情報処理学会論文誌　数理モデル化と応用　Vol. 3　No. 1　1–12 (Jan. 2010)

時間依存関係データ分析のための動的無限関係モデル

石黒勝彦[†1]　岩田具治[†1]　上田修功[†1]

近年，WWW や人間関係ネットワークなどオブジェクト間の関係を表す関係データの解析に大きな興味が寄せられている．関係データから構造を抽出するモデルとして，無限関係モデル（IRM）がある．IRM は，2 つのオブジェクト間で関係の有無が定義されたオブジェクト集合が与えられたとき，オブジェクト間の関係を最もよく要約した形になるようにそのオブジェクト集合を最適数のクラスタに分割する．しかし，IRM は静的なデータを対象としたモデルゆえ，関係が時間によって変化する，より現実的な関係データに対しては十分なモデルとはいえない．本論文では，IRM を拡張し，時間的に変化する関係データ解析のための新たなモデルを提案する．人工データおよび実データを用いた実験によりその有効性を確認する．

Dynamic Infinite Relational Model for Time-dependent Relational Data Analysis

Katsuhiko Ishiguro,[†1] Tomoharu Iwata[†1] and Naonori Ueda[†1]

Analysis of relational data such as the WWW and social networks' structures has drawn many attentions recently. The infinite relational model (IRM) is proposed as a model for this purpose. Given the relations between objects, IRM partitions the object set into the optimal number of clusters so that the relations between clusters well summarizes the relations between objects. Since IRM is a generative model for static relations, it is insufficient for dynamic relational data analysis where relations vary with time. In this paper, we extend IRM to a dynamic model to solve this problem. We show the usefulness of the model through experiments with synthetic and real world data sets.

†1 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

1. はじめに

近年，ICT 技術の発展に従ってインターネットは巨大なネットワークへと変化している．インターネット上で所望の情報を探し出す検索技術は，現在 PageRank [1] に代表されるようにホームページ間のリンク情報を用いたものが主流となっている．また，インターネット上で提供されるサービスも，SNS のように人と人のローカルなつながりによるネットワークに注目するものが多数提案されている．このように，近年では個々のデータだけでなく，ネットワーク構造に代表される “関係データ[⋆1]” の重要性が認識されている．実際，関係データの構造や生成過程のモデリングは，密な関係でつながっているコミュニティの抽出・発見[2] や，コミュニティの役割の解析，リンクの存在予測[3),4)] などに有用である．

関係データモデリングとして stochastic block model（SBM）[5] とその拡張である infinite relational model（IRM）[6] が著名である．これらのモデルは，オブジェクト間の関係データから，そのオブジェクト集合を，クラスタ間の関係を最も適切に表現するクラスタに分割する．SBM では，クラスタリングの際，クラスタ数を事前に決定しておく必要があるが，IRM では，ノンパラメトリックベイズの枠組みでデータから最適なクラスタ数が自動的に推定される．

現実の関係データ，たとえば人間関係や WWW ページのリンク関係は本質的に時間依存である．たとえば，WWW ページのリンクなどは，ある時話題になったページにはリンクが集中するが，ブームが過ぎればそれらのリンクは廃れてしまう．また，ブームを追いかける集団（ニュースサイトなど）は次のブームを追いかけて次々とリンク先を変化させるが，お互いの結合が強い SNS 上のコミュニティメンバは，流行に流されず，時間が経過しても高い確率でお互いをリンクするコミュニティに帰属し続ける可能性が高い．その一方で，コミュニティ自体が合併などによって消滅したり，新たに生成されたりすることもある．このようなリンク構造の時間変化によって，クラスタリング結果も時間とともに変化することが予想される．

このように，現実の関係データの中には，オブジェクト間の関係が時間的に変化するケースが数多く存在する．しかし，SBM，IRM などの従来モデルは，時間変化を直接モデル化していないため，時間変化する関係データへの適用は困難である．時間発展する関係データ

⋆1 本論文で述べる「関係データ」およびそのモデリング技術は，統計的機械学習の文脈におけるものを指し，データベースの研究分野で確立している「関係データ」およびその関連技術とは異なることに注意．

のためのモデルも提案されているが[7),8)]，これらの手法はクラスタ数を事前に決定・固定する必要があるなど，総じて既存の手法は関係性のダイナミックスをとらえることができないという点で十分とはいえない．

本論文では，IRM を，時間変化する関係データのために拡張した動的無限関係モデル（dynamic IRM，dIRM）を新たに提案する．前述したように，オブジェクトの帰属クラスタは時間とともに変化しうる．また，クラスタ数も時間とともに増減しうる．本モデルでは，ノンパラメトリックベイズおよびマルコフモデルに基づき，クラスタ数のダイナミックス，および，各オブジェクトの帰属確率のダイナミックスのための事前分布を導入することによりこれらの問題に対処している．

以下の本文では，まず 2 章で本モデルの基本となる IRM について説明する．続いて 3 章で提案モデルを説明する．4 章で人工データおよび実データによる実験結果を示す．5 章でまとめと今後の展望について述べる．

2. Infinite Relational Model（無限関係モデル）

本章では，提案モデルの基盤となる無限関係モデル（IRM）について説明する．

今，N 個のオブジェクトからなるオブジェクト集合 $D = \{1, 2, \ldots, N\}$ 上の二値の二項関係 $X : D \times D \to \{0, 1\}$ を考える（**図 1**）．本論文では説明を簡単にするため同一ドメインの二項関係（$D \times D$）を考えるが，IRM は，異なるドメイン間の関係（$D \times D'$）や三項関係（$D \times D \times D$）も取り扱うことができることに注意[6)]．

IRM は，N オブジェクト内で観測された関係データ $X = \{x_{i,j} \in \{0, 1\};\ 1 \le i, j \le N\}$ から，オブジェクト集合を複数のクラスタに分割する．これは SBM[5)] も同様だが，IRM ではこのクラスタ分割にノンパラメトリックベイズモデルの一種である Dirichlet Process Mixture（DPM）[9),10)] モデルを用いることで，そのクラスタ数も同時に推定可能としている点で SBM と異なる．観測データ $x_{i,j} \in \{0, 1\}$ は，オブジェクト i，j 間での関係の有無を表し，$x_{i,j} = 1(0)$ ならばオブジェクト i，j 間に関係が存在する（しない）ことを意味する．なお，この関係は有向，無向のどちらでもよく，無向ならば $x_{i,j} = x_{j,i}$ となる．

IRM は潜在クラスタを仮定する X の生成モデルで，以下のように表される．表記 $a|b \sim p$ は “b が所与の下で，a は分布 p から生成される” ことを意味する．なお $Z = \{z_i\}_{i=1}^{N}$，$H = \{\eta_{k,l}\}_{k,l=1}^{\infty}$ である．

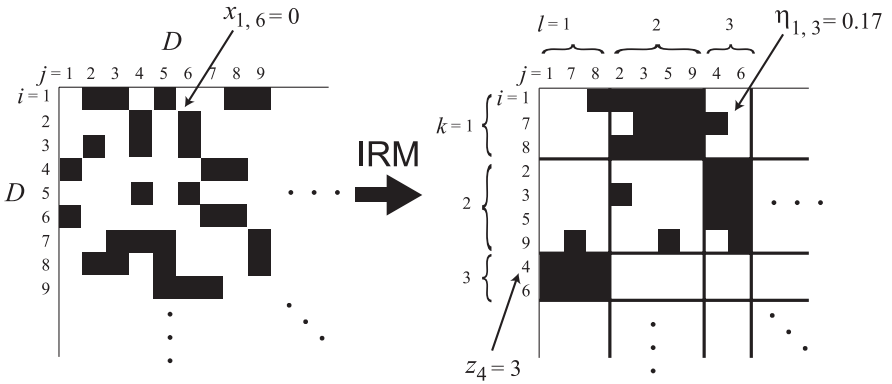

図 1　ブロックモデル（IRM）の例
Fig. 1　Example of block models (IRM).

$$\boldsymbol{\beta}|\gamma \sim \mathrm{Stick}\,(\gamma) \tag{1}$$

$$z_i|\boldsymbol{\beta} \sim \mathrm{Multinomial}\,(\boldsymbol{\beta}) \quad i = 1, 2, \ldots, N \tag{2}$$

$$\eta_{k,l}|\xi, \psi \sim \mathrm{Beta}\,(\xi, \psi) \quad k, l = 1, 2, \ldots, \infty \tag{3}$$

$$x_{i,j}|Z, H \sim \mathrm{Bernoulli}\left(\eta_{z_i, z_j}\right) \quad i, j = 1, 2, \ldots, N \tag{4}$$

まず，式 (1) では，無限次元のクラスタ混合比ベクトル $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ を生成する．これはオブジェクトが各クラスタに帰属する確率を表すベクトルである．右辺の Stick とは，DPM モデルを用いる際に利用する stick-breaking process[11)] を表す（**図 2**）．具体的には，長さ 1 の棒を，v_1 対 $1 - v_1$ の比で折り，v_1 に対応する部分の長さを β_1 とする．次いで，残された棒（$1 - v_1$ に対応する部分）をさらに v_2 対 $1 - v_2$ の比で折り，v_2 に対応する部分の長さを β_2 とする．以下この操作を繰り返すことにより $\beta_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$ を得る．ただし，v_k はベータ分布から生成されるものとする．すなわち $v_k \sim \mathrm{Beta}(1, \gamma)$．ここで γ（> 0）はパラメータである．元の棒の長さは 1 ゆえ $\sum_k \beta_k = 1$ が保証される．IRM ではこれを無限クラスタの混合比の事前分布として利用している．続いて，この混合比に従ってオブジェクト i が帰属するクラスタ $z_i = k$，$k = 1, 2, \ldots, \infty$ を多項分布よりサンプリングする（式 (2)）．

残りの 2 式は実際に観測された関係データを生成する過程である．式 (3) に従い，クラスタ k，l 間の関係の強さを表すパラメータ $\eta_{k,l}$ をサンプリングする．この値は，図 1 において，(k, l) で表されるブロック内の $x_{i,j}$ が 1 となる確率を表す．式 (4) では，$Z = \{z_i\}_{i=1}^{N}$，

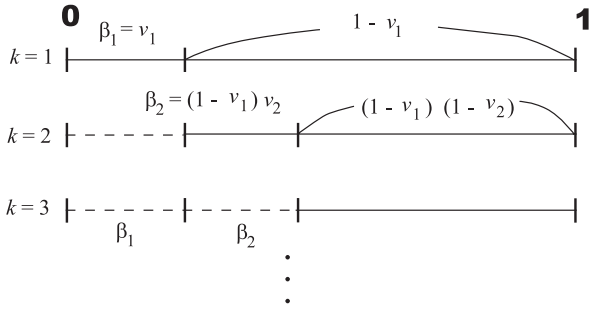

図 2　Stick-breaking 過程のイメージ図
Fig. 2　Stick-breaking process.

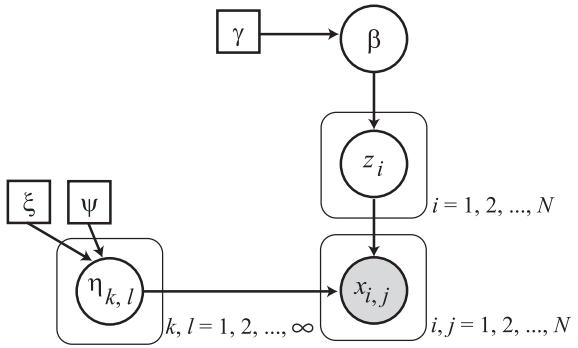

図 3　IRM のグラフィカルモデル（Eqs.1–4）．円ノードは変数，矩形ノードは定数を表す．影付きのノードは観測量を表す
Fig. 3　Graphical model of the IRM model (Eqs.1–4). Circle nodes are the variables, and the square nodes are constants. Shaded nodes indicate the observations.

$H = \{\eta_{k,l}\}_{k,l=1}^{\infty}$ が与えられたときに実際の関係データ $x_{i,j}$ をサンプリングする．$x_{i,j}$ の値は，各オブジェクト i, j の帰属するクラスタと，それによって規定されるブロックの関係の強さ η_{z_i,z_j} に従ってベルヌーイ分布で生成される．以上の過程のグラフィカルモデルを**図 3** に示す．グラフィカルモデルとは各変数の生成の依存関係を図示したもので，変数間にリンクがない場合は，その変数間は統計的に条件付き独立であることを意味する．逆に，変数 a から b へのリンクが存在する場合，b の生成は a の値に依存することを意味する．

3. 動的 IRM モデル

3.1　対象とする関係データ

本論文で対象とする時刻データを含む関係データは $X = \{x_{t,i,j} \in \{0,1\};\ i,j = 1,2,\ldots,N,\ t = 1,2,\ldots,T\}$ で表されるものとする．ここで，$x_{t,i,j} = 1(0)$ は時刻 t においてオブジェクト i，j に関係がある（ない）ことを意味する．時刻 t は離散時間とし，T 時刻までのデータが観測されているとする．N は総オブジェクト数．また，異なる時刻のオブジェクト間では関係は定義されないものとする．換言すれば，X は T 時刻での関係データの集合といえる．ただし，これらは独立ではなく，一般に何らかのダイナミックスを有する関係データを想定している．

たとえば，インターネット上のリンク関係が時間とともに変化することから各オブジェクトも時間とともに異なるクラスタ間を遷移すると予想される．しかし，その変化は完全なランダムではなく，たとえば，あるオブジェクトは隣接時刻では同じクラスタに高い確率で帰属しやすいであろう．これは，連続な時系列データを想定する場合には自然な仮定である．すなわち

［1］隣接時刻でのオブジェクトの帰属クラスタは高い相関を持つ

ことが予想される．

また，企業内の人間関係には部署などを反映したクラスタが構成されると思われるが，このようなクラスタは部署の合併や分裂などによって大きく変化しうる．ただし，合併や分裂はある時刻で突発的に起こりうる．インターネット上のリンク関係を大きく変化させるようなブームも同様である．すなわち

［2］オブジェクトのクラスタリングの時間発展は，絶対時間に依存し，非一様である

と考えられる．

また，クラスタの生成や消滅，合併・分裂などが発生するということは，

［3］クラスタ数は一定ではなく時間変化する

ということである．

したがって，以上 3 つの性質を有する時系列関係データを適切にモデル化することが必要となる．

3.2　ナイーブな拡張

提案モデルの説明の前に，まず，時間的に変化する関係データに対処するための IRM のいくつかのナイーブな拡張法について検討し，その問題点を整理する．

最も単純には，時刻データを含んだ関係データ X から時刻データを含まない関係データ $\tilde{X} = \{\tilde{x}_{i,j}\}$ を生成して通常の IRM を適用する方法が考えられる．たとえば

$$\tilde{x}_{i,j} = \begin{cases} 1 & \frac{\sum_{t=1}^{T} x_{t,i,j}}{T} > \sigma \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

とすればよい．ここで，σ は閾値である．この $\tilde{X}$ に対して IRM を適用し，そのクラスタリング結果を全時刻でのオブジェクトクラスタリングに適用する．しかし，この方法では関係性の時間変化が無視されるため，明らかに本論文の目的に整合しない．

IRM モデルでは異なるドメイン間の関係を表現できるので，$D^1 = \{1, 2, \ldots, N\}$ をオブジェクトのインデックス，$D^2 = \{1, 2, \ldots, T\}$ を時間のインデックスを表すとして，$X : D^1 \times D^1 \times D^2 \to \{0, 1\}$ に対して IRM モデルを適用する方法も考えられる．しかし，この方法ではオブジェクトクラスタリングの結果が全時刻で同一となる（$z_{1,i} = z_{2,i} = \cdots = z_{T,i}$）ため，クラスタリングの時間変化をモデル化することができない．

また，時系列関係データ X を時刻ごとの関係データ X_t に分解し，各時刻で独立に IRM を適用することも考えられる．この場合，確かに各時刻で異なるクラスタリング結果が得られるが，クラスタリングの時間変化情報が有効利用されないという問題がある．

クラスタリングの時間変化をモデル化する方法として，以下のように，通常の IRM でのオブジェクト i のクラスタ帰属変数 z_i を時刻 t 依存（$z_{t,i}$）とし，かつ，全時刻でその生成分布パラメータ $\boldsymbol{\beta}$ を共有化する簡易なモデル化も考えられる（**図 4**）．この共有化により，各時刻で独立に IRM を適用する場合と異なり，異なる時刻（隣接時刻とは限らない）のクラスタリング結果が高い相関を持つことが期待される．なお，Z_t，H はそれぞれ $Z_t = \{z_{t,i}\}_{i=1}^{N}$，$H = \{\eta_{k,l}\}_{k,l=1}^{\infty}$ を表す．

$$\boldsymbol{\beta}|\gamma \sim \text{Stick}\,(\gamma) \tag{6}$$

$$z_{t,i}|\boldsymbol{\beta} \sim \text{Multinomial}\,(\boldsymbol{\beta}) \quad t = 1, 2, \ldots, T, i = 1, 2, \ldots, N \tag{7}$$

$$\eta_{k,l}|\xi, \psi \sim \text{Beta}\,(\xi, \psi) \quad k, l = 1, 2, \ldots, \infty \tag{8}$$

$$x_{t,i,j}|Z_t, H \sim \text{Bernoulli}\left(\eta_{z_{t,i}, z_{t,j}}\right) \quad t = 1, 2, \ldots, T, i, j = 1, 2, \ldots, N \tag{9}$$

しかし，このモデルではすべての t，i についてクラスタインデックス $z_{t,i}$ が $\boldsymbol{\beta}$ 所与の下で条件付き独立となり，時刻 $t-1$ と時刻 t でのクラスタリングに直接の依存関係がモデル化されていない．換言すれば，時刻の順序が無視されたモデルゆえ，時間発展のモデル化としては適切ではない．

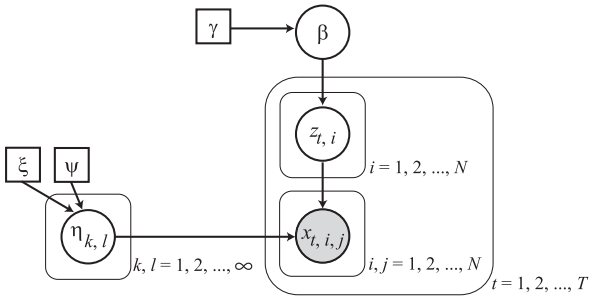

図 4　“tIRM” モデル (Eqs.6–9) のグラフィカルモデル．円ノードは変数を，矩形ノードは定数をそれぞれ表す．影付きのノードは観測量である

Fig. 4　Graphical model of “tIRM” model (Eqs.6–9). Circle nodes are the variables, and the square nodes are constants. Shaded nodes indicate the observations.

3.3　提案モデル

それでは，上述の問題をふまえて提案モデルの詳細を述べる．

前節までの考察に基づき，以下の性質を持つモデル化が必要といえる．

［1］クラスタリング結果が隣接時刻間で高い相関を有すること．

［2］クラスタリングの時間発展は絶対時間に依存し，一様でないこと．

［3］クラスタ数の時間変化を許容すること．

すなわち，性質［1］は，あるオブジェクトは隣接時刻では同じクラスタに高い確率で帰属しやすいことを意味し，本論文が対象とする関係データでは直観的に妥当といえる．また，前述したように，オブジェクトの部分集合であるクラスタは，突発的に生成や消滅，合併・分裂を起こしながら時間発展していくと考えられる．その意味で，性質［2］, ［3］も自然な要請といえる．

これらの性質を満たすべく，IRM を以下のように拡張する．なお，$\Pi_t = \{\boldsymbol{\pi}_{t,k} : k = 1, \ldots, \infty\}$ である．

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

$$\boldsymbol{\pi}_{t,k}|\alpha_0, \kappa, \boldsymbol{\beta} \sim \text{DP}\left(\alpha_0 + \kappa, \frac{\alpha_0 \boldsymbol{\beta} + \kappa \boldsymbol{\delta}_k}{\alpha_0 + \kappa}\right) \quad t = 1, 2, \ldots, T, k = 1, 2, \ldots, \infty \tag{11}$$

$$z_{t,i}|z_{t-1,i}, \Pi_t \sim \text{Multinomial}\left(\boldsymbol{\pi}_{t, z_{t-1,i}}\right) \quad t = 1, 2, \ldots, T, i = 1, 2, \ldots, N \tag{12}$$

$$\eta_{k,l}|\xi, \psi \sim \text{Beta}\,(\xi, \psi) \quad k, l = 1, 2, \ldots, \infty \tag{13}$$

$$x_{t,i,j}|Z_t, H \sim \text{Bernoulli}\left(\eta_{z_{t,i}, z_{t,j}}\right) \quad t = 1, 2, \ldots, T, i, j = 1, 2, \ldots, N \tag{14}$$

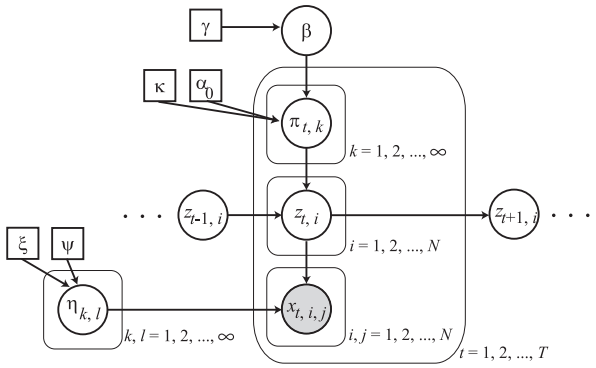

図 5　提案する dIRM モデルのグラフィカルモデル（Eqs.10–14）．円ノードは変数を，矩形ノードは定数を表す．影付きノードは観測量である

Fig. 5　Graphical model of the proposed dIRM model (Eqs.10–14). Circle nodes are the variables, and the square nodes are constants. Shaded nodes indicate the observations.

同時に上記モデルに対応するグラフィカルモデルを**図 5** に示す．IRM との違いは式 (11)，式 (12)，および，式 (14) にある．以下，これらについて説明する．

まず，$\boldsymbol{\beta}$ は全時刻での平均的な各クラスタへの帰属確率を表すベクトルである．$\boldsymbol{\pi}_{t,k} = (\pi_{t,k,1}, \pi_{t,k,2}, \ldots, \pi_{t,k,l}, \ldots)$ は，時刻 $t-1$ においてクラスタ k に帰属していたオブジェクトが，時刻 t で第 l クラスタに遷移する（$l = 1, 2, \ldots$）確率を表す．すなわち，オブジェクトの帰属クラスタに確率的なマルコフ遷移を導入し，時系列関係データクラスタリングのダイナミックスを表現する．明らかに，$\pi_{t,k,l} > 0$ かつ $\sum_{l=1}^{\infty} \pi_{t,k,l} = 1$．つまり，$\boldsymbol{\pi}_{t,k}$ は無限次元のディリクレ分布に相当し，実際には，ノンパラメトリックベイズで多用されるディリクレ過程（Dirichlet Process: DP）を用いて生成する．DP のより詳細な説明は文献 10) に譲り，ここでは，モデルの理解に必要な事項のみを簡単に説明する．

DP は確率分布に対する分布である．DP は基底分布 G_0 と正のパラメータ α_0 を用いて定義される．分布 G が DP に従うとき，$G \sim \mathrm{DP}(\alpha_0, G_0)$ と表記する．G は確率分布ゆえ，新たな確率変数 θ が G から生成されるとすると，DP では以下のような過程で θ_i が生成される．

1. $\theta_1 \sim G_0(\theta)$
2. $\theta_2 \sim \frac{1}{\alpha_0+1}\delta_{\theta_1}(\theta) + \frac{\alpha_0}{\alpha_0+1}G_0(\theta)$

   ...

i. $\theta_i \sim \frac{1}{\alpha_0+i-1}\delta_{\theta_1}(\theta) + \cdots + \frac{1}{\alpha_0+i-1}\delta_{\theta_{i-1}}(\theta) + \frac{\alpha_0}{\alpha_0+i-1}G_0(\theta)$

$\delta_x(y)$ は $x = y$ のとき 1，それ以外は 0 となるディラックのデルタ関数を表す．上記過程は，θ_i が，すでに生成された $\theta_1, \theta_2, \ldots, \theta_{i-1}$ から選択されるか，あるいは，基底分布（θ の事前分布）G_0 から新規に生成され，かつ，θ_j，$j = 1, \ldots, i-1$ については確率 $1/(\alpha_0+i-1)$ で生成され，新規の θ については確率 $\alpha_0/(\alpha_0+i-1)$ で生成されることを意味する．$i-1$ までに生成された θ_j（$j < i$）の種類が $\theta_{(1)}, \ldots, \theta_{(K)}$，すなわち，異なり数が K のとき，θ_i は，各パラメータの出現個数に比例した確率で $\theta_{(i)}$ を生成するが，それだけではなく，α_0 に比例した確率で新規の $\theta_{(K+1)}$ を生成しうるプロセスとなっている．これまで多く出た種類のパラメータほど，以後もよく出現し，パラメータのクラスタリングが自然に定義される．

翻って，式 (12) の場合，α_0 が $\alpha_0 + \kappa$ に，G_0 が $(\alpha_0\boldsymbol{\beta} + \kappa\boldsymbol{\delta}_k)/(\alpha_0 + \kappa)$ に相当する．$\boldsymbol{\delta}_k$ は第 k 要素の値が 1，他はすべて 0 のベクトルである．すなわち $(\alpha_0\boldsymbol{\beta} + \kappa\boldsymbol{\delta}_k)$ は κ（> 0）が $\alpha_0\boldsymbol{\beta}$ の第 k 要素に加算されていることを意味する．つまり，$(\alpha_0\boldsymbol{\beta} + \kappa\boldsymbol{\delta}_k)$ の第 j（$\neq k$）要素は $\alpha_0\beta_j$，第 k 要素は $\alpha_0\beta_k + \kappa$ となる．式 (11) の $\boldsymbol{\pi}_{t,k}$ が，時刻 $t-1$ においてクラスタ k に帰属していたオブジェクトが，時刻 t にどのクラスタにどの程度の確率で遷移するかを表すクラスタ遷移確率ゆえ，第 k 要素にバイアスを付与した無限次元多項分布：$(\alpha_0\boldsymbol{\beta} + \kappa\boldsymbol{\delta}_k)/(\alpha_0 + \kappa)$ を DP の基底分布とすることで，$\boldsymbol{\pi}_{t,k}$ の第 k 要素 $\pi_{t,k,k}$ の値が $\pi_{t,k,l}$（$l \neq k$）に比べ $\kappa/(\alpha_0 + \kappa)$ だけ確率値を加算させることになる．これにより，式 (12) での $z_{t,i}$ のサンプリングにおいて，$z_{t-1,i} = k$ のとき，$z_{t,i} = k$ とサンプリングされる確率を相対的に大きくすることができ，性質［1］を反映したモデル化が実現できる．パラメータ κ の値を大きくすればするほど，その性質が強められる．実際には，次節で説明するように，κ の最適値はデータから学習して決定する．また，非一様なクラスタリングの時間発展（性質［2］）は，マルコフモデルの観点からは遷移確率パラメータが時刻に依存して変化することに相当する．提案モデルではクラスタ遷移確率のパラメータである $\boldsymbol{\pi}_{t,k}$ は，平均混合比ベクトル $\boldsymbol{\beta}$ をパラメータとして時刻 t ごとにサンプリングされているため，性質［2］を満たすことが可能である．また，クラスタのインデックス k にも依存するため，オブジェクトの帰属するクラスタごとに特徴ある遷移パターンをモデル化することも可能となる．さらに，提案モデルは IRM を踏襲しているので，性質［3］を満たすことも明らかである．以上説明した提案モデルを，動的 IRM（dynamic IRM: dIRM）と呼ぶこととする．

ここで，既存のモデルと dIRM モデルとの関係について考察する．Teh ら[12] および Fox ら[13] によって提案された infinite HMM は，無限モデルに対するマルコフモデルという点で dIRM との関係が深い．infinite HMM は通常の HMM と同様に，推定すべき隠れ状態

系列 $\{s_{1:T}\}$ と対応する観測量が与えられたときに，自動的に隠れ状態数（dIRM ではクラスタ数に相当）を推定できる．しかし，一般に 1 つのデータシーケンスに対して推定すべき隠れ状態系列は 1 つのみである．一方，時系列関係データにおいては，1 つの関係データシーケンスから各オブジェクトの帰属クラスタ系列を同時に推定する必要がある．したがって，複数隠れ状態系列の同時推定が必要な関係データの解析には，infinite HMM を直接利用することができない．dIRM は infinite HMM における隠れ状態系列の数をオブジェクト数 N にまで拡張し，関係データへ適用可能にしたモデルと理解できる．

さらに，通常の (infinite)HMM ではクラスタ間の遷移確率が時間不変であるが，dIRM モデルは遷移確率 $\boldsymbol{\pi}_{t,k}$ が時刻 t にも依存する．したがって，前述したように，ブームによるクラスタの人気の変化や突発的なクラスタの合併や分裂などが考えられる場合，時刻によって遷移確率が変化する dIRM のようなモデル化が妥当といえる．ただし，前にも述べたように (infinite)HMM モデルとは対象とするタスクと適用範囲に違いがあるため，一概にこのことでモデル自身の優劣を示唆するものではない．

なお，関係データに対する時系列モデルは，提案モデルのほかにも文献 7)，8) などが最近提案されているが，いずれの手法もクラスタ数を事前に決定しておく必要があり，この点で我々の目的を満たさない．文献 8) は SBM モデルの時系列拡張モデルである．SBM の一般化モデルである IRM を時間拡張した dIRM モデルは文献 8) を包含したモデルといえる．一方，文献 7) は IRM や BPM とは異なるタスクのためのモデルであり，提案モデルとの直接的な関連性は薄いといえる．

以上より，クラスタ数などの事前知識なしに時間ダイナミックスを考慮した時系列関係データのクラスタリングが可能なモデルの提案は，IRM モデルが初めての試みであるといえる．

3.4 学　　習

本節では，観測された関係データ $X = \{x_{t,i,j}\}$ が与えられたときのモデルパラメータの学習方法について説明する．

ここでの学習とは $z_{t,i}$，$\boldsymbol{\pi}_{t,k}$，$\eta_{k,l}$，$\boldsymbol{\beta}$ およびハイパーパラメータ γ，κ，α_0，ξ，ψ を推定することである．具体的には Gibbs サンプリング手法に基づいて，他の変数がすべて所与のもとで各変数の事後分布を導出し，その分布を用いて変数のサンプリングを繰り返せばよい．

しかし，提案モデルのように，$z_{t,i}$ が $z_{t-1,i}$ に強く依存する場合，Gibbs サンプリングの収束が著しく遅くなるという問題がある．この種の問題に対し，近年 beam サンプリング法[14] が提案されている．本論文ではこの beam サンプリング法に基づいて学習アルゴリズムを導出する．ただし以下では紙面の都合上，詳しい導出は省略し，主要な結果のみについて概説する．

beam サンプリングでは，補助変数（auxiliary variable）$U = \{u_{t,i}\}$ を導入して，Z などと同時にサンプリングする．U を導入することで，無限個存在しうるクラスタ数を，U の条件付きの下で有限個に制限して学習を進めることが可能になる．したがって，$z_{t,i}$ のサンプリングにおいて通常の HMM における forward-backward アルゴリズムと類似した手法を用いて効率的なサンプリングが可能となる．さらに $\boldsymbol{\beta}$ や $\boldsymbol{\pi}_{t,k}$ のサンプリングに関しても，本来の無限次元分布ではなく，有限次元の分布からのサンプリングとなるため，アルゴリズムが簡潔になる．

3.4.1 $u_{t,i}$ のサンプリング

まず，$u_{t,i}$ のサンプリングについて説明する．

$u_{t,i}$ の事前分布は一様分布と仮定する．また，η，$\boldsymbol{\pi}$ が既知のとき，u，z と x の同時分布が次のようになるものと仮定する．

$$\begin{aligned} &p\left(x_{t,i,j}, u_{t,i}, u_{t,j}, z_{t-1:t,i}, z_{t-1:t,j}\right) \\ &\quad = \mathbb{I}\left(u_{t,i} < \pi_{t,z_{t-1,i},z_{t,i}}\right) \mathbb{I}\left(u_{t,j} < \pi_{t,z_{t-1,j},z_{t,j}}\right) x_{t,i,j}^{\eta_{z_{t,i},z_{t,j}}} \left(1 - x_{t,i,j}\right)^{1-\eta_{z_{t,i},z_{t,j}}} \end{aligned}$$

ここで，$\mathbb{I}$ は，続く条件式が満たされれば 1，そうでなければ 0 の値をとる．

上式をすべての時刻，オブジェクトに対して掛け合わせることで，$\boldsymbol{\beta}$，$\Pi = \{\boldsymbol{\pi}_{t,k}\}$，$H = \{\eta_{k,l}\}$ が所与の下での $X = \{x_{t,i,j}\}$，$Z = \{z_{t,i}\}$，$U = \{u_{t,i}\}$ の同時分布は次のようになる．

$$p\left(X, Z, U | \Pi, H, \boldsymbol{\beta}\right) = \prod_t \prod_i \mathbb{I}\left(u_{t,i} < \pi_{t,z_{t-1,i},z_{t,i}}\right) \prod_{i,j} x_{t,i,j}^{\eta_{z_{t,i},z_{t,j}}} \left(1 - x_{t,i,j}\right)^{1-\eta_{z_{t,i},z_{t,j}}}$$

また，$u_{t,i}$ を積分消去することで，$p\left(X, Z | \Pi, H, \boldsymbol{\beta}\right)$ も次のように求まる．

$$p\left(X, Z | \Pi, H, \boldsymbol{\beta}\right) = \prod_t \prod_i \pi_{t,z_{t-1,i},z_{t,i}} \prod_{i,j} x_{t,i,j}^{\eta_{z_{t,i},z_{t,j}}} \left(1 - x_{t,i,j}\right)^{1-\eta_{z_{t,i},z_{t,j}}}$$

この 2 式より，U の事後分布は次のように求まる．

$$p\left(U | X, Z, \Pi, H, \boldsymbol{\beta}\right) = \frac{p\left(X, Z, U | \Pi, H, \boldsymbol{\beta}\right)}{p\left(X, Z | \Pi, H, \boldsymbol{\beta}\right)} = \prod_t \prod_i \frac{\mathbb{I}\left(u_{t,i} < \pi_{t,z_{t-1,i},z_{t,i}}\right)}{\pi_{t,z_{t-1,i},z_{t,i}}}$$

したがって，$u_{t,i}$ は以下の事後分布からサンプリングすればよい．

$$u_{t,i} \sim \mathrm{Uniform}\left(0, \pi_{t,z_{t-1,i},z_{t,i}}\right) \leftrightarrow p\left(u_{t,i}\right) = \begin{cases} \frac{1}{\pi_{t,z_{t-1,i},z_{t,i}}} & 0 \le u_{t,i} \le \pi_{t,z_{t-1,i},z_{t,i}} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

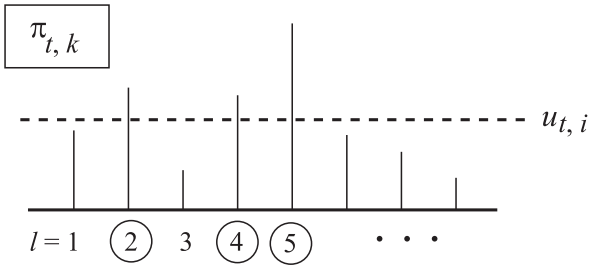

図 6　auxiliary variable $u_{t,i}$ によるクラスタの制限．この例ではクラスタ 2，4，5 のみが遷移先クラスタとしてサンプリング可能になる

Fig. 6　Limiting transition clusters by the auxiliary variable $u_{t,i}$. In this case the cluster 2, 4 and 5 are the only possible clusters to move.

なお，他の変数の事後分布も同様にして求めることができる．

前述のように，$u_{t,i}$ は実際に取り扱うクラスタ数を有限個にするために用いられる．具体的には，$z_{t,i}=k$ のサンプリングにおいて，$u_{t,i}>\pi_{t,z_{t-1,i},k}$ となる k を選択する確率を 0 にする（**図 6**）．この工夫により，$u_{t,i}$ が所与の下では遷移可能なクラスタ数が有限個に固定されるため，他の変数のサンプリングが簡単になる．

3.4.2　$z_{t,i}$ のサンプリング

時刻 t におけるオブジェクト i の帰属クラスタを表す $z_{t,i}$ のサンプリングを説明する．

HMM における forward-backward アルゴリズムと同様に，次のメッセージ変数を定義する．

$$p_{t,i,k}=p\left(z_{t,i}=k|X_{1:t},U_{1:t},\Pi,H,\boldsymbol{\beta}\right) \tag{16}$$

$z_{t,i}$ のサンプリングでは，まず上記のメッセージ変数を $t=1$ から $t=T$ まで計算する（forward filtering）．次に，この変数を用いて，$z_{t,i}=k$ の値を $t=T$ から $t=1$ までサンプリングする（backward sampling），という 2 段階のステップをとる．

forward filtering では次の式を計算する．

$$p_{t,i,k}\propto p\left(x_{t,i,i}|z_{t,i}=k,H\right)\prod_{j\neq i}p\left(x_{t,i,j}|z_{t,i}=k,H\right)p\left(x_{t,j,i}|z_{t,i}=k,H\right)\sum_{l:u_{t,i}<\pi_{t,l,k}}p_{t-1,i,l} \tag{17}$$

すなわち $z_{t,i}$ に関するすべての観測量の尤度（式 (14)）を既存のメッセージにかける．上記の式を $t=1$ から $t=T$ まですべての i，k について計算する．$t=1$ の場合には，右辺最終項を無視して計算する．

次に，backward sampling では，$z_{t,i}$ を次の事後分布よりサンプリングする．

$$p\left(z_{t,i}=k|z_{t+1,i}=l\right)\propto p_{t,i,k}\pi_{t+1,k,l}\mathbb{I}\left(u_{t+1,i}<\pi_{t+1,k,l}\right) \tag{18}$$

上記の式を，$t=T$ から $t=1$ まで，すべての i，k について計算して $z_{t,i}=k$ をサンプリングする．$t=T$ の場合には，右辺は第 1 項以外を無視する．$u_{t+1,i}$ によって，とりうる k の値が有限個に制限されることに注意．

結果，U が所与の下でサンプリングされた Z の値は，有限（K）種のバリエーションしか持たない．したがって，すべての $z_{t,i}$ のサンプリングが終了した時点で K の計算およびクラスタインデックス k の整理を行い，$z_{t,i}\in\{1,2,\ldots,K\}$ とする．

3.4.3　$\boldsymbol{\pi}_{t,k}$ のサンプリング

続いて，時刻 t においてクラスタ k から遷移するオブジェクトの遷移確率である $\boldsymbol{\pi}_{t,k}$ のサンプリングを説明する．

時刻 t において，$z_{t-1,i}=k$ かつ $z_{t,i}=l$ となるオブジェクト i の数を，クラスタ k から l に遷移したオブジェクト数 $m_{t,k,l}$ とする．U，Z が所与の下ではクラスタ数は有限 K 個なので，DP は通常の Dirichlet 分布となる．具体的には次式のようになる．ここで $\boldsymbol{\beta}$ は K 次元ベクトルになっていることに注意．

$$\boldsymbol{\pi}_{t,k}\sim\mathrm{Dirichlet}\left(\alpha_0\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k\right)$$

このとき，$\boldsymbol{\pi}_{t,k}$ の事後分布は，事前分布のパラメータにオブジェクトのクラスタリング結果 Z から得られる遷移情報 $m_{t,k,l}$ を加えた形になる．

$$\boldsymbol{\pi}_{t,k}\sim\mathrm{Dirichlet}\left(\alpha_0\beta_1+m_{t,k,1},\ldots,\alpha_0\beta_k+m_{t,k,k}+\kappa,\ldots,\alpha_0\beta_K+m_{t,k,K},\alpha_0\beta_u\right) \tag{19}$$

ここで，$\beta_u=1-\sum_{k=1}^{K}\beta_k$ である．この式は，時刻 $t-1$ において同じクラスタ k に帰属したオブジェクトの遷移結果を反映した形になっている．

3.4.4　$\eta_{k,l}$ のサンプリング

各クラスタ間の強さパラメータ $\eta_{k,l}$ のサンプリングについて説明する．

$z_{t,i}=k$，$z_{t,j}=l$ となる $x_{t,i,j}$ の数を全時刻にわたって加えたものを $N_{k,l}$，そのうち $x_{t,i,j}=1$ となった観測値の数を $n_{k,l}$ とする．$\eta_{k,l}$ の事前分布は式 (13) にあるとおりである．その事後分布は，事前分布のパラメータ ξ，ψ に，ブロックごとの観測量の値の情報を表す $n_{k,l}$，$N_{k,l}$ を加えた次式の形になる．

$$\eta_{k,l}\sim\mathrm{Beta}\left(\xi+n_{k,l},\psi+N_{k,l}-n_{k,l}\right) \tag{20}$$

3.4.5　$\boldsymbol{\beta}$ のサンプリング

クラスタの全体的な混合比を表す $\boldsymbol{\beta}$ のサンプリングについて説明する．

U，Z が所与の下ではクラスタ数は有限 K 個なので，$\boldsymbol{\beta}$ の事後分布も通常の Dirichlet 分

布となる．具体的には，次の式となる．

$$\boldsymbol{\beta} \sim \mathrm{Dirichlet}\left(\sum_{t,k}\hat{R}_{t,k,1}, \sum_{t,k}\hat{R}_{t,k,2}, \ldots, \sum_{t,k}\hat{R}_{t,k,K}, \gamma\right) \tag{21}$$

ここで $\hat{R}_{t,k,l}$ は次のように計算される．

$$\hat{R}_{t,k,l} = R_{t,k,l} - \delta_k(l)O_{t,k} \tag{22}$$

$R_{t,k,l}$ および $O_{t,k}$ は各々下記の式に従ってサンプリングする．

$$p\left(R_{t,k,l} = r | z_{t,i}, \theta\right) = s\left(m_{t,k,l}, r\right)\left(\alpha_0\beta_l + \kappa\delta_k(l)\right)^r \frac{\Gamma\left(\alpha_0\beta_l + \kappa\delta_k(l)\right)}{\Gamma\left(\alpha_0\beta_l + \kappa\delta_k(l) + m_{t,k,l}\right)} \tag{23}$$

$$O_{t,k} \sim \mathrm{Binomial}\left(R_{t,k,k}, \frac{\kappa}{\kappa + \alpha_0\beta_k}\right) \tag{24}$$

ここで，$s()$ は unsigned stirling number of the first kind とよばれる関数である．

これらの式は，他の変数と同様に事後分布の計算から自然に導出される．導出の詳細については文献 13) を参照のこと．

3.4.6　ハイパーパラメータのサンプリング

ハイパーパラメータ γ，κ，α_0，ξ，ψ も同様にサンプリングにより同時推定可能であるが，具体的な計算式については省略する．文献 12) などが参考となる．

最後に，学習アルゴリズムの全体のフローを**図 7** に掲載する．

1. INPUT hyper parmeters and the observation X
2. INITIALIZE $u_{t,i}$, $z_{t,i}$, $\pi_{t,k}$, $\eta_{k,l}$ and β

for loop=1:max_loop **do**

3. SAMPLE $u_{t,i}$ using Eq. (15)
4. SAMPLE $z_{t,i}$ using Eqs. (16-18)
5. SAMPLE $\pi_{t,k}$ using Eq. (19)
6. SAMPLE $\eta_{k,l}$ using Eq. (20)
7. SAMPLE β using Eqs. (21-24)
8. SAMPLE hyper parameters if needed

end for

9. OUTPUT estimated results

図 7　学習プロセスの擬似コード
Fig. 7　Inference process of dIRM model.

4. 実　　験

4.1　人工データによる評価

まず，クラスタリングの正解が既知の人工データを用いてモデルの定量的な評価を行った．人工データは 2 つ生成した．1 つ目の人工データ（synth1）は時間ステップ数 $T = 5$，オブジェクト数 $N = 16$，クラスタ数は $K = 4$ とした．全時刻で $K = 4$ 種類のクラスタが存在する．一部のオブジェクト（のべ 6%）は時間に応じてクラスタ間を遷移させた．2 つ目の人工データ（synth2）は時間ステップ数 $T = 10$，オブジェクト数 $N = 54$，クラスタ数は $K = 6$ とした．このデータでは，クラスタの一部が消滅もしくは新たなクラスタが発生するうえに，オブジェクトのクラスタ間遷移も synth1 データよりも頻繁に起こるデータとした（のべ 15%）．いずれのデータに関しても，クラスタ間の関係の強さを表す $\eta_{k,l}$ は positive なクラスタ間では $\eta = 0.9$，negative なクラスタ間では $\eta = 0.1$（synth1）あるいは $\eta = 0.05$（synth2）の 2 種類を用意した．クラスタ間の関係を positive とするか，あるいは negative とするかは事前に設定する．与えられた $z_{t,i}$，$\eta_{k,l}$ に従って各時刻の観測量 $x_{t,i,j}$ を生成した．各々の観測データの例を**図 8**（synth1），**図 9**（synth2）に示す．

実験では，上記の手続きに従って生成された観測データ X からオブジェクトのクラスタリングを行った．実験では以下の 3 モデルを比較評価した．

(1)　dIRM
(2)　tIRM（時間インデックスを導入した IRM）
(3)　通常の IRM

tIRM は，式 (6)，式 (7)，式 (8)，式 (9) の生成モデルで定義される時間インデックスを導入した IRM である．ただし，dIRM と違い時間方向の依存関係がモデル化されていない．通常の IRM では，式 (5) において，$\sigma = 0.5$ としてクラスタリングしたのち，IRM でのオブジェクト i のクラスタリング結果 z_i を各時刻でのオブジェクト i のクラスタリング $z_{t,i}$ と見なした．

本実験では，クラスタリングの評価尺度の 1 つである，Rand index[15] を利用した定量的評価を行った．Rand index とは，あるデータに対して 2 つのクラスタリング結果が与えられたとき，2 つのクラスタリング結果の類似度を測る指標である．Rand index は非負であり，2 つのクラスタリング結果が完全一致したときに最大値 1 をとる．本実験では，観測

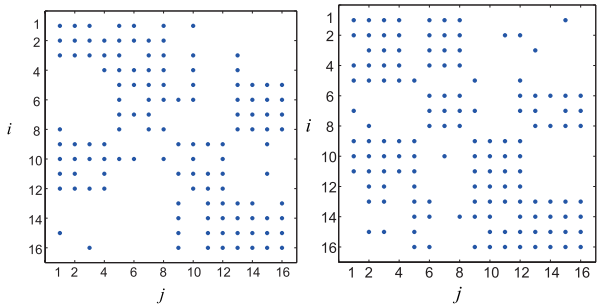

図 8　人工データ 1（synth1）の例．左は $t = 1$，右は $t = 4$ での観測結果である
Fig. 8　Examples of the synthetic data 1. Left: observations at $t = 1$. Right: observations at $t = 4$.

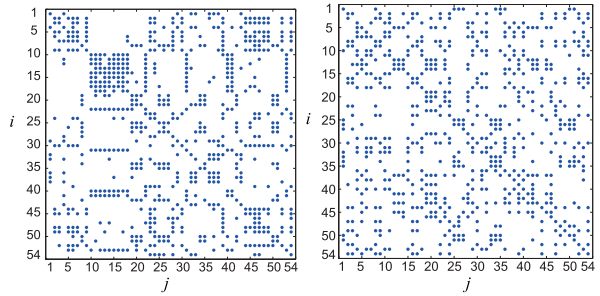

図 9　人工データ 2（synth2）の例．左は $t = 3$，右は $t = 8$ の観測結果である
Fig. 9　Examples of the synthetic data 2. Left: observations at $t = 3$. Right: observations at $t = 8$.

データ生成時に利用した正しいクラスタリング結果と，各モデルでの推定結果から得られたクラスタリング結果の間の Rand index を測る．より詳細には，各時刻 t においてクラスタリング推定結果 Z_t と正しいクラスタリング結果 $\hat{Z}_t$ の間で Rand index を計算し，T 時刻にわたる Rand index の算術平均をモデルの評価値とした．

各モデルによる Rand index の計算結果を**表 1** に示す．表より明らかなように，クラスタリングの時間変化を考慮しない IRM モデルや，tIRM のように時間ステップ間の依存関係をしないモデルに比べ，提案した dIRM はより良く時間変化する関係データをモデル化できることが確認された．

表 1　Rand index の計算結果
Table 1　Computed Rand indices for dIRM and other models.

Data	IRM	tIRM	dIRM (proposed)
synth1	0.7957	0.9462	**0.9819**
synth2	0.4331	0.7344	**0.8471**

4.2　実データによる評価

実データとして，Enron e-mail dataset [16] を用いて実験を行った．このデータは，破綻直前の Enron 社内のメールの一部を収集したもので，多くの研究で利用されている（e.g. 7), 17)）．文献 17) では，社員同士の関係をネットワークとしてとらえたときの特徴を分析し，重要なノード（社員）を発見する手法を提案している．文献 7) では，オブジェクト（社員）のクラスタリングではなく，時間変化する関係（リンク）のクラスタリングモデルを提案している．

本実験では，$N = 151$ 人の Enron 社員の 2001 年 1 月から 12 月のメールデータを用いた．データ全体を月ごと（$t = 1, 2, \ldots, 12$，$T = 12$）に分離し，ある月 t に社員 i から社員 j に対して 1 通でもメールが発信されていれば $x_{t,i,j} = 1$，メールがなければ $x_{t,i,j} = 0$ として観測データを生成した．

Enron 事件については，いくつかの重要な時間軸が存在する．まず，2001 年 7 月（$t = 7$）において Enron は 500 億ドルの売上げを報告した．しかし利益率は低く，海外事業の失敗やカリフォルニア電力危機への批判から経営を不安視された．8 月（$t = 8$）には株価急落を理由に CEO が辞任し，Founder の K. Lay が Enron の健全性についてアナウンスを出した．また，当初は辞任理由を隠していたため，メディアから経営姿勢などを問われることとなった．10 月に不正会計疑惑が報じられたことで，Enron はこれまでの財務状況を修正報告した．これを引き金に証券取引委員会の調査開始，株価の急落が続き，12 月（$t = 12$）に Enron は倒産した．

この時期の社内メールから生成した観測データに対し dIRM を適用・学習した結果，いくつかのオブジェクト（社員）クラスタとそれら時間変化に関する知見を得た．まず，**図 10** は学習されたクラスタ間の関係の強さ $\eta_{k,l}$ である．縦軸がメールの送信元，横軸がメールの受信元を表す．次に，全 12 ステップで各クラスタに帰属した延べオブジェクト数を**図 11** に示す．

図 10，図 11 より，サイズの大きいクラスタ 1 ～ 7 は，互いに他のクラスタとの関係を持たないことが分かる．つまり，これらのクラスタは主にそのクラスタ内のメンバだけでメー

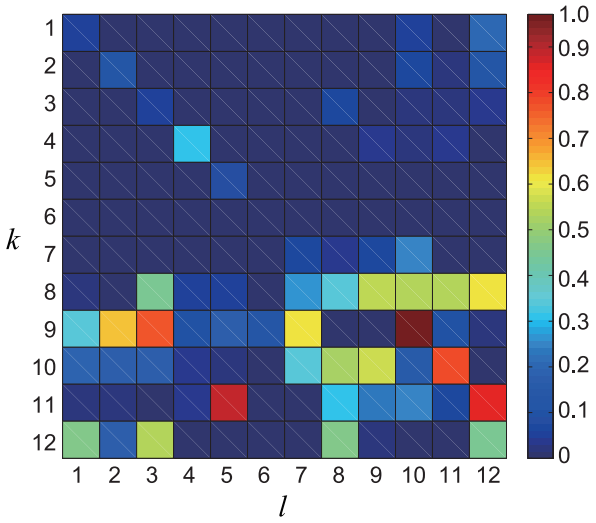

図 10　学習されたクラスタ間の関係強さパラメータ $\eta_{k,l}$
Fig. 10　Estimated $\eta_{k,l}$ (strength of relationship between clusters k, l).

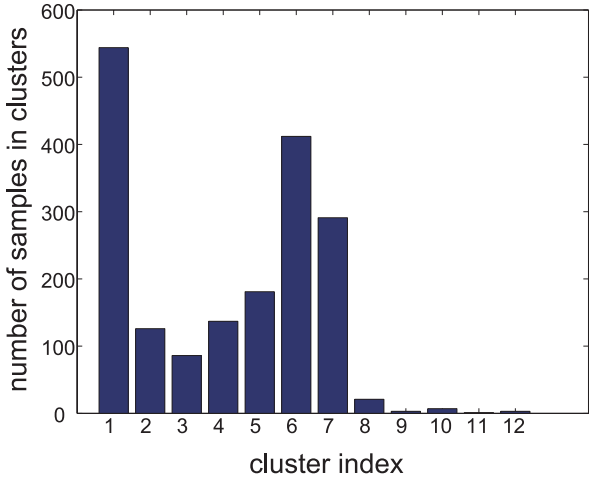

図 11　各クラスタに帰属したオブジェクトの総数
Fig. 11　Total number of items belong to the clusters.

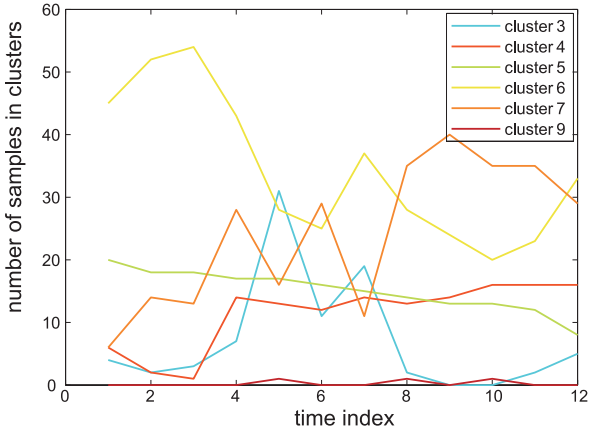

図 12　各時刻において各クラスタに帰属したオブジェクトの数（抜粋）
Fig. 12　Number of items belong to the clusters at each time steps.

ルの関係を構築しているコミュニティであると考えられる．特にその傾向が強いのがクラスタ 4 である．このクラスを解析すると，主に規制業種（エネルギー関連）やガス，パイプライン部門の Vice President などが帰属している．

クラスタ 5 は Trader や経理関係など，金融・財務部門のメンバが多く集まっている．クラスタ 6 は “non active” なクラスタ，つまりほとんどメールのやりとりをしていないメンバが集まっている．実際，図 10 のクラスタ 6 に関係する行および列の η はいずれも低い値をとっている．クラスタ 7 は主に CEO，COO，CFO，Vice president などに代表される上位役員たちが集中している．クラスタ 9 には全時刻を通してのべ 3 人（オブジェクト）だけが帰属している．ただし，これらのオブジェクトは，5 月，8 月，10 月（$t = 5$，$t = 8$，$t = 10$）という，Enron 事件において重要な月に特異的に多くのメンバにメッセージを発信している．5 月にクラスタ 9 に帰属したのは Enron America の CEO，8 月は Enron の Founder，そして 10 月は COO である．このクラスタは上位役員のクラスタであるクラスタ 7 にも多くのメッセージを出していることも納得できる．

クラスタ 9 の解析でも見たように，dIRM の 1 つの利点は，時刻ごとのクラスタ関係を追跡することが可能な点にある．**図 12** はその特徴が現れた図である．図 12 は各クラスタへの帰属オブジェクト数を時刻ごとにプロットした図である．この図から，各クラスタのサイズは刻々と変化することが分かる．

この図からは他にも興味深い点がいくつか散見される．まず，“non active” なクラスタ 6 は時間の経過とともに小さくなっている．これは，Enron がニュースを騒がせるとともに社内の連絡も活性化したのであろう．実際の観測データもメール数の増加を裏付けている．次にクラスタ 5 であるが，このグループの構成員数は他のクラスタと比較して安定していることから，経理・金融関連の社員の結びつきが強固であることがうかがえる．クラスタ 4 の帰属オブジェクト数も 4 月（$t = 4$）以降は安定していることが分かる．

また，クラスタ 3 は 5 月から 7 月にのみ多くのメンバを集めている．このクラスタのメンバには幾人かの管理職や Vice President，Trader たちが含まれていること，さらにクラスタ 9 との関連が非常に強いことから，8 月の CEO 辞任の引き金となった株価下落などに関連している可能性がある．

クラスタ 7 は特に 7 月以降に大きくメンバが増加している．クラスタ 7 は上位役員グループであることから，おそらく不正会計などによる全社的な影響に巻き込まれる社員数が増加したものと思われる．

以上の考察の信憑性は保証できないが，dIRM のクラスタリング結果から，上記のような直感的に妥当な関係性の時間的推移を分析することが可能となる．

5. ま と め

本論文では，関係データの解析モデルの 1 つである IRM を時間発展する関係データに適用できるよう拡張した動的 IRM（dynamic IRM，dIRM）モデルを提案した．そのアイデアは，各オブジェクトのクラスタ帰属確率を HMM によってモデル化することである．これによって，クラスタのサイズや各オブジェクトの帰属クラスタがダイナミックに変化する時系列関係データを自然にモデル化，解析できるようになる．本論文では dIRM モデルの確率的生成モデルを説明し，その学習方法の 1 つとして beam サンプリングを利用することを提案した．人工データを用いた実験で，既存の IRM モデルおよびそのナイーブな拡張モデルに比較して，時間依存関係データにおけるオブジェクトクラスタリングの精度が向上することを確認した．また実データによる実験で，関係性の時間的推移を分析することが可能であることを示した．

今後は他の時系列関係データに適用するとともに，観測モデルを Beta-Bernoulli 分布以外の組合せにするなど dIRM モデルの適用範囲を拡張してゆきたいと考えている．

謝辞　議論に参加していただいた，山田武士，持橋大地両氏に感謝する．

参 考 文 献

1) Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford InfoLab. (1999).
2) Saito, K., Yamada, T. and Kazama, K.: Extracting Communities from Complex Networks by the k-dense method, *Proc. ICDM2006 Workshop on Mining Complex Data*, pp.300–304 (2006).
3) Liben-Nowell, D. and Kleinberg, J.: The Link Prediction Problem for Social Networks, *Proc. International Conference on Information and Knowledge Management*, pp.556–559 (2003).
4) Clauset, A., Moore, C. and Newman, M.E.J.: Hierarchical Structure and the Prediction of Missing Links in Networks, *Nature*, Vol.453, pp.98–101 (2008).
5) Nowicki, K. and Snijders, T.A.B.: Estimation and Prediction for Stochastic Blockstructures, *Journal of the American Statistical Association*, Vol.96, No.455, pp.1077–1087 (2001).
6) Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T. and Ueda, N.: Learning Systems of Concepts with an Infinite Relational Model, *Proc. National Conference on Artificial Intelligence* (*AAAI*) (2006).
7) Fu, W., Song, L. and Xing, E.P.: Dynamic Mixed Membership Blockmodel for Evolving Networks, *Proc. International Conference on Machine Learning* (*ICML*) (2009).
8) Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R.: A Bayesian Approach toward Finding Communities and their Evolutions in Dynamic Social Networks, *Proc. SIAM International Conference on Data Mining* (*SDM*) (2009).
9) Ferguson, T.S.: A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, Vol.1, No.2, pp.353–355 (1973).
10) 上田修功，山田武士：ノンパラメトリックベイズモデル，応用数理，Vol.17, No.3, pp.196–214 (2007).
11) Sethuraman, J.: A Constructive Definition of Dirichlet Process, *Statistica Sinica*, Vol.4, pp.639–650 (1994).
12) Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet Process, *Journal of The American Statistical Association*, Vol.101, No.476, pp.1566–1581 (2006).
13) Fox, E., Sudderth, E., Jordan, M. and Willsky, A.: An HDP-HMM for Systems with State Persistence, *Proc. International Conference on Machine Learning* (*ICML*) (2008).
14) VanGael, J., Saatci, Y., Teh, Y.W. and Ghahramani, Z.: Beam Sampling for the Infinite Hidden Markov Model, *Proc. International Conference on Machine Learn-*

ing (*ICML*) (2008).
15) Hubert, L. and Arabie, P.: Comparing partitions, *Journal of Classification*, Vol.2, No.1, pp.193–218 (1985).
16) CALO Project (A Cognitive Assistant that Learns and Organizes): Enron Email Dataset (2005).
17) Shetty, J. and Adibi, J.: Discovering Important Nodes through Graph Entropy: The Case of Enron Email Database, *Proc. International Workshop on Link Discovery*, pp.74–81 (2005).

(平成 21 年 8 月 20 日受付)
(平成 21 年 9 月 29 日再受付)
(平成 21 年 10 月 15 日採録)

石黒　勝彦

平成 16 年東京大学工学部機械情報工学科卒業．平成 18 年東京大学大学院情報理工系研究科知能機械情報学専攻修士課程修了．同年 NTT 入社．現在，NTT コミュニケーション科学基礎研究所研究員．筑波大学大学院システム情報工学研究科博士後期課程在学中．機械学習，時系列データ解析，パターン認識の研究に従事．電子情報通信学会，IEEE 各会員．

岩田　具治（正会員）

平成 13 年慶應義塾大学環境情報学部卒業．平成 15 年東京大学大学院総合文化研究科広域科学専攻修士課程修了．同年 NTT 入社．平成 20 年京都大学大学院情報学研究科システム科学専攻博士課程修了．博士（情報学）．現在，NTT コミュニケーション科学基礎研究所研究員．機械学習，データマイニング，情報可視化の研究に従事．平成 16 年 FIT 船井ベストペーパー賞，平成 19 年 FIT ヤングリサーチャー賞等受賞．電子情報通信学会会員．

上田　修功（正会員）

昭和 57 年大阪大学工学部通信工学科卒業．昭和 59 年大阪大学大学院修士課程修了．工学博士．同年 NTT 入社．平成 5 年より 1 年間 Purdue 大学客員研究員．画像処理，パターン認識・学習，ニューラルネットワーク，統計的学習，Web 統計解析の研究に従事．現在，NTT コミュニケーション科学基礎研究所副所長，奈良先端科学技術大学院大学客員教授．電気通信普及財団賞受賞，電子情報通信学会論文賞，本会山下記念研究賞等受賞．電子情報通信学会，IEEE 各会員．