ノンパラメトリックベイズモデル

上田 修功 * 山田 武士 *
* 日本電信電話（株）ＮＴＴコミュニケーション科学基礎研究所

Introduction to Nonparametric Bayesian Models

Naonori Ueda* Takeshi Yamada*
*NTT Communication Science Laboratories

Abstract. This paper introduces nonparametric Bayesian models, in particular, Dirichlet process mixture (DPM) models and the infinite relational model (IRM) as an extension of DPM for multi-way data clustering. The nonparametric Bayesian modelling is a more flexible approach than a standard parametric Bayesian modelling in that a nonparametric prior distribution over model parameters is incorporated into the data gereration process. More specifically, DPM enables us to define distributions over the countably infinite sets by exploring their clustering structures. In this paper, we explain the basic idea of DPM modelling and its learning algorithms. We also illustrate practical usefulness of DPM modelling though experimental results using IRM.

1. はじめに

統計モデルとは，観測データの背後にあるデータ生成過程を確率モデルとして表現したものであり，生成モデル (generative model) とも呼ばれる．特に，観測されない変数を導入したより自由度の高い統計モデルを潜在変数モデル (latent variable model) と呼ぶ．代表例として，混合モデル (mixture model) では，各データは複数ある要素分布のうちの一つから生成されるが，どの要素分布から生成されたかは観測されない．そこで各データに対する，要素分布のインデックスを潜在変数とし，学習によってこれを推定する．要素モデルに正規分布を仮定する混合正規分布モデルは，単一の正規分布では表現できない複雑な確率分布の近似モデルとして位置づけられ，音声認識での隠れマルコフモデルにおけるデータの出力分布等，工学の幅広い応用分野で用いられている．

潜在変数モデルの実用上の問題点として，モデル選択問題がある．例えば，混合モデルの場合，基底となる要素数を表す混合数は本来，観測対象および観測データに応じて適切に選択されなければならない．一方，全ての未知量を確率変数と見なし，それらを観測データを得た下での事後分布としてベイズの定理を用いて推定するベイズアプローチがある．したがって，上記の混合数もモデルの事後分布からベイズ流に推定可能と言える．

しかし，ベイズアプローチでは，未知量に関する事前知識を事前分布として設定する必要がある．換言すれば，事前分布をどうモデル化するかという新たな問題が生じる．この

問題への有力な解決法が，本論文で紹介するノンパラメトリックベイズモデリングである．即ち，ノンパラメトリックベイズでは，統一的な枠組みで，自由度の高い事前分布が設定でき，同時に，上記モデル選択問題に対しても，無限混合モデルという形で対処可能である．

以下では，まず，ノンパラメトリックベイズモデリングの概要を説明する．次いで，その応用として，同じ，あるいは，異なるデータタイプ間での関係が与えられた時に，各タイプを同時にクラスタリングする Infinite Relational Models (IRM) [10] について説明し，実データでの適用例を通じてその有用性を示す．タイプ間の関係とは，例えば，論文引用データの場合，論文同士は「引用」関係，著者と論文は「執筆」関係にある．これらの関係データが多数与えられた時，IRM は，著者，論文を各々同時にクラスタリングし，類似著者，類似論文を各々見出すことができる．

2. ノンパラメトリックベイズモデル

2.1 パラメトリックベイズモデルとノンパラメトリックベイズモデルとの違い

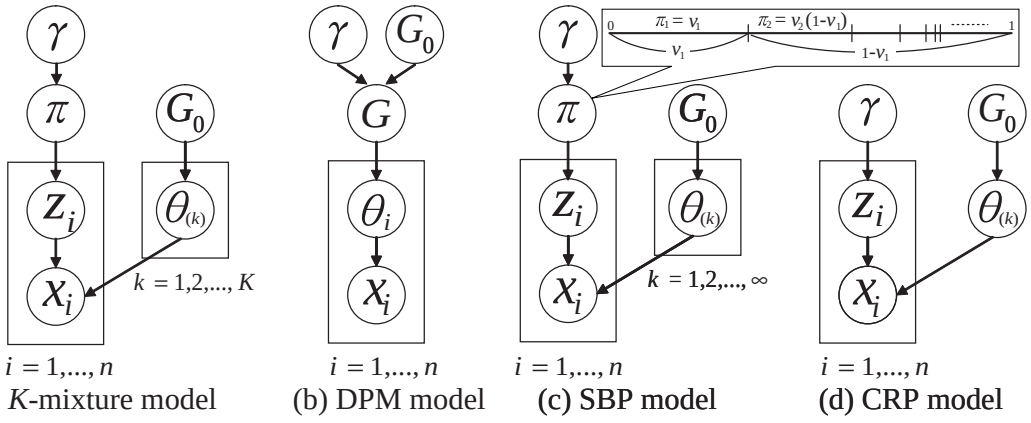

Fig. 1. パラメトリック／ノンパラメトリック混合モデルのグラフィカルモデル

混合モデルを用いて通常のパラメトリックベイズモデルについて復習する．混合モデルでは，n 個の観測データ $x_1, \ldots, x_n$ の生成過程は以下のようにモデル化される．

(1) $\theta_{(k)} \sim G_0$, for $k = 1, \ldots, K$
(2) $\pi = (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\pi;\ \gamma_1, \ldots, \gamma_K)$
(3) $z_i \,|\, \pi \sim \text{Multinomial}(z;\ \pi)$, and $x_i \,|\, \theta_{(z_i)} \sim p(x \,|\, \theta_{(z_i)})$, for $i = 1, \ldots, n$

ここで，表記 $y \sim p$ は y が分布 p から生成されることを意味する．また $y \,|\, x \sim p$ は，x が既知の下で y が分布 p から生成されることを意味する．

即ち，まず (1) で K 個の要素分布のパラメータ $\theta_{(k)}$, $k = 1, \ldots, K$ が共通の事前分布 G_0 から生成される[*1]．次に (2) で，要素分布の選択確率 $\pi = (\pi_1, \ldots, \pi_K)$ がディリクレ (Dirichlet) 分布により生成される．π_k は第 k 要素分布が選択される確率を表す．ディリクレ分布は，非負で総和が 1 ($\sum_k \pi_k = 1$) なる K 個の確率変数 (K 次元単体上の確率ベクトル) の同時分布で，$\mathrm{Dirichlet}(\pi;\ \gamma_1, \ldots, \gamma_K) = Z(\gamma)^{-1} \prod_{k=1}^{K} \pi_k^{\gamma_k - 1}$ で定義される．ただし，$Z(\gamma)$ は $\gamma_1, \ldots, \gamma_K$ にのみ依存する正規化項を表す[*2]．(3) では，x_i がどの要素分布から生成されるかを決める潜在変数 z_i が，π をパラメータとする多項分布 $\mathrm{Multinomial}(z;\ \pi)$ で確率的に決定される (π_k は $z_i = k$ となる確率に相当し，π_k の期待値は $\gamma_k / (\gamma_1 + \ldots + \gamma_K)$ である)．そして，z_i が決まれば対応する $\theta_{(z_i)}$ が決まり，x_i は要素分布 $p(x \mid \theta_{(z_i)})$ から生成される．上記は，一般の有限混合モデルのデータ生成過程で，応用に応じて，要素分布モデルを具体的に定め，また，それに応じ，パラメータの事前分布 G_0 も具体化される．なお (3) は θ が分布 $G(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_{(k)}}(\theta)$ に従って生成されると言い換えることができる．ただし，$\delta_u(v)$ はデルタ関数で，$u = v$ の時は 1，それ以外は 0 となる．

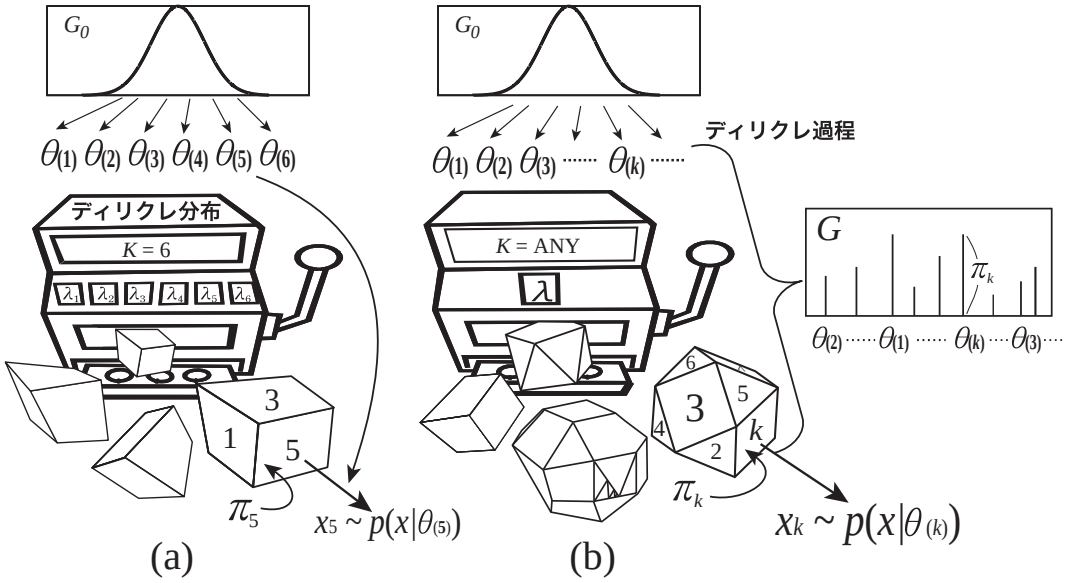

Fig. 2.　ディリクレ分布／ディリクレ過程サイコロ生成器

上記に対する読者の理解を深めるために，以下に，K 面サイコロを例にとって，ディリクレ–多項 (K 項) 分布モデルを直観的に説明する．Fig. 2(a) に $K = 6$ の場合のイメージ図を示す．

(1) の G_0 は各サイコロの目 k に対応するパラメータ $\theta_{(k)}$ を生成する分布である．即ち (1) で，まずあらかじめ $\theta_{(k)}$ ($1 \le k \le K$) が G_0 より生成される．(2) のディリクレ分布は，

[*1] 例えば，要素分布が正規分布の場合，$\theta_{(k)}$ はその平均と分散であり，5.1 節で説明するように，要素分布が 0 または 1 の二値をとるベルヌーイ分布の場合 $\theta_{(k)}$ は 1 の出現確率となる．

[*2] 具体的には，ガンマ関数 $\Gamma(x)$ を用いて $Z(\gamma) = (\prod_k \Gamma(\gamma_k)) / \Gamma(\sum_k \gamma_k)$．また，一般には，$K$ 個の異なるパラメータ $\gamma_1, \ldots, \gamma_K$ とするが，通常，それらが全て等しい (γ) とする対称ディリクレ分布が用いられる．

「K 面サイコロ生成器」に相当し，k の目が出る確率，つまりこの目の出やすさが π_k である不平等で偏った K 面サイコロを生成する．この不平等さはパラメータ $\gamma_1, \ldots, \gamma_K$ で制御される．このサイコロと各目に対応するパラメータ $\{\theta_{(k)}\}$ をあわせたものが上述の $G(\theta)$ に対応する．(3) では，(2) で生成されたサイコロを複数回振って目を出す．第 i 回目に出た目が $\theta_{(k)}$ の時，$\theta_{(k)}$ をパラメータとして持つ要素分布 $p(x|\theta_{(k)})$ から x_i が生成される．

上記 (1),(2),(3) による観測データ x_i の生成過程のグラフィカルモデルを Fig. 1(a) に示す．グラフィカルモデルとは，変数間の依存関係を有向グラフで表現したものである．例えば，図 1(a) で x_i は z_i からの矢印があるが，π からの矢印はない．もちろん，z_i は π に依存するため，x_i も間接的に π に依存するが，z_i が既知 (例えば，$z_i = k$) であれば，上記生成過程より明らかなように，x_i はもはや π に依存しない．即ち，x_i は z_i 既知の下で π と条件付き独立である．条件付き独立性より，x_i の確率分布は，矢印の起点の変数を条件として，$p(x_i \mid z_i = k, \theta_{(k)})$ と書ける．明らかに，これは，(3) の $p(x_i \mid \theta_{(z_i)})$ と等価表現である．

前節の (1) から明らかなように，パラメトリックベイズモデルでは，K 個のモデルパラメータ $\{\theta_{(k)}\}_{k=1}^{K}$ はデータを生成する前に K 種類全てが決められている．ただし，K の値は，モデル構造の尤度 $p(D|K)$ の最大化により選択される[*3]．ここに D は観測データ集合を表す．実際には，候補となる幾つかの K の各々に対して学習を行った後，それらの中で $p(D|K)$ の一番大きな K を選択することになる．各データ x_i をどの要素分布から生成するかは，多項分布に従う確率 $\pi = (\pi_1, \ldots, \pi_K)$ で決まる．

これに対し，ノンパラメトリックベイズモデル，即ち，Fig. 1(b) に示すディリクレ過程 (Dirichlet Process: DP) では，x_i 毎に θ_i が対応づけられている．つまり，データ数分の n 個のパラメータ，即ち最大 n 混合モデルまでが実現できる．もちろん，一つのデータに一つの要素分布というのは非現実的であり，実際，DP では，データが観測される毎に，要素分布数が必要に応じて増える柔軟なデータ生成過程となっている．換言すれば，$\theta_1, \ldots, \theta_n$ は全て異なるわけではなく，適応的に値の異なる K 個のパラメータ $\theta_{(1)}, \ldots, \theta_{(K)}$ から成る．即ち，$\theta_i \in \{\theta_{(1)}, \ldots, \theta_{(K)}\}$ (ここで表記に関し，$i \neq j$ に対し $\theta_i = \theta_j$ となり得るが，$\theta_{(i)} \neq \theta_{(j)}$ に注意)．つまり，$x_1, \ldots, x_n$ が有限個 (K 個) のクラスタに分割される．DP の場合，K の値が予め固定されているわけではなく観測データに応じて定まる．この性質を用いて混合モデルを構成するモデリングがディリクレ混合過程 (Dirichlet Process Mixture: DPM) モデルである．

2.2 ディリクレ混合過程

DP は確率分布に対する分布 (distribution over distributions) [5] で，基底分布 G_0 と正のパラメータ γ で定義される．G が DP に従う時，$G \sim \mathrm{DP}(\gamma, G_0)$ と表記する．そして，確率変数 θ が G から生成される時，$\theta \sim G$ (即ち $P(\theta|G) = G(\theta)$) と書く．Fig. 1(b) に，ディ

[*3] ベイズ推定の枠組みで，尤度関数の最大化としてパラメータ (本例では K の値) を点推定する方法を経験ベイズ法と呼ぶ．

リクレ混合過程 (Dirichlet Process Mixture: DPM) に従うデータ生成過程のグラフィカルモデルを図示する．DPM モデルによるデータ生成過程は以下で与えられる．

(1) $G \sim \mathrm{DP}(\gamma, G_0)$
(2) $\theta_i \,|\, G \sim G$, for $i = 1, \ldots, n$
(3) $x_i \sim p(x \,|\, \theta_i)$, for $i = 1, \ldots, n$

DP の厳密な定義等は紙面の都合上，参考文献に譲り，以下では，上記 DPM によるデータ生成過程について，再び「サイコロ」の例を用いてより直観的に説明する．

2.1 節で説明したように，ディリクレ分布は多項分布のパラメータの事前分布として用いられ，「K 面サイコロ生成器」を実現するものであった．一方，DP は直観的には K の値を固定しない，どんな面数のサイコロでも生成できる，ただし，面数の多いサイコロほど生成されにくい「任意面サイコロ生成器」を実現する．

Fig. 2(b) に示すように，まず，基底分布 G_0 からサイコロの目 k に対応するパラメータ $\theta_{(k)}$ として，十分多くの $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(k)}, \ldots$ を生成する．基底分布 G_0 は離散分布でも連続分布でもよい．次に，「任意面サイコロ生成器」によって不平等で偏ったサイコロが生成される．ここで，k の目の出る確率，つまりこの目の出やすさを π_k とする．このサイコロと，対応するパラメータ $\{\theta_{(k)}\}$ で定義される離散分布が G である．すなわち，以上が (1) による，確率分布に対する分布である DP からの，確率分布 G の生成に相当する．次節で説明するように，DP においては，このようなサイコロをいくつも生成すると，これらは平均して，k が大きいほど対応する面の出やすさ π_k が指数的に小さくなるという傾向を持ち，この傾向はパラメータ γ で制御される．したがって (1) の任意面サイコロ生成においては，各サイコロのいわば実効的な面数はサイコロごとに異なり，面数の多いサイコロほど生成されにくいことになる．次に (2) により，このサイコロを振って出た目を θ_i とし，(3) により，前節同様，x_i がパラメータ θ_i を持つ要素分布 $p(x \,|\, \theta_i)$ から生成される．

2.3 Stick-breaking 過程

前節 (1) の DP の直観的な説明として用いた「サイコロ生成器」は，どんな多くの面数を持つサイコロでも生成可能であった．理論的には，DP は，G_0 から生成される加算無限個の $\{\theta_{(k)}\}_{k=1}^{\infty}$ と，第 k 面の出やすさが π_k である加算無限面のサイコロを生成しているといえる．特に後者に着目すると，DP とは無限次元のディリクレ分布と見なすことができ，実際 G は，$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_{(k)}}(\theta)$ と書けることが示されている [18]．ただし，$\sum_{k=1}^{\infty} \pi_k = 1$ であり，したがって，有限個を除いたほとんどの k に対する π_k は限りなくゼロに近い．即ち G を生成することは，G_0 から生成される $\{\theta_{(k)}\}_{k=1}^{\infty}$ と，前節での面数の多いサイコロほど生成されにくいことに対応する，以下に示す一定の条件を満たす $\{\pi_k\}_{k=1}^{\infty}$ を生成することに相当する．以下，具体的な $\{\pi_k\}_{k=1}^{\infty}$ の構成法について説明する．

Sethuraman [18] は，$\pi_1, \pi_2, \ldots,$ の構成法として，$v_k \sim \mathrm{Beta}(v;\ 1, \gamma)$, および $\pi_k =$

$v_k \prod_{j=1}^{k-1}(1-v_j)$ からなる Stick-breaking 過程 (SBP) を示した．ここで，Beta() は区間 [0,1] で定義されるベータ分布で，ディリクレ分布における $K=2$ の特別な場合に相当する．ベータ分布は 2 つのパラメータ $(1,\gamma)$ を持ち (一つは 1 に固定)，γ の値が小さい (大きい) 程，v_k は 1 (0) に近い値が生成され易くなる．$\gamma=1$ の時は，ベータ分布は $[0,1]$ に渡る一様分布となる．

Fig. 1(c) に示すように，SBP では，長さ 1 の棒 (stick) を考える．まず，v_1 をベータ分布より生成し，v_1 対 $1-v_1$ の比で棒を折り，折り取った v_1 の比の方の棒の長さを π_1 とする．明らかに，$\pi_1=v_1$ である．次いで，v_2 を生成し，残っている長さ $1-v_1$ の棒に対し，v_2 対 $1-v_2$ の比で棒を折り，v_2 の比の方の棒の長さ $v_2(1-v_1)$ を π_2 とする．以下この操作を繰り返すと，第 k 操作では，$\pi_k = v_k \prod_{j=1}^{k-1}(1-v_j)$ となる．長さ 1 の棒を逐次折り取った長さを π_k としているので，無限回の操作による π_k の総和は 1 となり，$\sum_{k=1}^{\infty}\pi_k = 1$ が保証される．この過程は DP の Stick-breaking 過程と呼ばれる． 前節におけるサイコロの目の出やすさ π_k が，ここでの折り取った棒の長さに対応することに注意されたい．k が大きいほど棒の残りは短くなり，平均的に π_k は小さくなる．実際，SBP では $\pi_k = v_k \prod_{j=1}^{k-1}(1-v_j)$ で，各 v_k は独立かつ同じベータ分布に従うので，π_k の期待値は $E[\pi_k] = (1/(1+\gamma))(\gamma/(1+\gamma))^{k-1}$ となり，k に対し指数的に小さくなる．すなわち，前節で述べたように，面数の多いサイコロほど生成されにくい．

SBP は DP の一構成法を与えるが，無限個のパラメータを予め生成しておくというのは現実的ではない．より自然で扱いやすい DP 構成法として，次節で述べる Chinese restaurant 過程 (CRP) がある．

2.4 Chinese restaurant 過程

本節では，$\theta_{(k)}$ のインデックス k に着目して議論を進める．$z_i = k$ が決まれば G_0 から生成された $\theta_{(k)}$ が一意に対応する．前節で説明したように，Fig. 1(c) では，意味的には，Fig. 1(a) の K 項のディリクレ–多項分布モデルにおいて $K\to\infty$ とした極限と解釈できる．そこで，2.1 節の (2) のディリクレ分布のパラメータを $\gamma_k = \gamma/K$ とすると，$z_{1:i-1} = (z_1,\ldots,z_{i-1})$ が既知の下での z_i の事後分布は，

$$P(z_i = k \mid z_{1:i-1},\gamma,K) = \int P(z_i=k\mid\pi)p(\pi\mid z_{1:i-1},\gamma,K)d\pi = \frac{m_k+\gamma/K}{i-1+\gamma} \tag{2.1}$$

と計算できる [*4]．ここで，m_k は $z_j = k\,(j=1,\ldots,i-1)$ を満たす j の個数 ($x_1,\ldots,x_{i-1}$ の中で，クラス k に割り当てられた観測データの数) を表す．上記計算の際，$P(z_i=k\mid\pi)=\pi_k$

[*4] 慣例に従い，z_i など離散変数に関しては確率関数 (probability mass function) として大文字 P を用い，π など連続変数に関しては確率密度関数 (probability density function) として小文字 p を用いる．

およびベイズの定理，ディリクレ分布の共役性より[*5]，

$$p(\pi \mid z_{1:i-1}, \gamma, K) \propto P(z_{1:i-1} \mid \pi) p(\pi \mid \gamma, K)$$

$$\propto \left(\prod_{k=1}^{K} \pi_k^{m_k} \right) \left(\prod_{k=1}^{K} \pi_k^{\frac{\gamma}{K}-1} \right) = \prod_{k=1}^{K} \pi_k^{m_k+\frac{\gamma}{K}-1} \equiv \text{Dirichlet}\left(\pi;\ m_1+\frac{\gamma}{K}, \ldots, m_K+\frac{\gamma}{K} \right) \tag{2.2}$$

となることを用いている．ゆえに，Fig. 1(c) の場合は，式 (2.1) の極限として以下が得られる．

$$P(z_i = k \mid z_{1:i-1}, \gamma) = \lim_{K\to\infty} P(z_i = k \mid z_{1:i-1}, \gamma, K) = \frac{m_k}{i-1+\gamma} \tag{2.3}$$

一方，K クラスの中で，$x_{1:i-1}$ が一つも帰属しない空クラスの集合を $U = \{k \mid z_j \neq k,\ j = 1, \ldots, i-1\}$ とし，また，$k = 1, \ldots, K$ の中で，$x_{1:i-1}$ が帰属しているクラスの総数を K_{i-1} とする (明らかに $|U| = K - K_{i-1}$)．この時 $m_k = 0$ ($k \in U$) に注意すると式 (2.1) の極限として次式を得る．

$$P(z_i \in U \mid z_{1:i-1}, \gamma) = \lim_{K\to\infty} \sum_{k\in U} \frac{\gamma/K}{i-1+\gamma} = \frac{\gamma}{i-1+\gamma} \lim_{K\to\infty} \frac{K - K_{i-1}}{K} = \frac{\gamma}{i-1+\gamma} \tag{2.4}$$

式 (2.3), 式 (2.4) をまとめると，以下のように書ける．

$$P(z_i = k \mid z_{1:i-1}) = \begin{cases} \frac{m_k}{i-1+\gamma} & m_k > 0 \\ \frac{\gamma}{i-1+\gamma} & k \text{ は新規クラスタ } (U \text{ に属する}) \end{cases} \tag{2.5}$$

上記は，Chinese Restaurant Process (CRP) [1] と呼ばれ，DPM モデルの一構成法を与える．即ち，CRP による DPM モデルのグラフィカルモデルである Fig. 1(d) から明らかなように，CRP は有限混合モデルにおけるディリクレ–多項分布モデルにおけるパラメータ π を式 (2.1) のように積分消去し，かつ，混合数を無限極限にしたものと解釈できる．以降簡単のため $z \sim \text{CRP}(\gamma)$ などと記す．

式 (2.5) はレストランに無数のテーブルがあるとして, 既に $i-1$ 人の客がいずれかのテーブルに着席している時，i 番目の客は，既に m_k 人座っているテーブルに確率 $m_k/(i-1+\gamma)$ で着席し，誰も座っていないテーブルに確率 $\gamma/(i-1+\gamma)$ で着席することにたとえられる．即ち，$z_i = k$ は，客 i がテーブル k に着席することを意味する．γ の値が小さい (大きい) 程，既着テーブル (新規テーブル) に着席する可能性が高くなる．また，n が十分大きい時，クラスタ数はほぼ $\log n$ のオーダで増加する事が知られている [2].

CRP の重要な性質として，交換可能性 (exchangability) がある．即ち，式 (2.5) に従って n 個の観測値に対する潜在変数の同時分布 $P(z_{1:n})$ は $P(z_{1:n}) = P(z_1)P(z_2 \mid z_1) \cdots P(z_n \mid z_{1:n-1})$

[*5] データの分布に対し，そのパラメータの事前分布が共役事前分布であるとは，パラメータの事後分布と事前分布が同じ分布族となることである．その場合，解析的な取り扱いが容易となる．ディリクレ分布は多項分布の共役事前分布である．

として計算でき，その結果は，

$$P(z_{1:n} \mid \gamma) = \frac{\gamma^K \prod_{k=1}^{K} (m_k - 1)!}{\gamma(\gamma+1)\dots(\gamma+n-1)} \tag{2.6}$$

となり，右辺は m_k のみに依存するため，z_i の任意の i の入れ替えを行って出現順序を変えても結果は変わらない [11]．CRP におけるこの交換可能性は，次節で述べる，DPM における推論アルゴリズム (Gibbs サンプラー) において重要となる．

最後に，式 (2.5) を，θ_i の事後分布の形に書き直すとこれは $\theta_{1:i-1} = (\theta_1, \dots, \theta_{i-1})$ が既知の下で G を積分消去した，

$$P(\theta_i \mid \theta_{1:i-1}) = \int G(\theta_i) P(G \mid \theta_{1:i-i}) dG = \frac{\gamma}{i-1+\gamma} G_0(\theta_i) + \frac{1}{i-1+\gamma} \sum_{k=1}^{K} m_k \delta_{\theta_{(k)}}(\theta_i) \tag{2.7}$$

となる [2, 5, 9]．式 (2.7) は，θ_i は確率 $\gamma/(i-1+\gamma)$ で G_0 から新たに生成され，確率 $m_k/(i-1+\gamma)$ で既存の $\theta_{(k)}$ から選択されることを示している．たとえ最右辺第一項の G_0 が連続分布であっても，第二項より同じ $\theta_{(k)}$ が繰り返し選択され，$i \neq j$ に対し $\theta_i = \theta_j$ となり得るため，$\theta_{1:i}$ が $\{\theta_{(k)}\}$ にクラスタリングされる．

3. 各モデル間の関係

Fig. 1 を用いて，前節までに説明した各モデル間の関係を整理しておこう．まず，(a) の有限混合モデルと (c) の SBP モデルを比較する．(a) では，$\pi_1, \dots, \pi_K$ をパラメータとする K 項の多項分布から z_i を生成するのに対し，(c) は (a) の $K \to \infty$ の極限として，$\{\pi_k\}_{k=1}^{\infty}$ をパラメータとする無限項の多項分布から z_i を生成する．なお，$z_i = k$ の時，2.2 節の「任意面サイコロ」から出た目が k であることに相当する．同時に，各目に対応する無限個のパラメータ $\{\theta_{(k)}\}_{k=1}^{\infty}$ もあらかじめ G_0 より生成しておく．一方，(c) において $G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_{(k)}}(\theta)$ とおくと (b) と等価になる．したがって，(b) の具体化が (c) であるといえる．また (d) の CRP モデルでは，無限次元の G の生成を避けるため，$z_{1:i-1}$ が既知の下で，G を事後分布で積分消去し，式 (2.5) により直接 z_i を生成する．$\theta_{(k)}$ に関しても式 (2.7) に示すように，必要になった時点で G_0 から新たに生成する．

このように，モデルとしては (b), (c), (d) は等価である．次節以降では，最も取り扱いが容易な (d) の CRP モデルを用いる．

4. 学習アルゴリズム

DPM モデルにおける観測データの生成過程について説明したが，DPM モデルの学習とは，その逆問題，即ち，観測データ $D = x_{1:n} = \{x_1, \dots, x_n\}$ が与えられた下での種々の量を推定する問題である．例えば，x_{n+1} の値を予測したい場合は，事後予測分布 $p(x_{n+1} \mid x_{1:n})$

を推定することになる．また，4 節で紹介する無限関係モデルのように，観測データ集合 D のみを対象に D を分割したい場合は，$P(z_{1:n}\,|\,x_{1:n})$ を最大化する $z_{1:n}$ の組を求める問題となる．以下では，4 節との関係上，後者の問題に焦点を当てる．

4.1 ギブスサンプリング

$P(z_{1:n}\,|\,x_{1:n})$ を最大化する $z_{1:n}$ を求める際，単純に，$z_{1:n}$ の全ての可能な値で評価するのは計算量的に非現実的で，何らかの近似計算が必要となる．$P(z_{1:n}\,|\,x_{1:n})$ から $z_{1:n}$ をサンプリングできれば，確率的に，$P(z_{1:n}\,|\,x_{1:n})$ の値の大きな $z_{1:n}$ が求まることになる．このサンプリングを効率良く実現する手法がギブスサンプリングである．以下では，DPM に特化したギブスサンプリングアルゴリズム [9, 11, 14] について説明する．

ギブスサンプリングでは，$P(z_{1:n}\,|\,x_{1:n})$ から $z_{1:n}$ を同時にサンプリングするのではなく，以下に説明するように，逐次サンプリングにより効率的なサンプリングを行う．即ち，z_i をサンプリングする際，z_i 以外の z（便宜上，z_{-i} と書く）の値は既知とした，$P(z_i\,|\,z_{-i}, x_i, \Theta)$ により，z_i をサンプリングする．ここで Θ は現時点でのパラメータ，即ち，$\Theta = (\theta_{(k)}:\ k = z_i \text{ for some } i \in \{1, \ldots, n\})$ を表す．上記，$P(z_{1:n}\,|\,x_{1:n})$ の最大化問題の場合，サンプリングを十分繰り返し，$P(z_{1:n}\,|\,x_{1:n})$ が最大となる $z_{1:n}$ を求めれば良い．

4.2 事後確率の導出

ベイズの定理より以下が成り立つ[*6]．

$$P(z_i = k\,|\,z_{-i}, x_i, \Theta) = \frac{P(z_i = k, x_i\,|\,z_{-i}, \Theta)}{p(x_i\,|\,z_{-i}, \Theta)} \propto P(z_i = k\,|\,z_{-i}) p(x_i\,|\,z_i = k, \Theta) \tag{4.1}$$

一方，先に説明した CRP での交換可能性より，z_i をサンプリングする際，最新の z_i の値を一旦解除し，$z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$ をこれまでの $n-1$ 人の客が着席しているテーブルの割り当て，そして，z_i を新たな客 i に対するテーブルの割り当てと見なすことができる[*7]．これにより，$P(z_i = k\,|\,z_{-i})$ については式 (2.5) が適用でき，式 (4.1) の最右辺の $P(z_i = k\,|\,z_{-i})$ は次式となる．

$$P(z_i = k\,|\,z_{-i}) = \begin{cases} \frac{m_{-i,k}}{n-1+\gamma} & \text{if } k = z_l \text{ for some } l \neq i \\ \frac{\gamma}{n-1+\gamma} & \text{if } k \neq z_l \text{ for all } l \neq i \end{cases} \tag{4.2}$$

ここで，$m_{-i,k}$ は $z_l = k$ $(l \neq i)$ を満たす l の総数，即ち，z_i 以外の全ての $\{z_l\}$ のうち現時点でクラス k に帰属しているものの総数を表す．

更に，$p(x_i\,|\,z_i = k, \Theta)$ は，もし，$k = z_l$ となる l が存在すれば，x_i は既存のパラメータ $\theta_{(k)}$ から生成されたことになるので，$p(x_i\,|\,z_i = k, \Theta) = p(x_i\,|\,\theta_{(k)})$ と書ける．そうでない場

[*6] $p(x_i\,|\,z_{-i}, \Theta)$ は z_i に無関係な項なので，分子のみを分解して最右辺が得られることに注意．
[*7] z_i が i 単独のテーブルの場合は，テーブルそのものも解除する．

合，x_i に対する全ての可能なパラメータで積分消去した $p(x_i \mid z_i = k) = \int p(x_i \mid \theta) G_0(\theta) d\theta$ と書ける．G_0 が共役事前分布の場合，$p(x_i \mid \theta) G_0(\theta)$ が G_0 と同族の分布となるためこの積分は解析的に計算可能である．以上より，式 (4.1) は次式となる．

$$P(z_i = k \mid z_{-i}, x_i, \Theta) \propto \begin{cases} \{m_{-i,k}/(n-1+\gamma)\} p(x_i \mid \theta_{(k)}) & \text{if } k = z_l \text{ for some } l \neq i \\ \{\gamma/(n-1+\gamma)\} \int p(x_i \mid \theta) G_0(\theta) d\theta & \text{if } k \neq z_l \text{ for all } l \neq i \end{cases} \tag{4.3}$$

式 (4.3) の右辺の第 2 行目はもはや θ に依存しないので，第 1 行目についても $\theta_{(k)}$ の依存性をなくすくべく $\theta_{(k)}$ の事後分布 $P(\theta_{(k)} \mid x_{-i})$ で $\theta_{(k)}$ を積分消去すると次式を得る．

$$P(z_i = k \mid z_{-i}, x_i, x_{-i}) \propto \begin{cases} \{m_{-i,k}/(n-1+\gamma)\} \int p(x_i \mid \theta_{(k)}) P(\theta_{(k)} \mid x_{-i}) d\theta_{(k)} & \text{if } k = z_l \text{ for some } l \neq i \\ \{\gamma/(n-1+\gamma)\} \int p(x_i \mid \theta_i) G_0(\theta_i) d\theta_i & \text{if } k \neq z_l \text{ for all } l \neq i \end{cases} \tag{4.4}$$

ここで，$P(\theta_{(k)} \mid x_{-i})$ は，x_i を除く全ての観測データ x_l, $l \neq i$ が既知での事後分布を表し，ベイズの定理より次式で計算される．

$$P(\theta_{(k)} \mid x_{-i}) = \frac{\prod_{s:z_s=k, s\neq i} p(x_s \mid \theta_{(k)}) G_0(\theta_{(k)})}{\int \prod_{s:z_s=k, s\neq i} p(x_s \mid \theta) G_0(\theta) d\theta} \tag{4.5}$$

式 (4.4) では $P(\theta_{(k)} \mid x_{-i})$ を用いて積分消去を計算しているため，z_i のサンプリングは x_{-i} にも依存する．そこで式 (4.4) の左辺は x_{-i} 依存性を銘記している点に注意．また，式 (4.3), (4.4), (4.5) に現れる全ての積分計算は，G_0 の共役性より全て解析的に計算できる事に注意．結局，$p(z_{1:n} \mid x_{1:n})$ を近似的に最大化する $z_{1:n}$ は式 (4.4) を $i = 1, \ldots, n$ に対し繰り返し実行することにより求まる．式 (4.4) より明らかなように $z_{1:n}$ のサンプリング時に確率的に総クラス数が定まる．式 (4.4), (4.5) に基づく具体的なサンプリングアルゴリズムについては，次節以降の無限関係モデルの例で詳しく述べる．

5. 無限関係モデル

本節では，DPM の応用例として，Kemp 等によって提案された無限関係モデル (Infinite Relational Model: IRM) [10] に基づき，「関係データ」という構造化されたデータへの DPM の適用を考える．まず Fig. 3(a) の左側の行列で表現される 9 個の要素からなる $T = \{1, 2, \ldots, 9\}$ 上の二値の二項関係 $R : T \times T \to \{0, 1\}$ を考える（以下，二値であることが文脈から明らかな場合は省略して $R : T \times T$ もしくは単に $T \times T$ と記す）．$R(i, j) = 1\,(0)$ がそれぞれ黒（白）に対応する．これは例えば，T を 9 人の人間集合，$R(i, j)$ を i さんは j さんのことを好き（嫌い）という関係，のように解釈できる．IRM は T の要素を関係の観点からクラスタリングし，クラスタ毎に並べ替えることによって左側の行列から右側のブロック化された行列を得ることに対応する．

より複雑な例として，論文の引用関係を考える．ある論文がある論文を引用する（しない）という関係は，論文全体の集合を T^1 として，$R_1 : T^1 \times T^1$ で表現できる．さらに単語全体の集合を T^2，著者全体の集合を T^3 とすると，ある論文がある単語を含む（含まない）

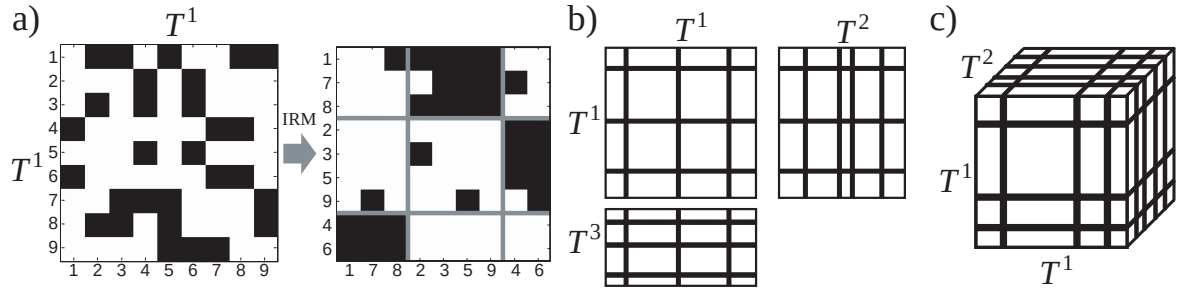

Fig. 3. a) $T^1 \times T^1$, b) $T^1 \times T^1, T^1 \times T^2, T^1 \times T^3$, c) $T^1 \times T^1 \times T^2$ 上の同時クラスタリングの例

という関係 R_2 および，ある論文がある人物を著者の一人として含む（含まない）という関係 R_3 についてもそれぞれ $R_2 : T^1 \times T^2$，$R_3 : T^1 \times T^3$ のような二項関係が定義できる．

IRM の目的は，T^1, T^2, T^3 といった一般に複数の要素集合（これらをタイプと呼ぶ）と，これらの上で定義される R_1, R_2, R_3 といった一般に複数の関係が与えられたとき，R_1, R_2, R_3 を利用して複数のタイプ T^1, T^2, T^3 を同時にクラスタリングすることである．このとき，前節の DPM を用いることで，あらかじめクラスタ数を固定しない柔軟なクラスタリングが可能となる．Fig. 3(b) にこれらのタイプを同時にクラスタリングした例を模式的に示す．また，Fig. 3(c) は 2 タイプ T^1, T^2 上の 3 項関係 $R : T^1 \times T^1 \times T^2$ を用いた同時クラスタリングの例である．いずれの例においても，複数の関係に共通するタイプ（この場合は T^1）についてはクラスタリング結果は常に共通であることに注意されたい．

このようなモデルは，社会科学分野において，stochastic block model として知られている．ここでは主に単一タイプ（人物）上の二項関係に限定し，同一のクラスタに属する人物はすべて他者との社会的関係，即ち社会的役割が確率的にすべて等しいと考える（stochastically equivalent と呼ぶ）[19]．以下に説明する IRM の生成モデルもこの考え方は共通である．また，Nowicki らはこの考え方に基づき観測されない隠れたクラスタ構造を関係データから学習する手法を提案しているが，クラスタ数は固定である [15].

数学的モデルとパラメータ学習法については次節で説明することにして，直観的に理解しやすいように，まず具体例として，50 種類の動物と，85 の特徴からなるデータ [16] に IRM を適用することを考える．動物は具体的な動物の名前，特徴は生息場所 (jungle, tree, coastal)，身体的特徴 (bulbous body shape, has teeth for straining food from the water) および，行動的特徴 (swims, slow) などから構成されている．一般にこのような「オブジェクト」の集合 T^1 とそれが有する「特徴全体」の集合 T^2 からなるデータは，「オブジェクト」が「特徴」を「有する」という二項関係 $R : T^1 \times T^2$ とみなせる．したがって，IRM の適用によって，オブジェクトである動物 (T^1) と，特徴 (T^2) の両方を同時にクラスタリングすることができる．Fig. 4 に実際に IRM を適用した結果を示す．図中のブロック化された行列は，IRM が 12 個生成した動物クラスタのうちの 7 個と，すべての特徴クラスタを示している．結果を見ると IRM は，関連の深い動物および関連する特徴をそれぞれクラ

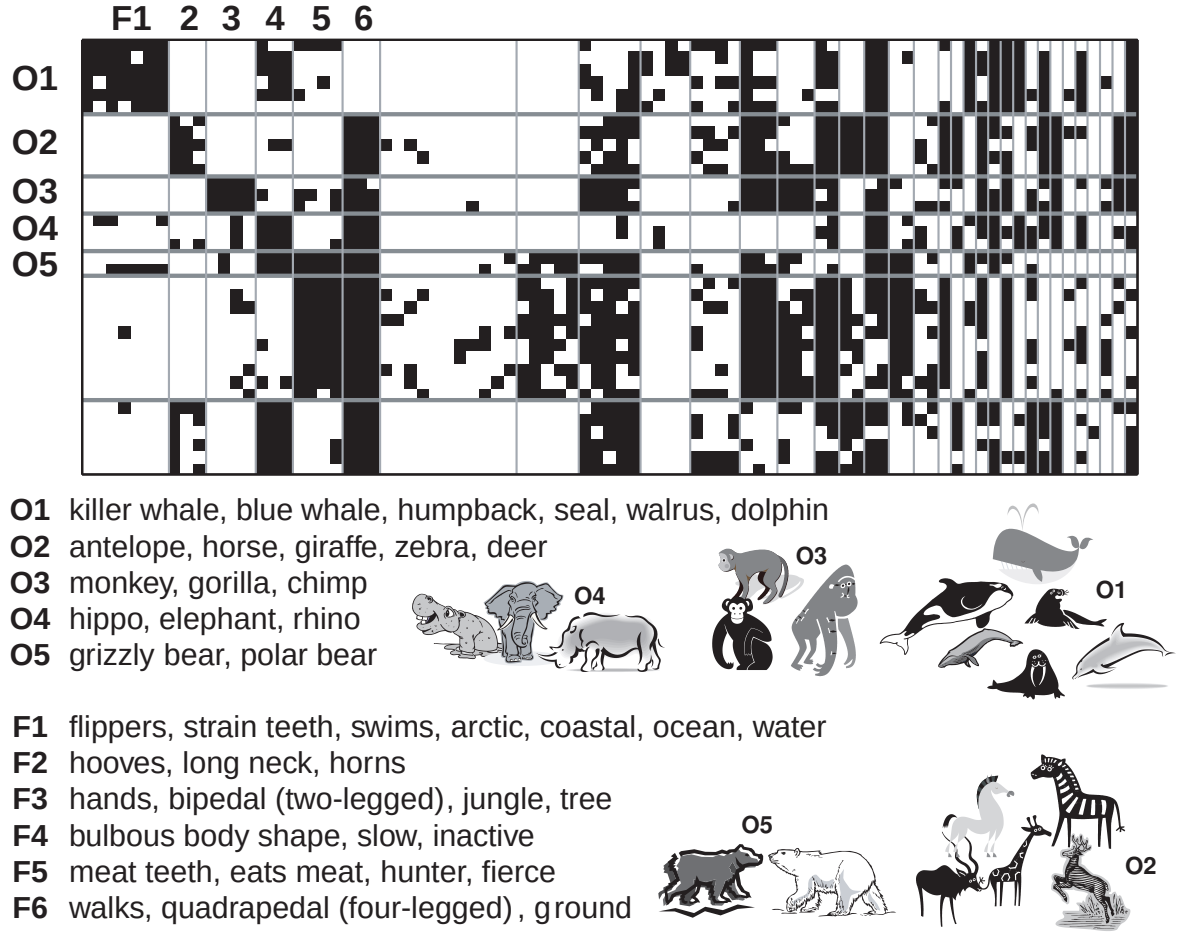

Fig. 4. 動物とその特徴の多重クラスタリング

スタに分割することに成功している．また，IRM は特徴クラスタ（例えば水棲動物 – 図中の **O1**）と動物クラスタ（例えば水に関係した特徴 – 図中の **F1**）との間の強い関係も発見している．

5.1 生成モデル

これ以降，関係データは常に二値とする．多値データや連続値データへの拡張は容易であるが，本稿では扱わない．簡単のため，まずは T^1, T^2 上の二項関係 $R : T^1 \times T^2$ を考える．任意の $i \in T^1, j \in T^2$ に対し，$R(i, j) = 1\,(0)$ は i から j へのリンクがある（ない）ことを意味する．

今，T^s $(s = 1, 2)$ を K^s 個からなるクラスタ $C^s = \{c_1^s, \ldots, c_{K^s}^s\}$ に分割するとし，T^s の要素 $1, \ldots, n_s$ が属するクラスタのインデックスを $z^s = (z_1^s, \ldots, z_{n^s}^s)$ とすると z^s の事前分布

として CRP を用いる IRM の生成モデルは以下のようになる．

$$
\begin{aligned}
&(1) \quad z^s \,|\, \gamma && \sim \mathrm{CRP}(\gamma) \;\; (s = 1, 2) \\
&(2) \quad \eta(k,l) \,|\, \beta && \sim \mathrm{Beta}(\eta(k,l);\, \beta, \beta) \;\; (k \in C^1,\; l \in C^2) \\
&(3) \quad R(i,j) \,|\, z^1, z^2, \eta && \sim \mathrm{Bernoulli}(R(i,j);\, \eta(z^1_i, z^2_j)) \;\; (i \in T^1,\; j \in T^2)
\end{aligned} \tag{5.1}
$$

ここで β, γ はハイパーパラメータである．ベルヌーイ分布 $\mathrm{Bernoulli}(x;\phi)$ は x が 0 または 1 の二値をとる確率変数 x の分布で，パラメータ ϕ は $x = 1$ となる確率を表し，$\mathrm{Bernoulli}(x;\phi) = \phi^x(1-\phi)^{1-x}$ で定義される．ここでは関係データは常に二値と仮定しているので，ベルヌーイ分布を用いるのは自然である．多値であれば多項分布，連続値であれば正規分布などを用いることになる．このモデルと前節で説明した通常の無限混合モデルとの大きな違いは，各 T^s それぞれが CRP でモデル化されること，観測されるのは各 T^s ごとの x_i ではなく $R(i,j)$ であること，モデルパラメータ $\eta(k,l)$ が前節の $\theta_{(k)}$ のように各クラスタごとではなく，クラスタの組合せ (k,l) に対して与えられることである．したがって $i \in k$ から $j \in l$ へのリンク存在確率 $\eta(k,l)$ は k, l のみに依存する（上述の stochastically equivalent に相当する）．また，ベータ分布 $\mathrm{Beta}(\eta;\, \beta,\beta)$ は前節の G_0 に相当する．なお，ベータ分布はベルヌーイ分布の共役事前分布である．$\beta > 1$ であれば，これは事前知識として，1, 0 のリンクが既にそれぞれ β 本づつ観測されているとすることに相当する．特に事前知識がない場合は $\beta = 1$ (一様分布)，もしくは $\beta = 0.5$ (0 付近 と 1 付近に分布が局在化) を用いる．β を固定せずに推定に含めることもできるが，ここでは省略する．共役事前分布であるベータ分布を仮定することにより，事後分布もベータ分布となり，計算が容易になる．

式 5.1 を Fig. 4 の例で説明すると，まず (1) により **O1**, **O2**,... および **F1**, **F2**,... といったクラスタが生成される．(2) によって各ブロックのリンク確率 η が生成される．さらに (3) によって実際にリンク確率 η に従ってリンクが生成される．例えば，**O1**×**F1** ブロックにおいて $\eta = 0.9$ であれば，結果的にこのブロックの黒が占める割合は 9 割程度となる．Fig. 4 の行列はこのような生成モデルで実際に生成された，と考えるわけである．

一般の場合として，$\mathcal{N}$ 個の異なるタイプ上の $\mathcal{M}$ 次元の関係 R を考える．この時，R の k 番目の引数となるタイプを d_k（例えば，$R : T^1 \times T^1 \times T^2$ の場合は，$d_1 = d_2 = 1, d_3 = 2$）とすると，式 (5.1) の (3) は

$$
R({i_1}^{d_1}, \ldots, {i_{\mathcal{M}}}^{d_{\mathcal{M}}}) \,|\, z^1, \ldots, z^{\mathcal{N}}, \eta \sim \mathrm{Bernoulli}(R({i_1}^{d_1}, \ldots, {i_{\mathcal{M}}}^{d_{\mathcal{M}}});\, \eta(z^{d_1}_{i_1}, \ldots, z^{d_{\mathcal{M}}}_{i_{\mathcal{M}}})) \tag{5.2}
$$

と書ける．関係が複数ある場合は，各関係 R^j ごとにパラメータ行列 η^j を定める．その他の議論は同様である．

5.2 目的関数

IRM では，未知データに対する予測というよりも，観測された R が与えられたもとでの事後確率を最大にするようなクラスタ分割 $z^1, \ldots, z^{\mathcal{N}}$ を求めることを考える．話を単純

化するために $T^1 \times T^2$ の例で説明する．式 (5.1) に示した，確率変数間の条件付独立性に注意すると，$P(R, z^1, z^2, \eta \mid \beta, \gamma) = P(z^1 \mid \gamma)P(z^2 \mid \gamma)P(R \mid z^1, z^2, \eta)P(\eta \mid \beta)$ と分解できる．したがって，$P(R, z^1, z^2 \mid \beta, \gamma) = P(z^1 \mid \gamma)P(z^2 \mid \gamma)\int P(R \mid z^1, z^2, \eta)p(\eta \mid \beta)d\eta$ となる．さらに，この積分の中の $P(R \mid z^1, z^2, \eta)$ は

$$
\begin{aligned}
P(R \mid z^1, z^2, \eta) &= \prod_{i,j} P(R(i,j) \mid z^1, z^2, \eta) = \prod_{i,j} \eta(z_i^1, z_j^2)^{R(i,j)}(1-\eta(z_i^1, z_j^2))^{1-R(i,j)} \\
&= \prod_{k \in C^1,\ l \in C^2} \eta(k,l)^{m(k,l)}(1-\eta(k,l))^{\overline{m}(k,l)} \qquad (5.3)
\end{aligned}
$$

となる．ただし，$m(k,l), \overline{m}(k,l)$ は (k,l) ブロック内でのリンク数，より正確には，$m(k,l) = \sum_{i:z_i^1=k}\sum_{j:z_j^2=l} R(i,j)$，$\overline{m}(k,l) = \sum_{i:z_i^1=k}\sum_{j:z_j^2=l}(1-R(i,j))$ である．即ち，R の事後確率は $m(k,l)$, $\overline{m}(k,l)$ で特徴付けられる．η の事前分布 $p(\eta \mid \beta)$ は，$B(x,y)$ をベータ関数とし，また添え字 (k,l) を簡略化すると $p(\eta \mid \beta)$=Beta$(\eta;\ \beta,\beta)$=$\eta^{\beta-1}(1-\eta)^{\beta-1}/B(\beta,\beta)$ である．一方 η の事後確率は，ベイズの定理より $P(\eta \mid z^1, z^2, R, \beta) \propto P(R \mid z^1, z^2, \eta)P(\eta \mid \beta)$ であるので，これに式 (5.3) を代入すると再びベータ分布 Beta$(\eta;\ m(k,l)+\beta, \overline{m}(k,l)+\beta)$ となる．これは共役事前分布の効用である．結局，最大化すべき z^1, z^2 の事後確率は以下となる．

$$
\begin{aligned}
P(z^1, z^2 \mid R, \beta, \gamma) \propto & P(z^1 \mid \gamma)P(z^2 \mid \gamma)P(R \mid z^1, z^2, \beta) \\
= & P(z^1 \mid \gamma)P(z^2 \mid \gamma)\int P(R \mid z^1, z^2, \eta)p(\eta \mid \beta)d\eta \\
= & P(z^1 \mid \gamma)P(z^2 \mid \gamma)\prod_{k \in C^1,\ l \in C^2} \frac{B(m(k,l)+\beta, \overline{m}(k,l)+\beta)}{B(\beta,\beta)} \qquad (5.4)
\end{aligned}
$$

なお $P(z^1 \mid \gamma), P(z^2 \mid \gamma)$ は式 (2.6) より計算できる．

5.3 ギブスサンプリングによるクラスタ割り当ての最適化

本節ではギブスサンプリングを用いて実際に事後確率を最大にするクラスタ割り当てを求める手順を説明する．IRM の場合，z_i^1（z_j^2 でも同様）以外を固定した条件付き事後確率 $P(z_i^1 = k^* \mid z_{-i}^1, z^2, R, \eta)$ は，式 (4.1) を適用することにより，以下となる（γ, β は省略）．

$$
P(z_i^1 = k^* \mid z_{-i}^1, z^2, R_i, \eta) \propto P(z_i^1 = k^* \mid z_{-i}^1)P(R_i \mid z_i^1 = k^*, z_{-i}^1, z^2, \eta) \qquad (5.5)
$$

ただし，R_i は R の i 行目のみを取り出したものとする．ここで k^* が既存のクラスタの場合，右辺第二項は $\prod_{l \in C^2} P(R_i \mid \eta(k^*, l), z^2)$ と書けて，これが式 (4.3) の $p(x_i \mid \theta_{(k)})$ に相当する．ここでもベータ分布がベルヌーイ分布の共役事前分布であることを活用すると，式 (4.4) は

$$
P(z_i^1 = k^* \mid z_{-i}^1, z^2, R) \propto \begin{cases} \dfrac{m_{-i,k^*}^1}{n-1+\gamma} \prod_{l \in C^2} \dfrac{B(m_{+i}(k^*,l)+\beta,\ \overline{m}_{+i}(k^*,l)+\beta)}{B(m_{-i}(k^*,l)+\beta,\ \overline{m}_{-i}(k^*,l)+\beta)} & m_{k^*,-i} > 0 \\ \dfrac{\gamma}{n-1+\gamma} \prod_{l \in C^2} \dfrac{B(m_{+i}(k^*,l)+\beta,\ \overline{m}_{+i}(k^*,l)+\beta)}{B(\beta,\ \beta)} & k^* \text{ は新規クラスタ} \end{cases} \qquad (5.6)
$$

となる．ただし，$m^1_{-i,k}$ は i を除いてカウントしたクラスタ $k \in C^1$ の要素数，$m_{-i}(k,l)$ は i を除いてカウントした k から l へのリンク数，$m_{+i}(k,l)$ は i をクラスタ k に新たに加えてカウントした k から l へのリンク数である[*8]．

以上をふまえ，ギブスサンプリングの概要は以下のようになる．

1. タイプ T^1, T^2 についてクラスタ割り当てを初期化する．
2. タイプ T^s $(s = 1, 2)$ およびその要素 i をランダム（もしくは順番）に選択する．
3. i のクラスタ割り当て $z^s_i = k$ を解除し，i を新たに到着した要素とみなす．ここで k が i 単独のクラスタの場合，k は消滅し，総クラスタ数が一つ減る．
4. 式 (5.6) に従って i に対する新たなクラスタ割り当て $z^s_i = k^*$（k^* は既存クラスタのひとつ，もしくは新規クラスタ）をサンプリングし受理する．
5. ステップ 2〜 4 を一定回数繰り返した後，得られたクラスタ割り当て z^1, z^2 のうちで式 (5.4) の事後確率が最大なものを出力して終了する．

実際の実装時には探索の効率化のため，温度パラメータの異なる複数のマルコフ連鎖を同時に更新し，一定の条件を満たす場合にマルコフ連鎖の間で状態の交換を行う，Metropolis Coupled MCMC [6] を用いたり，定期的にクラスタ割り当てのスプリット／マージ [9] などをあわせて行う．

ちなみに式 (5.4) において η は積分消去されるが，z^1, z^2 が得られた後，$p(\eta \mid R, z^1, z^2) \propto P(R \mid \eta, z^1, z^2)p(\eta)$ を最大にする η の MAP(maximum a posteriori) 推定値 $\bar{\eta}$（この場合 η の期待値でもある）は以下の式で得られる．

$$\bar{\eta}(k,l) = \frac{m(k,l) + \beta}{m(k,l) + \bar{m}(k,l) + 2\beta} \tag{5.7}$$

これはクラスタ k, l 間の関係の強さの尺度として用いることができる．例えば Fig. 4 の例の **O1**×**F1** ブロックでは $\bar{\eta}$ は大きく，**O1**, **F1** 間の強い関係を示している．

5.4 人工データによる評価

まず，$S1 : T^1 \times T^2$，$S2 : T^1 \times T^2, T^1 \times T^3, T^2 \times T^4$，$S3 : T^1 \times T^2 \times T^3$ の 3 つの異なる構成の人工データをランダムに生成し，これらに IRM を適用しその性能を検証した．Fig. 5(a) では，各タイプ 40 要素とし，真のクラスタ分割数 d を 2 から 10 まで変化させた[*9]．ただし，クラスタごとの要素数はほぼ等しい．また，各構成，各 d に対して β の値（式 (5.1) 参照）を 0.1 と 1.0 の二種類用意した．前述のように β が小さいと $\eta(k,l)$ の値は 0 もしくは 1 付近に局在化し，クラスタのブロック構造は明確化するが，逆に β が大きくなるとより曖昧になる．

[*8] もともと $i \in k$ なら $m(k,l)_{+i} = m(k,l)$，また，k が新規なら $m(k,l)_{+i}$ は i から l へのリンク数

[*9] Figs. 4, 5 は [10] より転載

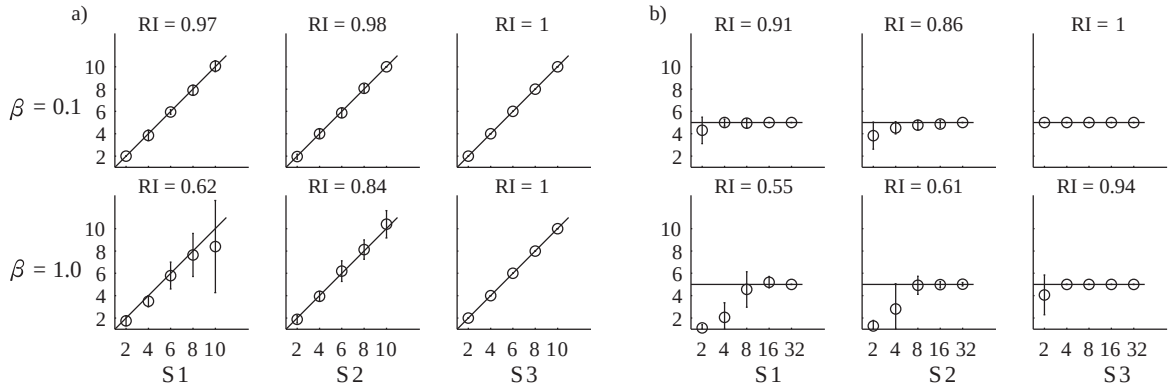

Fig. 5. 人工データによる IRM の性能の検証．x 軸は真のクラスタ数，y 軸は IRM が復元したクラスタ数を示し，各々乱数の初期値を変え 10 回実験した平均値を示す．

Fig. 5(a) の上段（下段）は $\beta = 0.1$ （$\beta = 1.0$）の場合の結果である．$\beta = 0.1$ の場合はいずれの構成も，IRM がほぼ完璧に真のクラスタ数を復元できていることがわかる．一方 $\beta = 1.0$ の場合はデータのノイズが増加するため，$S1$ など，真のクラスタ数を復元できない場合もあるが，概ね結果は安定している．特に $S3$ では $\beta = 1.0$ であっても常に 100% 正しいクラスタを復元している．$S3$ においてはさらに $\beta = 6$ まで変化させた予備実験の結果，ノイズが多い状況でも約 7 割で真のクラスタ数を復元できるほど結果は安定している．タイプ間の関係についての情報が増えるほど IRM はそれを有効活用し，真のクラスタ数の復元が容易になるといえる．

クラスタ割り当ての精度を評価するために計算した adjusted Rand index (RI) 指標 [8] の平均値を各グラフの上部に示す．RI は，与えられた二つの割り当て（一般に両者のクラスタ分割数は異なる）が（添え字の任意性を除いて）完全に一致する場合に値 1，前者に対して後者がまったくランダムである場合に（期待値として）値 0 となる指標である．この結果より，IRM はクラスタ数のみならず，割り当てそのものに関してもかなり正確に復元していることがわかる．

Fig. 5(b) はクラスタ数を $K = 5$ に固定したもとで要素数を 10 から 160 まで増加させた場合の IRM の性能の変化を比較したグラフである．2.4 節で述べたとおり，CRP 事前分布のもとでは要素数の増加に応じてクラスタ数の期待値は対数オーダで増加するが，この事前分布の影響はあまり大きくないことがわかる．実際，上段のグラフによると，データのノイズが少なければ IRM はほぼ正しいクラスタ数を復元している．また下段のグラフは，よりノイズの大きなデータであってもクラスタごとの要素数が統計的に不十分な場合を除いて確実に正しいクラスタ数を復元できることを示している．

既に Fig. 4 で紹介した動物とその特徴への適用例の他，生物医学オントロジ，オーストラリアの原住民族の血族関係，各国間の国際関係データなどへの適用については [10] を参照されたい．

6. おわりに

ノンパラメトリックベイズモデリングは，その基礎理論は数理統計学の分野で約 30 年以上前に提唱されていた [5] が，近年，その計算手法が再考案され，多方面の分野で研究が進められている．機械学習の分野では，多重特徴のモデル化の研究 [7]，変分ベイズ法に基づく DPM の効率的学習法 [4]，自然言語処理の分野では，無限のトピック遷移を実現する文書の生成モデル [3] 等も提案されている．さらに，数理心理学等の分野でも，個人差のモデリング [13] といった興味深い研究もなされている．各々の詳細は参考文献を参照されたい．

参考文献

[1] Aldous, D., Exchangeability and Related Topics, École dété de probabilités de Saint-Flour, **XIII**,(1983), 1–198.

[2] Antoniak, C.E., Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems, Annals of Statistics, **2**, 1152–1174, (1974).

[3] Blei, D. M., Griffiths, T., Jordan, M. I., and Tenenbaum, J., Hierarchical topic models and the nested Chinese restarant process. Advances in Neural Information Processing Systems, **16**, (2004).

[4] Blei, D. M., and Jordan, M. I., Variational inference for Dirichlet process mixtures, Bayesian Analysis, **16**,1, 121–144, (2005).

[5] Ferguson, T.S., A Bayesian analysis of some nonparametric problems. Annals of Statistics, **1**-2, 209–230, (1973).

[6] Geyer, C.J., Markov chain Monte Carlo maximum likelihood, Computing Science and Statistics, Proc. of the 23rd Symposium Interface, (1991), 156–163.

[7] Griffiths, T. L., and Ghahramani, Z., Infinite latent feature models and the Indian buffet process, Gatsby Computational Neuroscience Unit Technical Report GCNU TR 2005-001, (2005).

[8] Hubert, L., and Arabie, P., Comparing partitions, Journal of Classification, **2**, 193–218, (1985).

[9] Jain, S. and Neal, R.M., A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model, Jornal of Computational and Graphical Statistics, **13**, (2004), 158–182

[10] Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T. and Ueda, N., Learning systems of concepts with an infinite relational model, Proc. of the 21st AAAI conference, 381–388, (2006).

[11] MacEachern, S.N. and Müller, P., Estimating Mixture of Dirichlet Process Models, Journal of Computational and Graphical Statistics, **7**, (1998), 223–238.

[12] McCray, A.T., An Upper Level Ontology for the Biomedical Domain, Comparative and Functional Genomics, **4**, 80–84, (2003)

[13] Griffiths, T. L., Steyvers, M., and Lee, M.D., Modelling individual differences using Dirichlet processes, Journal of Mathematical Psychology, **50**, 101–122, (2006)

[14] Neal, R. M., Markov chain sampling methods for Dirichlet process mixture models,

Technical Report (No. 9815), Department of Statistics, University of Toronto, (1998)

[15] Nowicki, K and Snijders, T.A.B., Estimation and Prediction for Stochastic Blockstructures, Journal of the American Statistical Association, **96**, 1077–1087, (2001).

[16] Osherson, D.N., Default Probability, Cognitive Science, **15**, 251–269, (1991).

[17] Rasmussen, C., The Infinite Gaussian Mixture Model, in Advances in Neural Information Processing Systems, **12**, 554–560, (2000).

[18] Sethuraman, J., A constructive definition of Dirichlet priors, Statistica Sinica, **4**, 639–650, (1994).

[19] Wasserman, S., and Faust, K., Social Network Analysis: Methods and Applications, Cambridge University Press, (1994).

上田 修功 (非会員)　〒619-0237 京都府相楽郡精華町光台 2-4

昭和 33 年生．昭和 59 年 3 月大阪大学大学院通信工学専攻修士課程了．同年 NTT 入社．現在，NTT コミュニケーション科学基礎研究所　協創情報研究部長．主として統計的学習，パターン認識，データマイニングの研究に従事．博士（工学）．情処，信学会，日本神経回路学会，IEEE 各会員．

山田 武士 (非会員)　〒619-0237 京都府相楽郡精華町光台 2-4

昭和 39 年生．昭和 63 年 3 月東京大学理学部数学科卒業．同年 NTT 入社．現在，NTT コミュニケーション科学基礎研究所　創発環境研究グループリーダ．主として機械学習，最適化等の研究に従事．博士（情報学）．情処，信学会，ACM，IEEE 各会員．