第1回 Latent Dynamics Workshop
予稿集

Collection of Technical Reports of
the First Workshop on Latent Dynamics (LD-1)

- **主催：Latent Dynamics 研究会**
- 協賛:電子情報通信学会 情報論的学習理論と機械学習（IBISML）研究会
- 日時：2010年6月16日
- 場所：東京大学工学部

The articles in this publication have been printed without reviewing and editing as received from the authors, and the copyright of the articles belongs to the authors. Therefore, this publication shall not preclude any further submissions to other journals and conferences.

目的・背景

大量データが溢れる現在、 データの表層をみただけではわからない潜在的世界に注目し、 その変化を捉えることが重要性を増している。

例えば、SNS のような関係性を表すネットワークデータの背後には、 コミュニティや階層関係といった潜在的世界があり、 その構造的な変化や時間発展を知ることによって、新しい話題の潮流や人脈の形成、 新しい文化の出現などを知ることができる。 そのような知識は現実世界の変化の兆しとも関係している。 また、潜在的なダイナミクスについてのこのような構造的理解をもとに、 価値創出とリスク管理の能力の高いコミュニティへと再設計する可能性を 高めることもできる。

他にも

- 物理的なネットワークにおける障害検出
- コンピュータ操作履歴からの異常行動の検出と防御
- テロリスト出現の検知と抑制
- 新しい伝染病の発生予兆の気づきと抑制、あるいは事前の予防
- 食の安全に関わる情報のブレーク予測と、 その情報を安全社会デザインにフィードバックする技術

などの問題も同様な視点でとらえられる可能性を秘めている。

そこで、このような潜在的世界の構造的な変化を Latent Dynamics と仮称し、 ここから価値ある情報を引き出し、制御するための科学的方法論をめぐって議論するために 本ワークショップを開催する。本ワークショップでは Latent Dynamics の分析と制御という 両面の視点を含めて扱う。すなわち、

1. 潜在的な構造変化を理解するための分析的アプローチ
2. 変化を抑制している障害を除き発生させる、 あるいは変化の原因を的確に阻害し抑制するためのシステムデザイン的アプローチ

を実現する方法および方法論、 及び両者の相互作用を実現する技術・技法について様々な分野の視点から議論する。

また、本ワークショップでは、 Latent Dynamics のモデリングとして以下の２つに注目する：

- （A) 数理的モデルの視点：統計的モデリング（潜在変数モデル)、 機械学習、データマイニング、データ可視化、情報理論など
- （B) 認知的モデルの視点： 認知科学、チャンス発見、社会生態学、知識発見の認知プロセスなど

(A) と (B) は相補的であり、互いに刺激し合って発展していくことを期待する。

Latent Dynamics は決して新しい概念ではない。 古来、様々な分野で様々な方法論をもって扱われてきた。 しかし大量かつ多様なデータが扱われるようになった現在でこそ 注目されるべき概念であると考えられる。 従来の隠れマルコフモデルや状態空間モデルでも状態変数のダイナミックスが 扱われていたが、 そこでは扱われていなかった構造的な変化（潜在変数を規定する、 よりマクロな構造的な情報の変化）をも問題にしてきたい。

本ワークショップでは既存の手法を Latent Dynamics の概念のもとで見直し体系づけるとともに、 新しい数理的モデル及び認知的モデルを創出するための機会を作ることを目指す。 また、Latent Dynamics の基礎的原理のアカデミックな追求はもちろんのこと、 それのみならず、現場の事例に基づき、Latent Dynamics の実用的・ビジネス的な観点からの検討も行っていきたい。

2010 年 6 月 16 日
Latent Dynamics 研究会 発起人
山西健司
大澤幸生
井手剛

プログラム

- セッション 1（潜在ダイナミクスのモデリング） 座長：井手剛
  - 10:05-10:45 山西健司（東京大学）
    Tracking Latent Dynamics — 潜在的構造変化検出の情報論的学習理論
  - 10:45-11:25 石井信（京都大学）
    階層ベイズモデリングによる時系列からの再構成
  - 11:25-11:30 休憩
  - 11:30-12:10 矢入健久（東京大学）
    非線形次元削減と動的システムの学習について
- 12:10-13:30 昼休み
- セッション 2（テキストの世界の潜在変数） 座長：大澤幸生
  - 13:30-14:00 岩田具治（NTT コミュニケーション科学基礎研究所）
    潜在トピックモデルを用いたデータマイニング
  - 14:00-14:30 Xu Sun (東京大学)
    Decoding in Latent Conditional Models: A Practically Fast Solution for a NP-hard Problem
- 14:30-14:40 休憩
- セッション 3（人間行動と潜在世界） 座長： 井手剛
  - 14:40-15:10 大澤幸生（東京大学）
    潜在ダイナミクスとしての「都合」
  - 15:10-15:40 宮野廣（法政大学、日本保全学会 特別顧問）
    トラブルの経験と情報としての活用技術
- 15:40-15:50 休憩
- セッション 4（潜在世界の揺らぎ） 座長： 山西健司

- 15:50-16:20 前野義晴（Social Design Group）
  揺らぎと偏りから読み解く潜在構造
- 16:20-16:50 井手剛（IBM 東京基礎研究所）
  潜在的グラフ構造からの異常検知

目次

Technical Report of the 1st Workshop on Latent Dynamics (Jun 16, 2010, Tokyo, Japan)

潜在構造変化検出の情報論的学習理論

山西健司*
Kenji Yamanishi

Abstract: 本稿では、データが与えられたときにその背景にある潜在的構造の変化を捉える問題を考える。統計学や機械学習では、潜在情報を扱う手段として潜在変数を伴う確率モデルが発展してきた。ここでは、そのような確率モデルの構造そのものが時間的に変化するとき、それをいかに検知するかといった問題を考える。この問題は、近年、情報理論や情報論的学習理論の立場から「Tracking Best Experts」,「Switching 理論」,「動的モデル選択」などとして取り組まれ、１つの潮流を形成しつつある。さらにそれは、Novelty Detection（新規性の検知)、ネットワーク構造変化検知などのデータマイニングの新しい問題に対する有力なアプローチでもある。本稿はそういった理論の流れと広がる応用の世界を紹介する。

Keywords: latent dynamics, dynamic model selection, switching theory, data mining

1 まえがき

大量データが溢れる現在、与えられたデータの関係性や構造を抽出したいというニーズは高くなってきている。その際、例えば相関ルールのようなデータの表層的な関係性ではなく、クラスタリング構造のような潜在情報の抽出がより求められるようになってきている。そのような潜在情報は、従来、潜在変数を伴う確率モデルとして統計学の分野で扱われてきた。

さらに我々は潜在情報の「動き」や「変化」に注目する。なぜなら、潜在情報そのものもさることながら、その「動き」にこそ価値ある情報が内在するからである。統計学の分野では、状態空間モデルなど潜在情報のダイナミックスを扱う方法論が確立されてきた。その多くの議論の対象は潜在的情報（内部状態）の値の変化を論ずるものが多かった。

一方、潜在変数の世界をマクロに支配する構造（例えば、潜在変数の数や階層構造）の時間的変化をいかに捉えるか?といった問題–「潜在的な構造変化検出の問題」–は難しいとされてきた。しかし、それは重要な問題である。なぜなら、潜在的な構造変化こそがデータの背後に潜む大きな変化の兆しであったり、新規性 (Novelty) の発現であったり、カタストロフィ（破局）につながる可能性があるからである。本稿では、構造的、非構造的な潜在情報の変化を含めて（狭義の）**Latent Dynamics** と呼ぶことにしよう。

潜在的な構造変化の検出については、近年、計算論的学習理論の分野では”Tracking best experts”, ”Derandomization”として、 情報論的学習理論の分野では「Switching 理論」,「動的モデル選択」などとして取り組まれ、１つの潮流を形成しつつある。

本稿では、こうした流れとその応用展開を概括し、Latent Dynamics に対する１つの有望なアプローチとして位置づける。

2 潜在情報の数学モデル

今、コンピュータ操作のコマンドの時系列$x^n = x_1, \ldots, x_n$をデータとして、この確率モデルを考える。データの背後には、コマンドの操作の意図ともいえる潜在変数 Z が存在し、潜在変数 Z がどのような値をとるか（例えば、「プログラム作成」「メール作成」「資料作成」など）によって、各コマンド (顕在変数 X で表わす) の発生パタン $P(X|Z)$ が決まり、それに従って、確率的にコマンド X が発生するというモデルを考えるのが自然である。

そこで、k を潜在変数の総数として、$\mathcal{Z} = \{z_1, \ldots, z_k\}$ を潜在変数の集合とし、$\mathcal{Z}$ 上の確率分布を $P_k(Z)$ とするとき、X の発生確率分布は**有限混合モデル**を用いてモデル化できる。

$$P(X) \quad = \quad \sum_{Z \in \mathcal{Z}} P(X|Z) P_k(Z).$$

ここで、潜在変数は周辺化されて外には見えない。$P(X|Z)$ としては、上記のコマンド系列の場合には、多項分布、マルコフモデルなどを用いることができる。Z 自体がど

*東京大学情報理工学系研究科数理情報学専攻
〒 113-8656 東京都文京区本郷 7-3-1
e-mail yamanishi@mist.i.u-tokyo.ac.jp
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN

のような値をとっているのかは各データに対して、EM アルゴリズムなどを用いて推定することができる。

さらに、潜在変数 Z の時系列モデルを考えて、Latent Dynamics を考える研究は従来より行われている。典型的なモデルが**状態空間モデル**である [14]。これは、入力系列 $\{x_t\}$ と出力系列 $\{y_t\}$ が与えられたとして、潜在変数 Z の時刻 t における実現値 z_t のダイナミックスを以下のように表現するものである。

$$z_t = f(z_{t-1}, x_t),$$
$$y_t = g(z_t).$$

ここで、f,g はなんらかの関数であるとする。状態空間モデルでは潜在変数の値の変化を推定することができる。例えば、トレンド成分や周期成分の時系列を抽出し、その変化を捉えることを可能にする。

また、よく知られた**隠れマルコフモデル**では潜在的状態 z_t がマルコフ連鎖を描くというダイナミックスが仮定されている。

$$P(x^n) = \sum_{z^n} \prod_{t=1}^{n} P(x_t|z_t) \prod_{t=1}^{n-1} P(z_t|z_{t-1}) \cdot \gamma(z_1).$$

経済時系列データの分析においては**レジュームスイッチング**の理論 [3] が発展している。これは例えば、経済時系列 x_t のモデルとして 1 次の AR モデルを当てはめた場合：

$$x_t = \theta_1(s) + \theta_2(s)x_{t-1} + \sigma(s)\varepsilon_t.$$

に、係数 $\theta_1(s), \theta_2(s)$ 及び分散 $\sigma(s)$ は潜在状態 s(例えば、景気の状態）に依存すると仮定し、s の切り替わりのタイミングを推定するというものである。

また、データマイニングの分野では**隠れ変数マイニング** [5] という文脈の下で、潜在変数の分布の変化検知の問題が扱われている。そこでは、P_t を時刻 t における潜在変数 Z の事後確率分布として、何らかの距離関数 d の時系列 $\{v_t\}$ を考え、その変化点を知ることにより、潜在世界の異常を検知する手法が提案されている。

$$v_t = d(P_t(z), P_{t-1}(z)) = \sum_z P_t(z|x^t) \log \frac{P_t(z|x^t)}{P_{t-1}(z|x^{t-1})}.$$

さらに、**共分散構造解析** [15] の文脈においても潜在時系列の解析が研究されている。

以上はいずれも Latent Dynamics を扱うものであるが、いずれも潜在変数の値のダイナミックスを対象としたものであった。しかし、潜在変数の空間が $\mathcal{Z} = \{z_1\}$ であったものが、 $\mathcal{Z} = \{z_1, z_2\}$ に切り替わるなど、潜在変数を支配する構造（例えば、潜在変数の数、潜在変数の階層構造）の変化を検知する問題に対しては、新しい切り口が必要である。

3 動的モデル選択

潜在的な構造変化検知の問題は、計算論的学習理論の分野で **Tracking best experts** [4] あるいは **Derandomization** [11] の問題として研究されてきた。問題設定は、expert と呼ばれる予測器が複数用意され、各時刻で個別に予測を行うが、最良な予測を行う expert(best expert) が時間とともに変化する状況下で、expert を組み合わせて最良な予測器と同程度の予測精度を実現したい、というものである。ここで、best expert の時系列そのものが潜在変数と見なされる。しかし、最良な予測器がいつどのように切り替わったかということは陽には問題にされないできた。

また、隠れマルコフモデルで状態数が時間とともに切り替わるモデルとして、ノンパラメトリックベイズ学習の文脈で Infinite HMM なの概念が発達しているが [1]、無限の状態数に対する事前分布を用いた混合分布として捉えられているため、必ずしも状態数の変化そのもの検知には関心が払われていなかった。

また、情報論的学習理論の分野では**動的モデル選択** [13] の理論として研究されてきた。これは異なる複雑さをもった確率モデルが時間とともに切り替わる場合に、その系列をトラッキングするための理論であり、**MDL(Minimum Description Length) 原理**の枠組みの中で解かれてきた。これと独立に、**Switching 分布**の理論が生まれており、状況設定は動的モデル選択と同じながらも、目標はモデルの切り替わりの検出ではなく、最終的なモデルの収束性と収束速度の加速にあった。

上記の研究で共通するのは、モデル及びモデルの切り替わりの時系列情報も含めて潜在変数と見なしているということである。これは潜在的構造変化検知に固有の方法論である。以下、そのような方法論を踏まえた動的モデル選択の理論の骨子を紹介する。

今、k をモデルとする確率モデルのクラス

$$\mathcal{P}_k = \{P(x^n|\theta, k) : \ \theta \in \Theta_k\} \ \ (n = 1, 2, \cdots)$$

が与えられているとする。ここに、$\dim \Theta_1 < \cdots < \dim \Theta_k < \dim \Theta_{k+1} < \cdots$ であるとする。時刻 t におけるモデルを k_t として、x^{t-1} が与えられたもとでの x_t の予測分布を $P(x_t|x^{t-1} : k_t)$ で表わす。予測分布としては θ^{t-1} を x^{t-1} からの θ の最尤推定量として、これを代入した plug-in 分布:

$$P(x_t|x_{t-1} : k_t) = P(x_t|\hat{\theta}_{t-1} : k_t)$$

や、$P(\theta|x^{t-1})$ を x^{t-1} からの θ の事後確率密度関数として以下の形で与えられるベイズ予測分布:

$$P(x_t|x_{t-1} : k_t) = \int P(x_t|\theta)P(\theta|x^{t-1} : k_t)d\theta$$

や逐次的正規化最尤予測分布

$$P(x_t|x^{t-1}:k_t)=\frac{P(x_t\cdot x^{t-1}|\hat{\theta}(x_t\cdot x^{t-1}):k_t)}{\sum_x P(x\cdot x^{t-1}|\hat{\theta}(x\cdot x^{t-1}):k_t)}$$

などを用いることができる。

今、データ列 $x^n=x_1\cdots x_n$ に対して、m をモデルの変化点の総数、$\mathbf{t}=(t_0=1,t_1,t_2,\cdots,t_m)$ をモデルの変化点系列、$\mathbf{k}=(k_0,k_1,k_2,\cdots,k_m)$ を対応するモデル系列、として $\mathbf{s}=(m,\mathbf{t},\mathbf{k})$ を **Latent Dynamics** を表わす潜在変数とする。$P(\mathbf{s})$ を $\mathbf{s}$ の事前分布とする。このとき、潜在変数 $\mathbf{s}$ に付随する **Switching 分布** $P(x^n|\mathbf{s})$ を以下のように定義する。

$$P(x_i|x^{i-1}:\mathbf{s})=\begin{cases}P(x_i|x^{i-1}:k_0) & t_0\le t\le t_1\\ P(x_i|x^{i-1}:k_1) & t_1\le t\le t_2\\ P(x_i|x^{i-1}:k_2) & t_2\le t\le t_3\\ \cdots & \cdots\end{cases}$$

$$P(x^n|\mathbf{s})=\prod_{i=1}^{n}P(x_i|x^{i-1}:\mathbf{s})$$

さらに $\mathbf{s}$ に関して周辺化した x^n の確率分布を

$$P(x^n)=\sum_{\mathbf{s}}P(x^n|\mathbf{s})P(\mathbf{s})$$

で表わす。

与えられたデータ列から $\mathbf{s}$ を推定することを**動的モデル選択**と呼ぶ。以下では動的モデル選択のための規準を情報理論の立場から導出しよう。先ず、x^n のモデル系列に関する**確率的コンプレキシティ**を

$$-\log\sum_{\mathbf{s}}P(x^n|\mathbf{s})P(\mathbf{s})$$

として定める。これはモデル系列の全体のクラスに相対的なデータ系列の持つ情報量、または符号長と見なすことができる。ここで、

$$-\log\sum_{\mathbf{s}}P(x^n|\mathbf{s})P(\mathbf{s})\le\min_{\mathbf{s}}\{-\log P(x^n|\mathbf{s})-\log P(\mathbf{s})\}$$

の関係から、確率的コンプレキシティは左辺の値で近似できる。左辺の最小化すべき対象の第一項は $\mathbf{s}$ が与えられた下での x^n の符号長を、第二項は $\mathbf{s}$ 自身の符号長を表わす。つまり左辺はデータを Latent Dynamics を含めた 2 段階符号化の総符号長を意味する。

そこで、**MDL(Minimum Description Length) 原理**に基づいて、左辺の最小値を達成する $\mathbf{s}$ を用いて最適な Latent Dynamics と見なす。これを $\mathbf{s}_{opt}$ と記す。

$$\hat{\mathbf{s}}_{opt}=\arg\min_{\mathbf{s}}\{-\log P(x^n|\mathbf{s})-\log P(\mathbf{s})\}$$

つまり、$\mathbf{s}_{opt}$ はデータ圧縮の意味で最適な Latent Dynamics である。

$-\log P(x^n|\mathbf{s})$ 及び $-\log P(\mathbf{s})$ を**予測的符号長**を用いて計算する。今、$k^n=k_1,\ldots k_n$ の生成モデルとして α をパラメータとする 1 次マルコフモデルを仮定し、これを $P(k_t|k_{t-1}:\alpha)$ と表わす。そのとき、Latent Dynamics の推定問題は

$$\sum_{t=1}^{n}-\log P(x_t|x^{t-1}:k_t)+\sum_{t=1}^{n}-\log P(k_t|k^{t-1}:\hat{\alpha}_{t-1})\quad(1)$$

を最小にする $k^n=k_1,\cdots,k_n$ を求める事に帰着される。ここに、$\hat{\alpha}_{t-1}$ は k^{t-1} からの α の最尤推定量である。式 (1) を **DMS(Dynamic Model Selection) 規準**と呼ぶ。

以上の設定の下で重要な問題は以下の通りである。

- $\hat{\mathbf{s}}_{opt}$ をいかに効率的に求めるか？(計算論的問題)
- $\hat{\mathbf{s}}_{opt}$ はどのような性質をもつか？(情報論的問題)

4 動的モデル選択の諸性質

さて、Latent Dynamics の推定規準として DMS 規準が与えられたところで、これを具体的に達成するアルゴリズムの計算論的側面と情報論的側面に関して知られている幾つかの結果をやや大雑把な形で紹介しよう。

先ず、データが一括与えられた下で DMS 規準を最小化する Latent Dynamics を求めるアルゴリズムの性質について以下が成立する。

定理 1 *[13] Switching* 分布に対して一括型で *DMS* 規準 *(1)* を最小化する *Latent Dynamics* を計算量 $O(n^2)$ で出力する**一括型 DMS アルゴリズム**が存在し、その総記述長の上界は次式で抑えられる。

$$\min_{m}\min_{(t_0,k_0),\cdots,(t_m,k_m)}\left\{\sum_{j=0}^{m}\sum_{t_j+1}^{t_{j+1}-1}-\log P(x_t|x^{t-1}:k_i)\right.$$
$$\left.+nH\left(\frac{m}{n}\right)+\frac{1}{2}\log n+m+o(\log n)\right\}.\qquad(2)$$

定理 1 の一括型 DMS アルゴリズムは、モデルの遷移確率を Krishevsky and Trofimov 推定を用いて計算し、最適なモデル系列を動的計画法を用いて求めることで構成できる。

また、データが逐次的に与えられた下で DMS 規準を達成する Latent Dynamcis を逐次的に推定するアルゴリズムについて以下が成立する。

定理 2 *[16] Switching* 分布に対して逐次的に *DMS* 規準 *(1)* を最小化する *Latent Dynamics* を計算量 $O(n)$ で出力する**逐次型 DMS アルゴリズム**が存在し、そのときの総記述長は *(2)* よりも大きな上界をもつ。

定理 2 の逐次型 DMS アルゴリズムは、一定幅のウインドウを設けて、その中で最適なモデル系列を動的計画法を用いて求め、これを逐次的に接続していくアルゴリズムとして構成できる。

さらに、文献 [16] では、定理 1 の一括型 DMS アルゴリズムに対して、定理 2 の逐次型アルゴリズムの出力は計算オーダを減らすと共に、9 割以上同じ系列を出力し、情報論的な限界は大きく見劣りしないことが実験的に示されている。

さらに Switching 分布のバリエーションとして **Resetting 分布**をモデルの変化点で予測分布をリセットさせる分布として定義する。Switching 分布が各モデルで過去の予測分布を憶えている部分が異なることに注意する。

データが一括与えられた下での Resetting 分布に対して DMS 規準を最小化する Latent Dynamics の推定アルゴリズムについて以下が成立する。

定理 3 *Resetting* 分布に対して（各変化点で初期化）一括型で *DMS* 規準を最小化する *Latent Dynamics* を計算量 $O(n^3)$ で出力するアルゴリズムが存在し、そのときの総記述長は *(2)* よりも小さな上界をもつ。

以上が DMS 規準を用いた Latent Dynamics の推定アルゴリズムの計算論的及び情報論的側面であるが、仮説検定の観点から、その性能を調べることができる。これを以下に示そう。

今、モデルが 2 つ $\{M_1, M_2\}$ しかなくて、モデル遷移確率が一定の場合の Switching の問題を考えて、t^* をモデルの変化点として、以下の 2 つの仮説を考える。

$$\text{仮説 } H_0: \quad M_1 \qquad \text{for } x_1^n = x_1 \cdots x_n,$$

$$\text{仮説 } H_1: \quad \begin{cases} M_1 & \text{for } x_1^{t^*} = x_1 \cdots x_{t^*}, \\ M_2 & \text{for } x_{t^*+1}^n = x_{t^*+1} \cdots x_n. \end{cases}$$

DMS 規準によれば、

$$-\log P(x_{t^*+1}^n | x_1^{t^*} : M_1) + \log P(x_{t^*+1}^n | x_1^{t^*} : M_2) - \alpha < 0$$

であれば H_0 が採択され、そうでなければ H_1 が採択されることになる。ここに、$\alpha = \log(\omega/(1-\omega))$, $P(M_1|M_1) = P(M_2|M_2) = \omega > 1/2, P(M_2|M_1) = 1-\omega$ としてモデル遷移確率は既知とする。

この仮説検定問題に関して第一種の誤り確率及び第二種の誤り確率に関して以下が成り立つ。

定理 4 モデルクラスに関するある仮定の下で次式が成り立つ。

$$Prob[\text{モデルが変化しないが } H_1 \text{が採択}] \le 2^{-\alpha},$$

$$Prob[\text{モデルが変化するが } H_0 \text{が採択}] \le 2\exp\left(-Ch\beta^2\right).$$

ここに、$h \stackrel{\text{def}}{=} n - t^*$ はモデル変化検知の *delay* であり、$D_h(M_2||M_1) \stackrel{\text{def}}{=} \sum_{x_{t^*+1}^n} P(x_{t^*+1}^n | x_1^{t^*} : M_2) \log P(x_{t^*+1}^n | x_1^{t^*} : M_2)/P(x_{t^*+1}^n | x_1^{t^*} : M_1)$, $\beta \stackrel{\text{def}}{=} \frac{1}{h}(D_h(M_2||M_1) - \alpha)$ であり、C は定数である。

定理 5 定理 *4* の β の下界が h に関して一様に γ で抑えられる場合、*DMS* 規準に基づく動的モデル選択によるモデル変化点の検知の遅延時間の期待値は $O(1/\gamma^2)$ で与えられる。

5 データマイニング応用

Latent Dynamics の推定は、データマイニングの分野で広い応用可能性をもつ。特に、時系列データからの新規性の検出（Novelty Detection) や異常検出 (Anomaly Detection) においては既に、幾つかの実例を見ることができる。

例えば、文献 [13] にて、動的モデル選択はセキュリティの「なりすまし検出」問題に適用され、構造変化検知がなりすまし者の行動パタンの同定に結びついた事例が報告されている。また、文献 [12] にて, 動的モデル選択により Syslog からの新しい障害パタン発見が導かれた事例が報告されている。さらに、テキストストリームデータからのトピック分析において、潜在構造変化検知により新たなトピックの出現検知が可能であることが報告されている [8].

今後、特に興味深い応用として、グラフ時系列からのグラフ構造変化検出の問題が考えられる。非定常なグラフ時系列からの異常検出の問題は、ネットワーク異常検知やソーシャルネットワークにおけるコミュニティ分析をモチベーションとして最近、急速に発展している（例えば、[7],[6],[10] を参照されたい)。そこでも。潜在的構造変化検知は、新たに登場するネットワークのコミュニティや階層構造などの検出に有力な手段を与えるものと考えられる。

6 おわりに

本稿では、Latent Dynamics を扱う重要な問題が潜在的構造変化検知の問題であると説き、その解決方策の中心に MDL 原理を据えて、情報論的学習理論の立場から Latent Dynamics 推定のアルゴリズムの情報論的及び計算論的限界を示した。しかし、これは本ワークショップが本来議論しようとする Latent Dynamics の概念を限定した一部の見方にすぎない。とはいえ、Latent Dynamics とは何か？　を語る上で、明確な定義の下で 1 つの局面を切り出して行く操作は重要である。本稿で示した理論が、より多角的な数理的あるいは認知的方法

論と結びついて新たに成長を遂げ、Latent Dynamics の本質に少しでも近づければと願っている。

参考文献

[1] M.J. Beal, Z. Ghahramani and C.E. Rasmussen: The infinite hidden Markov model. *Advances in NIPS*, vol.14, MIT Press, pp:577-584, 2002.

[2] T. van Erven and P.D. Grunwald and S. de Rooij: Catching up faster in Bayesian model selection and model averaging. *Advances in Neural Information Processing Systems* 20, 2007.

[3] J.D. Hamilton: *Time Series Analysis*. Princeton University Press, 1994.

[4] M. Herbster and M. K. Warmuth. Tracking the best expert. *Journal of Machine Learning*, 30(2):151–178, 1998.

[5] S.Hirose and K.Yamanishi: Latent variable mining with its applications to abnormal behavior detection. *Statistical Analysis and Data Mining*, 2009.

[6] S.Hirose, K.Yamanishi, T.Nakata, R.Fujimaki: Network anomaly detection based on eigen equation compression. *Proc. of KDD2009*, 2009.

[7] T. Ide and H. Kashima. Eigenspace-based anomaly detection in computer systems. *Proc. of KDD2004*,ACM Press, 2004.

[8] S. Morinaga and K. Yamanishi: Tracking Dynamics of Topic Trends Using a Finite Mixture Model. *Proc. of KDD2004*, ACM Press, 2004.

[9] J. Rissanen: *Information and Complexity in Statistical Modeling*, Springer, 2007.

[10] J. Sun, P. S. Yu, S. Papadimitriou and C. Faloutsos: GraphScope: Parameter-free mining of large time-evolving graphs. Proceedings of KDD2007, 2007.

[11] V. Vovk: Derandomizing stochastic prediction strategies. *Machine Learning*, 35, pp:247–282, 1999.

[12] K. Yamanishi and Y. Maruyama: Dynamic syslog mining for network failure monitoring. *Proc. of KDD2005*, pp: 499-508, ACM Press, 2005.

[13] K. Yamanishi and Y. Maruyama: Dynamic model selection with its applications to novelty detection. *IEEE Trans. on Information Theory*,IT 53(6)：2180-2189, June, 2007.

[14] 北川源四郎: 時系列解析入門, 岩波書店, 2005.

[15] 豊田秀樹: 共分散構造分析 応用編—構造方程式モデリング. 朝倉書店 2000.

[16] 櫻井、山西: 逐次的動的モデル選択の線形時間アルゴリズム. 電子情報通信学会　情報論的学習理論と機械学習研究会　予稿集、2010.

Technical Report of the 1st Workshop on Latent Dynamics
(Jun 16, 2010, Tokyo)

階層ベイズモデリングによる時系列からの再構成

石井　信*

Abstract: 階層ベイズモデリングは、状態変数に複雑な確率依存性がある場合のデータ解析に有効な手法を提供する。ここでは、階層ベイズモデリングにより時系列データを取り扱った研究のいくつかを紹介する。すなわち、遮蔽除去超解像、明滅する信号源からの独立成分分析、隠れマルコフモデルによるヒト推論過程のモデル化である。

Keywords: superresolution, variational Bayes, hidden Markov model

1　発表の概略

階層ベイズモデリングは、状態変数に複雑な確率依存性がある場合のデータ解析に有効な手法を提供する。特に、ベイズ推定による事後分布の取り扱いが各種の不確実性を考慮した適切な推定を可能とする。本発表では、階層ベイズモデリングにより、時系列データの解析を行った応用研究をいくつか紹介する。

第一の課題は、遮蔽除去超解像 [1] である。画像の超解像とは、入力された画像よりも高い解像度の画像を出力する技術である。与えられた画像を単に補間拡大するのではなく、情報を増やすことで高い解像度の画像を出力する。これまでに、各画像が同一の撮影対象に対する異なる情報を持っていることを利用して高い解像度を得るマルチフレーム超解像の研究が行われている。このマルチフレーム超解像において、各々の低解像度画像に異なる遮蔽物が混入しているような状況を取り扱うのが遮蔽除去超解像である。まず、遮蔽物が低解像度画像ごとに独立に混入する状況を考え、さらに、それを画像間にあるダイナミクスを持って混入する状況へと拡張する。こうした遮蔽状況は、隠れ変数を持ち、したがって階層化された尤度関数を仮定することで取り扱うことができる。こうした尤度関数を用いて、複数の低解像度画像に基づき、遮蔽物とともに高解像度画像を推定する問題は、階層ベイズモデルに基づくベイズ推定として定式化できる。結果として、ピクセルごとに遮蔽されているかどうかを推定し、その推定の不確実性を考慮した高解像度画像の推定が実現できる。

複数の信号源からの信号が混合されている状況で、混合過程の知識があまりないままに、各信号源からの信号を分離する問題をブラインド信号分離といい、別々の信号源からの信号の独立性に基づき分離する独立信号分析(ICA)による解法が知られている。第二の課題は、このブラインド信号分離において、各信号源が明滅している、すなわち、ある時刻においては信号を生成しているが別の時刻においては信号を出してしないような状況で、そうした信号の明滅状況を推定しながら、信号源分離を可能とする非定常独立成分分析 [2] である。信号源の明滅についてマルコフモデルによりモデル化を行い、パラメトリック ICA により信号分離する。また、原信号については、混合正規分布を仮定する。この確率モデル（全体として隠れマルコフモデル）の推定を変分ベイズ法によって行った結果、信号の明滅を仮定しないモデルによる推定よりも良い結果が得られた。

我々人間は不確実性のある環境においても意思決定を行う必要がある。この際に、不確実性の解消は重要なステップである。第三の課題は、こうした人間の不確実性の解消過程のモデル化 [3] である。現在の観測だけからでは、どこにいるのか分からないような迷路（部分観測迷路）における最適意思決定過程を、隠れマルコフモデルと、貪欲方策でモデル化した。このモデルにより、人間の行動を高い確率で予測することができる。また、モデルを用いることで、不確実性の解消に関わる人間の認知負荷を見積もることができる。見積もった認知負荷と、課題実行中の核磁気共鳴図法による非侵襲脳活動計測データとの相関解析により、人間の不確実性の解消に関わる脳の処理基盤について論じることができる。

参考文献

[1] Kanemura, A., Maeda, S., Fukuda, W., Ishii, S. *Journal of Systems Science and Complexity*, 23(1), 116-136, 2010.

[2] Hiyarama, J., Maeda, S., Ishii, S. *IEEE Transactions on Neural Networks*, 18(5), 1326-1342, 2007.

[3] Yoshida, W., Ishii, S. *Neuron*, 50(5), 781-789, 2006.

*京都大学　大学院情報学研究科, 619-0011 京都府宇治市五ヶ庄, tel. 0774-38-3938,
e-mail: ishii@i.kyoto-u.ac.jp, URL: http://ishiilab.jp

Technical Report of the 1st Workshop on Latent Dynamics (Jun 16, 2010, Tokyo, Japan)

次元削減と動的システムの学習

矢入健久*
Takehisa Yairi

Abstract: 様々な動的なシステムのモデルを蓄積された観測データから推定・学習する問題は、制御工学 (システム同定) と機械学習の両分野に共通する興味対象である。特に近年では、センサー技術の発達等により観測データの高次元化が進んでおり、高次元の観測空間から低次元のシステムの本質的な状態遷移を見つけ出す「次元削減」の重要性が増している。本発表では、この観点からシステム同定・機械学習両分野でのトレンドを概観するとともに、発表者らの取り組みを紹介する。

Keywords: 動的システム学習, システム同定, 状態空間モデル, 次元削減, 部分空間同定法

1 まえがき

時間的に状態を変化させるシステム、すなわち、動的なシステムから発生するデータから、そのシステムのモデル構造・パラメータを推定する問題は、機械学習分野では動的システム学習 (learning of dynamical systems)、制御工学分野ではシステム同定 (system identification) と呼ばれ、それぞれの分野で近年まで互いにあまり干渉することなく独自の発展を遂げてきた。

本発表では、これら 2 つの分野における近年の研究動向を概観して両者の立場や興味、方法論の相違点を探るとともに、両者の融合による新たな動的システム学習理論の可能性を考察したい。

なお、以下の議論では、次式の (離散時間) 状態空間モデルによって表される動的システムを想定するものとする。

$$\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t) + \boldsymbol{v}_t \quad (1)$$

$$\boldsymbol{y}_t = g(\boldsymbol{x}_t) + \boldsymbol{w}_t \quad (2)$$

2 機械学習分野における動的システム学習研究の動向

2.1 スイッチング線形モデルの学習

未知の非線形システムをモデル化する最も基本的かつ現実的な方法は、複数の線形モデルによって局所的に近似することであろう。機械学習分野では、そのようなモデルは、スイッチング線形動的システム (Switching Linear Dynamical System: SLDS), スイッチングカルマンフィルタ (SKF) などと呼ばれている。SLDS モデルをデータから学習するには、通常、EM アルゴリズムや、それに類する反復的アルゴリズムが用いられるが、E ステップについては、どの局所モデルを選択するかを表す離散的確率変数 s_t と連続的状態変数 $\boldsymbol{x}_t$ の同時事後分布を厳密に推定することが困難であるため、変分近似を用いる方法 [6] や s_t の推定に Viterbi アルゴリズムを用いる方法 [19] などが提案されている。また、SLDS の学習では、そもそも局所線形モデルの数をいくつにするべきか、という問題が存在するが、最近の研究 [5, 3] では、ディリクレ過程混合モデルや変分ベイズ法を用いることによってモデル数を自動決定する手法が提案されている。

2.2 非線形潜在変数モデルの学習

非線形システムを学習するもう一つのアプローチは、状態空間モデル 1,2 において、状態遷移関数 f および出力関数 g を、非線形回帰や次元削減の手法を用いてデータから推定することである。

まず、カーネル以前の非線形アプローチの例としては、Ghahramani と Roweis による研究 [7] が有名である。これは、状態空間モデルにおける状態方程式および観測方程式を Radial Basis Function Networks (RBFN) によって表現し、そのモデルパラメータを EM アルゴリズムによって反復的に学習するというものである。

Kernel Kalman Filter (KKF) [20] は、「非線形なシステムであっても、カーネルによって写像された特徴空間では線形なモデルで表される」という考えに基づき、線形の状態空間モデルをカーネルトリックによって非線形化したものである。この手法では、特徴空間における状

*東京大学先端科学技術研究センター, 〒 153-8904 東京都目黒区駒場 4-6-1, e-mail yairi@space.rcast.u-tokyo.ac.jp,
RCAST, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo

態・観測方程式のモデルパラメータはやはり EM アルゴリズムによって学習される。ただし、学習した KKF を実際にフィルタリングや予測に用いるためには、特徴空間から元の観測空間への逆写像 (preimage) を求める必要がある。また、カーネルと後述の部分空間同定法を融合した例として、カーネル正準相関分析 (Kernel Canonical Correlation Analysis: KCCA) を用いた [11] が挙げられる。

また、カーネル行列を分散共分散行列とみなすことによって正規過程 (Gaussian Process) が導かれるが、この立場から動的システムを扱ったものとして、正規過程動的モデル (Gaussian Process Dynamical Models: GPDM)[28] が挙げられる。GPDM は、非線形次元削減法である正規過程潜在変数モデル (Gaussian Process Latent Variable Models: GPLVM)[15] を、隠れ状態のダイナミクスを扱えるように拡張したものである。最近、GPDM と類似の手法として、GPIL[24] が提案されているが、こちらでは擬似入力 (pseudo inputs) によるスパース化が図られている。また、これらと同様に、状態遷移関数 f, 出力関数 g を正規過程によってモデル化するものとして、GP-Bayes[12], GP-ADF[4] などもある。ただし、これらは訓練時において状態変数 $\{\boldsymbol{x}_t\}$ の真値が与えられると仮定しており、その学習は単に 2 つの GP 回帰に帰着される。

グラフベーズドな非線形次元削減、いわゆる多様体学習 (Manifold Learning) の非線形動的システムへの応用は今のところ多くないが、ラプラス固有写像 (Laplacian Eigenmap:LE)[2] を確率的な潜在変数モデルに拡張したものを非線形な観測モデルとして利用し、モーションキャプチャデータ等のモデル化に応用した例 [17] がある。ただし、この例では状態遷移には単純なランダムウォークモデルを利用している。

2.3 非定常 DBN の構造学習

動的ベイジアンネットワーク (DBN) は、確率的な動的システムを表現・モデル化する手段として機械学習では広く取り入れられているが、特に近年では、時間的にモデル構造やパラメータが変化するような、非定常 DBN の学習についての研究も行われている [21, 23]。

2.4 その他の話題

上で見たように、これまでのところ、機械学習における動的システムの学習では、「確率的な」手法であるガウス過程などが好まれて用いられている。これはフィルタリングや平滑化などの確率的推論との相性の良さに依るところが大きい。

しかし、ごく最近、この常識を変える可能性のある研究が Langford により発表されている [13]。その基本的なアイデアは、確率的な状態変数をそのまま扱う代わりに、その十分統計量を決定論的な状態変数として利用する、という比較的単純なものであるが、これにより、任意の「確率的でない」非線形回帰アルゴリズム (例えば SVR) を、確率的な動的システムの学習に用いることができるようになるという。

3 システム同定分野における動向

80 年代後半以降システム同定分野では、部分空間同定法 (subspace identification) と呼ばれる一連の方法が盛んに研究されている [18, 14, 27]。部分空間同定法では、システムの過去および未来の入出力データが張る部分空間上での、直交射影 (orthogonal projection) や斜交射影 (oblique projection) などの幾何学的演算によって状態ベクトルやモデルパラメータが求められる。

部分空間同定法は、線形動的システムに関する強固な理論に裏打ちされたものであると同時に、「状態とは、未来 (過去) を予測するために必要な過去 (未来) の情報を縮約したもの」[30] という直感的な解釈も可能であるところが魅力的である。その一方、部分空間同定法の非線形化に関しては、(あくまで筆者の印象ではあるが、) まだ限定的と言って良いであろう。また、前述の SLDS に似た考えとして、区分線形 (piecewise linear: PWL) システムに対する部分空間同定法 [26] もあるが、各時刻の観測がどの局所線形モデルから生じているかについて常に情報が得られるという強い仮定に基づいている。

なお、非線形モデルの学習という話題に関しては、[16] によれば、システム同定分野においてもカーネル [8] や多様体学習 (グラフ正則化)[9] など、様々な機械学習の手法や理論を取り入れようとする機運が高まっているようである。また、部分空間同定法以前の主流であった予測誤差法 (Prediction Error Minimization: PEM) は最尤推定と密接に関係しているが、意外にもシステム同定分野で EM アルゴリズムが認知されるようになったのは比較的最近のようである [22, 16]。

4 局所線形モデルの整列による非線形システムの学習法

上では、機械学習およびシステム同定における動的システムの学習の従来研究を概観したが、これら 2 つの方法論を融合する 1 方法として、著者らは局所線形動的モデルの整列による非線形システムの学習法に取り組んでいる [10]。

これは、CCA に基づく (線形の) 部分空間同定法を、CCA の確率的解釈 [1] と局所線形モデルの整列法 [25] によって拡張したものである。詳細については [10] を参考にされたい。

5 おわりに

本稿では、状態空間モデルによって表現される動的システムのデータからの学習法に関して、機械学習分野およびシステム同定分野におけるトレンドを概観した。また、著者らのグループが行っている両分野の要素技術を取り入れた新しい非線形動的システムの学習法を紹介した。今後も、機械学習とシステム同定の両分野が交流を持ち、相互の発展につながることを期待したい。

謝辞

本稿における動的システム学習に関するサーベイの大部分は、井手剛氏 (IBM) との共著解説 [29] に基づいている。また、部分空間同定法と機械学習との関連性については、河原吉伸氏 (現・大阪大学) の博士論文研究から得た情報に基づいている。なお、著者らのグループの取り組みとして紹介した「局所線形モデルの整列による非線形システムの学習法」は、河原氏および東大院修士課程の上甲昌郎氏との共同研究である。これらの関係者の協力に厚くお礼を申し上げる次第である。

参考文献

[1] Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, 2001.

[3] Silvia Chiappa, Jens Kober, and Jan Peters. Using bayesian dynamical systems for motion template libraries. In *Advances in Neural Information Processing Systems 21*. MIT Press, 2009.

[4] Marc Peter Deisenroth, Marco F. Huber, and Uwe D. Hanebeck. Analytic moment-based gaussian process filtering. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 225–232, 2009.

[5] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. Nonparametric bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems 21*. MIT Press, 2009.

[6] Zoubin Ghahramani and Geoffrey E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12:963–996, 1998.

[7] Zoubin Ghahramani and Sam Roweis. Learning nonlinear dynamical systems using an em algorithm. In *Advances in Neural Information Processing Systems 11*, pages 431–437. MIT Press, 1999.

[8] Ivan Goethals, Kristiaan Pelckmans, Johan A. K. Suykens, and Bart De Moor. Subspace identification of hammerstein systems using least squares support vector machines. *IEEE Trans. on Automatic Control*, 50:1509–1519, 2005.

[9] Lennart Ljung Henrik Ohlsson, Jacob Roll. Manifold-constrained regressors in system identification. In *Proc. 47st IEEE Conference on Decision and Control*, pages 1364–1369, 2008.

[10] Masao Joko, Yoshinobu Kawahara, and Takehisa Yairi. Learning non-linear dynamical systems by alignment of local linear models. In *20th International Conference on Pattern Recognition (ICPR)*, page (to appear), August 2010.

[11] Yoshinobu Kawahara, Takehisa Yairi, and Kazuo Machida. A kernel subspace method by stochastic realization for learning nonlinear dynamical systems. In *Advances in Neural Information Processing Systems 19*, pages 665–672. MIT Press, 2007.

[12] Jonathan Ko and Dieter Fox. Gp-bayesfilters: Bayesian filtering using gaussian process prediction and observation models. *Autonomous Robots*, 27(1):75–90, 2009.

[13] John Langford, Ruslan Salakhutdinov, and Tong Zhang. Learning nonlinear dynamic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 593–600, 2009.

[14] Wallace E. Larimore. Canonical variate analysis in identification,filtering, and adaptive control. In *Proceedings of the 29th IEEE Conference on Decision and Control*, pages 596–604, 1990.

[15] Neil D. Lawrence. Gaussian process latent variable models for visualisation of high di mensional data. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[16] Lennart Ljung. Perspectives on system identification. In *the IFAC Congress*, 2008.

[17] Zhengdong Lu, Miguel Carreira-Perpinan, and Cristian Sminchisescu. People tracking with the laplacian eigenmaps latent variable model. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1705–1712. MIT Press, Cambridge, MA, 2008.

[18] Peter Van Overschee and Bart De Moor. N4sid: Subspace algorithms for the identification of combined deteministic-stochastic systems. *Automatica*, 31:75–93, 1994.

[19] Vladimir Pavlovic, James M. Rehg, and John Maccormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing Systems 13*, pages 981–987. The MIT Press, 2001.

[20] L. Ralaivola and F. d'Alche Buc. Time series filtering, smoothing and learning using the kernel kalman filter. In *Proc. of IEEE Int. Joint Conference on Neural Networks*, pages 1449–1454, 2005.

[21] Joshua W Robinson and Alexander J Hartemink. Non-stationary dynamic bayesian networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1369–1376. MIT Press, 2009.

[22] Thomas B. Schn, Adrian Wills, and Brett Ninness. Maximum likelihood nonlinear system estimation. In *In Proceedings of the 14th IFAC Symposium on System Identification*, pages 1003–1008, 2006.

[23] Le Song, Mladen Kolar, and Eric Xing. Time-varying dynamic bayesian networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1732–1740. MIT Press, 2009.

[24] Ryan Turner, Marc Deisenroth, and Carl Rasmussen. State-space inference and learning with gaussian processes. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 868–875, 2010.

[25] Jakob Verbeek. Learning nonlinear image manifolds by global alignment of local linear models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(8):1236–1250, 2006.

[26] Vincent Verdult and Michel Verhaegen. Subspace identification of piecewise linear systems. In *Proceedings of 43rd IEEE Conference on Decision and Control*, pages 3838–3843, 2004.

[27] Michel Verhaegen. Identification of the deterministic part of mimo state space models in innovations form from inputoutput data. *Automatica*, 30(1):61–74, 1994.

[28] Jack Wang, David Fleet, and Aaron Hertzmann. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems 18*, pages 1441–1448. MIT Press, 2006.

[29] 井手 剛 and 矢入 健久. 機械学習技術の最近の発展とシステムモデリングへの応用. **計測と制御**, 49(7):(to appear), 2010.

[30] 片山 徹. **システム同定: 部分空間法からのアプローチ**. 朝倉書店, 2004.

Technical Report of the 1st Workshop on Latent Dynamics (Jun 16, 2010, Tokyo, Japan)

潜在トピックモデルを用いたデータマイニング

岩田具治*
Tomoharu Iwata

Abstract: 近年，文書や購買履歴などの離散データを解析する手法として， トピックモデルが注目されている． トピックモデルとは，文書が潜在意味（トピック）に基づいて生成される過程を確率的に表現したモデルである． トピックモデルを用いることにより，多様なデータに内在する隠れた構造を抽出できる．本稿では，基本となるモデルについて解説した後，トピックモデルの応用として，時間変化する購買履歴データの解析のためのトピック追跡モデルを紹介する．

Keywords: トピックモデル，生成モデル，ギブスサンプリング，購買行動解析

1 まえがき

近年，文書や購買履歴などの離散データを解析する手法として，bag-of-words 表現された文書の生成過程を確率的にモデル化したトピックモデルが注目されている．トピックモデルの代表例として，Probabilistic Latent Semantic Analysis (PLSA)[10] や Latent Dirichlet Allocation (LDA)[6] があり，情報検索 [10]，音声認識 [22]，可視化 [15]，画像認識 [25, 17]，推薦システム [11, 13] など，様々なデータマイニング分野に適用されている．トピックモデルの特徴は，一つの文書が複数のトピックの混合として表現されることである．一つの文書が一トピックで表される混合多項分布に比べ，トピックモデルは高い精度で文書をモデル化できることが確認されている [6].

2 トピックモデル

文書 d の出現単語集合を $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ とする．ここで w_{dn} は文書 d の n 番目の単語，N_d は文書 d の単語数を表す．トピックモデルでは，各文書が固有のトピック比率 $\boldsymbol{\theta}_d$ を持ち，単語 w_{dn} は，$\boldsymbol{\theta}_d$ に従いトピック z_{dn} を選択した後，そのトピックに固有の単語分布 $\boldsymbol{\phi}_{z_{dn}}$ に従って生成される，と仮定する．文書集合を学習データとして推定したトピック比率 $\hat{\boldsymbol{\theta}}_d$ は，例えば，類似文書検索や文書分類 [6]，可視化 [12] に用いることができる．また，推定した単語分布 $\hat{\boldsymbol{\phi}}_k$ から，トピック毎に特徴的な単語を知ることができる．具体的には，トピックモデル (LDA) では，文書集合 $\boldsymbol{W} = \{\boldsymbol{w}_d\}_{d=1}^{D}$ は以下の過程で生成される．

(1) For each topic $k = 1, \cdots, K$:
  (a) Draw word distribution,
    $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta)$,
(2) For each document $d = 1, \cdots, D$:
  (a) Draw topic proportion,
    $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$,
  (b) For each word $n = 1, \cdots, N_d$:
    (i) Draw topic,
      $z_{dn} \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$,
    (ii) Draw word,
      $w_{nm} \sim \mathrm{Mult}(\boldsymbol{\phi}_{z_{dn}})$,

ここで K はトピック数，D は文書数，$\boldsymbol{\phi}_k$ はトピック k の単語分布，$\boldsymbol{\theta}_d$ は文書 d のトピック比率，z_{dn} は文書 d の n 番目の単語の潜在トピックを表す．また $\mathrm{Dir}(\cdot)$ はディリクレ分布，$\mathrm{Mult}(\cdot)$ は多項分布を表す．

トピックモデルにおける文書集合 $\boldsymbol{W}$ とトピック集合 $\boldsymbol{Z} = \{\{z_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}$ の完全尤度は下式で表される．

$$P(\boldsymbol{W}, \boldsymbol{Z}|\alpha, \beta) = P(\boldsymbol{Z}|\alpha)P(\boldsymbol{W}|\boldsymbol{Z}, \beta). \quad (1)$$

第一因子は $P(\boldsymbol{Z}|\alpha) = \prod_{d=1}^{D} \int P(\boldsymbol{z}_d|\boldsymbol{\theta}_d)P(\boldsymbol{\theta}_d|\alpha)d\boldsymbol{\theta}_d$ であり，$\{\boldsymbol{\theta}_d\}_{d=1}^{D}$ を積分消去することにより，以下の Polya 分布で表される．

$$P(\boldsymbol{Z}|\alpha) = \left(\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K}\right)^D \prod_d \frac{\prod_k \Gamma(N_{kd} + \alpha)}{\Gamma(N_d + \alpha K)}, \quad (2)$$

ここで $\Gamma(\cdot)$ はガンマ関数を表す．また第二因子も同様に Polya 分布，

$$P(\boldsymbol{W}|\boldsymbol{Z}, \beta) = \left(\frac{\Gamma(\beta V)}{\Gamma(\beta)^V}\right)^K \prod_k \frac{\prod_w \Gamma(N_{kw} + \beta)}{\Gamma(N_k + \beta V)}, \quad (3)$$

*NTT コミュニケーション科学基礎研究所，〒 611-0237 京都府相楽郡精華町光台 2-4，e-mail iwata@cslab.kecl.ntt.co.jp
NTT Communication Science Laboratories, 2-4, Hikaridai, Seikacho, Sorakugun, Kyoto

で表される．ここで V は語彙数である．

トピック集合 $\boldsymbol{Z}$ は，文書集合 $\boldsymbol{W}$ を入力とし，Collapsed ギブスサンプリング [9] を用いることで効率的に推定できる．文書 d の n 番目を生成する単語のトピック z_j，$j=(d,n)$，のサンプリング確率は下式により計算できる．

$$P(z_j=k|\boldsymbol{Z}_{\setminus j},\boldsymbol{W}) \propto \frac{N_{dk\setminus j}+\alpha}{N_{d\setminus j}+\alpha K}\cdot\frac{N_{kw_j\setminus j}+\beta}{N_{k\setminus j}+\beta V}, \quad (4)$$

ここで N_{dk} は文書 d におけるトピック k が割り当てられた単語数，N_{kw} はトピック k における単語 w の出現回数，$N_k=\sum_{k=1}^{K}N_{kw}$，$\setminus j$ は文書 d の n 番目の単語を除いたときの回数もしくは変数を表す．上式は，文書 d でのトピック k の割合と，トピック k での単語 w_j の割合の積で表されている．ディリクレ分布のパラメータ α および β は，不動点反復法 [19] を用いて完全尤度 (1) を最大化することによりデータから推定できる．例えば α は下式で更新される．

$$\alpha^{(\mathrm{new})} \leftarrow \hat{\alpha}\frac{\sum_d\sum_k[\Psi(N_{dk}+\alpha)-\Psi(\alpha)]}{K\sum_d[\Psi(N_d+\alpha K)-\Psi(\alpha K)]}, \quad (5)$$

ここで $\Psi(\cdot)$ はディガンマ関数 $\Psi(x)=\frac{\partial\log\Gamma(x)}{\partial x}$ を表す．全単語に対する潜在トピックのサンプリング (4) とパラメータの最適化 (5) を収束するまで繰り返すことによりモデルを学習できる．文書毎のトピック比率 $\boldsymbol{\theta}_d$ およびトピック毎の単語分布 $\boldsymbol{\phi}_k$ の推定値は下式により計算できる．

$$\hat{\boldsymbol{\theta}}_d=\frac{N_{dk}+\alpha}{N_d+\alpha K}, \quad (6)$$

$$\hat{\boldsymbol{\phi}}_k=\frac{N_{kw}+\beta}{N_k+\beta V}. \quad (7)$$

他の推論手法として変分ベイズ法 [6]，Collapsed 変分ベイズ法 [24]，期待伝搬法 (EP)[20]，パーティクルフィルタ [7] などが提案されている．文献 [2] では複数の推論手法の比較実験が行われている．

3 応用

トピックモデルは拡張性が高く，多様な情報を統合することを可能にする．例えば，著者 [21]，時間 [5, 26, 13, 14]，アノテーション情報 [3, 16] を統合したモデルが提案されている．

トピックモデルの一応用例として，時間発展する購買履歴データのためのトピック追跡モデル [13] を紹介する．トピック追跡モデルを用いることにより，ユーザの興味を予測し推薦システムやパーソナライズド広告に応用できるとともに，トピック毎の流行の時間発展を解析できる．購買履歴データにおけるユーザと商品は，文書データにおける文書と単語に対応する．つまり，時刻 t においてユーザ d が n 番目に購入する商品 w_{tdn} は，ユーザ固有のトピック比率 $\boldsymbol{\theta}_{t,d}$（興味を表す）に従ってトピック z_{tdn} を選択した後，トピック固有の商品分布 $\boldsymbol{\phi}_{t,z_{tdn}}$（流行を表す）に従って生成される．ここで，興味 $\boldsymbol{\theta}_{t,d}$ および流行 $\boldsymbol{\phi}_{t,k}$ は時間依存であることに注意．LDA ではこれらの多項分布パラメータは対称ディリクレ分布から生成されると仮定されているが，トピック追跡モデルではダイナミクスを考慮するために，過去のパラメータに依存するように拡張する．具体的には，興味の平均は，新たなデータが観測されない場合，その一時刻前の興味と同じであると仮定し，以下のディリクレ分布を興味 $\boldsymbol{\theta}_{t,d}$ の事前分布として用いる．

$$\boldsymbol{\theta}_{t,d}\sim\mathrm{Dir}(\alpha_{t,d}\hat{\boldsymbol{\theta}}_{t-1,d}), \quad (8)$$

ここで，平均は一時刻前の興味 $\hat{\boldsymbol{\theta}}_{t-1,d}$，精度（分散の逆数）は $\alpha_{t,d}$ である．精度 $\alpha_{t,d}$ は，直感的には，ユーザ d の時刻 $t-1$ と t 間での興味の一貫性を表す．興味の一貫性はユーザおよび時間に依存するため，精度 $\alpha_{t,d}$ を各ユーザ，各時刻でデータから推定する．精度を逐次推定することにより，変化する興味を柔軟に追跡できるようになる．興味と同様に，流行も一時刻前の興味に依存した以下のディリクレ分布から生成されると仮定する．

$$\boldsymbol{\phi}_{t,k}\sim\mathrm{Dir}(\beta_{t,k}\hat{\boldsymbol{\phi}}_{t-1,k}), \quad (9)$$

ここで平均は一時刻前の流行 $\hat{\boldsymbol{\phi}}_{t-1,k}$，精度は $\beta_{t,k}$ である．

トピック追跡モデルでは，新たに得られた購買履歴データと，過去に推定した興味・流行を用いて，現在の興味・流行を逐次的に推定する．すなわち，過去のデータはモデル推定に不要であり，保持する必要もないため，計算コストと記憶容量を低く抑えることができる．共役事前分布であるディリクレ分布を用いるため，ダイナミクスを考慮しない LDA と同様，Collapsed ギブスサンプリングによる効率的な潜在トピック推論が可能である．またハイパーパラメータである $\alpha_{t,d}$ や $\beta_{t,k}$ は，不動点反復法 [19] により完全尤度を最大化することによりデータから推定できる．

実購買履歴データを用いた実験により，トピック追跡モデルは，従来法に比べ購買行動をより高い精度で予測でき，かつ，大規模データでも効率的に扱うことができることを確認している．

トピック追跡モデルでは，LDA における事前分布に時間依存性を導入することで，時間変化する購買履歴データにも適用可能なように拡張している．文書データ，購買履歴データの以外にも，画像 [3, 8, 25]，ネットワーク [1]，音楽 [27] など，様々なデータでトピックモデル

の有効性が確認されている．その他のトピックモデルの発展として，ディリクレ過程を用いたトピック数の自動推定 [23]，トピック間相関の導入 [4]，トピック階層構造の導入 [18] などがある．

参考文献

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, 2008.

[2] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *UAI '09: Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

[4] D. M. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120, 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] K. R. Canini, L. Shi, and T. L. Gri ths. Online inference of topics with latent Dirichlet allocation. In *AISTATS '09: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009.

[8] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of IEEE Intern. Conf. in Computer Vision (ICCV).*, 2007.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl 1:5228–5235, 2004.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *UAI '99: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[11] T. Hofmann. Collaborative filtering via Gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 259–266. ACM Press, 2003.

[12] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. L. Griffiths, and J. B. Tenenbaum. Parametric embedding for class visualization. *Neural Computation*, 19(9):2536–2556, 2007.

[13] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI '09: Proceedings of 21st International Joint Conference on Artificial Intelligence*, pages 1427–1432, 2009.

[14] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Online multiscale dynamic topic models. In *KDD '10*, 2010.

[15] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 363–371. ACM, 2008.

[16] T. Iwata, T. Yamada, and N. Ueda. Modeling social annotation data with content relevance using a topic model. In *NIPS '09*, pages 835–843, 2009.

[17] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2036–2043, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[18] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 577–584, New York, NY, USA, 2006. ACM.

[19] T. Minka. Estimating a Dirichlet distribution. Technical report, M.I.T., 2000.

[20] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *UAI '02: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[21] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[22] Y.-C. Tam and T. Schultz. Correlated latent semantic model for unsupervised language model adaptation. In *ICASSP '07: Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume IV, pages 41–44, 2007.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[24] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

[25] X. Wang and E. Grimson. Spatial latent Dirichlet allocation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1577–1584, Cambridge, MA, 2008. MIT Press.

[26] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD '06*, pages 424–433, 2006.

[27] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):435–447, 2008.

Decoding in Latent Conditional Models: A Practically Fast Solution for a NP-hard Problem

Xu Sun[†] **Jun'ichi Tsujii**[†‡§]
[†]Department of Computer Science, University of Tokyo, Japan
[‡]School of Computer Science, University of Manchester, UK
[§]National Centre for Text Mining, Manchester, UK
{sunxu, tsujii}@is.s.u-tokyo.ac.jp

Abstract

Latent conditional models have become popular recently in both natural language processing and vision processing communities. However, establishing an effective and efficient inference method on latent conditional models remains a question. Actually, inference in graphical models, even in a linear chain case (the case discussed in this work), is NP-hard. In this paper, we describe the latent-dynamic inference (LDI), which is able to produce the optimal label sequence on latent conditional models by using efficient search strategy and dynamic programming. Furthermore, we describe a straightforward solution on approximating the LDI, and show that the approximated LDI performs as well as the exact LDI, while the speed is much faster. Our experiments demonstrate that the proposed inference algorithm outperforms existing inference methods on a variety of natural language processing tasks.[1]

1 Introduction

When data have distinct sub-structures, models exploiting latent variables are advantageous in learning (Matsuzaki et al., 2005; Petrov and Klein, 2007; Blunsom et al., 2008). Actually, discriminative probabilistic latent variable models (DPLVMs) have recently become popular choices for performing a variety of tasks with sub-structures, e.g., vision recognition (Morency et al., 2007), syntactic parsing (Petrov and Klein, 2008), and syntactic chunking (Sun et al., 2008). Morency et al. (Morency et al., 2007) demonstrated that DPLVM models could efficiently learn sub-structures of natural problems, and outperform several widely-used conventional models, e.g., support vector machines (SVMs), conditional random fields (CRFs) and hidden Markov models (HMMs). Petrov and Klein (Petrov and Klein, 2008) reported on a syntactic parsing task that DPLVM models can learn more compact and accurate grammars than the conventional techniques without latent variables. The effectiveness of DPLVMs was also shown on a syntactic chunking task by Sun et al. (Sun et al., 2008).

DPLVMs outperform conventional learning models, as described in the aforementioned publications. However, inferences on the latent conditional models are remaining problems. In conventional models such as CRFs, the optimal label path can be efficiently obtained by the dynamic programming. However, for latent conditional models such as DPLVMs, the inference is not straightforward because of the inclusion of latent variables.

In this paper, we propose a new inference algorithm, latent dynamic inference (LDI), by systematically combining an efficient search strategy with the dynamic programming. The LDI is an exact inference method producing the most probable label sequence. In addition, we also propose an approximated LDI algorithm for faster speed. We show that the approximated LDI performs as well as the exact one. We will also discuss a post-processing method for the LDI algorithm: the

[1]Technical Report of the 1st workshop on Latent Dynamics (Jun 16 2010, Tokyo, Japan). Materials of this Technical Report are from a published conference paper in proceedings of European association of computational linguistics 2009 (EACL 2009). For more details of the work, refer to "Sequential Labeling with Latent Variables: An Exact Inference Algorithm and Its Efficient Approximation", Xu Sun and Jun'ichi Tsujii, EACL 2009.

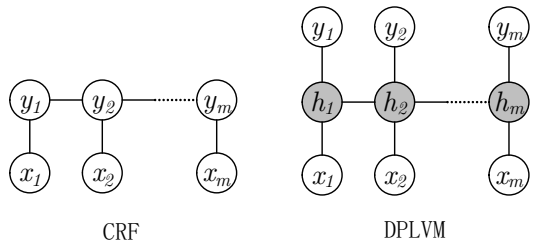

Figure 1: Comparison between CRF models and DPLVM models on the training stage. x represents the observation sequence, y represents labels and h represents the latent variables assigned to the labels. Note that only the white circles are observed variables. Also, only the links with the current observations are shown, but for both models, long range dependencies are possible.

minimum bayesian risk reranking.

The subsequent section describes an overview of DPLVM models. We discuss the probability distribution of DPLVM models, and present the LDI inference in Section 3. Finally, we report experimental results and begin our discussions in Section 4 and Section 5.

2 Discriminative Probabilistic Latent Variable Models

Given the training data, the task is to learn a mapping between a sequence of observations $\mathbf{x} = x_1, x_2, \ldots, x_m$ and a sequence of labels $\mathbf{y} = y_1, y_2, \ldots, y_m$. Each y_j is a class label for the j'th token of a word sequence, and is a member of a set $\mathbf{Y}$ of possible class labels. For each sequence, the model also assumes a sequence of latent variables $\mathbf{h} = h_1, h_2, \ldots, h_m$, which is unobservable in training examples.

The DPLVM model is defined as follows (Morency et al., 2007):

$$P(\mathbf{y}|\mathbf{x}, \mathbf{\Theta}) = \sum_{\mathbf{h}} P(\mathbf{y}|\mathbf{h}, \mathbf{x}, \mathbf{\Theta}) P(\mathbf{h}|\mathbf{x}, \mathbf{\Theta}), \quad (1)$$

where $\mathbf{\Theta}$ represents the parameter vector of the model. DPLVM models can be seen as a natural extension of CRF models, and CRF models can be seen as a special case of DPLVMs that employ only one latent variable for each label.

To make the training and inference efficient, the model is restricted to have disjointed sets of latent variables associated with each class label. Each h_j is a member in a set $\mathbf{H}_{y_j}$ of possible latent variables for the class label y_j. $\mathbf{H}$ is defined as the set of all possible latent variables, i.e., the union of all $\mathbf{H}_{y_j}$ sets. Since sequences which have any $h_j \notin \mathbf{H}_{y_j}$ will by definition have $P(\mathbf{y}|h_j, \mathbf{x}, \mathbf{\Theta}) = 0$, the model can be further defined as:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{\Theta}) = \sum_{\mathbf{h} \in \mathbf{H}_{y_1} \times \ldots \times \mathbf{H}_{y_m}} P(\mathbf{h}|\mathbf{x}, \mathbf{\Theta}), \quad (2)$$

where $P(\mathbf{h}|\mathbf{x}, \mathbf{\Theta})$ is defined by the usual conditional random field formulation:

$$P(\mathbf{h}|\mathbf{x}, \mathbf{\Theta}) = \frac{\exp \mathbf{\Theta} \cdot \mathbf{f}(\mathbf{h}, \mathbf{x})}{\sum_{\forall \mathbf{h}} \exp \mathbf{\Theta} \cdot \mathbf{f}(\mathbf{h}, \mathbf{x})}, \quad (3)$$

in which $\mathbf{f}(\mathbf{h}, \mathbf{x})$ is a feature vector. Given a training set consisting of n labeled sequences, $(\mathbf{x}_i, \mathbf{y}_i)$, for $i = 1 \ldots n$, parameter estimation is performed by optimizing the objective function,

$$L(\mathbf{\Theta}) = \sum_{i=1}^{n} \log P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{\Theta}) - R(\mathbf{\Theta}). \quad (4)$$

The first term of this equation represents a conditional log-likelihood of a training data. The second term is a regularizer that is used for reducing overfitting in parameter estimation.

3 Latent-Dynamic Inference

On latent conditional models, marginalizing latent paths exactly for producing the optimal label path is a computationally expensive problem. Nevertheless, we had an interesting observation on DPLVM models that they normally had a highly concentrated probability mass, i.e., the major probability are distributed on top-n ranked latent paths.

Figure 2 shows the probability distribution of a DPLVM model using a L_2 regularizer with the variance $\sigma^2 = 1.0$. As can be seen, the probability distribution is highly concentrated, e.g., 90% of the probability is distributed on top-800 latent paths.

Based on this observation, we propose an inference algorithm for DPLVMs by efficiently combining search and dynamic programming.

3.1 LDI Inference

In the inference stage, given a test sequence $\mathbf{x}$, we want to find the most probable label sequence, $\mathbf{y}^*$:

$$\mathbf{y}^* = \text{argmax}_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \mathbf{\Theta}^*). \quad (5)$$

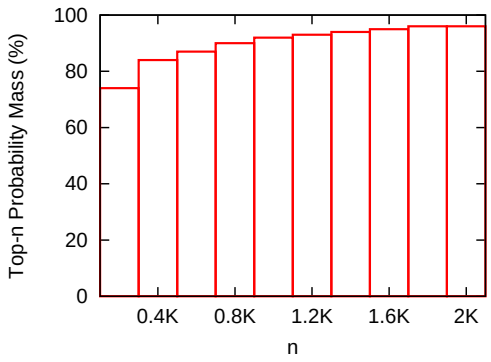

Figure 2: The probability mass distribution of latent conditional models on a NP-chunking task. The horizontal line represents the n of top-n latent paths. The vertical line represents the probability mass of the top-n latent paths.

For latent conditional models like DPLVMs, the $\mathbf{y}^*$ cannot directly be produced by the Viterbi algorithm because of the incorporation of latent variables.

In this section, we describe an exact inference algorithm, the latent-dynamic inference (LDI), for producing the optimal label sequence $\mathbf{y}^*$ on DPLVMs (see Figure 3). In short, the algorithm generates the best latent paths in the order of their probabilities. Then it maps each of these to its associated label paths and uses a method to compute their exact probabilities. It can continue to generate the next best latent path and the associated label path until there is not enough probability mass left to beat the best label path.

In detail, an A^* search algorithm[2] (Hart et al., 1968) with a Viterbi heuristic function is adopted to produce top-n latent paths, $\mathbf{h}_1, \mathbf{h}_2, \ldots \mathbf{h}_n$. In addition, a forward-backward-style algorithm is used to compute the exact probabilities of their corresponding label paths, $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_n$. The model then tries to determine the optimal label path based on the top-n statistics, without enumerating the remaining low-probability paths, which could be exponentially enormous.

The optimal label path y^* is ready when the following "exact-condition" is achieved:

$$P(\mathbf{y}_1|\mathbf{x}, \mathbf{\Theta}) - (1 - \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x}, \mathbf{\Theta})) \geq 0, \quad (6)$$

[2] A^* search and its variants, like beam-search, are widely used in statistical machine translation. Compared to other search techniques, an interesting point of A^* search is that it can produce top-n results one-by-one in an efficient manner.

Definition:
$\mathrm{Proj}(\mathbf{h}) = \mathbf{y} \Longleftrightarrow h_j \in \mathbf{H}_{y_j}$ *for* $j = 1 \ldots m$;
$P(\mathbf{h}) = P(\mathbf{h}|\mathbf{x}, \mathbf{\Theta})$;
$P(\mathbf{y}) = P(\mathbf{y}|\mathbf{x}, \mathbf{\Theta})$.
Input:
weight vector $\mathbf{\Theta}$*, and feature vector* $F(\mathbf{h}, \mathbf{x})$.
Initialization:
$\mathrm{Gap} = -1$*;* $n = 0$*;* $P(\mathbf{y}^*) = 0$*;* $\mathbf{LP}_0 = \emptyset$.
Algorithm:

while $\mathrm{Gap} < 0$ **do**
  $n = n + 1$
  $\mathbf{h}_n = \mathrm{HeapPop}[\mathbf{\Theta}, F(\mathbf{h}, \mathbf{x})]$
  $\mathbf{y}_n = \mathrm{Proj}(\mathbf{h}_n)$
  if $\mathbf{y}_n \notin \mathbf{LP}_{n-1}$ **then**
    $P(\mathbf{y}_n) = \mathrm{DynamicProg} \sum_{\mathbf{h}:\mathrm{Proj}(\mathbf{h})=\mathbf{y}_n} P(\mathbf{h})$
    $\mathbf{LP}_n = \mathbf{LP}_{n-1} \cup \{\mathbf{y}_n\}$
    if $P(\mathbf{y}_n) > P(\mathbf{y}^*)$ **then**
      $\mathbf{y}^* = \mathbf{y}_n$
    $\mathrm{Gap} = P(\mathbf{y}^*) - (1 - \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k))$
  else
    $\mathbf{LP}_n = \mathbf{LP}_{n-1}$

Output:
the most probable label sequence $\mathbf{y}^*$.

Figure 3: The exact LDI inference for latent conditional models. In the algorithm, HeapPop means popping the next hypothesis from the A^* heap; By the definition of the A^* search, this hypothesis (on the top of the heap) should be the latent path with maximum probability in current stage.

where $\mathbf{y}_1$ is the most probable label sequence in current stage. It is straightforward to prove that $\mathbf{y}^* = \mathbf{y}_1$, and further search is unnecessary. This is because the remaining probability mass, $1 - \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x}, \mathbf{\Theta})$, cannot beat the current optimal label path in this case.

A simple proof

Given the *exact condition*

$$P(\mathbf{y}_1|\mathbf{x}, \mathbf{\Theta}) - (1 - \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x}, \mathbf{\Theta})) \geq 0, \quad (7)$$

suppose there is a label sequence $\mathbf{y}'$ with a larger probability,

$$P(\mathbf{y}'|\mathbf{x}, \mathbf{\Theta}) > P(\mathbf{y}_1|\mathbf{x}, \mathbf{\Theta}), \quad (8)$$

then it follows that $\mathbf{y}' \notin \mathbf{LP}_n$, because otherwise it will happen that

$$P(\mathbf{y}'|\mathbf{x}, \mathbf{\Theta}) \leq P(\mathbf{y}_1|\mathbf{x}, \mathbf{\Theta}) = \max_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x}, \mathbf{\Theta}). \quad (9)$$

It follows that

$$
\begin{aligned}
&P(\mathbf{y}'|\mathbf{x},\mathbf{\Theta}) + \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x},\mathbf{\Theta}) \\
&> P(\mathbf{y}_1|\mathbf{x},\mathbf{\Theta}) + \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x},\mathbf{\Theta}) \\
&\geq (1 - \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x},\mathbf{\Theta})) + \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x},\mathbf{\Theta}) \\
&= 1.
\end{aligned}
\tag{10}
$$

Therefore, we have

$$
P(\mathbf{y}'|\mathbf{x},\mathbf{\Theta}) + \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x},\mathbf{\Theta}) > 1, \tag{11}
$$

which is impossible, therefore the assumption of $\mathbf{y}'$ is impossible.

3.2 An Approximated Version of the LDI

By simply setting a threshold value on the search step, n, we can approximate the LDI, i.e., LDI-Approximation (LDI-A). This is a quite straightforward method for approximating the LDI. In fact, we have also tried other methods for approximation. Intuitively, one alternative method is to design an approximated "exact condition" by using a factor, α, to estimate the distribution of the remaining probability:

$$
P(\mathbf{y}_1|\mathbf{x},\mathbf{\Theta}) - \alpha(1 - \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x},\mathbf{\Theta})) \geq 0. \tag{12}
$$

For example, if we believe that at most 50% of the unknown probability, $1 - \sum_{\mathbf{y}_k \in \mathbf{LP}_n} P(\mathbf{y}_k|\mathbf{x},\mathbf{\Theta})$, can be distributed on a single label path, we can set $\alpha = 0.5$ to make a loose condition to stop the inference. At first glance, this seems to be quite natural. However, when we compared this alternative method with the aforementioned approximation on search steps, we found that it worked worse than the latter, in terms of performance and speed. Therefore, we focus on the approximation on search steps in this paper.

References

Phillip Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. *Proceedings of ACL'08*.

P.E. Hart, N.J. Nilsson, and B. Raphael. 1968. A formal basis for the heuristic determination of minimum cost path. *IEEE Trans. On System Science and Cybernetics*, SSC-4(2):100–107.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic CFG with latent annotations. *Proceedings of ACL'05*.

Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. *Proceedings of CVPR'07*, pages 1–8.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07)*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2008. Discriminative log-linear grammars with latent variables. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 1153–1160, Cambridge, MA. MIT Press.

Xu Sun, Louis-Philippe Morency, Daisuke Okanohara, and Jun'ichi Tsujii. 2008. Modeling latent-dynamic in shallow parsing: A latent conditional model with improved inference. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, pages 841–848.

潜在ダイナミクスとしての「都合」

第１回ワークショップ事務局*

Abstract: 人の行動の背景にある意図と、その意図を成就する為の前提制約、そして行動によってもたらされる新たな派生制約をあわせて「都合」と呼んでいる。多様な文書群や発想過程の時系列データを可視化し、背景にある人の様々な関心を読みとる技法は普及に至ったが、そこからサービスや製品を具体化するためには、漠然とした「関心」を、意図と制約から結果へと至る潜在的なダイナミクスすなわち「都合」を捉えなければならない。ここでは、様々な都合の絡み合いに意識を払い続けるチャンス発見プロセスとその効果について話す。

Keywords: Tsugoes, Social constraint network,

1 この障害は何故起きたのか？

ある日、A先生は早く呑みに行きたいという思いに耐え切れず、早々に職場を去ってしまった。実はこの日は午後７時から会議の予定であって、６時ごろからA先生が不在となってしまったせいで、会社に対して様々な迷惑をかけてしまう結果になった。

このような組織活動における障害は、図１の様に様々な意図をメンバーが実現しようとする時に発生する制約間で不整合が起きるために生じてしまうものである。障害そのものは単純であっても、その背景には様々な都合が言語化されないまま潜んでおり、都合間には相互接触が存在しているのである。すなわち都合は、そして都合間の相互作用は、組織活動における様々な現象 – 停滞、紛糾、新事業の展開など- を支配する潜在ダイナミクスに他ならない。

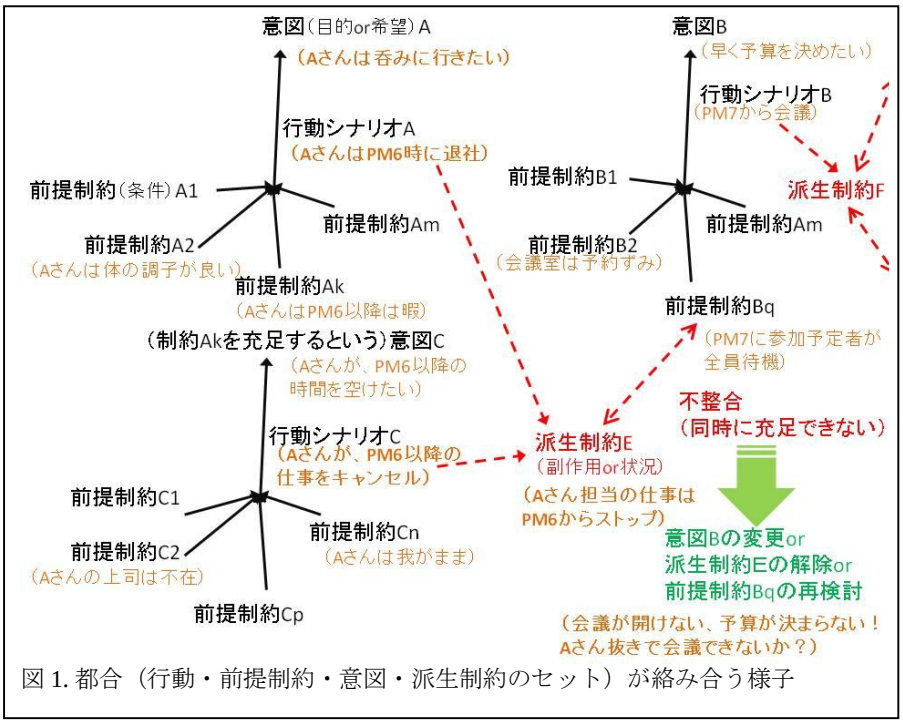

図 1. 都合（行動・前提制約・意図・派生制約のセット）が絡み合う様子

2 「都合」：障害の潜在因子

都合とは、意図と、意図の実行前後における制約からなる複合体である。これが日本特有の概念であることは、「都合」を和英辞典（www.goo.ne.jp）で調べると次のように様々な単語で訳されており、丁度該当する英単語が無いということからわかる。

都合：《事情》circumstances;《便宜》convenience;《機会》(an) opportunity;《理由》reason;《繰合せ》arrangement(s);…

都合という言葉は日本語において例えば、「きょう、ちょっと都合で行けなくなりました」の様にほとんど情報のない修辞として用いられたり、「A先生が大学の都合で、来られなくなりました」のように一部だけ垣間見せる表現をとるために用いられたりする。一般に一つの都合は、次の３つの要素の組で表すことができる。

[意図] 達成しようとする目標と、それに至る行動のおおざっぱなシナリオ。図１でいえば、A先生の意図は早く呑みに行くという目標と、そのために一刻も早く職場を去るというシナリオからなる。

[前提制約] 意図の実現を阻害する可能性のある制約。意図実現のために満たさなければならない条件と、その条件を阻害する状態からなる。図１のA先生の意図は、体調やA先生の当日の予定によって制約されるのである。

[派生制約] 意図の実行により生まれる、（他者又は本人の）他の意図の実行を阻害する制約。ある都合の派生制約が、他の都合の前提制約になることもある。A先生の意図

を実行してしまうと、午後６時からＡ先生は不在となり、会社に対して様々な派生制約を及ぼしてしまう。会社からすれば、これは前提制約が増えたとみることもできる。

実際に人が認知あるいは発現する都合においては、上記の３つ組のうち一部しか表現されないことが多いせいで、都合は潜在してゆくことになる。この潜在性が後から齟齬や不信感を生み、チームワークの手戻りや人間関係の崩壊の原因となることも多い。

言いかえれば、「ご都合主義」と言われるように個別の都合だけを満たす行動では乱されてしまう、多様な都合のネットワークの動作が組織としての集団（あるいはコミュニティ、社会）の活動である。都合に関する表現と認知の不完全性を見直し、改善することによって個人にとっても組織にとっても健全な活動のダイナミクスを取り戻す手法の研究都合学である。該当する英語がないことから、Tsugologyと英訳している（[1,2]）。

3　グループワークの潜在ダイナミクスとしての都合ネットワーク

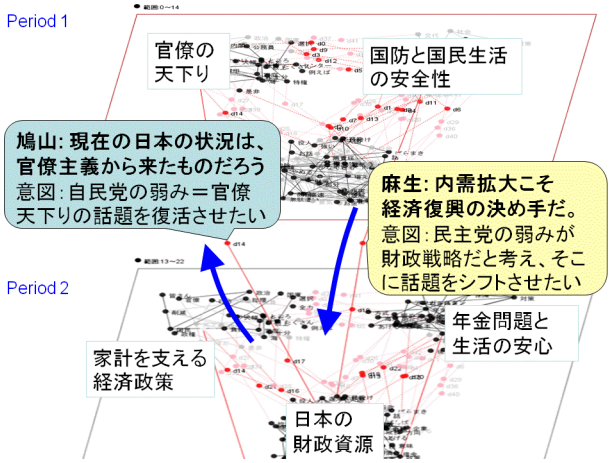

図２　麻生氏と鳩山氏の論戦における意図の構造

図２は、二人の議論者がいて双方が戦略的に言葉にしていない意図が存在しているとき、ある時区間（図で横に繋がっている各平面）と次の時区間の中間に出現するダミーノードとして意図の存在を可視化したものである（[3]と同様の技術と用いた）。このノードの位置を見て、隠された意図や制約をユーザが解釈する。

しかし、このグラフの解釈は必ずしも平易ではない。ダミーノードは数字で名をふられているだけであり、内容に相当する単語はそもそも会話に登場しないためデータに入っていないからである。ダミーノードが意図を表しているのか、前提制約あるいは派生制約を表しているのかが併せて可視化されるだけでもこの問題は大きく改善されよう。

本質的な情報をデータに含むためにデータ収集段階から工夫することが重要であることは、この例からも分かる。ここで必要となるのは人間の営みの背景を探って記録するという人手作業かもしれず、技術・手法と呼ぶには素朴な工夫となるかも知れないが、なおかつ困難な課題であることは事実である。都合学の基本はこの工夫を編み出すことにある。

4　都合ネットワークの可視化から整合的シナリオを生み出すプロセス

原子力施設については市民、事業者、メーカー、政府など様々なカテゴリのステークホルダーが関わっている。この多様なステークホルダーの意図と制約の絡み合いを容易に書き下ろすことは非常に難しいので、各種の技術や高経年化対策案に関わる立案者の意図と制約をカード化し、これらを統合しながら新たな対策を生み出すためのイノゲーションゲーム（Innovators Market Game [4]）を行っている。このゲームの本来の狙いは新規性と有用性を兼ね備えたアイデアの発想であるが、その前に、普段は話合わないような本音の意図と制約について深堀するコミュニケーションが実現でき、衝突する意見も抵抗なく話し合い否定的な意見も前向きに取り入れる場が得られることが分かっている（図 3）。

図 3. 原子力高経年化対策のためのイノベーションゲームの１シーン

同じ企業においても部署間で都合は異なることが多い。例えば、Ｂ社では全社の技術知識を俯瞰するマップを導入しようと提案された。この案を具体化する方法とその意図、そして関連する制約を社員から聴取する調査をＢ社は全社的に行った。ところが、この結果を前提制約、意図、派生制約という３区分で整理したところ、特に派生制約についてはインタ

ビュー形式で調査した場合でさえも回答が少なくなりがちであることが明らかとなった。

これでは、様々な構成員の都合が相互に作用し合うような関係を検討することが困難となる。都合の相互作用は組織全体の動きの停滞と推進の双方を司る隠れた力となるので、正確な都合を調査するためには都合と制約について調査するヒアリングを一層改善する必要がある。製品・サービスの設計における要求獲得法[5-7]の中でも、都合という視点に接近したヒアリング技法[7]が存在し、実現における隠れがちな問題を早期に解決する効果を発揮している。

一方、ここでは回答で欠落した部分を調査者が推察によって補てんすることを試みた。この作業は、図４のように組織的な活動において多様な構成員の間で多様な衝突が生じており、潜在的にどのようなダイナミクスがこのような衝突を支配しているのか不明である場合に、これを客観視する立場から解釈することに当たる。

この結果、回答セットは図５～６のように可視化された（紙芝居 KeyGraph[8]を利用）。図５は同社におけるマップ導入推進者の手で補填された制約・意図の内容を可視化したものであり、図５に表れている単語を見ると、マップの作成における各種の制約とその解決に焦点が当たっていることが分かる。したがって、現時点で推進者の直面している、すなわちマップ導入段階での難しさや問題意識を具体的に表し社内で共有する効果は高いと考えられる。

しかしながら、マップ導入後には利用者にとって価値の高い情報が提供できているかについて社内で評価・批判されることになるであろうから、高価値な情報提供を行うための指針も必要となる筈である。したがって、この指針を考えるために欠かせないチェックポイントはマップを作る前から明らかにしておくべきであろう。

これに対し、図６ではこのマップ導入活動について概要を聞いた外部者（知識経営に関連する専門家）が同様に補填した制約・意図の内容を可視化したものであるが、図６中の単語を見ると、新規市場開発に向けた投資を効果的にするなど導入後のマップの利用者にとっての効果とその発揮方法について考えていることが分かる。

このように、都合記述を調査者あるいは調査協力者が補てんする場合に、補てんする者の視点の持ち方が多様であることを上手に利用すると、多視点からの評価が必要な組織活動を推進する効果的な戦略を捉えることもできるようになる。この例の場合には補てん者の視点を大きく４つに分けることができ

① 利用するに向けたユーザの動機付け
② ユーザの目的に合わせた技術開発
③ ユーザにおける導入
④ ユーザにおける活用

という４フェーズうち別々のフェーズに該当することが分かったので、各フェーズについて改めて調査を行うべきという戦略を得ることになった。

このようにして、組織活動の潜在ダイナミクスである個々の都合と、都合間の絡み合いにおいて齟齬を来さぬように（齟齬によって推進すべき活動が止まってしまう状態もまた、静止という一つの状態を説明するダイナミクスである）解きほぐしてゆくべきポイントを明らかにしてゆくことは、新しいダイナミクスを組織に与えるために大切な役割を果たすことになる。

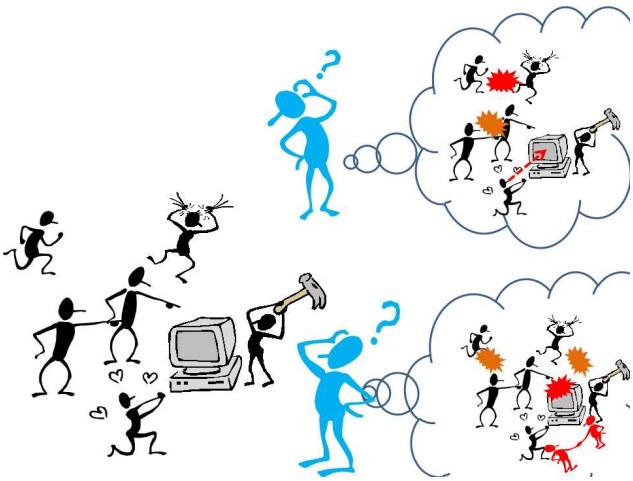

図 4. 組織活動の潜在ダイナミクスを考える人

参考文献

[1] 大澤幸生, 西原陽子他「都合学に取り組む 3 つの理由」人工知能学会第二種研究会・ことば工学研究会資料, SIG-LSE-A903-4, pp. 29--36 (2010) (2010)

[2] Ohsawa, Y., Keynote Lecture: "Tsugology: Structuring of Intentions and Constraints for Innovative Communication" in The 11th International Symposium on Knowledge and Systems Sciences, Xian, China (2010)

[3] Nitta, K, Zeze, K., et al, "Scenario Extraction System Using Word Clustering and Data Crystallization“, *Proceedings of Juris Informatics 2009* (JURISIN 2009)

[4] Yukio Ohsawa, Innovators Market Game as a Tool of Chance Discovery, in Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (LNCS 5908), Springer, pp.59-66 (2009)

[5] Carrol, J.M., "Making Use:Scenario-based design of Human-computer interactions", The MIT press (2000)

[6] 熊澤努, 玉井哲雄「ユースケースによる安全性要求分析のための想定外シナリオ抽出法の提案」, 情報処理学会研究報告 2007-SE-155, 191-198 (2007)

[7] 久代, 大澤「多次元ヒアリングと階層的要求統合プロセスによる要求獲得手法」, 情報処理学会論文誌, 47 (10), 2909 - 2916 (2006)

[8] Ohsawa, Y., Ito, T., and Itakura, K.M., Kamishibai KeyGraph: Tool for Visualizing Structural Transitions for Detecting Transient Causes, *New Mathematics and Natural Computation* 6 (2), pp.1-15 (2010)

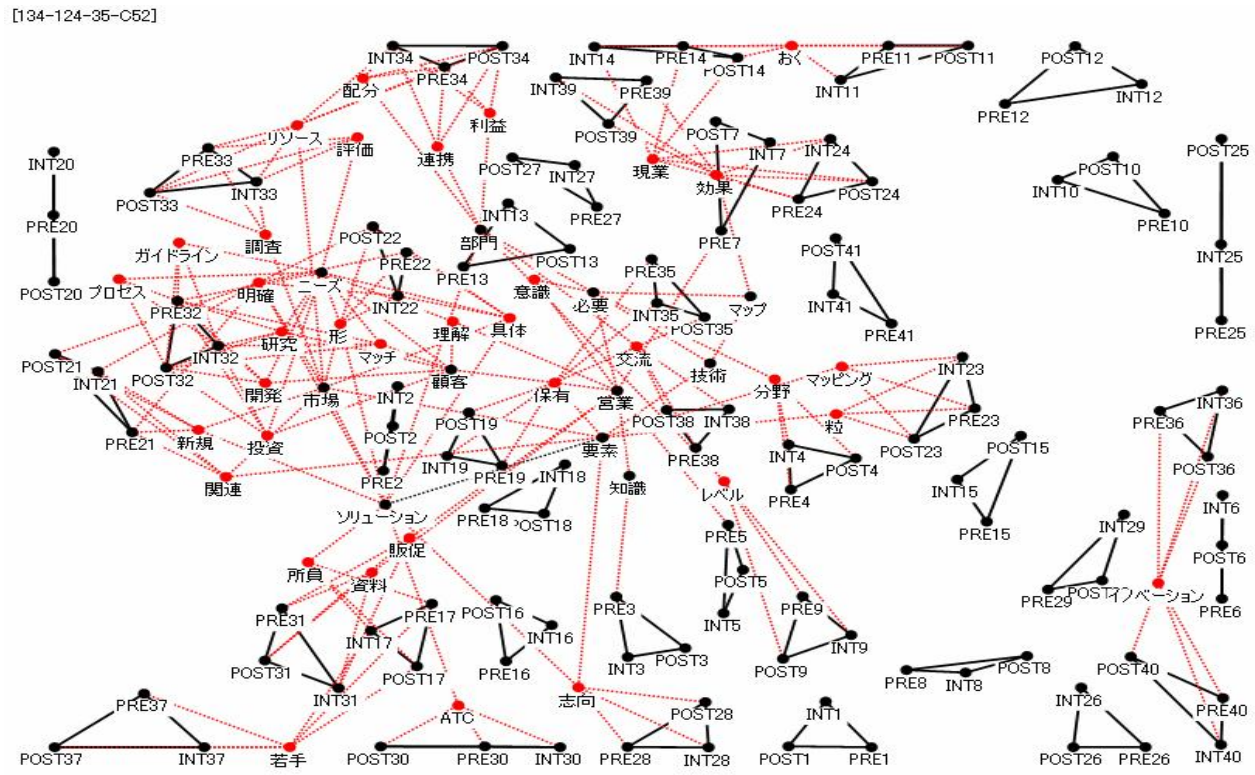

図 5. A社における技術知識マップ導入について社内構成員らの都合（希望する行動、前提制約、意図、派生制約）を聴取し、後で記述不足部分を推進者が補てんした内容を可視化したキーグラフ

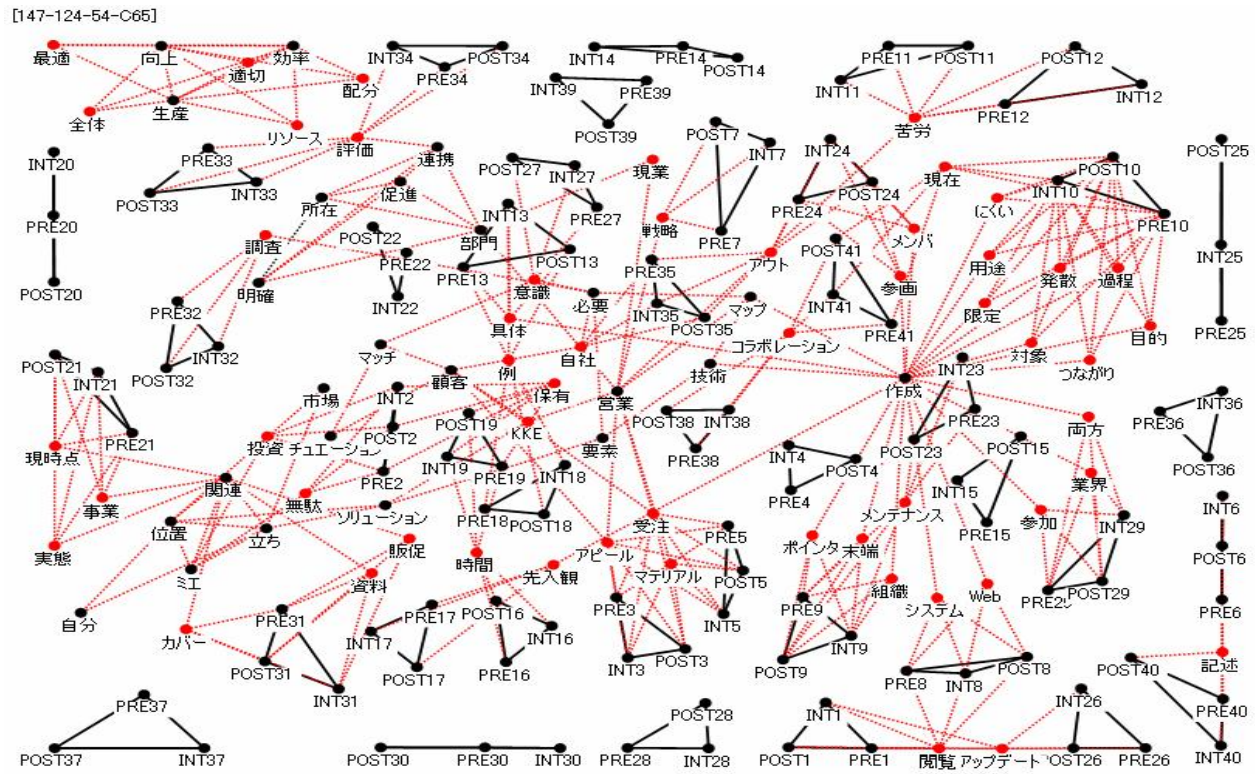

図 6. 図 5 と同様に行った都合調査に、社外者が行った補てん結果を可視化したキーグラフ

トラブルの経験と情報としての活用技術

宮野　　廣*
Hiroshi MIYANO

Abstract: 原子力発電所では、多くのトラブル経験をしている。これらの経験は、すでに国、事業者、メーカなど個々の技術団体はもちろんNUCIA原子力施設情報公開ライブラリーとして蓄積されており、運転管理に役立っているものと考える。この情報の活用は、技術者個人の技量に依存して知恵として活用されているものである。このようなものは、すでに世界にも多くもあり米国の情報ライブラリーなどの情報も日本に提供されるようになっている。それらを統合して役立つ情報基盤を構築しようとの動きもある。このように多くの情報、知識の基盤をすでに有しているが、十分に活用されているかについては課題がある。わが国の原子力界では、醸成された情報をさらに有効に活用して「原子力安全」を確保する仕組みを構築することが今求められている。これまでのトラブル対応での経験や上記の状況を踏まえて、これからの情報の活用の在り方の一つとして構築すべき情報基盤を提案する。

Keywords: Information Infrastructure, Database, Information, Knowledge, Wisdom.

1　はじめに

わが国に原子力発電所が導入されてから約 40 年になる。今年、運転 40 年目を向かえるプラントのこれからの運用の是非を確認する高経年化対応の技術評価が実施された。導入初期から大小数々のトラブル、不具合を経験してきたが、このような仕組みに集約されたことは、対応してきたことが成果として確認された感を覚える。

これらの経験は、以下に示すように国内ではデータベースとして様々な形で残される仕組みができ、整備されてきている。

（公開データ）

- 国の審議会での検討資料類
- 原子力施設運転管理年表（JNES 発行）
- 原子力施設公開情報ライブラリー（NUCIA）
- 日本原子力学会標準「高経年化対策実施基準」AESJ-SC-P005 のデータベース（PLM 標準）
- 国内外の学術講演会の論文、講演集など

（非公開データ）

- 電気事業者のデータベース（電事連、原子力技術協会のデータ）
- プラントメーカのデータベース
- 機器メーカ・ベンダーのデータベース

一方、国際的には原子力学会で策定したPLM標準のデータベースを受けてIAEAにおいては、データベースの国際化が進められ、参照事例（C0mmendable Practice）として集約されつつあるーiGALL: International Generic Ageing Lessons Learned Guideline）。また、個々の専門分野においてはSCC−応力腐食割れ事象やケーブルの劣化事象での国際データベースの構築作業にも取り組まれている。

このようにデータベースの基本は、トラブルデータの蓄積にあると言っても過言ではない。すなわち、トラブルの経験を生かして、運用中のプラントでは同様のトラブルを起こさないようにするためであり、新規設計においては、同様のトラブルが起きない設計とするためにデータベースを構築、活用するものである。しかし、トラブルの経験は、設計の基準が策定されて初めて生かされるものである。例えば“もんじゅの熱電対の 流力振動による損傷”への対応については、日本機械学会の「配管円柱状構造物の流力振動評価指針」JSME S012-1998 の制定により、また“敦賀 2 号機での再生熱交換器の高サイクル熱疲労損傷事故”への対応については、「配管の高サイクル熱疲労に関する評価指針」JSMES017-2003 の制定により、これらの事象は設計時に配慮される仕組みに組み込まれるようになった。

このように現状でも、過去のトラブルの経験は情報として十分に整備される仕組みが構築されている。しかし一方、個々の技術者にとっての技術の習得は、個々の技術者のトラブルの経験によるところが大きく、単に知識としての情報では、情報そのものの不十分さと同時に、技術として感覚的に捉えることができず、使える知識になりにくい面がある。知識とは、簡単には実際の設計や評価、検討に即座に生かされるものとはならないと言うことである。汎用的な知識として周知されたものになって初めて経験の共有が図れると言うことであろう。設計者や技術者の技量、センスにより規格基準を参照するだけで、劣化事象を予測して、それに対応することは、極めて難しいものである。

2　情報の知識化に向けてーその課題

トラブルの情報の流れの例を図 2-1 に示す。その上で、情報の問題点を提示する。この問題点は、特に原子力発電に係わることから生じているものではなく、一般的にトラブルの対応として指摘されるであろう点を示した。

- トラブルの発見者によるが、少なからず事態の状況は脚色される
- 原因の同定においては、判断ミスや間違いで発生したのではなく、予測不可能な事態により偶発的に発生した、止むを得ないものに落とし込みがちになる
- 他への波及を押しとどめる無意識の判断が働く
- 全ての情報が網羅されない
- 予測シナリオが優先され、それに係わる情報のみが残る可能性がある
- 検討の過程が省略され、結論のみが残りがちになる

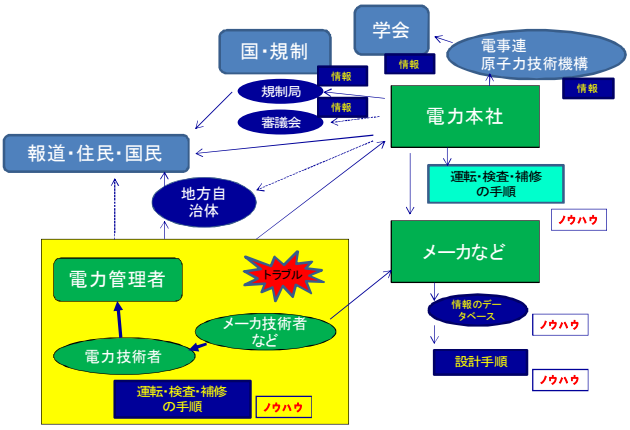

図 2-1　情報の伝達・流れと蓄積

従って、どのように生のデータを取り込み、将来の第三者の分析や新たな知見を加えた判断ができる状況を残しておくか、工夫が求められる。

更に、トラブルの経験を生かして一般的に規格基準化とする場合には、様々な状況を踏まえた判断ができない、また複合事象を判断するような基準化は難しい、という課題もある。

米国ワシントン州のタコマ橋の「風による振動で一瞬のうちに破壊してしまったこと」は有名であり、多くの技術者に流力振動の怖さを知識として植え付けている。しかし、このような情報だけで、知識とするには難しい事例もある。例えば“もんじゅの熱電対の流力振動による損傷”の場合（図 2-2　参照）、既に ASME の規格においても同様の振動が発生する恐れは参考として指摘されてはいたものではあるが、専門家の知識として形式化されてはいたものの、一般的には十分な経験がなく、一般の技術者の間では知識として形成されてはいなかったものであった。いわゆるカルマン渦による流力振動については、多くの人の知るところであり、一般的に技術者が知恵として活用できるまでに昇華したものとなっているが、“もんじゅ”の熱電対で発生したいわゆる双子渦による振動については、過去の事例では大型海洋構造物で発生した振動例しか報告されたものはなく、熱電対のような細い小さな構造物が振動で破損してしまうことは、机上では容易に予測することができない事象であった。もちろん流力振動の専門家は感覚的にそのような設計は避けたであろうが。従って、専門家の関与しない設計においては、このような広く認知されていないような事象については、どこかに参考となる基準があったとしても、十分に注意しなければならない。しかし、このような詳細な検討が必要であるという認識を共通して持つことは難しいものである。この事例は、知恵というのは個人の技量の問題だけではなく、知識を生かす仕組みにあると言うことを的確に示した例であったと言える。

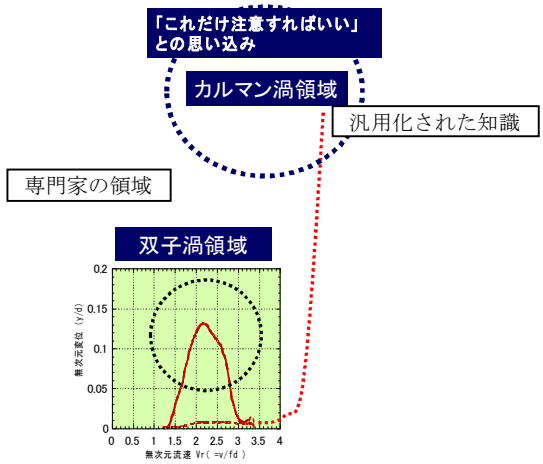

図 2-2　情報のレベル事例（カルマン渦と双子渦）

これまで構築してきた情報基盤は、図 2-3 のような基本的にはデータベースであり、得られたものの単なる蓄積である。それが報告書の形となっているに過ぎないものか、もしくは知識の溜め込みであり、体系的に活用できるものとはなっていないものが多い。では、どのような情報基盤を構築すればいいのか。

図 2-3　現状での情報の集約

3　情報の知識化に向けて－情報基盤の構築

情報基盤は、図 3-1 に示すような構造化された体系を持つと考える。様々な生のデータから情報が構築され、情報が汎用化されることで知識となる。知識が獲得されて、頭脳で様々に組み合わされ知恵となって、実用に活用される。それが情報基盤の全体構成である。

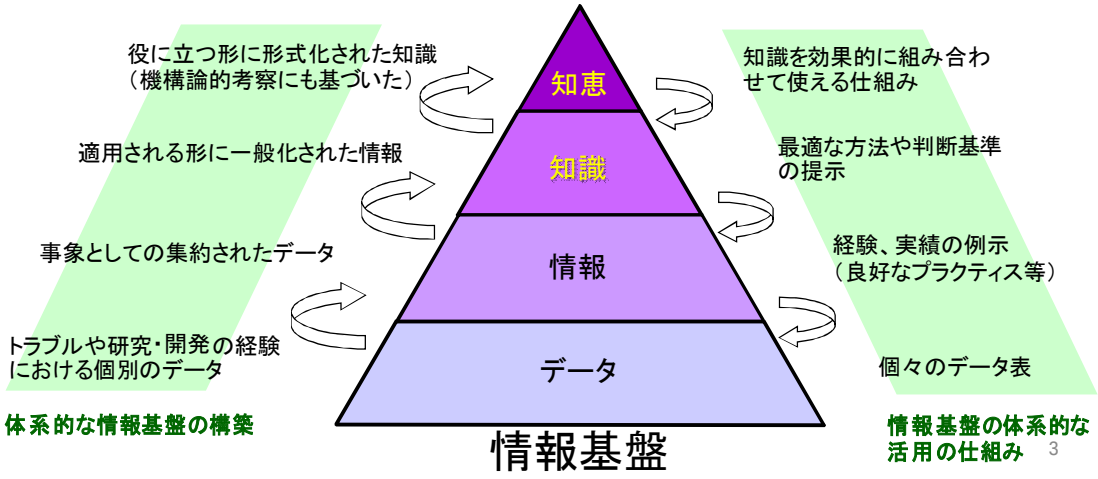

図 3-1　情報基盤の構成

その中で課題となるのは、知識の集約である知恵とする部分であり、どのように「知恵の構築」をすればいいのかである。またそれぞれの「要素をつなぐ仕組み」であり、データから情報、知識、知恵とつなぎ知恵として作り上げる仕組みと構成された知恵から知識に落とし込み、管理する情報、データに落とし込む仕組みを作り上げることが、この“情報基盤”の最も大きな課題である。

4　情報基盤、その体系

情報基盤の一つの解の構成を以下に提示する。

トラブル情報は、“保全”を適切に行うために必要な情報である。「事故」、「故障」と言われる不具合が明確な場合には事故対応として元に復する行為を指し、また通常の劣化部品を取り替えるなどの行為や劣化部分の補修、修繕などの行為を指すものである。従って、トラブルという事象には、設計、建設時の条件やデータを含めて扱うということが含まれ、それらを含めて「保全」という。公開で残されている情報は、顕在化したトラブル対応の情報が基本であることから、情報基盤の対象としてはトラブル情報を基本としてきた。しかし、上述のように、「保全」の行為そのものが、通常の劣化も扱うとすると、劣化事象全体を情報基盤構築の対象の事象とすることが望ましいと考える。特に、最近では材料や機器の劣化、損傷などの現象や状態は、経験や研究により、その多くが捉えられるようになってきており、また計測の精度がよくなり、その劣化の状態を定量的に扱えるようになってきたことから、多くの劣化事象を精度良く捉えることができるようになったと考える。

一般に装置や設備、発電プラントを設計し、製造・建設して、運用する流れを図 4-1 に示す。

“もの”を作る基本は、まずニーズ、社会的要請からスタートする。人や社会が何を欲するか、これが全ての根本であり、これに技術的な条件と経済的な条件が加わり、作られる“もの”の基本仕様が定まる。

これに事業性が加味され、コスト、納期などの仕様が決められ、総合的な基本仕様となる。“もの”はこれに従い設計、製造、建設されることになる。

この“もの”設計、製造、建設においては、様々な技術的、社会的に様々な制約が与えられている。いわゆる規制や規格基準というものである。規制や規格基準は、技術的、工学的な理論やデータにより裏付けられることは言うまでもない。しかし、第一には社会に受け入れられるものでなければならない。社会からの受容性には、経済的な要素も含まれる。技術的事項だけではなく、人文･社会科学的な観点からの検討が求められる。トラブル情報が反映されるのは、単なる技術的な判断の見直しではない。トラブルには社会的な判断や経済的な判断を含んだものが多く、総合的な見直しができる。

科学や工学には“決まり”や“関係”と言うものがある。“決まり”や“関係”には、法則や理論がある。法則とは、一つの、もしくは複数の簡単な式などの関係により現されるものであり、それらの関係を複雑だが定まった論理体系により現したものを理論と言う。このような関係は、産業における工学

の取り扱いや技術の確立においても、同様の定まった“関係”で成り立ち、人との関わりの社会においても、このような論理体系に加えて、個人、企業、社会の判断や、時間的、経済的効果などの要素も加わり関係が成り立つ。それらが統合されて、法則や理論のような“決まり”が、規制や規格基準として定められる。これら規制や規格基準とは一律なものではなく、上述のような様々な要素により定まり、国や時代、また“決まり”の対象によっても様々に違ったものとなる。

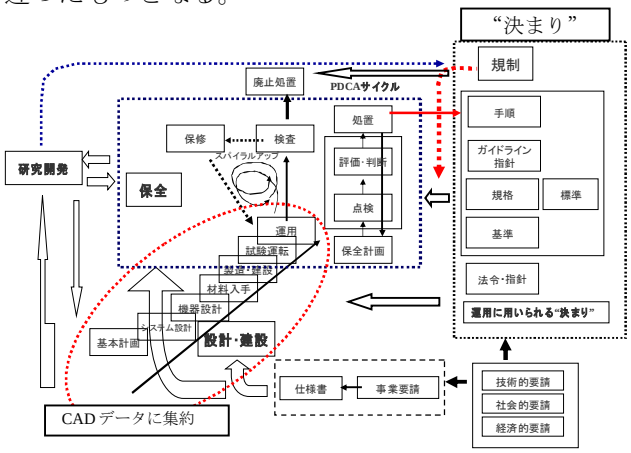

図 4-1　プラントの設計・建設・運転・保全の流れ

図 4-1 は、原子力発電プラントのような発電用設備などのプラントの設計・建設から運転、検査、保修の保全の循環に至るまでの流れを示している。どの工程においても、要求される仕様のほか、公的な法令・指針などの規制や手順書からガイドライン、規格、標準、基準などの、民間での自主規制を含めて様々な“決まり”を定めて、各工程の作業を支援している。すなわち設計・建設から運用・保全までのそれぞれの役割を分担する人の作業が、安全に、要求される機能を満足する品質を確保し、適正なコストで行えるように、ガイドするもので、この作業への要求を技術的要請や、社会的要請、経済的要請に答えるように“決まり”として与えることで、社会として安定した管理が行えるようにする「社会の仕組み」を構築しているのである。

この各工程に適用される“決まり”、すなわち規制や規格基準というものは、技術の進展や経験の蓄積により変化するものであり、また研究開発により大きく変わる。また常に向上させようとするものであり、これは社会情勢や経済情勢などにも大きく左右される。従って、この“決まり”にはこれといった普遍的なものはなく、上述の技術、社会、経済の３要素の変化により、最適化するように、これらの“決まり”も変化するものである。

*法政大学　大学院ｼｽﾃﾑﾃﾞｻﾞｲﾝ研究科　客員教授,
〒102-8160　東京都千代田区富士見 2-17-1,tel. 090-4072-2679,
e-mail:hiro.miyano@hosei.ac.jp

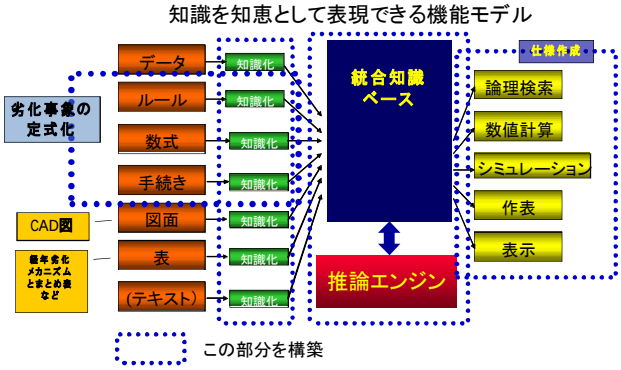

図 4-2　知識の機能モデル例

トラブル情報、すなわち劣化事象の情報をどのように知識化し、知恵として使える形にするかについて、一つの提案を行った。図 4-2 に示すように、知識ベースとして構築するものは、個別のデータの知識化のルールであり、その利用方法のルールである。頭脳としての推論エンジンがそれらを運用する。すなわち、①“要請”に答える、②規制・基準の決まりに従う、③CAD データを活用する、④劣化事象を選択する、⑤劣化予測（シミュレーション）を行う、⑥健全性を判断する、この評価を個別に行い、それらを組み合わせて最終判断する仕組みを構築するものである。それぞれに選択肢があり、「どれを選択することが妥当」か、もしくは「どれを選択すれば、結果がどうなる」、ということを提示できるようにすることも必要であろう。

5　おわりに

トラブルの少ない時代になってきた。このような時代では情報の知識化、知恵化が重要となってくる。「体感できないトラブルをどのように体感し、身につけていくか」は大きな課題である。一方、時代は高速コンピュータの時代である。多くの情報を適切に処理することで、単なるデータが知識となり知恵とすることができる時代である。全ての劣化事象をシミュレーションできる仕組みを提案した。多くの記述的知識を予め組み込み知識化しておくことで、データとして取り込んだ手続き的知識（ノウハウ）が知恵として活用できるようになるものと考える。

これからの開発に期待するものである。

参考文献

[1] 宮野　廣、保全学の構築と体系化-保全活動と規格基準，Vol.6,No.1,April,2007,JSM

Technical Report of the 1st Workshop on Latent Dynamics (Jun 16, 2010, Tokyo, Japan)

揺らぎと偏りから読み解く潜在構造

前野 義晴*
Yoshiharu Maeno

Abstract: 成長しながら伝播し拡散する特色を持つ多様な社会・物理現象を論じる．理論と観測から，このような現象を支配する隠れたメカニズムを理解する新たな方法を述べる．

2009 年にメキシコから世界じゅうに広がった新型インフルエンザは記憶に新しい．日本でも多くの感染者が出て，社会問題にもなっている．2010 年に入っても，危機が去ったとは言えない状況にある．2003 年には，重症急性呼吸器症候群（SARS）のウイルスが，瞬く間に中国から世界じゅうに広がった．このような感染症の大流行は有史以来数多くあり，不幸にして 100 万人規模の死者が出た例さえある．現代の感染症の大流行は，さらに危険なものだ．というのは，各大陸の多くの都市に張り巡らされたグローバルな航空機のネットワークを介して，数日の内にウイルスが世界じゅうに広まってしまうのだ．潜伏期間の長いウイルスであれば，ホストの人間が発症する前に，ホストと共に感染の発生していない地域へ移動してしまう可能性が高まる．地球の裏側からでも飛んでくるグローバル時代のウイルスは，我々の日常生活を脅かす，特に身近なリスクと言えよう．疫学では，感染症が発生した地域から隣接する地域への拡散をシミュレーションで再現して，拡散を封じ込めるために有効な対処法を研究し各国当局の対応に役立てている．

一般に，このようなシミュレーションでは，地域間の人の移動量や地域内での人と人との接触頻度についての統計的な知見が重要な役割を果たす．しかし，人の移動量や人と人との接触頻度についての実測値や統計的な知見が存在するとは限らない．たとえ存在していても，精度の不充分さや最新の実情との乖離が懸念されることが多い．そのため，往々にして，シミュレーションでは再現できない，遠隔地への思わぬ飛び火が起こったりする．例えば，2004 年に発表されたハフナゲルの研究は，世界の主要都市間の旅客機ネットワークの輸送量にもとづくシミュレーションによって，SARS の感染拡大をおおむね再現できることを示した．しかし，香港で感染者が急増し始めた 2003 年 2 月 19 日から 90 日後の 5 月 20 日の予測値を見ると，実際の値からかなりはずれている場合がある．シミュレーションでのカナダの感染者数（の累積値）の期待値は 42 人・最大値は 107 人と予測されたが，実際の感染者は 140 人で，10 日後の 5 月 30 日には 188 人に達している．逆に，日本の感染者数の期待値は 60 人・最小値は 27 人と予測されたが，実際に感染者は発生しなかった．オランダやバングラディッシュでも感染の発生が予測されたが，実際に感染者は発生しなかった．シミュレーションの予測は，大きく誤る場合がある．つまり，地域間の人の移動量や地域内での人と人との接触頻度を既知とはできない場合があることを頭に入れておく必要がある．むしろ，これらを推定する問題を避けて通れないということだ．

地域をノード，人の移動をリンクとして問題を抽象化した上で一般化すると，ネットワークのリンクに沿ってノードからノードへ伝播しながら成長する現象について，ノードで観測されたデータから，直接的に観測できないネットワークのトポロジや伝播，拡散，成長に係わるパラメータの大きさを推定する問題となる．数学的に，不均一な空間での反応拡散過程という視点で見てみると，理論的にも興味深い問題である．ウイルスの拡散に留まらず，人・モノ・金・情報の流れを理解するには，その背後にある未知のネットワークの解明が重要な鍵となる．地球上の物理的な空間だけでなく，サイバースペースでの情報の流れを理解する際にも有用であろう．

*Yoshiharu Maeno, Ph.D. is a founder management consultant and scientist at Social Desing Group, and a principal researcher at NEC Corporation. E-mail maeno.yoshiharu@socialdesigngroup.com.

参考文献

[1] Y. Maeno: Profiling of a network behind an infectious disease outbreak, *e-print* http://arxiv.org/abs/0905.3582.

[2] Y. Maeno: Node discovery in meta-population network behind infectious disease outbreak, *e-print* http://arxiv.org/abs/1006.2322.

Technical Report of the 1st Workshop on Latent Dynamics
(Jun 16, 2010, Tokyo, Japan)

潜在的グラフ構造からの異常検知

井手剛*
Tsuyoshi Idé

Abstract: 潜在世界のダイナミックスを異常検知などの実応用につなげる時に問題になるのが、潜在構造の安定性という問題である。もし同定した潜在構造がわずかなデータの揺らぎにより、あるいは、反復アルゴリズムの初期値のような非本質的なパラメターによりがらりとその様相を変えるとしたら、見出された潜在構造が実世界の何らかの反映だと主張することは難しい。本稿では、共分散構造解析の限界を打破したとして一躍有名になった Meinshausen-Bühlmann 理論が、やはり多重共線性の下で困難を持つことを指摘し、それへの対処策について検討する。

1 Introduction

ネットワークやグラフからの知識発見は、データマイニングにおける最近の中心的な課題のひとつである。従来多くの研究は、グラフ構造もしくはそのデータベースを所与とし、それに対して何らかの機械学習的なタスクを行うことに注力きたが、ここ 2-3 年、グラフ構造の学習のための技法が急速に発展している。実用上の要請を考えても、グラフに対する詳細な知識が事前に得られることはむしろまれであり、グラフ構造それ自体をいわば潜在構造として扱い、潜在構造の学習もまた問題の一部であると捉える方が多くの場合自然である。

我々のグループではこれまで、変数間の依存関係が強い状況での、複数のセンサーデータからの異常検出・解析という問題に取り組んできた [9, 8, 12, 11, 10]。実用上、ノイジーなセンサーデータからの異常発見問題においては、以下のような条件が要請されることが多い。

1. データの非定常な変動に対応できること。特に、ノイズによる値のぶれに頑強であること。
2. システムのモジュール構造、もしくは変数のヘテロ性に対応できること。
3. 系全体が異常か否かのみならず、どの変数がどの程度異常かの情報が得られること。

我々は上記のような要請を満たす手法として、グラフィカル・ガウシアン・モデル（GGM）のスパース構造学習に基づく異常検知の手法を最近提案した [10]。我々の貢献のひとつは、上記のような実用上の要請の下で、望ましい構造学習の手法を同定したことである。とりわけ、伝統的な共分散構造解析の限界を打破したとして一躍有名になった Meinshausen-Bühlmann（MB）の算法 [14] が、実際上非常に不安定な結果を与えることを指摘し、Friedman らが提案した [7] のグラフィカル Lasso（以下 gLasso と呼ぶ）と呼ばれる算法が、実用上非常に優れていることを実験的に示した。

本稿では、GGM および MB の算法、gLasso などの理論を比較的詳細に解説し、多重共線性の下での構造不安定性について実験結果を紹介する。

図 1 に我々の問題設定をまとめておこう。ノイジーな多変量のセンサーデータを想定し、複数箇所で窓を取る。問題を簡単にするため、システムの正常稼動時 A と、異常が疑われる状況 B という 2 つの窓でデータを観測したとする。すなわち、2 つのデータセット

$$\mathcal{D}_{\rm A} \equiv \{\boldsymbol{x}_{\rm A}^{(n)} | \boldsymbol{x}_{\rm A}^{(n)} \in \mathbb{R}^M, n = 1, 2, ..., N_{\rm A}\}$$
$$\mathcal{D}_{\rm B} \equiv \{\boldsymbol{x}_{\rm B}^{(n)} | \boldsymbol{x}_{\rm B}^{(n)} \in \mathbb{R}^M, n = 1, 2, ..., N_{\rm B}\}.$$

が与えられたと考える。我々は主にセンサーデータに興味を持つので、インデックス n は典型的には時刻を表す離散値に対応する。$\mathcal{D}_{\rm A}$ および $\mathcal{D}_{\rm B}$ において、測定の回数 $N_{\rm A}$、$N_{\rm B}$ は一般には異なっていてもよい。データセット $\mathcal{D}_{\rm A}$ と $\mathcal{D}_{\rm B}$ が与えられた時、それぞれのデータにおいて変数間の依存関係を表すグラフの相違にどれだけ寄与したかを表す異常度を、各変数について計算せよ、というのが我々の問題である。この問題は統計学における 2 標本検定の問題と似ているが、知りたいのが個々の変数のスコアであるという点で異なる。

*IBM 東京基礎研究所, 242-8502 大和市下鶴間 1623-14（LAB-S7B）, e-mail: goodidea@jp.ibm.com,
IBM Research – Tokyo, 1623-14 Shimo-Tsuruma, Yamato-shi, Kanagawa 242-8502, Japan
Submitted Jun 15, 2010, revised Jun 17, 2010.

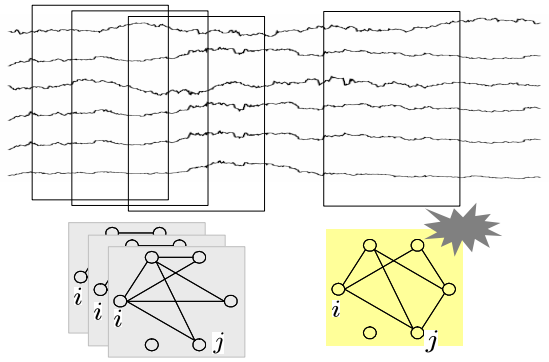

図 1: 問題設定。ノイジーなセンサーデータにおいて、正常稼動時と目される状況と、異常が疑われる状況のそれぞれに対して、(1) その共分散行列に基づいて疎なグラフを学習する。 (2) 次に、その 2 つの疎なグラフを比較して**それぞれの**変数の異常度を求める。

ここで、物理系にはよくあるように、一般に測定系はある冗長性があり、従って、多重共線性は前提と考える必要があることに注意されたい。また、データはノイジーで定常性を持たず、状態空間モデルを用いた時系列モデリング（これはある意味で Latent Dynamics を考える王道である）は簡単ではないことにも注意されたい。今のところ我々はシステムのダイナミックスを明示的に取り入れることはせず、滑走窓の形で系の非定常性を取り込む。

2 グラフィカル・ガウシアン・モデルの構造学習

本節では図 1 におけるステップ 1、すなわち、データからいかに疎なグラフを学習するかについて考える。このステップはデータ A と B に共通なので、以下しばらく両者を区別する添え字を落とし、どちらかを表すデータを、$\mathcal{D} = \{\boldsymbol{x}^{(n)} | n = 1, ..., N\}$ と書くことにする。データ $\mathcal{D}$ における M 個の変数はそれぞれ、平均ゼロ、標準偏差 1 に標準化されていると仮定する。この仮定の下、標本共分散行列 S は

$$\mathsf{S}_{i,j} \equiv \frac{1}{N} \sum_{n=1}^{N} x_i^{(n)} x_j^{(n)} \tag{1}$$

のように与えられる。これはデータの相関係数行列と同じものとなる。

2.1 精度行列と条件付き独立性

グラフィカル・ガウシアン・モデル（GGM）で考えるグラフは、M 個の変数のそれぞれを頂点とするグラフである。一般に、グラフィカル・モデルにおいて、頂点（もしくは変数）x_i と x_j をつなぐ辺が欠けている時、両者は、他のすべての変数を固定した時に条件付き独立である。逆も真である。頂点間の辺の有無を定義するために、GGM では次の M 次元正規分布

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \Lambda^{-1}) = \frac{\det(\Lambda)^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \Lambda \boldsymbol{x}\right) \tag{2}$$

を考える。ここで、det は行列式、$\Lambda \in \mathbb{R}^{M\times M}$ は精度行列を表す。$\mathcal{N}(\cdot|\boldsymbol{\mu}, \Sigma)$ は平均 $\boldsymbol{\mu}$、共分散行列 Σ の正規分布を表す記号である。先に述べたように、精度行列は共分散行列の逆行列である。

正規分布の仮定の下、x_i と x_j をつなぐ辺を欠く条件は下記のように書かれる。

$$\Lambda_{i,j} = 0 \quad \Rightarrow \quad x_i \perp\!\!\!\perp x_j \,|\, \text{other variables} \tag{3}$$

ここで $\perp\!\!\!\perp$ は統計的独立を示す。この条件 (3) は条件付き分布を明示的に書き下すことにより容易に理解することができる。これを以下示す。$(x_i, x_j)^\top$ をまとめて $\boldsymbol{x}_a$ と表し、これら以外の変数をやはりまとめて $\boldsymbol{x}_b$ と表しておく。中心化されたデータに対して、正規分布のよく知られた分割公式（例えば [2] の Sec. 2.3 参照）を用いて、求める条件付き分布は

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a| - \Lambda_{aa}^{-1}\Lambda_{ab}\boldsymbol{x}_b, \Lambda_{aa}^{-1}) \tag{4}$$

のようになる。ここで、$\boldsymbol{x}_a$ と $\boldsymbol{x}_b$ の分割に対応して、

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \tag{5}$$

と置いた。この場合、Λ_{aa} は 2×2 行列に過ぎないから、その逆行列は容易に求められ、(1,2) 成分は $\Lambda_{i,j}$ に比例する。したがって、もし $\Lambda_{i,j} = 0$ ならば、x_i と x_j は、他の変数を条件付けたときに統計的に独立である。

したがって、スパースなグラフを（GGM の範囲で）求めることは、スパースな精度行列を求めることと等価である。

2.2 共分散構造選択

共分散選択 [4] は疎構造学習のための標準的な手法である。簡単に言えばこれは、精度行列においてある小さい行列要素を 0 とおき、その条件を考慮した上で他の行列要素を推定し直す、という過程を繰り返す。しかしながら実用上は、まず標本共分散行列の逆行列を求めねばならないという問題に加え（実データは共分散行列はしばしばランク落ちする）、計算コストが高いこと、統計的検定の観点で必ずしも最適ではないことなどの欠点が知られていた。Drton と Perlman は統計的検定の最適

性の問題を詳しく検討し [5]、SIN と呼ばれる新しいアルゴリズムを提案した。ただしこれは、共分散行列が正則でなければならない要請を取り除いたわけではない。我々は測定系に冗長性があり、それゆえいくつかの変数が強い相関をもつという状況に興味があるので、SIN は我々の問題には有用とは言えない。

2.3 Meinshausen-Bühlmann の方法

MB の方法では [14]、ひとつの変数をターゲットにし、他の変数を入力とした L_1 正則化つきの回帰問題を解く。すなわち、ある変数 x_i に対し、

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\mathsf{Z}_i \boldsymbol{\beta} - \boldsymbol{y}_i||^2 + \mu ||\boldsymbol{\beta}||_1 \right\} \tag{6}$$

を解く。ただし、$\boldsymbol{y}_i \equiv (x_i^{(1)}, ..., x_i^{(N)})^\top$ と定義し、データ行列を

$$\boldsymbol{z}_i^{(n)} \equiv (x_1^{(n)}, .., x_{i-1}^{(n)}, x_{i+1}^{(n)}, ..., x_M^{(n)})^\top \in \mathbb{R}^{M-1} \tag{7}$$

に対して $\mathsf{Z}_i \equiv [\boldsymbol{z}_i^{(1)}, ..., \boldsymbol{z}_i^{(N)}]^\top$ と置いた。これに対応して、精度行列と分布のパラメターとしての共分散行列を、

$$\Lambda = \begin{pmatrix} L & \boldsymbol{l} \\ \boldsymbol{l}^\top & \lambda \end{pmatrix} \quad \Sigma \equiv \Lambda^{-1} = \begin{pmatrix} W & \boldsymbol{w} \\ \boldsymbol{w}^\top & \sigma \end{pmatrix} \tag{8}$$

と分割しておく。ここで行列の行と列は、x_i に関係する要素が最後の行と列と来るように適当に並び替えられているとする。これらの表現において、$W, L \in \mathbb{R}^{(M-1)\times(M-1)}$、$\lambda, \sigma \in \mathbb{R}$、$\boldsymbol{w}, \boldsymbol{l} \in \mathbb{R}^{M-1}$ である。

式 (6) は通常のいわゆる Lasso と同じであり、MB 理論の主張は、各変数に Lasso を解いてまとめれば、統計学的に一致性を持つ構造学習が行える、というものである。より詳しく書けば、まず、式 (6) を解いて、係数 $\boldsymbol{\beta}$ を求める。この係数は、「ターゲット変数」x_i を $\boldsymbol{\beta}^\top \boldsymbol{z}_i$ の形で予測するものであるから、ガウス分布の分割公式 (4) を眺めると、精度行列の対応する 1 列が、

$$\lambda = \frac{1}{\tilde{\sigma_i}^2}, \quad \boldsymbol{l} = -\frac{\boldsymbol{\beta}}{\tilde{\sigma_i}^2},$$

で与えられることが分かる。ただし、$\tilde{\sigma_i}^2$ は予測分散の推定値であり、最尤推定量を使う場合、

$$\tilde{\sigma_i}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_i^{(n)} - \boldsymbol{\beta}^\top \boldsymbol{z}_i^{(n)})^2.$$

のように与えられる。全ての変数について Lasso 回帰の問題を解くことにより精度行列の全要素を求めることができる。

3 グラフィカル Lasso

3.1 ラプラス事前分布による MAP 推定

GGM では、構造学習は多変量正規分布（式 (2)）の精度行列 Λ を求めることに帰着される。まず、疎な構造を得るための工夫は脇に置いて、データ $\mathcal{D}$ からどのように Λ を求めればよいか考えてみよう。最も自然な方法は、次の対数尤度を最大化することである。

$$\ln \prod_{t=1}^{N} \mathcal{N}(\boldsymbol{x}^{(t)} | \boldsymbol{0}, \Lambda^{-1}) = \text{const.} + \frac{N}{2} \{\ln \det(\Lambda) - \mathrm{tr}(\mathsf{S}\Lambda)\}$$

ここで tr は行列の対角和を表す。また、よく知られた恒等式 $\boldsymbol{x}^{(t)\top} \boldsymbol{x}^{(t)} = \mathrm{tr}(\boldsymbol{x}^{(t)} \boldsymbol{x}^{(t)\top})$ と式 (1) を使った。行列の微分に関するよく知られた公式

$$\frac{\partial}{\partial \Lambda} \ln \det(\Lambda) = \Lambda^{-1}, \quad \frac{\partial}{\partial \Lambda} \mathrm{tr}(\mathsf{S}\Lambda) = \mathsf{S} \tag{9}$$

を使えば、直ちに $\Lambda = \mathsf{S}^{-1}$ が最尤解であることが分かる。しかしながら、すでに何度か述べたように、標本共分散行列が正則であることは実用上はまれで、また、仮に正則であったとしても精度行列が疎になるということはほとんどありえない。このため、この解は実用的な価値に乏しい。

われわれも GGM に基づく構造学習を志向するが、上記の限界に基づき、解くべき問題を拡張する。すなわち、式 (2) を、精度行列 Λ が与えられた時の条件付き分布 $p_{\mathrm{G}}(\boldsymbol{x}|\Lambda)$ と見なし、Λ については、事前分布として、要素ごとに同一のラプラス分布を付す。すなわち、

$$p(\Lambda) = \prod_{i,j=1}^{M} \frac{\lambda}{2} \exp\left(-\lambda |\Lambda_{i,j}|\right) \tag{10}$$

である。この式から明らかに分かるように、この事前分布は、Λ の要素の値を 0 付近に束縛する効果を持つ。

そうして GGM の隣接行列 Λ^* を、事後確率最大（MAP: Maximum a posteriori）原理に従って求める。

$$\Lambda^* = \arg\max_{\Lambda} \left\{ \ln p(\Lambda) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}^{(n)} | \boldsymbol{0}, \Lambda) \right\} \tag{11}$$

それゆえ、ただの最尤推定を行うのではなく、次の L_1 制約項付きの最尤方程式を解くことにする。

$$\Lambda^* = \arg\max_{\Lambda} f(\Lambda; \mathsf{S}, \rho), \tag{12}$$

$$f(\Lambda; \mathsf{S}, \rho) \equiv \ln \det \Lambda - \mathrm{tr}(\mathsf{S}\Lambda) - \rho ||\Lambda||_1 \tag{13}$$

ここで $||\Lambda||_1$ は $\sum_{i,j=1}^{M} |\Lambda_{i,j}|$ により定義される。罰金項の重み ρ は入力パラメターとなるが、我々の文脈では、これは異常検知性能を最大化するように決定することができる。

3.2 ブロック勾配法

式 (12) は凸計画問題であり [1]、劣勾配法によって手軽に解くことができる。最近、Friedman、Hastie、および Tibshirani [7] は、グラフィカル Lasso （以下 gLasso と表す）と呼ばれる効率のよい劣勾配アルゴリズムを提案した。gLasso はまず、式 (12) の問題を、ブロック勾配法 [1, 6] という技術を用いて、L_1 制約付き回帰問題の集まりに帰着させる。「ブロック」というのは、上記行列方程式の特定の変数に着目して式変形を行うことに由来する。公式 (9) を用いると、式 (12) の勾配が

$$\frac{\partial f}{\partial \mathsf{\Lambda}} = \mathsf{\Lambda}^{-1} - \mathsf{S} - \rho\,\mathrm{sign}(\mathsf{\Lambda}) \tag{14}$$

と与えられることがわかる。ただし行列 $\mathrm{sign}(\mathsf{\Lambda})$ は、$\mathsf{\Lambda}_{i,j} \neq 0$ に対してはその (i,j) 要素が $\mathrm{sign}(\mathsf{\Lambda}_{i,j})$ で、また、$\mathsf{\Lambda}_{i,j} = 0$ に対しては $\in [-1,1]$ で与えられると定義する。

方程式 $\partial f/\partial \mathsf{\Lambda} = 0$ をブロック勾配法で解くために、ある特定の変数 x_i に着目し、$\mathsf{\Lambda}$ とその逆行列が (8) のように分割されているものとする。この x_i による分割に対応して、標本共分散行列 S も同様に分割するものとし、

$$\mathsf{S} = \begin{pmatrix} S^{\setminus i} & \boldsymbol{s} \\ \boldsymbol{s}^\top & s_{i,i} \end{pmatrix} \tag{15}$$

のように書いておく。

ここで方程式 $\partial f/\partial \mathsf{\Lambda} = 0$ の解を求めよう。$\mathsf{\Lambda}$ は正定値であるため、容易に証明できるように、その対角要素は正でなければならない。したがって、対角要素に関しては、勾配ゼロの条件は

$$\sigma = s_{i,i} + \rho \tag{16}$$

と書かれる。

$\boldsymbol{w}$ および $\boldsymbol{l}$ で表される非対角要素に関しては、他の変数をすべて固定したという条件の下での最適解は、

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||W^{\frac{1}{2}}\boldsymbol{\beta} - \boldsymbol{b}||^2 + \rho\,||\boldsymbol{\beta}||_1 \right\} = 0 \tag{17}$$

を解くことで求められる。ただし、$\boldsymbol{\beta} \equiv W^{-1}\boldsymbol{w}$、$\boldsymbol{b} \equiv W^{-1/2}\boldsymbol{s}$、$||\boldsymbol{\beta}||_1 \equiv \sum_l |\beta_l|$ である。

上式を示そう。分割公式 (8) に基づいて、方程式 $\partial f/\partial \mathsf{\Lambda} = 0$ の右上部分は直ちに

$$\boldsymbol{w} - \boldsymbol{s} - \rho\,\mathrm{sign}(\boldsymbol{l}) = 0 \tag{18}$$

と書かれる。$\mathsf{\Sigma\Lambda} = \mathsf{I}_M$ であるから、

$$\mathsf{\Sigma\Lambda} = \begin{pmatrix} WL + \boldsymbol{w}\boldsymbol{l}^\top & W\boldsymbol{l} + \lambda\boldsymbol{w} \\ \boldsymbol{l}^\top W + \lambda\boldsymbol{w}^\top & \boldsymbol{w}^\top\boldsymbol{l} + \sigma\lambda \end{pmatrix} = \begin{pmatrix} \mathsf{I}_{M-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 1 \end{pmatrix}. \tag{19}$$

を得る。この恒等式の右上部分を使うと、

$$\boldsymbol{l} = -\lambda W^{-1}\boldsymbol{w} = -\lambda\boldsymbol{\beta}, \tag{20}$$

であることが分かる。ただし、$\boldsymbol{\beta} \equiv W^{-1}\boldsymbol{w}$ である。$\mathsf{\Lambda}$ は正定であるから、λ は正でなければならない。従って、$\mathrm{sign}(\boldsymbol{l}) = -\mathrm{sign}(\boldsymbol{\beta})$ が成り立つ。これを用いると、式 (18) は次と等価であることが分かる。

$$\frac{\partial}{\partial\boldsymbol{\beta}} \left\{ \frac{1}{2}\boldsymbol{\beta}^\top W\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{s} + \rho\,||\boldsymbol{\beta}|| \right\} = 0 \tag{21}$$

$W^{-1/2}\boldsymbol{\beta}$ を $\boldsymbol{b}$ とおけば、この式が式 (17) と等価であることが分かる。この最適化問題をどう解くかについては Appendix を参照されたい。

さて、これを解いて $\boldsymbol{\beta}$ を得たとすれば $\mathsf{\Lambda}$ の対応する列を

$$\lambda = \frac{1}{\sigma - \boldsymbol{\beta}^\top W\boldsymbol{\beta}}, \quad \boldsymbol{l} = -\frac{\boldsymbol{\beta}}{\sigma - \boldsymbol{\beta}^\top W\boldsymbol{\beta}} \tag{22}$$

によって更新できる。ただしここで、式 (19) $\boldsymbol{w}^\top\boldsymbol{l} + \sigma\lambda = 1$ の右下部分と、式 (20) を用いた。また、式 (19) の右上部分を用いて、$\boldsymbol{w}$ を

$$\boldsymbol{w} = -W\boldsymbol{l}/\lambda.$$

のように更新することができる。ここで σ は式 (16) のために一定に保たれることに注意。したがって、グラフィカル Lasso アルゴリズムにおいては、$\mathsf{\Sigma} = \mathsf{\Lambda}^{-1}$ は $\mathsf{\Lambda}$ の副産物として得られ、明示的な逆行列の計算は不要である。

最終的な解 $\mathsf{\Lambda}^*$ を得るため、式 (17) を $x_1, x_2, ..., x_M, x_1, ...$ について解くことを収束するまで繰り返す。式 (16) のため、行列 W は必ず正則となることに注意。この点はこの算法の数値的安定性を示唆する。

3.3 MB の方法との関係

gLasso により導かれた座標ごとの最適化問題（式 (17)）には、Lasso に基づく構造学習法 (6) との明らかな類似がみられる。S の定義（式 (1)）を用いれば、もし条件

$$W = S^{\setminus i} \quad \text{and} \quad \rho = M\mu \tag{23}$$

が成り立てば、この問題が式 (17) と等価であることが分かる。W は $\mathsf{\Lambda}^{-1}$ の主対角行列であるので、ρ が小さい時には W と $S^{\setminus i}$ の間に何らかの密接な関係があることが推察されるが、$\rho > 0$ の時は両者は等しくなることはない。

さらに深刻なのは、多重共線性がある時の振る舞いである。Lasso においては、変数に多重共線性がある時、

強く相関した変数のグループのどれかが、例えば実装上の変数の順序付けの違いのようなほとんど偶然の要因で選択される。これは学習された構造が偶然の要因で大幅に変わるということを意味する。

結局、MB の算法は、gLasso と異なり、MAP 最適性のような明確な大局的最適性を持たず、実用上は非常に使いにくい、というのが結論になる。

4 相関異常度のスコアリング

本節では、論文 [10] で与えた相関異常度の定義を要約する。前節で論じた方法に基づいて、二つの疎な GGM $p_{\mathrm{A}}(\boldsymbol{x})$ および $p_{\mathrm{B}}(\boldsymbol{x})$ を得たとしよう。$\mathcal{D}_{\mathrm{A}}$ と $\mathcal{D}_{\mathrm{B}}$ の間の相違に対し、いかに個々の変数が寄与しているかを表すスコアを計算したい。確率モデル $p_{\mathrm{A}}(\boldsymbol{x})$ および $p_{\mathrm{B}}(\boldsymbol{x})$ が与えられている時、最も自然な相違度の尺度は、Kullback-Leibler（KL）距離である。しばらくの間、特定の変数 x_i に着目しよう。量

$$d_i^{\mathrm{AB}} \equiv \int \mathrm{d}\boldsymbol{z}_i \, p_{\mathrm{A}}(\boldsymbol{z}_i) \int \mathrm{d}x_i \, p_{\mathrm{A}}(x_i|\boldsymbol{z}_i) \ln \frac{p_{\mathrm{A}}(x_i|\boldsymbol{z}_i)}{p_{\mathrm{B}}(x_i|\boldsymbol{z}_i)} \quad (24)$$

は $p_{\mathrm{A}}(x_i|\boldsymbol{z}_i)$ と $p_{\mathrm{B}}(x_i|\boldsymbol{z}_i)$ の間の KL 距離の期待値を、分布 $p_{\mathrm{A}}(\boldsymbol{z}_i)$ によって計算したものである。$\boldsymbol{z}_i$ の定義は式 (7) を参照。式 (24) において A と B を入れ替えることで、d_i^{AB} の定義も得る。上式に現れる分布は正規分布のみであるから、この積分は解析的に実行できる。結果は

$$\begin{aligned} d_i^{\mathrm{AB}} &= \boldsymbol{w}_{\mathrm{A}}^{\top}(\boldsymbol{l}_{\mathrm{B}} - \boldsymbol{l}_{\mathrm{A}}) \\ &+ \frac{1}{2}\left\{ \frac{\boldsymbol{l}_{\mathrm{B}}^{\top}\mathsf{W}_{\mathrm{A}}\boldsymbol{l}_{\mathrm{B}}}{\lambda_{\mathrm{B}}} - \frac{\boldsymbol{l}_{\mathrm{A}}^{\top}\mathsf{W}_{\mathrm{A}}\boldsymbol{l}_{\mathrm{A}}}{\lambda_{\mathrm{A}}} \right\} \\ &+ \frac{1}{2}\left\{ \ln\frac{\lambda_{\mathrm{A}}}{\lambda_{\mathrm{B}}} + \sigma_{\mathrm{A}}(\lambda_{\mathrm{B}} - \lambda_{\mathrm{A}}) \right\} \end{aligned} \quad (25)$$

となる。ここで、Λ_{A} およびその逆行列 Σ_{A} をそれぞれ次のように分割した（式 (8) 参照）。

$$\Lambda_{\mathrm{A}} = \begin{pmatrix} L_{\mathrm{A}} & \boldsymbol{l}_{\mathrm{A}} \\ \boldsymbol{l}_{\mathrm{A}}^{\top} & \lambda_{\mathrm{A}} \end{pmatrix} \; \Sigma_{\mathrm{A}} \equiv \Lambda_{\mathrm{A}}^{-1} = \begin{pmatrix} W_{\mathrm{A}} & \boldsymbol{w}_{\mathrm{A}} \\ \boldsymbol{w}_{\mathrm{A}}^{\top} & \sigma_{\mathrm{A}} \end{pmatrix} \quad (26)$$

同様の分割は Λ_{B} および Σ_{B} にも適用される。定義 d_i^{BA} もまた、A と B を入れ替えることで得られる。式 (25) は、よく知られた分割公式 (4) を使えば容易に導出できる。

異常度の定義 (25) の各項は次のような明確な解釈を持つ。GGM の定義から、$\boldsymbol{l}_{\mathrm{A}}$ における非ゼロ要素の数は、頂点 x_i の次数と同じである。この意味で、第 1 項は主に近傍の生成および消滅に関する異常を検知する。第 2 項は、重み付きグラフとしての近傍グラフの「緊密さ」を表している。すなわち、仮に x_i が単一の辺を j に対して持つとすれば、この項は、対応する相関係数の間の差を、単一の変数に対する精度 λ_{A} および λ_{B} で割ったものに比例する。第 3 項は、変数間の関係の変化というよりは各変数ごとの精度もしくは分散の変化に結び付けられる。

最終的な異常度は、A と B の立場を入れ替えたスコアとあわせて、次のように定義するのが自然である。

$$a_i \equiv \max\{d_i^{\mathrm{AB}},\, d_i^{\mathrm{BA}}\} \quad (27)$$

5 実験

この節では、共線形性が強い場合の構造の安定性という切り口で、異なるいくつかの構造学習手法を比較する。

5.1 構造学習手法の比較

相関が強い変数を持つデータに対しては伝統的な共分散構造選択の手法の適用が難しいという事実を考えれば、最近新たに提案された L_1 制約付きの学習手法の安定性を調べてみるのは興味ある研究課題である。我々は gLasso （改めて `Glasso` と表記する）を他の 2 つの構造学習手法と比較した。

最初の比較対象は、Meinshausen と Bühlmann [14] により提案された手法（`Lasso` と表記する）である。彼らの手法は、各変数を目的変数としそれ以外を説明変数とする Lasso 回帰の問題を M 個独立に解くものである。彼らは、この手法がある種の統計的一致性を満たすことを示した。しかし実際上は、過剰に近傍を取り込む傾向があることが知られている [15, 3]。

そこで、もうひとつの比較対象として、適応 Lasso（adaptive lasso）[16] を使う構造学習手法を取り上げる。適応 Lasso（以下 `AdaLasso` と表記する）は 2 段階の線形回帰の手法であり、最初の回帰の結果を 2 度目の回帰の結果に使うことで「オラクル性」という性質を満たすようにする。理論的詳細は原論文 [16] を参照されたい。ここでは、文献 [3] において優れた結果を示した、2 段とも Lasso を用いる手法を使う。

我々はいくつかの変数が強い相関をもつ状況に興味があり、それゆえ S がランク欠損を起こしているというのが前提であるため、S の逆行列の存在を明示的に仮定する伝統的な共分散選択手法とその拡張 [13, 5] は比較の対象としない。

データと評価指標。求められたグラフ構造の安定性を調べる目的で、データにガウスノイズを印加する前後での構造の変化を調べた。用いたデータは第 **??**節で詳しく説明した *Actual spot rates* データである。時間軸を重複がないように 25 個に分け、連続した 100 日を含む小データを作った。そして罰金項の係数をいろいろと変えながら、それぞれの小データに対して何度も構造学習を

行った。その結果に対して、疎度（sparsity）を

$$(\text{疎度}) \equiv \frac{N_0}{M(M-1)}$$

で定義する。ここで、N_0 は Λ の非対角要素におけるゼロ要素の数である。

第 1 回目の構造学習の後、各小データに対して $x_i \leftarrow x_i + \epsilon_i$, のようにガウスノイズを加えた。ただし、$\epsilon_i$ は、平均ゼロの独立同一分布に従うガウスノイズを表す。ノイズの印加により、新たに生ずる辺と、消滅してしまう辺があるので、それらの数を数えて、「辺のフリップ確率」を形式的に

$$(\text{フリップ確率}) \equiv N_1/N_0,$$

で定義する。ここで N_1 は生成または消滅した辺の数である。

結果。図 2 に結果を示す。これは疎度の関数としてフリップ確率を示したものである。ガウスノイズの標準偏差は、平均 0、分散 1 に標準化した後のデータを対象に 0.1 とした。図から、`Lasso` および `AdaLasso` が、極めてノイズに脆弱であることが分かる。これらの方法だと、疎度が 0.5 の時に、フリップ確率は実に 50%にも及ぶ。これは要するに、推定されたグラフが、少なくともノイジーなデータに対しては、ほとんど**まったく信用できない**ことを示す。「真の」グラフ構造を求めることが必要な用途、例えばバイオインフォマティックスにおけるネットワーク推定問題などでは、この点に対して慎重な考察が必要であろう。一方、 `Glasso` はこれらよりははるかにノイズに対し安定である。

`Lasso` および `AdaLasso` の不安定性の大きな理由は、ある程度相関の強い変数がある時、その中のひとつの変数だけが選択されるという Lasso の傾向に帰すことができる。*Actual spot rates* データでは実際、BEF、FRF、DEM、NLG といった欧州通貨は互いに強く相関しあっている。この中のどれが説明変数として選ばれるかはほとんど偶然による。この種の変数節約の傾向は、汎化性能の観点から回帰問題では有用なものであるが、構造学習においては実用上深刻な欠陥と言える。

以上まとめると、変数ごとに独立した回帰問題を解くという `Lasso` および `AdaLasso` の構造学習手法は、データに強い相関を持つ変数群が含まれる場合は、安定した結果を与えない。対照的に `Glasso` は妥当に安定した結果を与える。

6　まとめ

相関異常の検知に対し疎な構造学習を用いるという手法を提案した。我々の問題は、2 つのデータセットの比較に基づいて、個々の変数の異常度を計算するというものであり、この意味で、データセット全体の相違度を求める 2 標本検定の枠組みのひとつの一般化になっている。

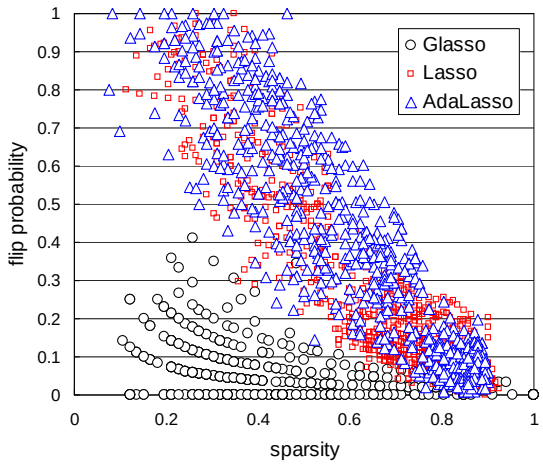

図 2: 疎度の関数としてプロットしたフリップ確率。`Lasso` と `AdaLasso` における著しい不安定性に注目。

我々は、最近提案された疎構造学習の手法のいくつかが共線形性の下で著しく不安定になり、したがって多くの場合、実センサーデータの解析には実用性が乏しいことを指摘した。しかしながら、gLasso アルゴリズムはこの深刻な問題に直面することなく、構造を学習できることを実験的に示した。

参考文献

[1] O. Banerjee, L. E. Ghaoui, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proc. Intl. Conf. Machine Learning*, pages 89–96. Press, 2006.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] P. Bühlmann. Variable selection for high-dimensional data: with applications in molecular biology. 2007.

[4] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[5] M. Drton and M. D. Perlman. A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

[6] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[8] T. Idé and K. Inoue. Knowledge discovery from heterogeneous dynamic systems using change-point correlations. In *Proc. SIAM Intl. Conf. Data Mining*, pages 571–575, 2005.

[9] T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. In *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pages 440–449, 2004.

[10] T. Idé, A. C. Lozano, N. Abe, and Y. Liu. Proximity-based anomaly detection using sparse structure learning. In *Proceedings of 2009 SIAM International Conference on Data Mining*, 2009.

[11] T. Idé, S. Papadimitriou, and M. Vlachos. Computing correlation anomaly scores using stochastic nearest neighbors. In *Proc. IEEE Intl. Conf. Data Mining*, pages 523–528, 2007.

[12] T. Idé and K. Tsuda. Change-point detection using krylov subspace learning. In *Proc. 2007 SIAM Intl. Conf. Data Mining*, pages 515–520, 2007.

[13] S. L. Lauritzen. *Graphical Models*. Oxford, 1996.

[14] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[15] R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8(Suppl.2):S3, 2007.

[16] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

A 劣勾配法による Lasso の解法

L_1 正則化項付きの 2 次計画問題 (17) は、変数ごとに劣勾配法を適用することで解くことができる。式 (17) の代わりに、等価な表現 (21) を考えよう。β_i について微分すると、

$$\sum_m W_{i,m}\beta_m - s_i + \rho\,\mathrm{sign}(\beta_i) = 0.$$

を得る。$\beta_i > 0$ に対して、この方程式に対する形式的な解は、

$$\beta_i = \frac{1}{W_{i,i}}(A_i - \rho),$$

で与えられる。ただし、

$$A_i \equiv s_i - \sum_{m\neq i} W_{i,m}\beta_m \tag{28}$$

と定義した。

$W_{i,i} > 0$ であるため、この解は $A_i > \rho$ を満たさなければならない。もしこの条件が満たされなければ、この目的関数の最小は $\beta_i = 0$ において得られる。なぜなら、この場合勾配が正であるからである。同様に、$\beta_i < 0$ の場合を考えると、各 i に対して次のような更新式を得る。

$$\beta_i \leftarrow \begin{cases} (A_i - \rho)/W_{i,i} & \text{for } A_i > \rho \\ 0 & \text{for } -\rho < A_i < \rho \\ (A_i + \rho)/W_{i,i} & \text{for } A_i < -\rho \end{cases}$$

この式を収束するまで繰り返すことで解を得る。