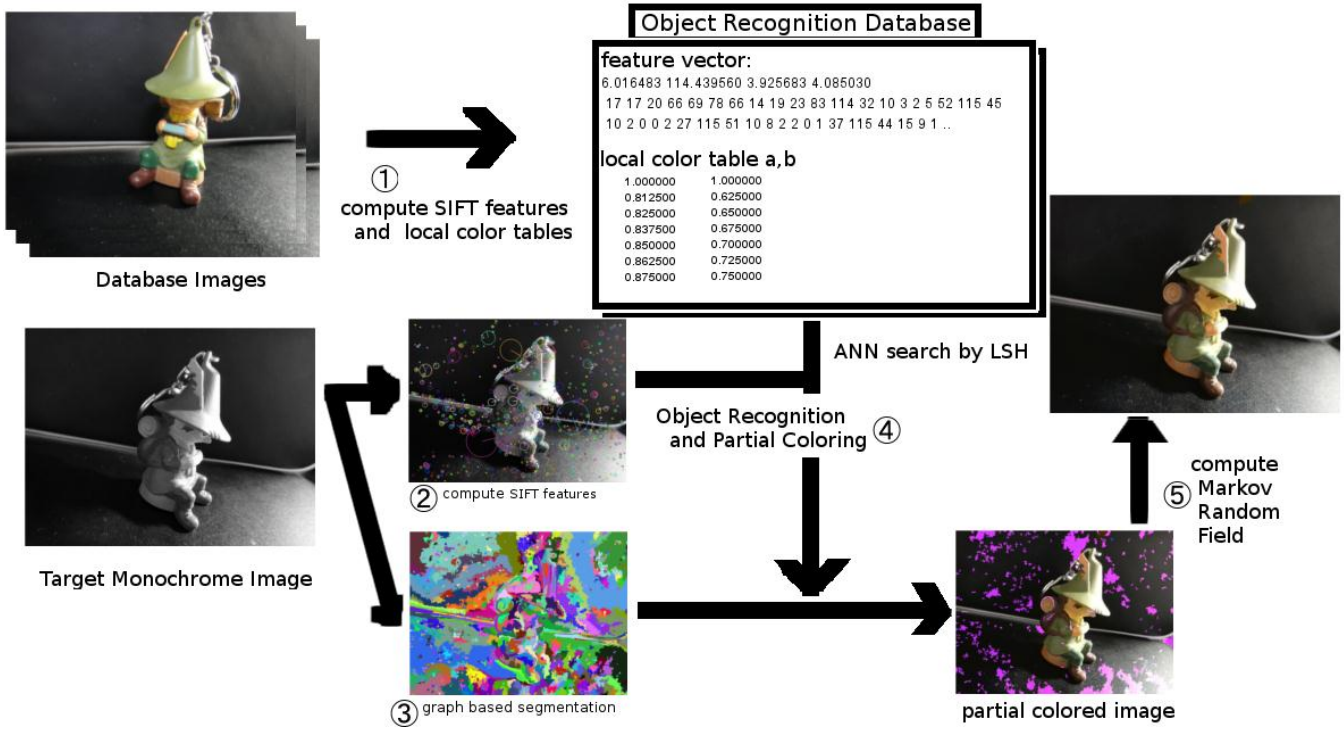
母基底からの基底行列の生成 $\mathbb{D}$，基底行列の更新 $\mathbb{U}$，母基底の更新 $\mathbb{M}$ の繰り返しで，母基底の空間 $\mathcal{A}$，構造化した基底の空間 $\mathcal{S}$，基底行列全体の空間 $\mathcal{D}$ を遷移しながら更新を行う．

T-51 特定物体認識とマルコフ確率場によるモノクロ画像の自動彩色法

松尾恒（東京大学教養課程）、牧野貴樹（東京大学生産技術研究所）

提案手法

入力のモノクロ画像について、参照画像を用いた特定物体認識の結果を考慮してマルコフ確率場による領域単位の色成分推定を行うことで自動的に彩色する



手法の概略

①参照画像のSIFT特徴量と周辺の色情報の組を抽出してデータベースに格納 [事前処理]

②彩色対象のモノクロ画像のSIFT特徴点を検出

③対象画像の領域分割

④参照画像の絞込み＋対応点検出で、領域に色成分と信頼度を設定

⑤マルコフ確率場で色の補完と補正を行い合成する

T-51 特定物体認識とマルコフ確率場によるモノクロ画像の自動彩色法

松尾恒（東京大学教養課程）、牧野貴樹（東京大学生産技術研究所）

色の補完、補正のマルコフ確率場モデル

$$argmin_{C_{r_{1..,n}}} \sum_i \left\{ k_2 \left| C_{r_i} - \overline{C_{r_i}} \right| * conf(i) + \left[\sum_{j \in n(i)} \frac{k_1 \left| C_{r_i} - C_{r_j} \right|}{Dif(i,j) * Dif(i,j)} \right] \right\}$$

subject to $0 \leq C_{r_i} \leq 255 \ \ (1 \leq i \leq n)$

C_{r_i} 領域 r_i の色成分

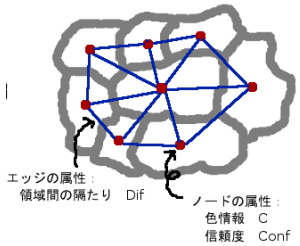
$\overline{C_{r_i}}$ 物体認識で推定された領域 r_i の色成分

$conf(i)$ 領域 r_i の色成分の推定の信頼度

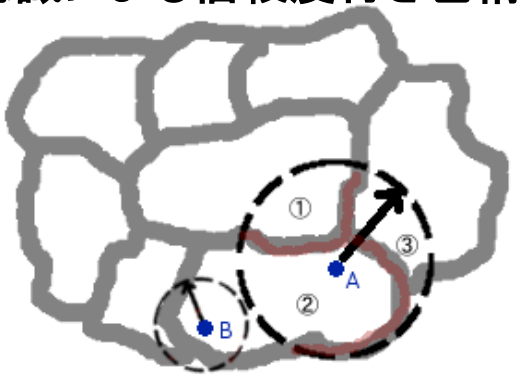
$n(i)$ 領域 r_i に近接する領域の集合

$Dif(i,j)$領域 r_i と領域 r_j の境界上でのモノクロ成分の最大の隔たり

k_1, k_2 パラメーター



物体認識による信頼度付き色情報推定



入力画像の特徴点周りのヒストグラムと
参照画像で対応する特徴点周りの
ヒストグラムをマッチさせることにより
領域単位の色成分推定を行う

適用後

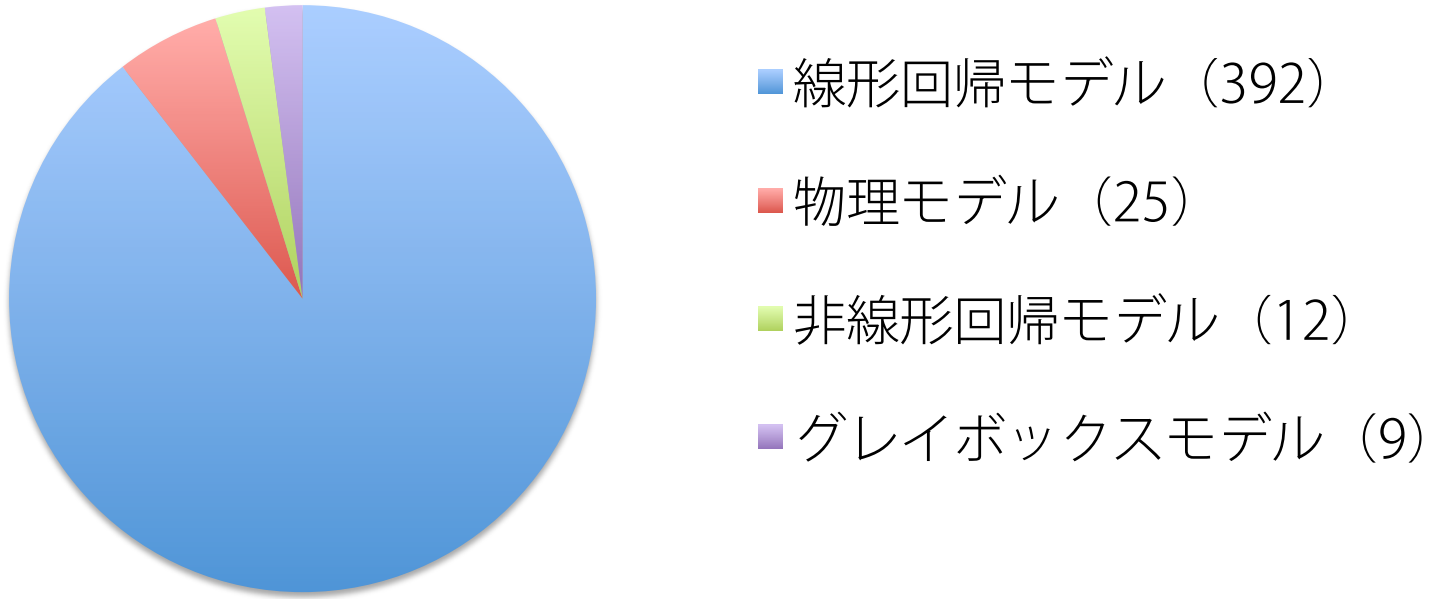
まとめ

- 高速かつ領域に忠実で物体認識を考慮した彩色が可能
- 同一カテゴリーの画像を参照できないと良好な画像を得られない

T-52

線形回帰モデルにおけるNCスペクトラルクラスタリングを用いた入力変数選択

京都大学 藤原幸一



産業界における数理モデルの利用（JSPS143委員会調査）

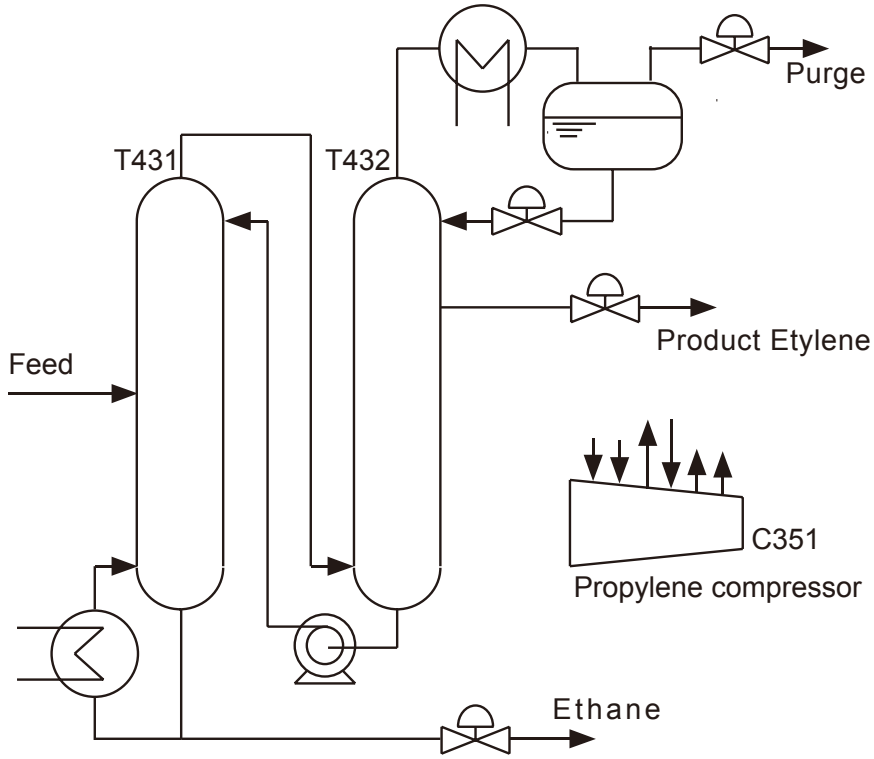
生産性向上のために線形回帰モデルの効率的構築が必要

入力変数のクラスタリング・選択

1. 入力変数を相関関係に従ってクラスタリング

2. 出力予測の寄与の降順に，変数グループを選択

3. 選択された変数を用いて，線形回帰モデルを構築



	RMSE	R
PLS	28.7	0.88
物理知識	23.8	0.88
Lasso	28.1	0.88
Stepwise	26.6	0.89
提案法	18.0	0.92

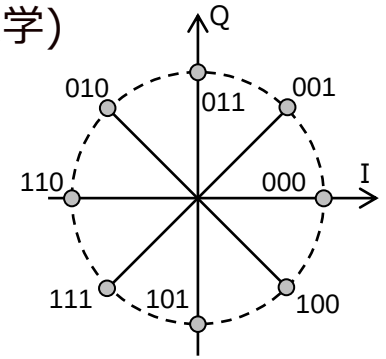
昭和電工エチレン精留塔のエタン濃度予測モデル

T-53　M元位相偏移変調によるCDMA通信のレプリカ解析

加藤 弘之
(関西大学大学院)

岡田 真人
(東京大学，理研)

三好 誠司
(関西大学)



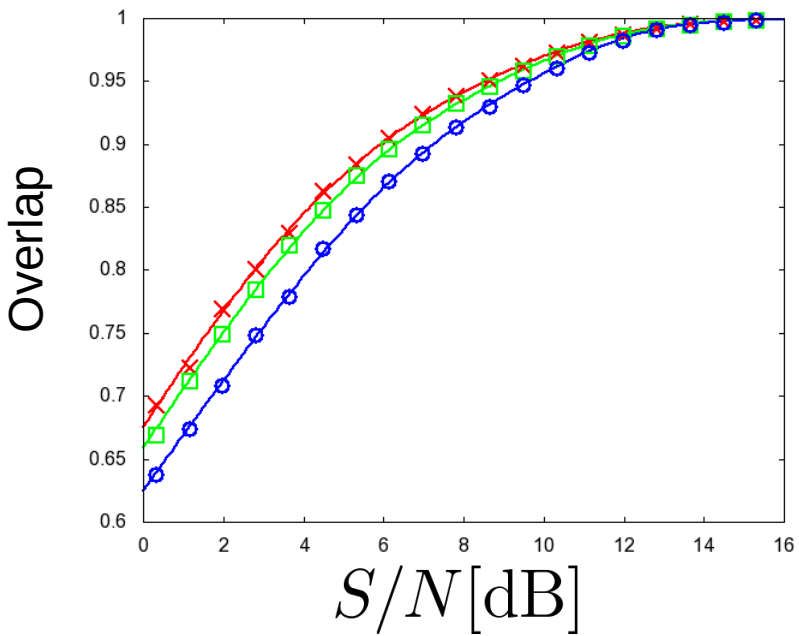
M元位相偏移変調
(M=8)

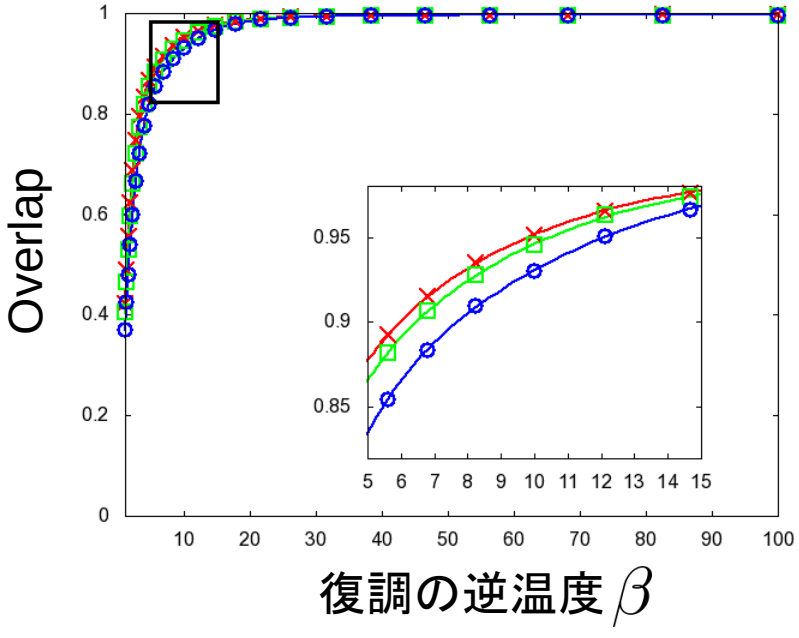
・Q状態連想記憶モデルのレプリカ法による解析　(J.Cook, 1989)

・BPSK/CDMAのレプリカ法による解析　(T.Tanaka, 2002)

→M元位相偏移変調によるCDMA通信のレプリカ解析

RS解





T-54

クラスタリングに基づく局所的半教師付きガウス過程回帰

カクシンロ（神戸大） 安村 禎明（芝浦工大） 上原 邦昭（神戸大）

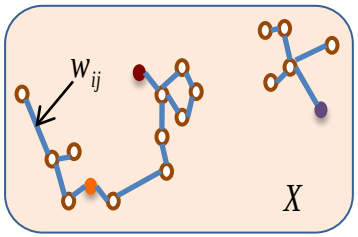
背景

- 回帰とは，既存の訓練データに基づいて，新たな入力に対する出力値を予測することである
- グラフを用いた半教師付き手法が分類問題に適用
- 半教師付き回帰に関する研究が注目

目的

- ラベルなしデータを利用するために隣接グラフを定義
- 隣接グラフの情報をガウス過程の枠組みに組み入れ半教師付き回帰で用いる分布を導く
- ガウス過程の計算効率問題を解決するために，クラスタリングの枠組みを導入

提案手法



$$w_{ij} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|^2}{2\delta}\right) & \text{隣接} \\ 0 & \text{非隣接} \end{cases}$$

$$D_{ii} = \sum_j w_{ij}$$

$$L = D - W$$

$$\Rightarrow \quad \Delta = \lambda \cdot L^{\upsilon}$$

ラベルなしデータから学習するために隣接グラフを利用

目的関数：

$$p(y \mid G) = \mathrm{N}(0, \Sigma)$$

$$p(f_* \mid x_*, X, y, G) = \mathrm{N}(\hat{\mu}, C)$$

勾配下降法によって最適パラメータを求める

なお，出力の予測平均値と標準偏差

$$\hat{\mu} = \tilde{k}_*^T \Sigma^{-1} y$$

$$C = \tilde{k}_{**} - \tilde{k}_*^T \Sigma^{-1} \tilde{k}_*$$

ここで，

$$\Sigma = \left(K^{-1} + \Delta\right)^{-1} + \sigma^2 I$$

グラフの隣接情報を目的関数に組み入れた

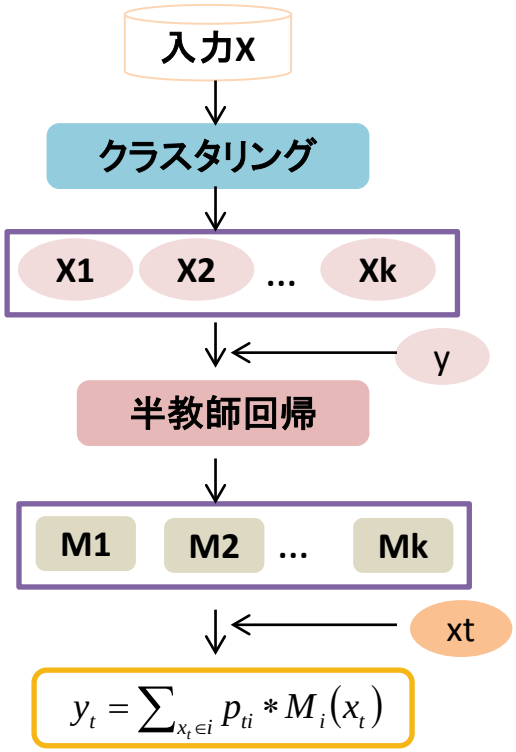
$$k_{ij} = c \cdot \exp\left(-\sum_{d=1}^{D} b_d (x_i^d - x_j^d)/2\right)$$

$$\tilde{k}_{ij} = k_{ij} - k_i^T (I + \Delta K)^{-1} \Delta\, k_j$$

T-54 クラスタリングに基づく局所的半教師付きガウス過程回帰

カクシンロ（神戸大） 安村 禎明（芝浦工大） 上原 邦昭（神戸大）

クラスタリング



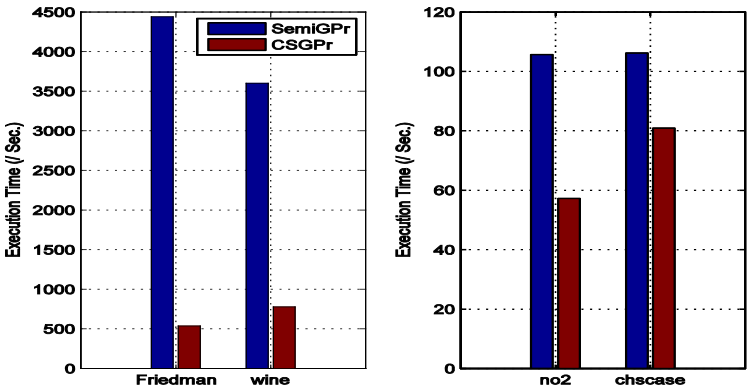
まとめ

数値実験

提案手法の回帰誤差

Dataset	GPr	SemiGPr	Improv.
Friedman	0.0113	0.0101	10.62%
	0.0114	0.0102	10.53%
Wine	0.0196	0.0190	**3.06%**
	0.0205	0.0199	**2.93%**
chscase	0.0273	0.0264	3.30%
	0.0268	0.0265	1.12%
no2	0.0180	0.0161	**10.56%**
	0.0183	0.0164	**10.38%**
kin8nm	0.0136	0.0131	**3.68%**
	0.0134	0.0132	1.49%
bodyfat	0.0026	0.0026	0%
	0.0061	0.0027	55.74%
pyrim	0.0524	0.0359	**31.49%**
	0.0544	0.0495	9.01%
triazines	0.1215	0.0843	**30.62%**
	0.1205	0.0925	**23.24%**

クラスタリングを導入した場合の実行時間



クラスタリングを導入した場合の回帰誤差

Dataset	Friedman	Wine	chscase	no2
SemiGPr	0.0101	0.0190	0.0264	0.0161
	0.0102	0.0199	0.0265	0.0164
CSGPr	0.0106	0.0220	0.0256	0.0154
	0.0107	0.0232	0.0261	0.0160

- ラベルなしデータには，モデルの訓練に役に立つ情報が含まれ，この情報を有効に利用している
- 従来手法よりも上か同等であり，多くのデータセットにおいて，精度が改善（平均 12.98%）
- クラスタリングの枠組みを導入すれば，計算時間を最低2/3に短縮，精度向上する可能性もある

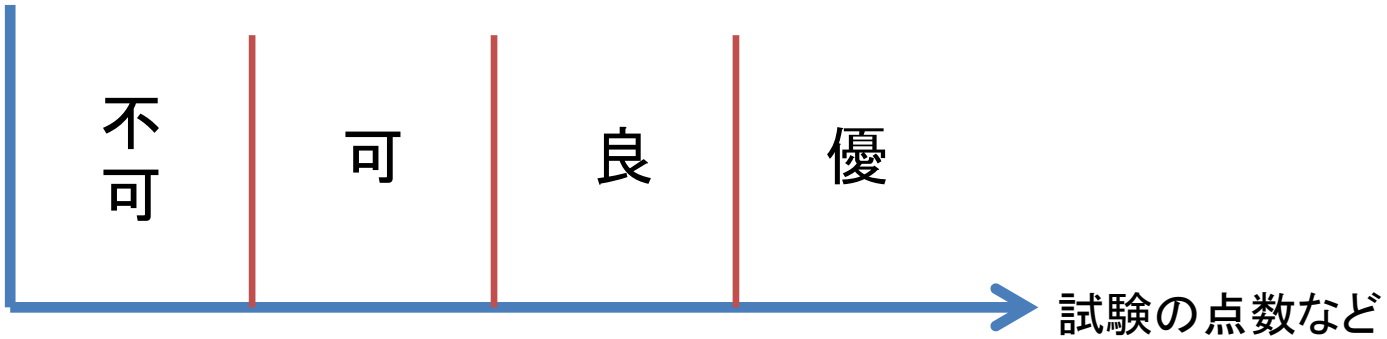
T-55

ロジスティック回帰モデルを組み合わせた順序回帰モデルと高速な疎Bayes学習

長島主尚，井上真郷
（早稲田大学）

- 順序データとは
  - 出力ラベル間に既知の順序関係があるデータ

  例）レビューのランク，学校の成績，出世魚など



- 順序回帰問題とは
  - 順序データのマイニングや予測を行う問題設定

T-55

ロジスティック回帰モデルを組み合わせた順序回帰モデルと高速な疎Bayes学習

長島主尚，井上真郷
（早稲田大学）

Automatic Relevance Determination（ARD）との相性

- 既存モデル
  - $P(\mathbf{t}|\mathbf{X}) \equiv \prod_{n=1}^{N}(f_1(x_n) - f_2(x_n))$
  - 解析が困難
    - ARD事前分布の導入は可能だが、高速な解法は得られない

- 提案モデル
  - $P(\mathbf{t}|\mathbf{X}) \equiv \prod_{n=1}^{N} g_1(x_n)g_2(x_n)$
  - 解析が容易
    - ARD事前分布の導入と簡単な近似により、高速に解ける

T-56 経路アノテーションからの学習による無線LAN位置推定の簡易な構築

川尻亮真，下坂正倫，福井類，佐藤知正 (東大)

背景: 無線LAN位置推定のための訓練データ収集コストが高い

- 機械学習を用いて実測したデータから推定する研究が主流

提案: 経路アノテーションを用いた無線LAN位置推定の簡易な構築

貢献:

- 経路アノテーションによる簡易な訓練データを収集手法を示した
- 経路アノテーションのための機械学習をSelf-Trainingとして定式化

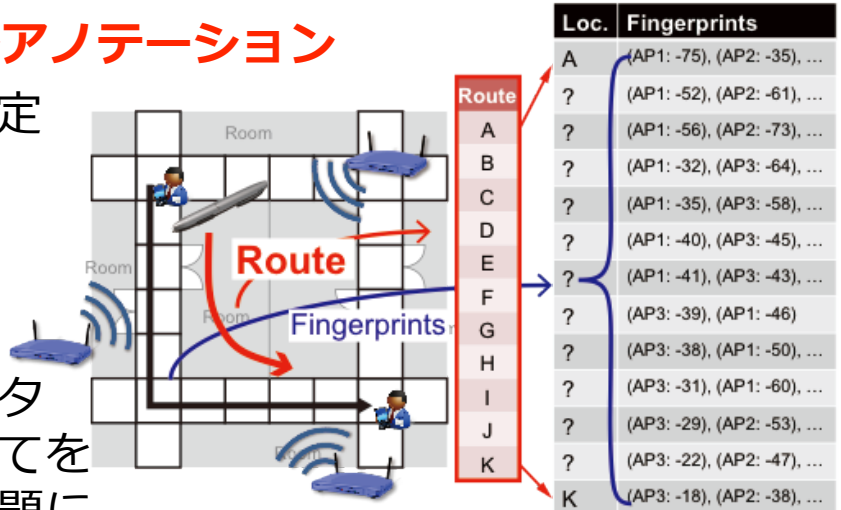
従来手法

全ての電波情報に位置情報を付与
時間がかかる

提案手法: 経路アノテーション

・移動しながら電波測定
・ときどき移動した
　経路を描くだけ

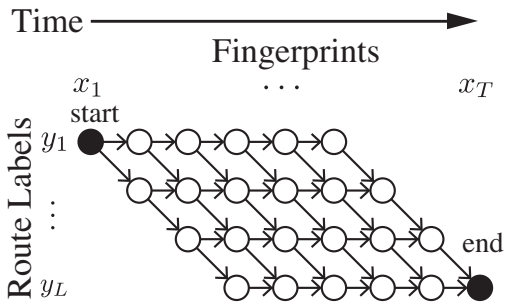
・短時間に多量のデータ
・経路ラベルの割り当てを
　推定・学習する問題に



隠れ変数を持つ構造推定の学習として定式化

1. 隠れラベル推定

経路制約の下で
高速かつ安定に解ける



繰り返し

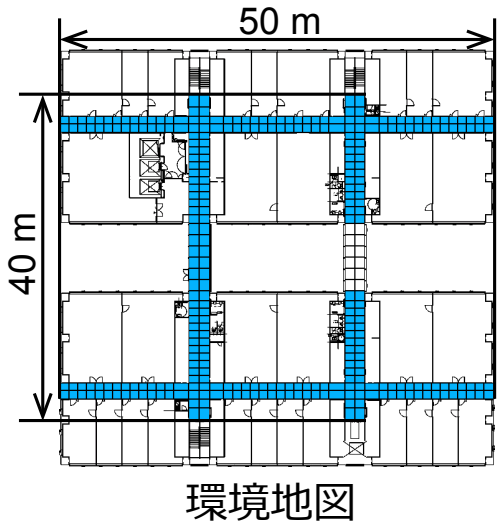
2. パラメータ更新

距離誤差を組み込んだ
コスト考慮型最大マージン法

$$\xi(y,x) = \max\left[0, \max_{\hat{y}\in\mathcal{Y}\setminus y}\{\Delta(y,\hat{y})\,(1 - \boldsymbol{w}^{\intercal}(\phi(y,x) - \phi(\hat{y},x)))\}\right]$$

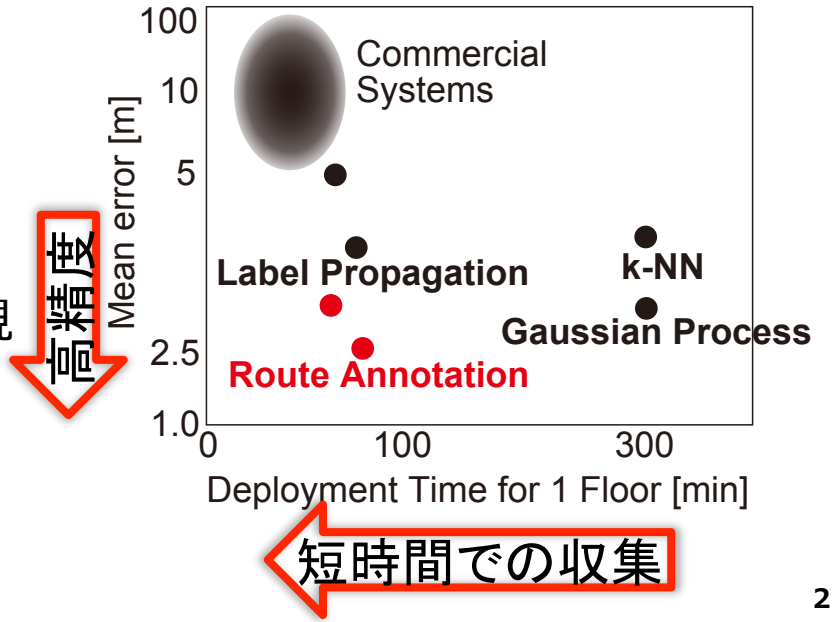
損失　　距離誤差　　マージン

実験



環境地図

提案手法により，
短い時間で収集した
データにおいても
高精度での推定を実現

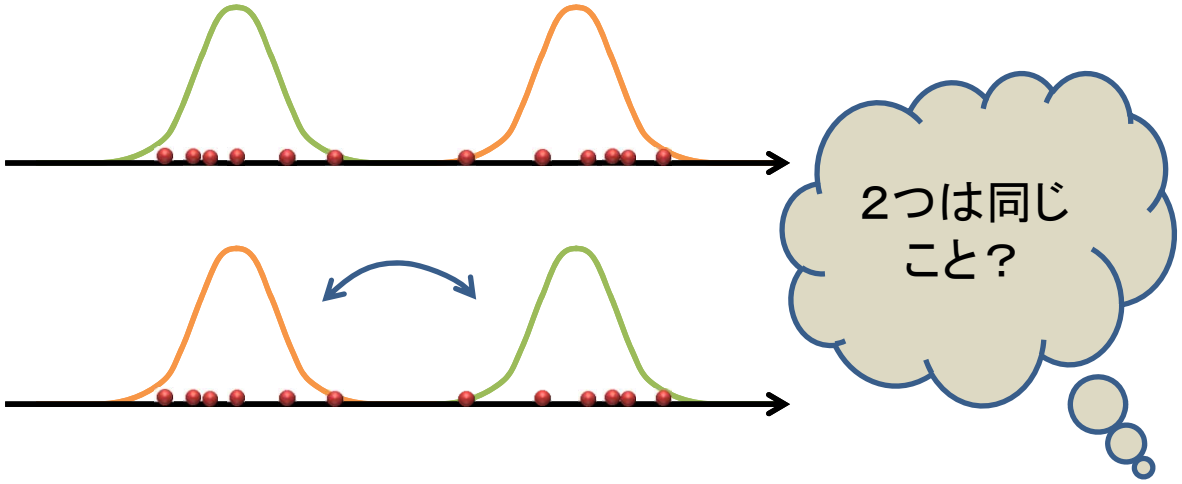


T-57 解の多重度を考慮した混合分布モデルはより良い解をもたらす

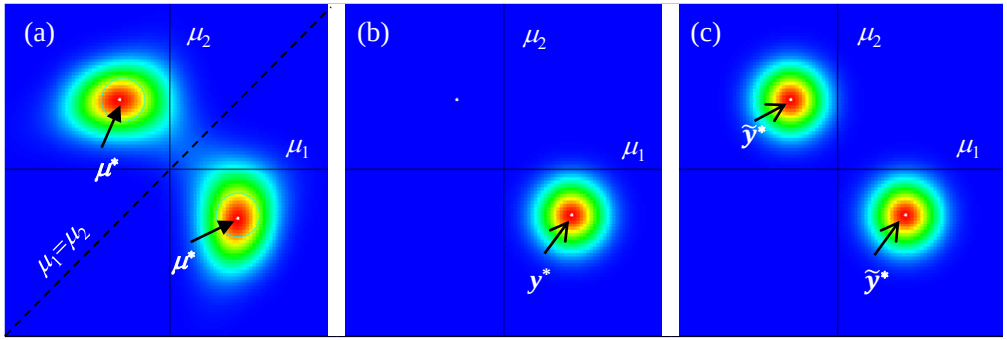
古川徹生（九州工業大学）

何が問題か

ラベルを付け替えたものは
本当に同じと思って良いのか？



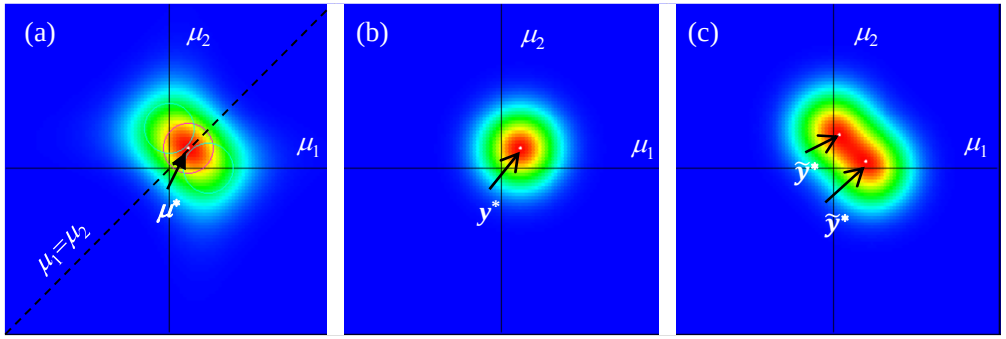
ケース1



真の事後分布　　変分ベイズ　　多重度考慮変分ベイズ

多重度を考慮しなくても推定結果は同じ

ケース2



真の事後分布　　変分ベイズ　　多重度考慮変分ベイズ

多重度を考慮すると推定結果が変わる

T-57 解の多重度を考慮した混合分布モデルはより良い解をもたらす

古川徹生（九州工業大学）

提案手法

ラベル置換を表現する潜在変数πを追加

⇒ 逆温度パラメータβを見かけ上大きくすることに相当

通常の責任率

$$R_{nk} = Ce^{-\beta E_{nk}}$$

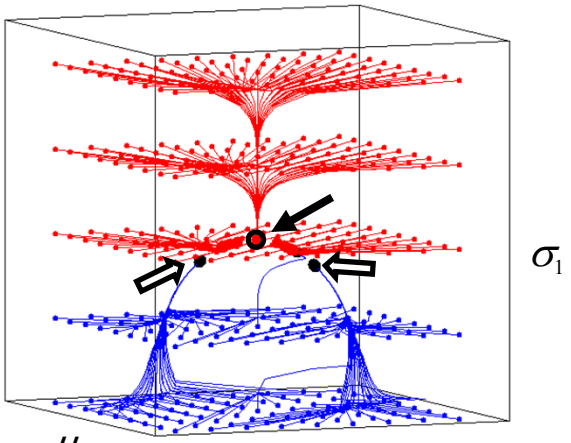
多重度考慮型責任率

$$\hat{R}_{nk} = \frac{C}{G_k} e^{-\beta \kappa_{nk} E_{nk}}$$

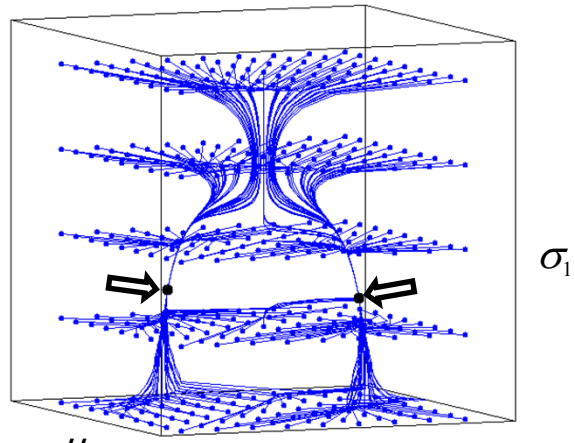
$\kappa_{nk} = 2 - R_{nk}$

数値実験

さまざまな初期値から計算.
局所解（赤）に陥るケースが
なくなった.



従来法

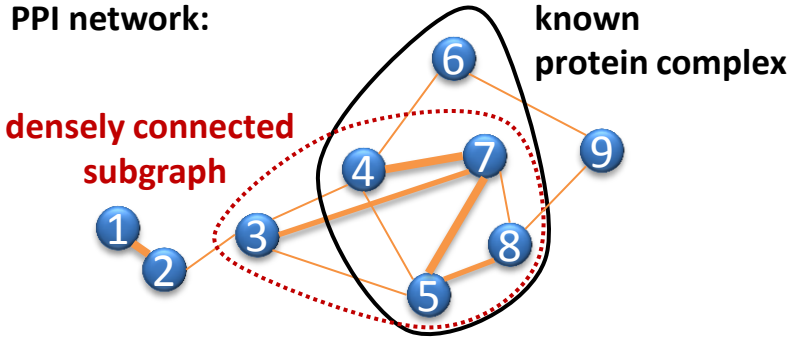


提案法

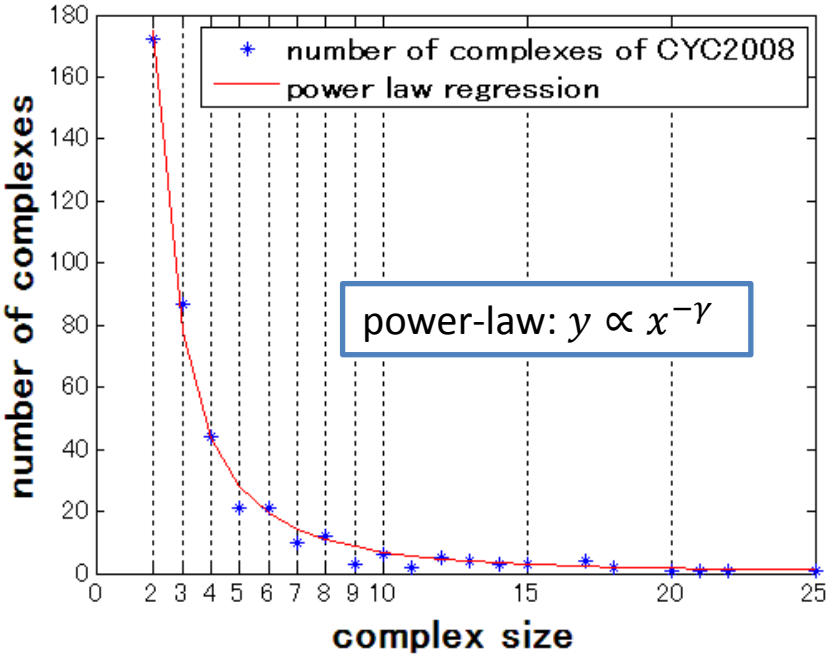
MCMC Strategy for Protein Complex Prediction Using Cluster Size Frequency

Daisuke Tatsuke and Osamu Maruyama (Kyushu Univ.)

We propose a sampling method, called **PPSampler**, for predicting protein complexes using the following two prior knowledge:



Well-known observation: Dense subgraphs and known complexes often overlap each other.



New observation: The distribution of the frequency of protein complexes w.r.t. complex size is power-law.

scoring subfunction $f_1(C)$

$f_1(C) \equiv \sum_{d \in C} f_1(d)$ where C is a partition of a given set of proteins.

$w(u,v)$: weight of the interaction between proteins u and v.
N: upper bound on the size of a cluster

$$f_1(d) \equiv \begin{cases} 0 & if\ |d| = 1 \\ -\infty & else\ if\ |d| > N\ or\ \exists u \in d, \forall v(\neq u) \in d, w(u,v) = 0 \\ \sum_{u,v(\neq u) \in d} w(u,v) & otherwise \end{cases}$$

MCMC Strategy for Protein Complex Prediction Using Cluster Size Frequency

Daisuke Tatsuke and Osamu Maruyama (Kyushu Univ.)

scoring subfunction $f_2(C)$

$\psi_C(i)$: relative frequency of clusters of size i in C
$\psi(i)$: predefined target number for relative frequency for size i

$$f_2(C) \equiv \prod_{i=2}^{N} \frac{1}{1+i^2\cdot\left(\psi(i)-\psi_C(i)\right)^2}$$

scoring subfunction $f_3(C)$

$s(C)$: number of proteins within clusters of size 2 or more in C
λ : predefined target number for $s(C)$

$$f_3(C) \equiv \frac{1}{1+\frac{(s(C)-\lambda)^2}{10^3}}$$

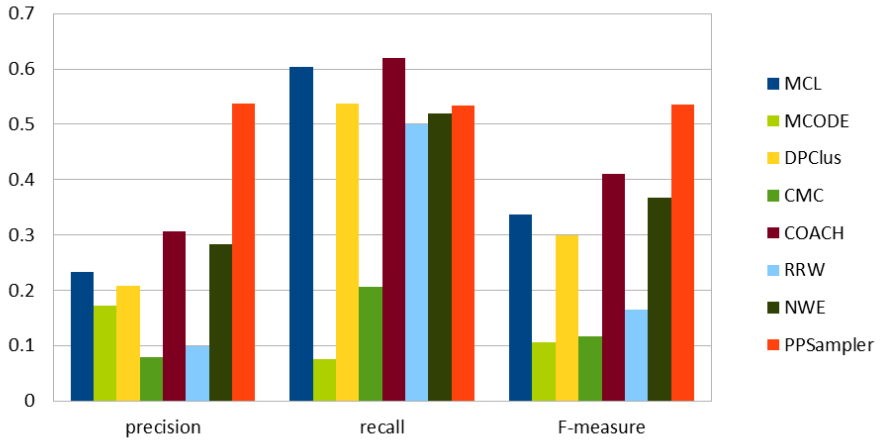
scoring function and its probability

$$f(C) \equiv -f_1(C)\cdot f_2(C)\cdot f_3(C)$$

$$P(C) \propto \exp\left(-\frac{f(C)}{T}\right)$$

Partitions of proteins are generated as samples from P(C) by the Metropolis-Hastings algorithm.

Results:



- Precision of PPSampler is 70% higher than the second one.
- F-measure of PPSampler is 30% higher than the second one.

Concluding Remarks:

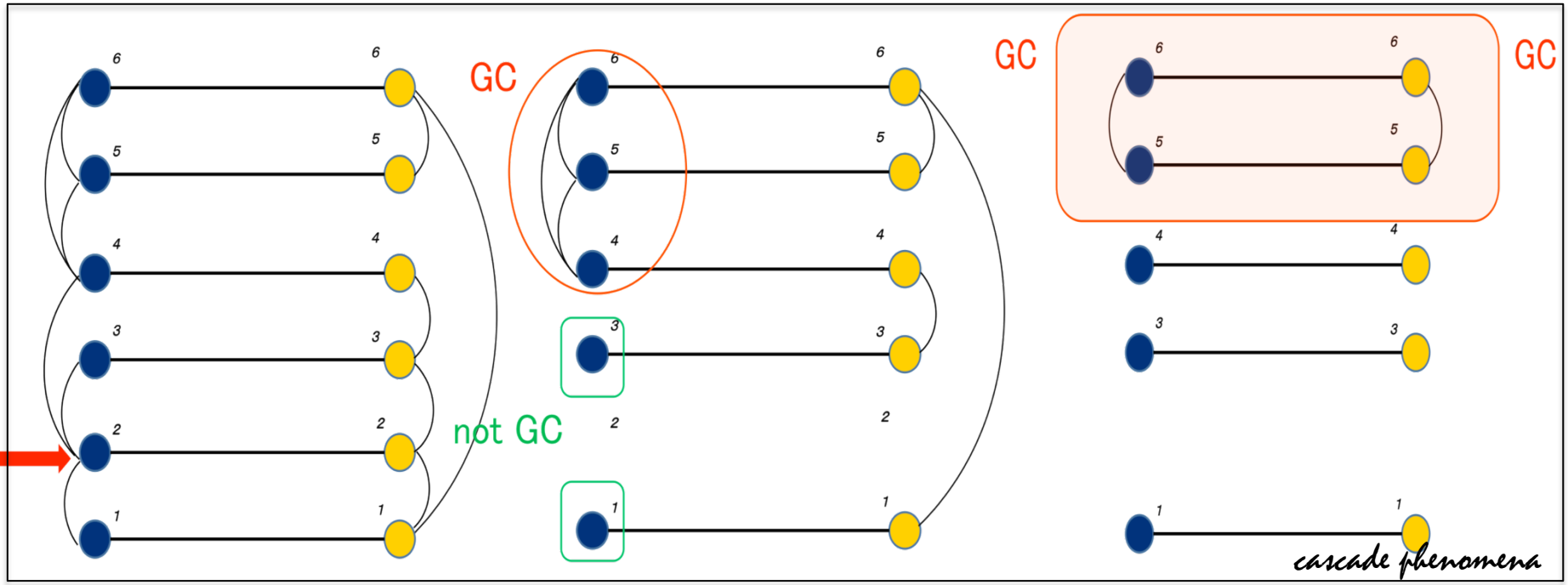
- About half of predicted clusters not matched with any known complexes are statistically significant wrt Gene Ontology.

T-60

Cavity法による次数相関のある相互依存型ネットワークの解析

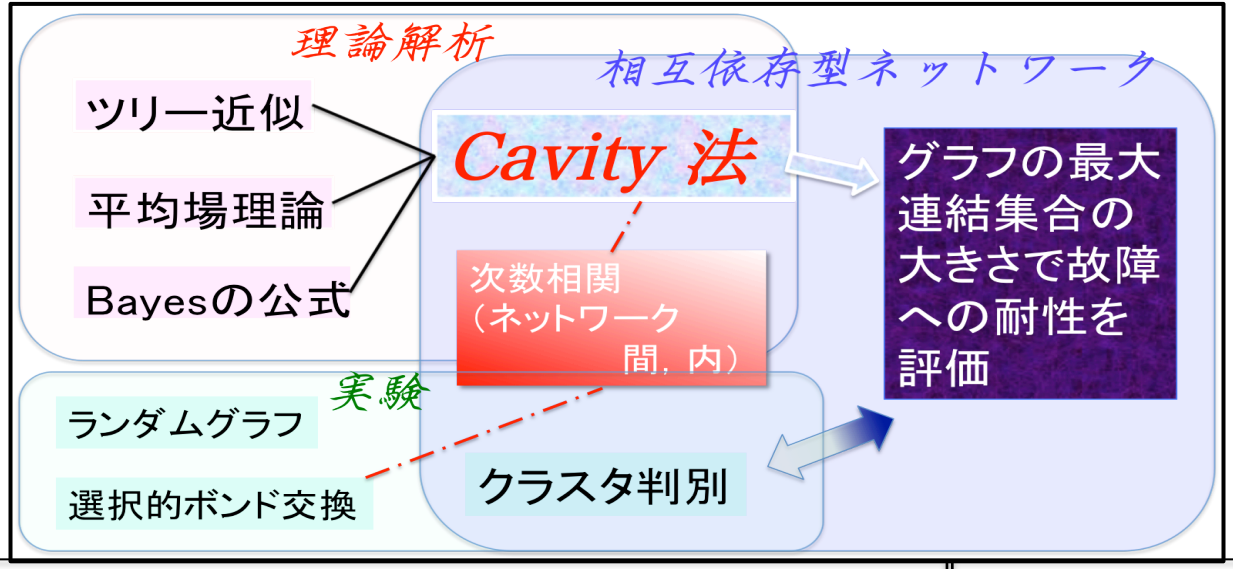
渡辺駿介　樺島祥介　東京工業大学大学院

2010年，S.V.Buldyrevらは，少数のサイトの故障がネットワーク全体に連鎖する現象の数理モデルとして相互依存型ネットワークを提案した．（Nature, vol464, pp1025~1028, 2010）.

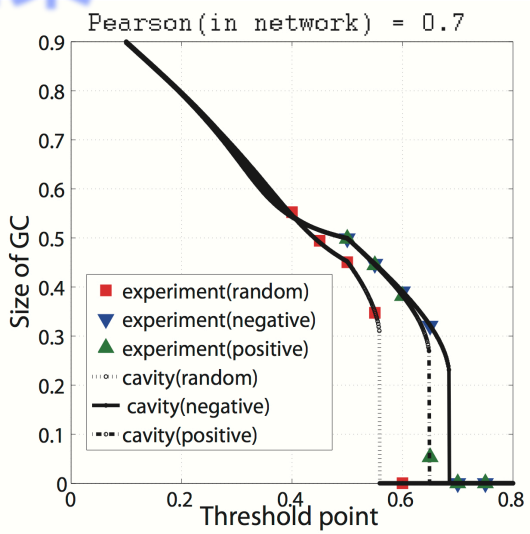


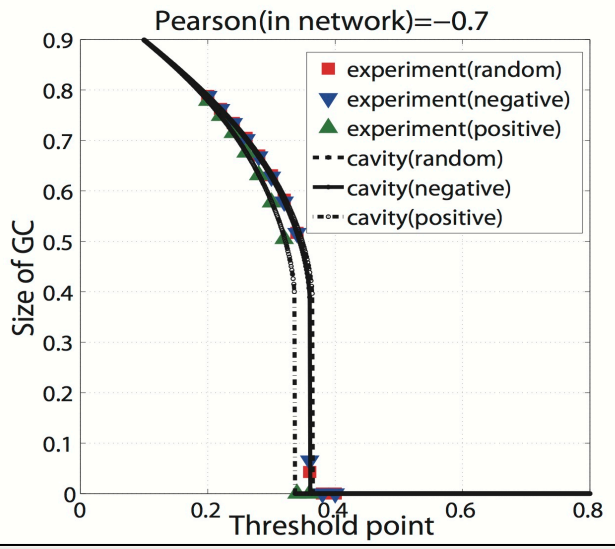
本研究では，次数（各サイトが持つ結合数）の相関がこの現象に与える影響を調べる

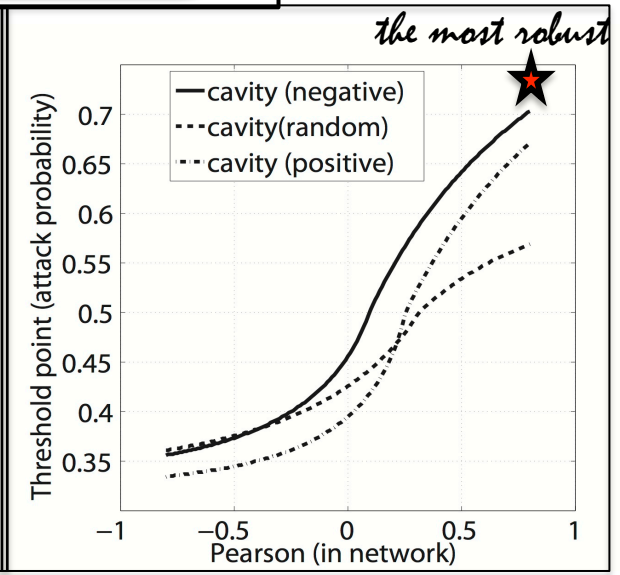
概要



結果







結論

“内”での相関が正に高く，“間”での相関が負に高いシステムが最もロバストである．