Technical Report of the 2nd Workshop on Latent Dynamics
(Jun 22, 2011, Tokyo, Japan)

時系列関係データにおける非定常な潜在構造の推定

石黒 勝彦*
Katushiko Ishiguro

Abstract: SNS に代表される関係ネットワークデータの潜在構造解析は多くの研究者の注目を集めている。最近では、関係データの時間変化に着目した新しい技術がいくつか提案されている。本稿では、そのなかでもネットワーク内のコミュニティ抽出やノードのクラスタリングを目標にした生成モデル手法を紹介する。また、その一例として非定常・非連続な構造変化に着目した「動的無限関係モデル」を提案する。

Keywords: 時系列関係データ、ネットワーク、時間変化、ノンパラメトリックベイズ

1 まえがき

近年、インターネットのハイパーリンクや SNS 上の友人関係、あるいは Twitter でのリツイート情報のようなアイテム・オブジェクト間の関係をまとめた関係データの解析が注目を集めている。このような関係データはその多くがデジタル表現されており、まだ情報量が多いため、統計的な手法を利用した自動的な解析手法が多くの場合採用される。これまでにも [10, 2, 18, 11] を始めとして多くの手法が研究者によって提案されている。

特に最近では、関係データの時間変動の解析手法が精力的に研究されている [1, 14, 17, 16, 7, 5, 9]。これは、現実に関係というものが時間の経過とともに変化することからも自然な要請である。例えば、インターネット上のハイパーリンクは新しいホームページの出現や古いページの削除などに従って不可避的に構造が変化する。また、SNS 上の友人関係なども時間に依存する。例えばユーザの職場が変わった場合、つながり関係が大きく変化することが予想される。Twitter 上のリツイートを例にとれば、あまり注目されていなかった発現もハブとなる人物の推薦によってネットワーク内を急速に拡散する。このとき、リツイートに基づく関係ネットワークは大きな構造変化を起こしているだろう。このように、関係ネットワークにおいて時間変化は重要な要素であり、それを正確に解析することは様々な知見の発掘や応用上の利点に役立つものと思われる。

本稿では、このような時系列性をもった関係データの潜在構造推定の手法の中でも、特にネットワーク内のコミュニティ抽出やノードのクラスタリングを目的とした確率的な生成モデルについていくつか紹介する。

*NTT コミュニケーション科学基礎研究所, 619-0247 京都府相楽郡精華町光台 2-4, e-mail ishiguro.katsuhiko@lab.ntt.co.jp,
NTT Communication Science Laboratories, 2-4 Hikrai-dai Seika-cho Soraku-gun Kyoto 619-0237 Japan

1.1 データについて

時系列関係データには様々な定義が考えられ得る。また実際に論文の著者によってその定義は異なる。

本稿で最も単純な設定の一例を考える。すなわち、データは時間発展するノード間のネットワークとして表現される。各ノードは関係を結ぶ主体 (アイテム、ユーザ、HP など) を表す。ノード間のリンク (エッジ) は関係の有無 (共起関係、友人関係、ハイパーリンクなど) を表す。各ノードのもつ情報は観測できるリンク情報のみとし、その他のノード特有の特徴量は考えない。また、ノード間のリンク観測量は $\{0, 1\}$、すなわちリンクの有無だけを表現し、関係の強さは考慮しないものとする。

T をデータのもつ時間ステップ数、$t = \{1, 2, \ldots, T\}$ を時間のインデックスとする。また、N をノードの総数とし、$i, j = \{1, 2, \ldots, N\}$ をノードのインデックスとする。$x_{t,i,j} = \{0, 1\}$ を時刻 t におけるノード i から j への有向関係の有無を表す観測量とする。なお、無向グラフを考える場合は $x_{t,i,j} = x_{t,j,i}$ とする。時間ステップを越えたノード間の関係は認めない。すなわち、時刻 t におけるノード i と時刻 $t' \neq t$ におけるノード j の間にリンクは定義されないものとする。

2 連続的な時間変化を仮定したモデル

まず、連続的な時間発展モデルの例として、Mixed Membership Stochastic Block (MMSB) model [3] に基づく重複有りクラスタリングモデル [16] を紹介する。

このモデルでは、ユーザ間のインタラクションに潜在的な複数種の関係の種類を仮定する。たとえば、あるユーザ間にメールの送受信関係が観測された場合、その

関係を実現した潜在要因は「会社の同僚」、「趣味の仲間」といった異なる原因に起因すると考えらえるとするモデルである。論文 [16] では各ノードは actor、そして潜在関係は role と呼ばれ、ノード間のそれぞれの観測陸を各 actor 同士がネットワーク内でどのような role を演じているかによって説明する、と述べられている。提案された Dynamic MMSB (dMMSB) モデルは、これら actor ノード間の潜在的な role 関係を時間発展する関係データから自動的に推定することを目的としている。

各ユーザ i の時刻 t における K 種類の潜在 role の混合割合を $\boldsymbol{\pi}_{t,i} \in \mathbb{R}^K$ とする。この $\boldsymbol{\pi}_{t,i}$ を推定することにより、各時刻で各ユーザがどのような潜在 role のクラスタに属しているかを重複有りクラスタリング (あるいは soft clsuterig) することが可能となる．また、各観測リンク $x_{t,i,j}$ に対して、ユーザ i の role を表す 1-of-K ベクトル $\boldsymbol{z}_{t,i\to j}$ とユーザ j の role を表す 1-of-K ベクトル $\boldsymbol{z}_{t,i\leftarrow j}$ を推定することで、どのような潜在関係によって個々のリンクが構成されたのかを推測することも可能である。

生成モデルは次のように記述される。

$$\boldsymbol{\mu}_1 \sim \text{Normal}(\boldsymbol{\nu}, \boldsymbol{\Phi}) \tag{1}$$

$$\boldsymbol{\mu}_t \sim \text{Normal}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}, \boldsymbol{\Phi}) \quad t > 1 \tag{2}$$

$$\eta_1 \sim \text{Normal}(\iota, \psi) \tag{3}$$

$$\eta_t \sim \text{Normal}(b\eta_{t-1}, \psi) \quad t > 1 \tag{4}$$

$$\beta_{t,k,l} \sim \text{LogisticNormal}(\eta_t, S_t) \tag{5}$$

$$\boldsymbol{\pi}_{t,i} \sim \text{LogisticNormal}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \tag{6}$$

$$\boldsymbol{z}_{t,i\to j} \sim \text{Multinomial}(\boldsymbol{\pi}_{t,i}) \tag{7}$$

$$\boldsymbol{z}_{t,i\leftarrow j} \sim \text{Multinomial}\left(\boldsymbol{\pi}_{t,j}\right) \tag{8}$$

$$x_{t,i,j} \sim \text{Bernoulli}\left(\boldsymbol{z}_{t,i\to j}\boldsymbol{B}_t\boldsymbol{z}_{t,i\leftarrow j}\right) \tag{9}$$

まず、観測量 $x_{t,i,j}$ はリンク i, j に対するユーザ i, j それぞれの role を表す $\boldsymbol{z}_{t,i\to j}$ と $\boldsymbol{z}_{t,i\leftarrow j}$ の距離を comatibility matrix $\boldsymbol{B}_t = \{\beta_{t,k,l}\}_{k,l} \in \mathbb{R}^{K\times K}$ で調整したパラメータで決定する。$\boldsymbol{z}_{t,i\to j}$ の事前分布はユーザ i の role 混合確率 $\boldsymbol{\pi}_{t,i}$ に支配されており、その分布は時刻依存のパラメータ $\boldsymbol{\mu}_t$ によって制御される。一方、comatibility matrix $\boldsymbol{B}_t$ の各要素も時刻依存のパラメータ η_t によって制御される。

このモデルの特徴は大きく二点ある。まず、式 (2), 式 (4) にあるようにパラメータ空間で Gaussian diffusion に基づく時間発展を仮定している点である。このことで、関係データ全体でパラメータあるいは隠れ変数が緩やかに時間変化に追従できる。次に、式 (5), 式 (6) にあるように Logistic 正規分布を利用している点である。このことで要素間の共分散関係を表現することができる。

3 非定常な変化に対応する重複無しクラスタリング

次に、非連続・非定常な時間発展に適したモデルの例として、Infinite Relationao Model (IRM) [8] に基づく重複無しクラスタリングモデル [7] を紹介する。

このモデルでは、ネットワーク内のコミュニティなどのクラスタ形成に非定常な変化が起こることを仮定している。例えば組織のリストラや買収などによってメンバー間の接点が大きく変化することを考える。このような変化は繰り返しは発生しない、単発の大きな変化となる。また、インターネット上では日々新しい記事と話題が発生し、古い記事と話題が消滅していく。このようなデータを考える際、先に説明した連続的な変化を仮定するモデルは必ずしもふさわしくないと思われる。そこで、より離散的に、また突然の変化にも対応可能なモデルを考案する必要がある。

離散的な時間発展のモデルとしては隠れマルコフモデル (HMM) [12] による離散クラスタ間の遷移モデルが有名である。ただし、上のような関係データのダイナミクスをとらえる上ではいくつかの工夫が必要である。[7] では必要な条件について以下のように考察している。

- (A) 時刻ステップごとにクラスタ間の遷移確率は変化する
- (B) 連続する時刻ステップの間は高い相関を持つ
- (C) クラスタ数の事前決定は避けるべきである

条件 (A) は突然のクラスタの分割・統合や、各時刻ステップにおけるクラスタの新規生成・消滅などを表現するために必須である。先ほどの例でいえば、組織の合併や買収などによる人間関係の大きな変化は特定の期間にしか発生しない。したがって、そのようなクラスタの変化を表す遷移確率はその時刻においてのみ表現されるべきである。一方、条件 (B) は多くの時系列データに見られる特徴である。組織内の突然で急激な変化はあるものの、日々の人間関係の変化は緩やかなものであり、その点で隣接する時刻ステップにおいてクラスタ構造の相関は高いと考えられる。条件 (C) は付随的なものである。条件 (A) でみるようなクラスタの生成や消滅を仮定する以上、クラスタ数を事前に固定することはモデル設計の前提にそぐわないと思われる。

以上のような条件を踏まえて、[7] では Dynamic IRM(動的無限関係モデル) と呼ばれるノードの重複無しクラスタリング (あるいは hard clustering) アルゴリズムを提案している。このモデルでは、各時刻 t におけるユーザあ

るいはアイテム i の所属するクラスタを $z_{t,i} = k$ と一つに定める。このクラスタ所属変数は HMM の状態遷移確率に従って毎時刻変遷する。各アイテムの時刻ごとのクラスタ所属変数を推定することで、時間変化する関係データ内のクラスタ変化を追跡することが可能となる。また、HMM はノンパラメトリックベイズによる無限状態数拡張 [15, 4] を適用する。これによって、データに内在するクラスタの総数は自動的に推定される。

生成モデルは次のように記述される。

$$\boldsymbol{\beta}|\gamma \sim \mathrm{Stick}(\gamma) \tag{10}$$

$$\boldsymbol{\pi}_{t,k}|\alpha_0, \kappa, \boldsymbol{\beta} \sim \mathrm{DP}\left(\alpha_0 + \kappa, \frac{\alpha_0\boldsymbol{\beta} + \kappa\boldsymbol{\delta}_k}{\alpha_0 + \kappa}\right) \tag{11}$$

$$z_{t,i}|z_{t-1,i}, \{\boldsymbol{\pi}\} \sim \mathrm{Multinomial}\left(\boldsymbol{\pi}_{t,z_{t-1,i}}\right) \tag{12}$$

$$\eta_{k,l}|\xi, \psi \sim \mathrm{Beta}(\xi, \psi) \tag{13}$$

$$x_{t,i,j}|Z_t, \{\eta\} \sim \mathrm{Bernoulli}\left(\eta_{z_{t,i},z_{t,j}}\right). \tag{14}$$

まず、観測量 $x_{t,i,j} \in \{0, 1\}$ は時刻 t においてアイテム i からアイテム j に向かう有向関係の有無を表現する観測量である。そのパラメータは時刻 t におけるアイテムの所属するクラスタ $z_{t,i} = k, z_{t,j} = l$ によって選択される。$\eta_{k,l}$ はクラスタ k に所属するアイテムからクラスタ l に所属するアイテムへリンクが張られる確率である。前章で紹介した dMMSB は観測リンクの存在確率の計算に compatibility matrix $\boldsymbol{B}$ による K 次元ベクトルの距離計算が発生したが (式 (9))、排他的クラスタリングを仮定する dIRM では各アイテムの所属クラスタインデックスに対応する Bernoulli パラメータを参照するだけである (式 (14))。

一方、クラスタへの所属を表す変数は式 (12) にあるように、前の時刻に所属したクラスタ $z_{t-1,i}$ に依存してクラスタ遷移確率 $\boldsymbol{\pi}_{t,z_{t-1,i}}$ を切り替えてサンプリングされる。この遷移確率は無限次元のベクトルであり、$\pi_{t,k,l}$ は時刻 $t-1$ においてクラスタ k に所属したアイテムが時刻 t にクラスタ l に所属する確率を表す。このことで、dMMSB と異なり、離散的かつ時刻ごとに特異なクラスタ変動を許容する時系列関係データ解析を行うことができる。

このモデルの特徴は、先に述べたとおり 3 つの異なる条件を全て満たしている点にある．まず、各時刻における遷移確率は無限次元のディリクレ分布に相当する Dirhchelt Process(DP) を利用してサンプリングされる。ここで各遷移確率ベクトル $\boldsymbol{\pi}_{t,k}$ が時刻 t および前時刻の所属クラスタ k ごとに i.i.d にサンプリングされる。このことで、ある時刻で大きなクラスタの変化が起こってもそれに応じた遷移確率を個別にサンプリングし、非定常な変化をモデリングすることができる (条件 (A))。

次に、式 (11) では、sticky iHMM [4] と呼ばれるモデルを踏襲している。このモデルでは、通常の DP に $\kappa > 0$ という sticky parameter が加えられている。$\boldsymbol{\delta}_k$ は k 番目の要素が 1 でそれ以外の要素は全てゼロとなる無限次元ベクトルである。すなわち、各時刻 t にクラスタ k を遷移元とする遷移確率のパラメータにこの sticky 項が加わることで、クラスタ k から同じクラスタ k への遷移確率を $\frac{\kappa}{\alpha+\kappa}$ 分だけ大きくするという効果をもつので ([4]), 各アイテムは前時刻と同じクラスタへ所属する確率が高まる。これは条件 (B) を満たすための要素である。

最後に、式 (10) は全時刻を通してみたときのクラスタのサイズ（メンバーシップ）の比を表す無限次元ベクトルである。Stick() は stick breaking process [13] と呼ばれる、無限次元のベクトルを生成する確率プロセスである。このモデルにおいては、$\boldsymbol{\beta}, \boldsymbol{\pi}$ など全ての部分においてクラスタ数 K を固定せず、無限個のクラスタを仮定している。データに則した最適なクラスタ数は事後確率から自動的に決定される。これによって条件 (C) も担保されている。

4 おわりに

以上、時間発展に対する仮定にもとづいて大きく 2 種類の手法を紹介した。これらの手法の間には優劣はなく、解析の目的やデータの特性に応じて使い分けることが望ましい。また、生成モデルによらない手法 [14, 1] や、データマイニングとは違う応用分野での研究例 [6] なども数多く存在し、本稿は時系列関係データ解析のごく一部をカバーしたにすぎない。本稿が読者の皆様に幾許かの有益な情報を提供できれば幸いである。

参考文献

[1] A. Ahmed and E. P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 106(29):11878–11883, July 2009.

[2] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[3] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(Suppl 1):5220–5227, 2004.

[4] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, Helsinki, Finland, July 2008.

[5] F. Guo, S. Hanneke, Fu. W., and E. P. Xing. Recovering temporally rewiring networks: a model-based approach. In *Proceedings of the 24th international conference on Machine learning (ICML)*, 2007.

[6] O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, D. S. Chamock-Jones, C. Print, and S. Miyano. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, 24(7):932–942, 2008.

[7] K. Ishiguro, T. Iwata, N. Ueda, and J. Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23 (NIPS)*, 2010.

[8] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, Boston, MA, jul 2006.

[9] M. S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. In *Proceedings of the 35th International Conference on Very Large Data Bases (VLDB)*, volume 2, 2009.

[10] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 556–559. ACM, 2003.

[11] S. A. Myers and J. Leskovec. On the convexity of latent social network inference. In *Advances in Neural Information Processing Systems 23 (NIPS)*, 2010.

[12] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[13] J. Sethuraman. A constructive definition of dirichlet process. *Statistica Sinica*, 4:639–650, 1994.

[14] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 677–685, 2008.

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

[16] E. P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566, 2010.

[17] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. A Bayesian approach toward finding communities and their evolutions in dynamic social networks. In *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2009.

[18] S. Zhu, K. Yu, and Y. Gong. Stochastic relational models for large-scale dyadic data using mcmc. In *Advances in Neural Information Processing Systems 21 (NIPS)*, 2009.