大規模次元観測時系列のダイナミクスモデルに関する研究

Nguyen Viet Phuong†　鷲尾　隆†
大阪大学産業科学研究所†

1. はじめに

近年，センシングネットワークなどの技術発展により，計算機のみならず様々な情報機器，家電，自動車などがネットワークで結ばれつつある．その際，センサーをはじめネットワークに設置された膨大なデータ収集装置群全体から，逐次膨大な変数からなる大規模時系列トランザクションデータが創出される．巨大なネットワーク化社会システムについて，円滑な制御，管理，異常診断などを行うためには，上記観測大規模次元時系列トランザクションデータが表す対象のダイナミクスモデルの構築が必要となる．しかし，大規模次元データが表す状態組み合わせ爆発[1]の扱いが困難であるため，このようなモデル化手法の研究は世界的にも十分ではない．

ある状態から次状態への遷移を確率的に表す隠れ状態マルコフ連鎖モデルや高次マルコフ連鎖モデルがよく知られているが，これらは上記の状態組み合わせ爆発問題により信頼性が低く計算コストが高い[2]．この困難を軽減するため，近似を導入して状態空間および状態遷移を縮めた Mixture Transition Distribution (MTD)[1] や Variable Length Markov Chain (VLMC)[2] などのモデリング手法がある．しかしながら，これらも適用できる状態数は，数十個止まりであることが知られている．このような背景から，我々は HIgh-order Substate Chain (HISC)モデルを提案した[3]．HISC は，データ中に頻繁に出現する複数の変数及びその値の組み合わせを対象システムの主要部分状態とし，それら状態間の確率遷移によりデータに埋め込まれた主要ダイナミクスを同定する．これにより，状態爆発問題を解決できる．しかし，多頻度な変数・値の組み合わせで定義される各部分状態は，互いにオーバーラップした非独立状態であり，確率定量的扱いが困難である．

そこで，本研究では，上記従来手法の問題を克服しつつ，HISC のような個々の部分状態遷移規則知識のみならず，データを創出する大規模システム全体のダイナミクスを定量的に表す確率モデルを，大規模観測時系列データから同定する手法を提案する．

Title: Modeling Dynamics of Massive Dimensional Systems
Nguyen Viet Phuong†, Washio Takashi†
† The Institute of Scientific and Industrial Research, Osaka University.

2. 提案モデルの定式化

2.1 トランザクションおよび状態の定義

時系列トランザクションデータにおいて，各変数およびそのある値のペアを1個のアイテムとする．例えば，計測された距離<距離:2km>は一つのアイテム(*item*)となる．トランザクション時系列データは

$$\mathbf{D} = X_1 X_2 ... X_T$$

と記述される．但し，$X_t (t = 1,2,...,T)$は時刻tに観測されたアイテムからなるトランザクション $X_t = \{item_1^t, ..., item_{m_t}^t\}$である．このデータに出現可能な全アイテム数を$N$とすれば，時刻$t$に観測されるトランザクションは$N$次元バイナリベクトル$\mathbf{Q}_t = (q_t^i)_{i=1\sim N}$で表現できる．アイテム$i$が時刻$t$に観測されれば$q_t^i = 1$であり，そうでなければ$q_t^i = 0$である．これにより，時刻$t$の対象システム状態を，状態ベクトル$\mathbf{X}_t = (x_t^i)_{i=1\sim N}$で表す．各$x_t^i (-\infty < x_t^i < \infty)$はアイテム$i$の出現可能性を決める測度である．$x_t^i$が大きければアイテム$i$の出現可能性が高く、逆に小さければ低い．

2.2 モデルの定式化

提案モデルは次の2つ方程式から成る．

$$\mathbf{X}_t = \mathrm{B}_m(\mathbf{A} \times \mathbf{X}_{t-1}) \quad (1)$$

$$\mathbf{Q}_t = \mathrm{H}(Sm(\mathrm{E}_r(\mathbf{X}_t))) \quad (2)$$

式(1)はシステム状態遷移過程を表す．行列$\mathbf{A} = (a_{ij})_{N\times N}$は$N \times N$状態遷移行列であり，$\mathrm{B}_m$はパラメータ$m \in R$を持つ状態遷移のノイズ・バイアス線形写像である．$\mathrm{B}_m(x) = mx$であり，$m \in [-1,+1]$のとき$|\mathbf{B}_m(x)| \le |x|$となり，$\mathrm{B}_m$はアイテムの出現事象の情報エントロピー(あいまいさ)を大きくするノイズ作用を表す．一方，$m \in [-\infty,-1] \cup [1,\infty)$のとき，$\mathrm{B}_m$は出現事象の情報エントロピーを小さくするバイアスの役割を担う．式(2)はシステム状態$\mathbf{X}_t$から観測アイテム出現ベクトル$\mathbf{Q}_t$への写像を表す観測過程である．ここで，$Sm()$はシグモイド関数で，

$$Sm(x)=\frac{1}{1+\exp(-x)} \quad (x\in R, Sm(x)\in[0,1])$$

である．Hは[0, 1]から{0, 1}への写像である．$\mathrm{H}(p)$は[0, 1]で一様乱数hを生成し，$0\le h\le p$の場合$\mathrm{H}(p)=1$，$p<h\le 1$の場合$\mathrm{H}(p)=0$を出力する．E_rはB_mと同じくパラメータrを持つ観測ノイズ・バイアス線形写像である．

3 モデリング手法

以上より，提案モデルはシステム状態遷移行列$\mathbf{A}=(a_{ij})_{N\times N}$および状態測度ベクトル系列$\mathbf{X}_t(t=1,2,..,T)$で構成される．観測時系列データのモデリングは，$\mathbf{Q}_t=(q_t^i)_{=1\sim N}\ (t=1,2,...,T)$からの$\mathbf{A}$と$\mathbf{X}_t(t=1,2,..,T)$の推定である．

3.1 状態遷移行列から状態測度系列の推定

システム状態遷移行列$\mathbf{A}$が与えられた時，時刻tにおいて観測データ$\mathbf{Q}_{t-1},\mathbf{Q}_t$下でシステム状態測度ベクトル$\mathbf{X}_t$は事後確率$P(\mathbf{X}_t\mid\mathbf{Q}_{t-1},\mathbf{Q}_t)$を最大化することにより決定される．確率計算により，$P(\mathbf{X}_t\mid\mathbf{Q}_{t-1},\mathbf{Q}_t)$は以下となる．

$$P(\mathbf{X}_t\mid\mathbf{Q}_{t-1},\mathbf{Q}_t)=\frac{P(\overline{\mathbf{X}}_{t-1})R_{t-1}(\overline{\mathbf{X}}_{t-1})R_t(\mathbf{X}_t)}{P(\mathbf{Q}_t\mid\mathbf{Q}_{t-1})P(\mathbf{Q}_{t-1})} \qquad (3)$$

但し，

$$R_t(\mathbf{X}_t)=\prod_i^N\left(q_t^i Sm[E_r(x_t^i)]+(1-q_t^i)\{1-Sm[E_r(x_t^i)]\}\right),$$

$$R_{t-1}(\overline{\mathbf{X}}_{t-1})=\prod_i^N\left(q_{t-1}^i Sm[E_r(\overline{x}_{t-1}^i)]+(1-q_{t-1}^i)\{1-Sm[E_r(\overline{x}_{t-1}^i)]\}\right)$$

$$\overline{x}_{t-1}^i=\sum_j^N\left[a'_{ij}(x_t^j/m)\right],\quad (a'_{ij})_{N\times N}=\mathbf{A}^{-1}.$$

定理 1：時刻tにおいて，状態測度ベクトル$\mathbf{X}_t$の各要素x_t^iが独立で$(-\infty,\infty)$の値を取り，x_t^iのp.d.fが$P(x_t^i)=f(x_t^i)$である場合，式(3)の$P(\mathbf{X}_t\mid\mathbf{Q}_{t-1},\mathbf{Q}_t)$が最大となる$\mathbf{X}_t$は以下の非線形連位方程式の解である．

$$r\left(q_t^k-Sm[rx_t^k]\right)+\sum_{i=1}^N(ra'_{ik}/m)\left(q_{t-1}^i-Sm[\sum_{j=1}^N a'_{ij}x_t^j/m]\right)+$$

$$+\sum_{i=1}^N\frac{a'_{ik}/m}{f(\overline{x}_{t-1}^i)}f'(\overline{x}_{t-1}^i)=0\quad(k=1,2,...,N) \qquad (4)$$

（証略）

定理 2：式(4)の解はユニークであり，常にNewton法より収束計算可能である．（証略）

定理 1, 2 により，状態遷移行列$\mathbf{A}$，観測データ系列$\mathbf{Q}_t$から状態ベクトル系列$\mathbf{X}_t$を導出できる．

3.2 状態遷移行列の導出

状態測度ベクトル系列$\mathbf{X}_t(t=1,2,...,N)$が与えられた時，一義的には予測誤差

$$L=\sum_{t=1}^T\left\|\mathbf{X}_t-\mathrm{B}_m(\mathbf{A}\times\mathbf{X}_{t-1})\right\|_2^2$$

を最小化することにより$\mathbf{A}$を推定可能である．$\|\bullet\|_2^2$は2次ベクトルのノルムである．しかし，一般に状態間の影響は局所的であることが多く，状態遷移行列$\mathbf{A}$はスパースであるという現実的仮定ができる．そこで，L1-正則化法[4]を用い行列$\mathbf{A}$を導出する．L1は

$$L_1=\sum_{t=1}^T\left\|\mathbf{X}_t-\mathrm{B}_m(\mathbf{A}\times\mathbf{X}_{t-1})\right\|_2^2+\lambda\|\mathbf{A}\|$$

である．$\lambda>0$は正則化パラメータである．

定理 3：状態系列が正確であれば，時系列の長さTが大きくなるほど，L_1を最小化する推定遷移行列$\hat{\mathbf{A}}$は真の遷移行列に近づく．（証略）

定理 4：L_1を最小化する推定遷移行列$\hat{\mathbf{A}}$は

$$\hat{\mathrm{A}}=\left(\sum_{t=2}^T\mathbf{S}_t-\frac{\lambda}{2}\mathbf{Sign}\right)\times\left(\sum_{t=2}^T m\mathbf{S}_{t-1}\right)^{-1} \qquad (5)$$

で計算できる．但し，$\mathbf{S}_t=(s_{ij}^t)_{N\times N}=(x_t^i m x_t^j)_{N\times N}$であり，$\mathbf{Sign}=(sign_{ij})_{N\times N}=(sign(a_{ij}))_{N\times N}$である．

定理 3, 4 により大規模次元状態遷移モデルを効率的かつ正しく推定できることが保証される．

3.3 観測データのモデリング手法

Step1：初期の遷移行列$\mathbf{A}$を仮定する．

Step2：観測時系列データ$\mathbf{Q}_t=(q_t^i)_{=1\sim N}$から定理2により測度状態ベクトル系列を推定する．

Step3：推定した状態測度ベクトルにより定理4で遷移行列を更新する．

Step2 と Step 3 を収束するまで繰り返し行えばモデルが得られる．

4. まとめ

本研究では大規模次元観測時系列データの確率定量的状態遷移モデルとモデリング手法を提案した．また，それらのモデリング収束性の保証を示した． 今後は実用化に向けて，提案手法をより高度化する予定である．

参考文献

[1] A.Berchtold, A.E.Raftery: The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. Statistical Science, 17(3). (2002)

[2] G.Bejerano, G.Yona: Variations on probabilistic suffix trees - a new tool for statistical modeling and prediction of protein families. Bioinformatics, 17(1). (2001)

[3] V.P.Nguyen, T.Washio: Modeling Dynamic Substate Chains among Massive States for Prediction. Proc. of ICDM Workshops 2006, 484 - 48917(1).

[4] N.Meinshausen, P.Buhlmann: High-Dimensional Graphs and Variable Selection with the LASSO. The Annuals of Statistic, 34(3). (2006)