F-036 Some Pitfalls in Infinite Relational Data Analysis

中野允裕
NTT コミュニケーション科学基礎研究所
〒 243-0198 神奈川県厚木市森の里若宮 3-1, masahiro.nakano.pr@hco.ntt.co.jp

Abstract： 本論文では，無限関係データ解析に利用される配列の分割モデルが直面する問題を明らかにし，その解決策を模索する中で浮上したノンパラメトリックベイズモデルの構成法について議論する．潜在的に無限のデータを仮想的に無限の計算機資源を使って解析する機械学習手法としてノンパラメトリックベイズ法と呼ばれる一分野が 20 年以上に渡って発展を続けている．特に，一次元配列（系列）モデルに関してはその存在，表現法，推論法について様々な知見が得られているとともに実応用にも幅広く利用されている．一方，多次元配列モデルに関しては未だ多くの課題を残しており，一次元配列モデルで得られた知見がそのまま多次元へ適用できない場合が少なくない．本論文では，特に二次元配列の長方形分割モデルにおける存在，表現法に着目し，一次元配列の分割モデルとの間のギャップを明らかにする．まずはじめに，ノンパラメトリックベイズモデルの持つ重要な性質として交換可能性・射影性を復習し，それらを同時に満たす一次元配列の分割モデルの表現法を見直す．次に，二次元配列の長方形分割モデルに関して交換可能性・射影性を持つものの存在を構成的に示す．最後にその表現法に関して，有限次元のパラメータ化に纏わる困難について述べる．

1. はじめに

統計的機械学習は，多くの観測データを十分な計算資源を持つコンピュータ上のアルゴリズムの入力とすることで，その出力を通してデータの統計的な性質を明らかにすることを目的としている．素朴な直観としては，観測データは多ければ多いほどそのデータの真の統計的な性質に近くなることが期待され，コンピュータの資源は多ければ多いほどより正確にデータ解析が行われるものと想像される．この素朴な直観が正しいことの裏付けを持った統計的機械学習手法の一つがノンパラメトリックベイズ法と呼ばれ，20 年以上に渡る発展とともに機械学習の一分野を築いてきた．ノンパラメトリックベイズ法は，無限のパラメータを持つ確率モデルを人手で設計し，それを潜在的に無限の観測データにフィッティングさせてモデルの事後確率分布を推論することによってデータ解析を行う．しかし，現実の世界では無限のデータを計算機に入力することはおろか，言うまでもなく無限のパラメータを計算機上に実装することも出来ない．肝となるのは，ある種の相似性（仮定）を利用することによって，無限次元の確率モデルによる無限のデータの解析を，有限のパラメータによる有限のデータの解析を通して実現することにある．

ノンパラメトリックベイズ法が仮定するある種の相似性とは**交換可能性**と**射影性**の 2 つの性質によって説明される．**交換可能性**とは，インデックス付けされた観測の確率がインデックスの付け方によらず不変であることを指す．例えば大量の画像をクラスタリングする問題を考え，インデックス付けられた画像を入力としてそれらをクラスタリングした結果に対して確率値を返すような確率モデルを想像する．画像を例えば $1, 2, 3, \ldots,$ や $2, 1, 3, \ldots,$ などのようにどのようにインデックス付けたとしてもその返す確率値が不変である場合，その確率モデルは交換可能性を持つと言う．これは Kallenberg の表現定理 [7, 8] によって課せられる要件でもあり，より詳細な定義は次節にて述べる．これによって，そのいかなる有限の部分データも潜在的に無限のデータが持つ統計的な性質を反映することが仮定される．次に**射影性**とは，観測を一部の部分に制限する操作が確率の周辺化の操作に一致することを指している．これは Kolmogorov の拡張定理 [3] により課せられる要件であり，有限次元のパラメータを持つ確率モデルが族が一意に無限次元の確率モデルへ拡張されるため必要条件に当たる．以上のことから，交換可能性と射影性を同時に満たす（**無限交換可能性を持つ**とも呼ばれる）確率モデルによって，潜在的に無限のデータを無限のパラメータを持つ確率モデルによる解析が実現される．

ノンパラメトリックベイズ法による統計的機械学習においては，ある確率過程によるデータ解析手法の確立のために，次の 4 点が相互に強い関係を持つ形で研究対象となる：

- 確率過程の存在.
- 確率過程の表現法.
- 確率過程の実データへの適用法（実データのモデリング）.
- 確率過程の持つパラメータの事後確率の推論法.

例えば，より効率的な推論法のために新しい表現法が模索されることもあれば，ある実データのモデリングを動機として新しい確率過程が見いだされその存在の保証が必要となることもある．本論文での中心的な話題は確率過程の存在と表現法に対応している．

本論文ではまず 2 章にて，無限交換可能なノンパラメトリックベイズモデルの最も基礎的なものの一つである一次元配列の分割モデルについて，その存在と表現法の観点から既存の知見を復習する．具体的にはディリクレ過程混合モデルを例とし，その表現方法を紹介する．次に，一次元配列の分割モデルの多次元拡張の典型的な一例として，二

次元配列の長方形分割モデルを取り上げる．関連する従来研究を述べるとともに，技術的な課題をここで明らかにする．3 章では無限交換可能性を持つ長方形分割モデルの存在を構成的に示す．4 章ではその長方形分割モデルの表現法において，一次元配列の分割モデルにはない技術的なギャップを明らかにする．

2. 無限交換可能な一次元配列分割モデル

はじめに 1 章で述べた交換可能性と射影性に対してより正確な定義を与えるため，必要となる記法を導入する．世界のランダムな出来事の全ては共通の抽象確率空間 $(\Omega, \mathcal{A}, \mathbb{P})$ に支配されているものとし，Ω が点集合，$\mathcal{A}$ は σ-加法族，$\mathcal{P}$ を確率測度とする．Ω は世界のランダムな出来事全てを記述するのに十分な情報を持っており，もしも $\mathcal{P}$ に従う世界の状態を指す一点 $\omega \in \Omega$ を知ることが出来たとすればそれはすなわち世界のランダムな出来事の全ての結果を知ることが出来ることを意味している．当然そのようなことが叶う訳もなく，我々は世界の一部を切り出した観測可能な空間 $\mathcal{X}$ を導入して，確率変数 $X : \Omega \to \mathcal{X}$ とその分布 $\mu_X := X(\mathcal{P})$ を介して世界の状態 ω の手がかりの一端を得ているのである．さらに，ベイズ統計では μ_X を直接捉えるのではなく，何らかのパラメータとなる確率変数 Θ を導入し，計算機上取扱いやすいパラメトリックなモデル $\mu_X(X|\Theta)$ を介することで ω の手がかりを探る戦略が用いられる．ここで条件付き確率は以下のように定義される: 任意の可測な集合 $B \in \mathcal{B}$ に対して $\mu_X(B|\Theta)(\omega) := \mu_X(B|\sigma(\Theta))(\omega) = \mathbb{E}[\mathbb{I}[B]|\sigma(\Theta)](\omega)$.

2.1 交換可能性と射影性

交換可能性: 集合 A と自然数 $k \in \mathbb{N}$ に対して，$A^{(k)}$ を A の部分集合でサイズ（要素数）が k のものの集合とする．確率変数の配列 $X := (X_e)_{e \in \mathbb{N}^{(k)}}$ が交換可能性を持つとは，任意の $\mathbb{N}$ の置換 π に対して，確率測度 μ_X が

$$\mu_X((X_e)_{e \in \mathbb{N}^{(k)}}) = \mu_X((X_{\sigma(e)})_{e \in \mathbb{N}^{(k)}}), \qquad (2..1)$$

を満たすことを指す．Kallenberg の表現定理 [7, 8] は交換可能な配列に対する必要十分条件を与えるもので，任意の配列 $(X_e)_{e \in \mathbb{N}^{(k)}}$ が交換可能であるならば, 可測関数 $f : [0,1] \times [0,1]^k \times [0,1]^{\binom{k}{2}} \times [0,1]^{\binom{k}{k-1}} \times [0,1] \to \mathcal{X}$ が存在し，

$$(X_e)_e = f(U, (U_a)_{a \in e}, (U_a)_{a \in e^{(2)}}, \ldots, (U_a)_{a \in e^{(k-1)}}, U_e),$$

と表せることが必要十分である．ここで，U と $(U_a)_{a \subseteq \mathbb{N}, |a| \leq k}$ は独立な一様分布 Uniform$([0,1])$ に従うものとする．配列の次元に相当する k がそれぞれ $k = 1$ と $k = 2$ の場合は de Finetti の定理（図．1，左）と Aldous-Hoover 表現定理（図．1，右）[1, 2, 6] として知られている．

射影性: 射影系とはあるインデックス集合 E でインデックス付けられた集合の族を指す．例えば E は自然数の集合の集合であり，$I \in E$ は自然数の集合 $I = \{1,2,3,4,5\}$ などを想像して頂きたい．まず $\langle \mathcal{X}_I, \mathcal{B}_I, Q_{J,I} \rangle_{I \preceq J \in E}$ を可測空間の射影系とし，任意の $I \in E$ に対して $\mathcal{X}_I$ は位相空間，$\mathcal{B}_I$ はその $\sigma-$ 加法族，$Q_{J,I}$ は $\mathcal{X}_J$ から $\mathcal{X}_I$ への写像，射の原像は集合 $B_I \subset \mathcal{X}_I$ に対して $Q_{J,I}^{-1} B_I = \{X_J \in \mathcal{X}_J | P_{J,I} X_J \in B_I\}$ と表せるものとする．例えば，添え字集合の要素のペア $I \prec J \in E$ に関して $I = \{1,2,\ldots,n\}$，$J = \{1,2,\ldots,n'\}$ $(n < n')$ とし，$\mathcal{X}_I$ として I の分割の集合を考えてみる．例えば射 $Q_{J,I} : x_J \mapsto x_I$ としては，J のある分割 x_J に関して，I の部分の分割を保存し，$J \setminus I$ の分割は全て削除するものなどが用いられる．次にこの可測空間上の確率測度の族 $\langle \mu_X^I \rangle_{I \in E}$ を考える．$\langle \mu_X^I \rangle_{I \in E}$ が射影性を持つとは，任意のペア $I \preceq J \in E$ に対して

$$Q_{J,I} \mu_X^J)(B_I) := \mu_X^J(P_{J,I}^{-1} B_I) = \mu_X^I(B_I), \qquad (2..2)$$

が任意の $B_I \in \mathcal{B}_I$ に関して成り立つことをいう．Kolmogorov の拡張定理 [3] は，射影性を持つ確率測度の族は，位相空間の射影極限 $\mathcal{X}_E$ 上の確率測度の射影極限 μ_X^E の一意の存在を保証するものである．

2.2 無限交換可能な一次元配列分割モデルの 4 つの表現法

一次元配列の分割モデルで無限交換可能性を満たすものはディリクレ過程混合モデルによって実現できることが知られている．ここではその 4 つの表現法を紹介し，無限交換可能なモデルの構成に対する既知の戦略を明らかにする．

入力の一次元配列として自然数 $\mathbb{N}$ でインデックス付けられた系列を考える．分割はバイナリ関係 $x : \mathbb{N} \times \mathbb{N} \to \{0,1\}$ を用いて表すものとし，$x(i,j) = 1$ は系列の i 番目と j 番目の要素が同じクラスタに属することを指すとする．いま確率変数の系列 $z_1, z_2, z_3, \ldots$ を導入し，各 z_i $(i \in \mathbb{N})$ は $\{1,2,3\ldots,\}$ の値をとるものとする．直観的には z_i は入力系列の i 番目の要素が属するクラスタ ID を意味している．入力系列の分割 $\mathbf{x} := (x(i,j))_{i \in \mathbb{N}, j \in \mathbb{N}}$ は $\mathbf{z} := z_1, z_2, z_3, \ldots$ から直ちに得られ，具体的には，もし $z_i = z_j$ ならば $x(i,j) = 1$ とし，さもなければ $x(i,j) = 0$ とすればよい．このことから，以下の分割モデルの表現においては確率変数 $\mathbf{z}$ の確率的生成モデルの記述を目指すものとする．ただし，無限交換可能性を求められるのは確率変数 $\mathbf{x}$ であって，系列の分割 $\mathbf{x}$ 自体には陽にクラスタ ID の情報は不要であることも念のため強調しておきたい．

無限次元のパラメータ化 - 一つ目の表現は，各 z_i が独立に無限次元の離散分布から生成されるものと見なし，その無限次元の重みが棒折り過程 [15] から生成されたと見なす階層モデルである．棒折り過程とはその名の通り，$[0,1]$ を逐次的に分割していく確率モデルである：ある自然数 $d \in \mathbb{N}$ と正の実数（一般的にはチューニング可能なハイパーパラメータ）$\beta > 0$ が与えられた時, 棒の分割 $(s_1, s_2, \ldots, s_d)$ が

$$\begin{aligned} s'_l &\sim \mathrm{Beta}(1,\beta)\ (m = 1,\ldots,d-1), \\ s_1 &= s'_1, \\ s_m &= s'_m \prod_{l=1}^{m-1} (1 - s'_l)\ \ (m = 2,\ldots,d-1), \\ s_d &= 1 - \sum_{m=1}^{d-1} s_m, \end{aligned} \qquad (2..3)$$

に従うものとする．簡単のため以降では $(s_1, s_2, \ldots, s_d) \sim \mathrm{SBP}(d, \beta)$ と書くことにする． 有限の棒折り過程の族 $\mathrm{SBP}(d, \beta)$

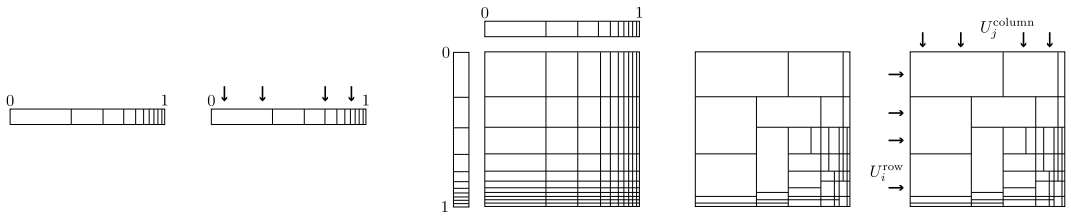

図 1: 左： de Finetti の表現定理に従う一次元配列の分割．$[0,1]$ 上の棒折り過程を生成し，その上の一様分布確率変数が同じ区画に属したものを同一のクラスタに割り当てることにより確率的に分割を得る． **右：Aldous Hoover の表現定理に従う二次元配列の長方形分割モデル．**

($d \in \mathbb{E}$) は無限の棒折り過程 SBP(∞, β) に一意に拡張されることが知られている [15]．まとめると，各 z_i は以下のような確率的生成モデルに従う：

$$\begin{aligned}(s_1, s_2, \dots) &\sim \text{SBP}(\infty, \beta) \\ z_i &\sim \text{Categorical}\big(\{1, 2, \dots\} \mid (s_1, s_2, \dots)\big)\end{aligned}$$

に従う．この表現法における一番の注意点は，仮に我々の興味が（本来無限の長さを持つ）入力系列の有限の部分系列にしか興味がないとしても，それらの分割を表現するためには必ず無限次元のベクトル $(s_1, s_2, \dots)$ を要することにある．このことからこの表現法は無限次元のパラメータ化と捉えることが出来る．

回顧的表現 - 先述の表現法に常に無限次元のベクトル $(s_1, s_2, \dots)$ を要する意味で，その表現を直接用いたアルゴリズムの実装は素朴には不可能である．しかし幸運なことに，もし我々の興味が入力系列の有限の部分に対する分割 $\{x(i,j) | i, j \in \{1, \dots, n\}\}$ だけにある場合には（そして計算機で表現する意味においてこれは常に成り立つ場合であると考えられるが）この無限次元のベクトルを回避するための工夫が知られている [12, 16]．まず独立な $[0,1]$ 上の一様分布に従う確率変数の系列 $U_1, \dots, U_n$ を導入する：$U_d \sim \text{Uniform}([0,1])$ $(d = 1, 2, \dots, n)$．次に，ベクトル $(s_1, s_2, \dots s_{n'}) \sim \text{SBP}(m, \beta)$ を逐次的にサンプリングし，これを $\max\{U_1, U_2, \dots, U_n\} < s_1 + s_2 + \cdots + s_{n'-1}$ を満たすまで繰り返す．念のため，この逐次的なサンプリングは確率 1 で有限回で終わることを強調しておきたい．最後に各クラスタ ID は $z_i = \sum_m m \cdot \mathbb{I}\Big[\sum_{l=0}^{m-1} s_l < U_i \leq \sum_{l=1}^{m} s_l\Big]$ として得られる．この表現法の肝は，$U_1, \dots, U_n$ を先に陽にサンプリングすることによって，離散分布の重みベクトルのうち，不必要なクラスタ ID に対応する重み全てをあらかじめ排除することができるため，重みベクトルはもはや無限次元とはならないことにある．

周辺化表現 - 最も驚くべき事実は，先述と等価なモデルがパラメータを介さずに観測そのものだけから直接表現できてしまうことであり，これは中華料理店過程 [2, 13] と呼ばれる確率過程に対応している．無限のデータを無限のクラスタに逐次的に割り当てていく処理を中華料理店で人がテーブルに割り当てられていく様に比喩して表現することを考える．いまデータを客，テーブルをクラスタだと見なすことにする．これは以下の再帰的な表現で記述できる：現在，n 人の客がテーブルに割り当てられているとする．ここで j 番目のテーブルに座っている客の人数を n_j とする．このとき，$n+1$ 人目の客は n_j に比例する確率で j 番目のテーブルを選び，β に比例する確率で新しいテーブルを選ぶ．

2.3 確率過程構成のための戦略

はじめに無限次元のパラメータ化と周辺化表現の違いから垣間見える無限交換可能なモデル構成法に対する 2 つの戦略から明らかにしたい．後者は交換可能性と射影性を同時に満たすモデルを直接構成しているのに対し，前者は棒折り過程を射影性のみの観点から設計し，それを de Finetti の表現定理の中間関数として用いることによって観測される系列の分割の交換可能性と射影性を同時に獲得していると見なすことが出来る．このことから，我々は無限交換可能なモデルの構成において以下の大きく二つの戦略を持っていると捉えることが出来る：

- 交換可能性と射影性を同時に考慮したモデルを直接構成する．
- 射影性のみ考慮したモデルを構成し，それを交換可能モデルの表現定理と組み合わせることで交換可能性と射影性を獲得する．

先述の一次元配列の分割モデルの場合，幸運にも前者の（無限の）中間確率変数を全て周辺化することによって後者と等価なモデルが導かれた．しかし，この中間確率変数の周辺化は一般には簡単に実現できるとは限らず，その意味において後者の表現法が得られることは非常に幸運であることを強調しておきたい．次の章では二次元配列の長方形分割モデルに対して，その存在と表現法について述べる．

3. 無限交換可能な二次元配列の長方形分割モデル

二次元配列の長方形分割に対するノンパラメトリックベイズモデルの歴史は，無限関係モデル [9] にまで遡る．これは先述の一次元配列の分割モデルである中華料理店過程を行と列の分割として用いその直積によって二次元配列の分割を作るもので，中華料理店過程の性質がそのまま引き継がれるために交換可能性と射影性のどちらも満たすモデルであった．しかし，行と列の分割によって生成されうる分割は全ての長方形分割の集合のうちの非常に制限された部分集合であり，無限関係モデルの台は *regular grid* と呼ばれる長方形分割クラスに制限されてしまう短所を抱えていた．

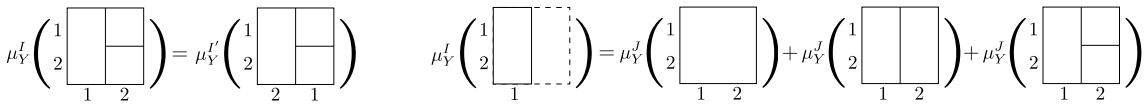

図 2: 左： 交換可能性．任意のサイズの配列に関する全ての長方形分割に関して，その行と列のインデックス付けを変更してもその確率は変わらない．これは **Aldous-Hoover** の表現定理に要する条件である． 右：射影性．行と列を削除する射が周辺化に対応している．これは **Kologorov** の拡張定理に要する条件である．

この問題を解消するため，長方形分割のより広いクラスである *hierarchical* にまで台を拡大したものがモンドリアン過程 [14] として提案された．モンドリアン過程による長方形分割表現は先述の**無限次元のパラメータ化**の戦略に基づいており，モンドリアン過程自体は $[0,1]\times[0,1]$ 上の分割モデルで射影性のみを持つものであり，これを Aldous-Hoover の表現定理と組み合わせことではじめて交換可能性と射影性を実現している．しかし，依然としてモンドリアン過程による長方形分割モデルの台も制限された長方形分割クラスであり，これを任意の長方形分割にまで広げるべく提案されたのが長方形分割過程 [11] である．その構成法については次節にて詳細に述べる．ここで一点強調しておきたいのは，無限関係モデル以降に登場した（長方形に限らない）多次元配列の分割モデルで無限交換可能性を実現したものは全て [14, 10, 5, 4]，その表現法において無限次元のパラメータ化の戦略を用いている．我々が知る限り例外としては，起源となった無限関係モデル自体が周辺化表現の方法を用いていることのみである．

ここでは入力の二次元配列として，行と列が自然数でインデックス付けられたものを考える．添え字集合 E として，配列 $I = I^{(r)}\times I^{(c)}$ $(I\in E)$ の集合を考える．ただし，$I^{(r)}$ と $I^{(c)}$ は自然数の集合で，それぞれ行と列のインデックスの集合に対応している．ここで，行と列の並び順が固定されていないことに注意されたい．添え字集合の要素のペア $I = I^{(r)}\times I^{(c)}$, $J = J^{(r)}\times J^{(c)}$ に対して，半順序 $I\preceq J$ を包含関係として定義する，すなわち $I^{(r)}\subseteq J^{(r)}$，$I^{(c)}\subseteq J^{(c)}$ として定義する．例えば $I^{(c)} = \{1,2,6,7\}\subset J^{(c)} = \{1,2,5,6,7\}$ はこの包含関係を満たしている．記法の簡単のため，(v,h)-セルと書いて，I の要素でその行と列のインデックスがそれぞれ $v\in I^{(r)}$ と $h\in I^{(c)}$ であるものを指すとする．

長方形分割 - 配列 I の全てのセルをクラスタリングすることを考え，もしある行 $I^{(r)}$ と列 $I^{(c)}$ の並び順に関して，全てのクラスタが長方形の形になるようなものが存在する時，そのようなクラスタリングを長方形分割と呼ぶことにする．配列 I の全ての長方形分割の集合を $\mathcal{Y}_I$ と書くことにする．長方形分割のサンプル $y_I\in\mathcal{Y}_I$ 等価関係 $y_I : I^{(r)}\times I^{(c)}\times I^{(r)}\times I^{(c)}\to\{0,1\}$ によって表すことができ，$y_I(i,j,i',j') = 1$ は (i,j)-セルと (i',j')-セルが同じクラスタに属することを表すものとする．

本章の目的は，$(\mathcal{Y}_I, \mathbf{2}^{\mathcal{Y}_I})$ 上の確率測度 μ_Y^I で，その族 $\langle\mu_Y^I, Q_{J,I}\rangle_{I\prec J\in E}$ が以下のような交換可能性と射影性を満たすものを構成することにある：

(C1) **交換可能性**．任意の配列 $I = I^{(r)}\times I^{(c)}\in E$ と置換 $\sigma : I^{(r)}\to I^{(r)}$，$\sigma' : I^{(c)}\to I^{(c)}$ を考え，$I' = \sigma(I^{(r)})\times\sigma'(I^{(c)})\in E$ とする．任意の長方形分割 $y_{I^{(r)}\times I^{(c)}}\in\mathcal{Y}_I$ に対して，$\mu_Y^I(y_{I^{(r)}\times I^{(c)}}\in\mathcal{Y}_I) = \mu_Y^{I'}(y_{\sigma(I^{(r)})\times\sigma(I^{(c)})}\in\mathcal{Y}_{I'})$ が成り立つ．

(C2) **射影性**．任意の配列のペア $I\prec J\in E$ に対して，$\mu_Y^I(y_I) = \mu_Y^J(Q_{J,I}^{-1}y_I)$ が成り立つ．

3.1 長方形分割過程 [11] の構成

ここでは元論文 [11] の構成法をより洗練されたものを紹介する．まずは非常に小さなサイズの配列に着目する．入力の配列 I として $|I^{(r)}||I^{(c)}| = 2$ を満たすものを考える，つまり I は二つのセルしか持たない場合を考える．この時，考えられる I の分割は二通りで，二つのセルを同一のクラスタへ割り当てるか別々のクラスタへ割り当てるかのいずれかである．そこで，ある正の実数 $0<p<1$ をハイパーパラメータとして，二つのセルが同一のクラスタへ割り当てられる確率を p，別々のクラスタへ割り当てられる確率を $(1-p)$ とする．この簡単なモデルは確かに I の長方形分割上の確率測度だと捉えることが出来る．次に行と列がそれぞれ二つの要素を持つような場合として，入力の配列 $J\in E$ で $|J^{(r)}| = 2$，$|J^{(c)}| = 2$ を満たすものを考える．J の長方形分割は 8 通り考えられるが，以下のようにそれぞれの確率を定めるとする（図．3）：

- 4 つのセルが全て同じクラスタへ属する確率を p^2,
- 2 つのクラスタへ属する確率を $(1/2)p(1-p)(p+q)$,
- 3 つのクラスタへ属する確率を $(1/2)p(1-p)(2-p-q)$,
- 4 つのクラスタへ属する確率を $(1-p)^3(1-q)$,

ただし， $q = p/(p^2-p+1)$ とする．可読性のため，長方形分割過程を $\mathrm{RTP}(I,p)$ で書き，その確率測度を ϖ_Y^I で表すものとする．この特別な入力配列のペア $I\prec J$ に対して，確かに射影性 $\varpi_Y^I(y_I) = \varpi_Y^J(Q_{J,I}^{-1}y_I)$ が成り立つことは簡単に確かめることが出来る．例えば，図 3 において 2 行 2 列の配列の上の行を削除する射を考えてみる．残った下の行の二つのセルが同じクラスタに属するパターンというのは，上段左，上段左から 3 番目，下段左から 2 番目の 3 つである．これらの確率を合計すると，

$$\begin{aligned} p^2 + \frac{1}{2}p(1-p)(p+q) + \frac{1}{2}p(1-p)(2-p-q) \\ = p^2 + \frac{1}{2}p(1-p)\{(p+q)+(2-p-q)\} \\ = p^2 + p(1-p) = p, \quad (3..1) \end{aligned}$$

となるので，確かに 1 行 2 列の入力配列の 2 つのセルが同一のクラスタに属する確率 p と一致する．同様に，図 3 にお

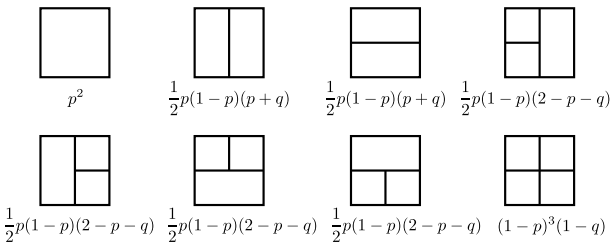

図 3: 2 行 2 列の入力配列に対する 8 通りの長方形分割とそれらの確率.

いて 2 行 2 列の配列の下の行，左の列，右の列を削除するような射に対しても射影性が成立することが確かめられる．

次に，入力配列がより大きなサイズの場合を考える．ここで肝となるのが先の 2 行 2 列の場合における確率モデルである．簡単な整理によって，4 つのセルのうち 3 つのセルのクラスタ割り当てが与えられた時，最後の 4 つ目のクラスタ割り当てに関する条件付き確率を得ることが出来る．したがって，もし行と列を逐次的に追加していきながら 2 行 2 列の範囲に対する条件付き確率を生成モデルとして適用していくことで，より大きな入力配列に対しても，長方形分割の確率的生成モデルを得ることが出来る．

3.2 長方形分割過程の持つ落とし穴

一見すると前節のモデルは確かに任意のサイズの入力配列に対して長方形分割の確率的生成モデルとして機能しているように思われる．しかし残念なことに，このモデルは任意の入力配列のペア $I \preceq J \in E$ に対して射影性 (C2) を満たす訳ではないことが分かる．その代り，以下のように入力配列に特別な要件を課すことによって辛うじて射影性を回復することが出来る．$E^{(\mathrm{RTP})}$ $(\subset E)$ を二次元配列の集合で，行と列は全順序を持つような自然数の集合であるものとする．特に，もし $I \in E^{(\mathrm{RTP})}$ であれば, $I = \{1, 2, \ldots, |I^{(r)}|\} \times \{1, 2, \ldots, |I^{(c)}|\}$ と表されるものと仮定する．ただし，$|A|$ は集合 A の要素数を表している．この制限された添え字集合 $E^{(\mathrm{RTP})}$ に対して, 長方形分割過程は射影性 (C2): $\varpi^I_Y(y_I) = \varpi^J_Y(Q^{-1}_{J,I} y_I)$ を満たすことが入力配列のサイズに関する数学的帰納法から確認できる．

3.3 無限交換可能な長方形分割の (R1) 表現

それでは無限交換可能な長方形分割の確率的生成モデルを構成的に述べていく．パラメータとなる確率変数 $\Theta_I : \Omega \to [0,1]^{|I^{(r)}|} \times [0,1]^{|I^{(c)}|} \times \mathcal{Y}_{\{1,\ldots,|I^{(r)}|\} \times \{1,\ldots,|I^{(c)}|\}}$ を導入する． 可読性のため，以下 $\Theta_I : \Omega \to \mathcal{Z}_I$ と書くことにする．パラメータのサンプル θ_I は以下のように構成される：

$$\begin{aligned} S_I &\sim \mathrm{SBP}(|I^{(r)}|, \beta) \times \mathrm{SBP}(|I^{(c)}|, \beta), \\ z_I &\sim \mathrm{RTP}(\{1, \ldots |I^{(r)}|\} \times \{1, \ldots |I^{(c)}|\}, p). \end{aligned} \quad (3..2)$$

パラメータ Θ_I 上の確率測度を $\nu^I_\Theta : \mathbf{2}^{\Theta_I} \to \mathbb{R}_+$ とする． その族 $\langle \nu^I_\Theta, Q_{J,I} \rangle_{I \prec J \in E}$ は，棒折り過程の射影性と長方形分割過程の持つ射影性から直ちに自身も射影性 (C2) を満たすことが確認できるため，Kologorov の拡張定理からその射影極限 ν^E_Θ が一意に存在することが分かる．パラメータ

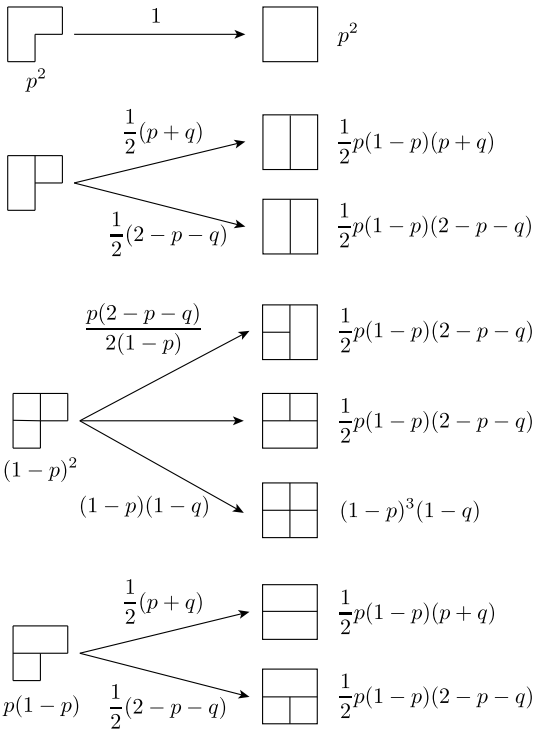

図 4: 2 行 2 列の入力配列に対する長方形分割過程から導かれる連鎖規則．3 つのセルのクラスタ割り当てが与えられた時に最後の一つのセルのクラスタ割り当ての条件付き確率を表す.

$\langle \Theta_I \rangle_{I \in E}$ の射影極限 Θ_E で条件付けた Y_I のパラメトリックモデル $\varpi^I_Y(Y_I | \Theta_E)$ を表すものとし, その生成モデルを次のように構成する:

$$\begin{aligned} U^{\mathrm{row}}_i &\sim \mathrm{Uniform}([0,1]) \ (i \in I^{(r)}), \\ U^{\mathrm{column}}_j &\sim \mathrm{Uniform}([0,1]) \ (i \in I^{(c)}), \\ y_I(i,j,i',j') &= z_E\Big(\gamma(U^{(r)}_i, \mathbf{s}^{(r)}_K), \gamma(U^{(c)}_j, \mathbf{s}^{(c)}_K), \\ &\qquad \gamma(U^{(r)}_{i'}, \mathbf{s}^{(r)}_K), \gamma(U^{(c)}_{j'}, \mathbf{s}^{(c)}_K)\Big), \end{aligned} \quad (3..3)$$

ただし，$\gamma(U, (s_1, s_2, \ldots)) := \sum_m m \cdot \mathbb{I}\Big[\sum_{l=0}^{m-1} s_l < U \leq \sum_{l=1}^{m} s_l\Big]$． ここでパラメータを周辺化した確率測度 $\mu^I_Y : \mathbf{2}^{\mathcal{Y}_I} \to \mathbb{R}_+$ $(I \in E)$，すなわち $\mu^I_Y(\cdot) = \int_{\mathcal{Z}_E} \varpi^I_Y(\cdot | \Theta_E = \theta_E) \nu^E_\Theta(\theta_E) d\theta_E$ を考える．この確率測度の族 $\langle \mu^I_Y, Q_{J,I} \rangle_{I \prec J \in E}$ は以下の性質を持ち，無限交換可能でかつ任意の長方形分割の集合を台に持つ：

命題 3..1 $(\mathcal{Y}_I, \mathbf{2}^{\mathcal{Y}_I})$ $(I \in E)$ 上の確率測度 λ^I_Y を上記の生成モデルによって与える．その族 $\langle \lambda^I_Y, Q_{J,I} \rangle_{I \prec J \in E}$ は

- 任意の $I \in E$ に対して交換可能性 *(C1)* を満たす.
- 任意の $I \prec J \in E$ に対して射影性 *(C2)* を満たす.
- 任意の $I \in E$，$y_I \in \bar{\mathcal{Y}}_I$ に関して $\lambda^I_Y(y_I) > 0$.

以上により無限交換可能な長方形分割モデルに対して無限次元のパラメータ化を用いた表現が得られた．では，中間確率変数の周辺化した表現は得られるのだろうか．一次

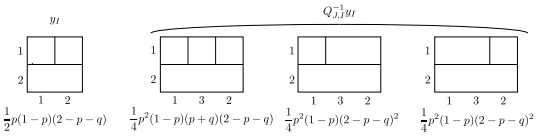

図 5: 中華料理店過程が $I \prec J \in E$ において射影性を満たさないことの例．この例において $\mu_X^I(y_I) \neq \mu_X^J(Q_{J,I}^{-1}y_I)$.

元配列の分割モデルにおいては中華料理店過程によってそれが実現されていた．では，中華料理店過程の多次元拡張にも相当するような表現法が可能なのであろうか．

4. おわりに

前章で得た無限交換可能な長方形分割の無限次元のパラメータ化表現は，その中間確率変数として棒折り過程による $[0,1]$ の無限の棒折りを 2 本と，長方形分割過程による無限に大きなサイズの配列の長方形分割を持っている．これらの確率変数を周辺化してしまうことは果たして可能なのであろうか．著者らが知る限り，これを実現する方法が未だ発見されていない．

ここではその難しさの一端を知るために，以下のような中間確率変数を有限次元に制限したものの族を考えてみる．有限次元のパラメータ化 μ_Y^I として，パラメータ Θ_I を条件付けた Y_I $(I \in E)$ のパラメトリックモデルを $\tilde{\mu}_Y^I(\cdot) = \int_{\mathcal{Z}_I} \varpi_Y^I(\cdot|\Theta_I = \theta_I)\nu_\Theta^I(\theta_I)d\theta_I$ により与える．再掲しておくと，前章で得た表現は $\mu_Y^I(\cdot) = \int_{\mathcal{Z}_E} \varpi_Y^I(\cdot|\Theta_E = \theta_E)\nu_\Theta^E(\theta_E)d\theta_E$ としてパラメータの次元が無限になっていたことに注意されたい．

Remark 4..1 確率測度の族 $\langle \tilde{\mu}_Y^I, Q_{J,I}\rangle_{I\in E}$ は射影性 *(C2)* を満たさない．

図 6 の例から射影性が崩れることが確認できる．入力配列のペア $I \prec J \in E$ として $|I^{(r)}| = |J^{(r)}| = 1$, $|I^{(c)}| = 2$, $|J^{(c)}| = 3$ の場合を考える．$y_I \in \bar{\mathcal{Y}}_I$ を 2 つのセルが同一のブロックに属するパターンとする．このとき

$$
\begin{aligned}
&\mu_Y^I(y_I) \\
&= \int ds_1' \mathrm{Beta}(s_1'; 1, \beta) \cdot \Big\{ s_1'^2 + (1-s_1')^2 + 2ps_1'(1-s_1') \Big\} \\
&\qquad = \frac{2}{(1+\beta)(2+\beta)} + \frac{\beta}{2+\beta} + 2p\frac{\beta}{1+\beta}.
\end{aligned}
$$

一方で，

$$
\begin{aligned}
\mu_Y^J(Q_{J,I}^{-1}y_I) = \int\int & ds_1' ds_2' \mathrm{Beta}(s_1'; 1, \beta)\mathrm{Beta}(s_2'; 1, \beta) \\
&\cdot \Big[p^2 + s_1'^2 + (1-s_1')^2 s_2'^2 + (1-s_1')^2(1-s_2')^2 \\
&\quad + p(1-p)\Big\{ \{1-(1-s_1')(1-s_2')^2\}^2 \\
&\quad + \{(1-s_1')(1-s_2')^2\}^2 + s_1'^2 + (1-s_1')^2 \Big\} \Big] \\
&= p^2 + \frac{4}{(2+\beta)^2} + p(1-p)\Big\{ 3 - \frac{\beta^2}{(1+\beta)^2} \\
&\quad + 2\frac{\beta^2}{(2+\beta)^2} - \frac{2}{1+\beta} + \frac{4}{(1+\beta)(2+\beta)} \Big\}.
\end{aligned}
$$

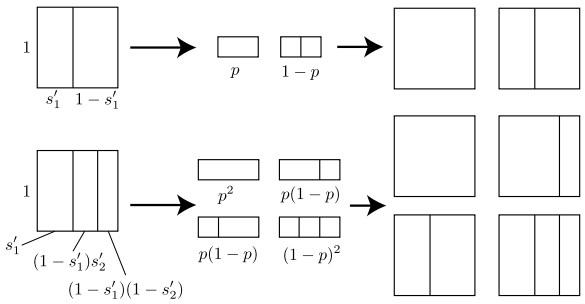

図 6: 確率測度の族 $\langle \tilde{\mu}_Y^I, Q_{J,I}\rangle_{I\in E}$ が射影性を満たさないことの例.

以上から $\mu_Y^I(y_I) \neq \mu_Y^J(Q_{J,I}^{-1}y_I)$ となることが分かる．このことから，もし前章で得た無限交換可能なモデルの周辺化表現を得ることが出来るとすれば，それは本当に無限の中間確率変数を周辺化しなければならないことを意味しており，非常に難しい問題であることが分かる．一次元配列における分割モデルの中華料理店過程の多次元化に相当する確率過程は果たして構成できるのか否か，その結論には今後の研究の発展が求められる．

参考文献

[1] D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11:581–598, 1981.

[2] D. J. Aldous. Exchangeability and related topics. *École d'Été St Flour 1983*, 1985.

[3] S. Bochner. *Harmonic analysis and the theory of probability*. University of California Press, 1955.

[4] X. Fan, B. Li, and S. A. Sisson. The binary space partitioning-tree process. In *International Conference on Artificial Intelligence and Statistics*, pages 1859–1867, 2018.

[5] X. Fan, B. Li, Y. Wang, Y. Wang, and F. Chen. The Ostomachion Process. In *AAAI Conference on Artificial Intelligence*, 2016.

[6] D. N. Hoover. Relations on probability spaces and arrays of random variables. Technical report, Institute of Advanced Study, Princeton, 1979.

[7] O. Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989.

[8] O. Kallenberg. Symmetries on random arrays and set-indexed processes. *Journal of Theoretical Probability*, 5(4):727–765, 1992.

[9] 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 National Conference on Artificial Intelligence*, 2006.

[10] M. S. Mackisack and R. E. Miles. Homogeneous rectangular tessellation. *Advances on Applied Probability*, 28:993, 1996.

[11] M. Nakano, K. Ishiguro, A. Kimura, T. Yamada, and N. Ueda. Rectangular tiling process. In *International Conference on Machine Learning*, 2014.

[12] O. Papaspiliopoulos. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

[13] J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *IBM Journal of Research and Development*, 2(25):855–900, 1997.

[14] D. M. Roy and Y. W. Teh. The Mondrian process. In *Advances in Neural Information Processing Systems*, 2009.

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

[16] S. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics Simulation and Computation*, 36(1):45–54, 2007.