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.