言語類型論的特徴からの潜在的2値パラメータの獲得

村脇 有吾
京都大学大学院情報学研究科
murawaki@i.kyoto-u.ac.jp

1 はじめに

生得的な言語能力については人間集団間で差が認められない一方で、世界の諸言語は多種多様な特徴を示す。ある言語は動詞・目的語語順が VO 型であり、また別の言語は OV 型である。同様に、形容詞・名詞語順 (AN 型か NA 型)、接置詞型 (前置詞型と後置詞型)、声調の有無といった言語類型論的特徴について、世界の諸言語は様々な値をとる。なぜこのような多様性があるかは、普遍性と表裏一体の疑問である。

この問題への手掛かりとして、特徴間の依存関係が知られている。例えば、OV 型の言語は AN 型である傾向がみられる。このようにいくつかの特徴対に普遍的に成り立つ関係を Greenberg は提示した [3]。

Greenberg は表層的特徴の説明に終始したのに対し、普遍文法を追究する生成文法の一派は、潜在的なパラメータの存在を主張する。この枠組みは「原理とパラメータ」とよばれ、言語を普遍的な原理と可変のパラメータによって説明する。言語の多様性を説明するのは後者であり、複数あるパラメータについて、値を一通り決めると個別の言語があらわれる。パラメータは 2 値であり、一般に複数の表層的特徴の値を一度に、決定的に決める。例えば、主要部方向性パラメータは主要部先行、主要部後置のいずれかの値をとる [1]。この潜在的パラメータの値を主要部先行とすれば、表層的特徴の値は VO 型、NA 型、前置詞型と定まる。反対に主要部後置であれば、OV 型、AN 型、後置詞型となる。

本稿では、このような潜在的パラメータ列[1]を表層的特徴列から計算的に獲得することを試みる。そのために、パラメータ列から特徴列を生成するノンパラメトリック・ベイズのモデル (IBP-BP) を提案する。ただし、実データを扱う都合上、IBP-BP は原理とパラメータの枠組みとはいくつかの点で異なる。第一に、生成文法ではパラメータと特徴の関係は決定的だが、IBP-BP は確率モデルである。実データには例外がつきものであり、決定的な関係を獲得するのは困難だからである。また、表層的特徴のなかには、これまでにパラメータによって説明されたことのないものも少なくない。これに関連して、第二に、生成文法では表層的特徴は一つの潜在的パラメータによって決定されるが、IBP-BP では複数のパラメータの組み合わせによって表層的特徴の生成確率が決まる。第三に、IBP-BP も 2 値パラメータを採用するものの、モデル化の都合上、オンとオフが非対称的であり、オンとなったパラメータだけが表層的特徴の生成確率に作用する。そのため、例えば、主要部方向性パラメータは、IBP-BP では 1 つとは限らず、主要部先行と主要部後置に対応する 2 つ (あるいはそれ以上) のパラメータによって実現されるかもしれない。

本稿と似た動機から、筆者は以前別のモデル (AE-CP) を提案した [7] が、IBP-BP は大きく 3 つの点で異なる。第一に、AE-CP が獲得する潜在的パラメータは連続値だが、IBP-BP は 2 値パラメータを採用する。2 値パラメータは原理とパラメータの枠組みと共通するだけでなく、実際的な利点もある。得られたパラメータに対して、言語や生物 (特に遺伝子) のデータを分析する従来手法 (例えば [10]) をそのままか、あるいはわずかな変更のみで適用できる。こうしたデータは、多値の、特に 2 値の特徴列で表現されることが多かったからである。第二に、欠損値の扱いが異なる。既存の言語類型論のデータベースには欠損値が非常に多く、その扱いは挑戦的な課題であり続けている。自己符号化器に基づく AE-CP は欠損値に弱いため、あらかじめ別の手法 [6] で欠損値補完を行ったデータを用いていた。これに対し、IBP-BP は不確実性に頑健なベイズモデルであり、パラメータと欠損値の同時推論を実現する。第三に、AE-CP はパラメータの数を事前に決める必要があったが、我々は最適なパラメータ数について事前知識を持たない。IBP-BP はノンパラメトリック・ベイズのモデルであり、データからのパラメータ数の自動推定を試みる。

実験では欠損値補完の精度を評価したところ、IBP-BP はベースライン手法よりも高い性能を示した。したがって、IBP-BP は表層的特徴列に潜む規則性を捉えることに成功していると判断できる。

[1] 便宜的に列として扱うが、順番に意味ない。

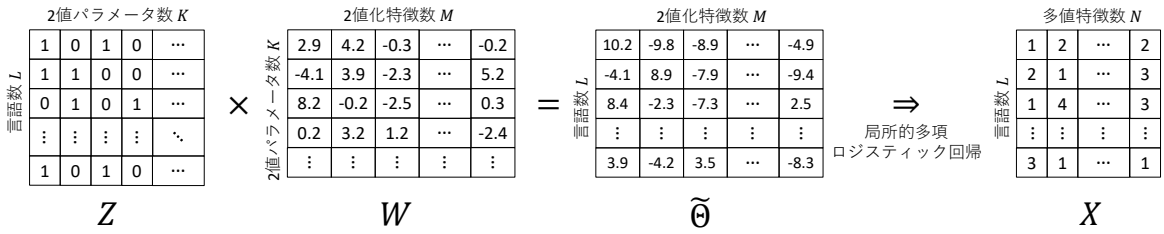

図 1: インディアンビュッフェ過程に基づくパラメータ列からの特徴列の確率的生成

L	言語数
K	パラメータ数 (無限)
K^+	有効パラメータ数
K°	推論時の追加パラメータ数
M	2 値化特徴数
N	多値特徴数
$Z \in \{0,1\}^{L\times K}$	2 値パラメータの行列
$W \in \mathbb{R}^{K\times M}$	重み行列
$\tilde{\Theta} \in \mathbb{R}^{L\times M}$	特徴スコア行列 (正規化なし)
$\Theta \in [0,1]^{L\times M}$	特徴選択確率の行列
$X \in \mathbb{N}^{L\times N}$	多値特徴行列

表 1: 本稿で用いる表記

2 データと前処理

言語類型論のデータベースとして *World Atlas of Language Structures* (WALS)[4] を用いた。2017 年時点で、WALS は 2,679 言語、192 特徴を収録する。ただし、言語と特徴からなる行列には欠損値が多く、被覆率は 15%を下回る。

元データに対して、以下の前処理を行った。まず手話、ピジン、クレオールを除去し、次に論理的に決定可能な特徴値を補完した。次に、手話に関する特徴や言語被覆率が 10%を下回る特徴を除去した。この結果、言語数 $L = 2,607$、特徴数 $N = 104$ を得た。被覆率は 26.9%に上昇したが、依然として低い。

特徴は一般に多値である。例えば基本語順の特徴 81A は `SVO`, `SVO`, `VSO`, `VOS`, `OVS`, `OSV`, `No dominant order` の 7 値のいずれかをとる。これらの特徴を 1-of-F_i 方式で 2 値化 (値の異なり数 F_i の特徴を合計で F_i 個で、そのうち 1 個しか 1 をとらないような 2 値特徴列に展開) すると、2 値化特徴数 $M = 723$ を得た。

3 提案手法

3.1 インディアンビュッフェ過程 (IBP)

表 1 に本稿で用いる表記を示す。ベイズ統計と生成文法という 2 つの異なる分野の用語が混じるため注意を要する。ベイズ統計の文献で特徴とよばれる潜在変数を、本稿では生成文法に従ってパラメータとよぶ。一方、本稿でいう特徴は観測変数 (ただし欠損値を含む) である。

表層的特徴列は、元の多値特徴列と 2 値化特徴列の 2 種類の表現を持つ。潜在的パラメータは 2 値化特徴列に対して働きかける。主要部方向性パラメータの例が示すように、ある特徴のある値 (動詞・目的語語順が VO 型) と別の特徴のある値 (形容詞・名詞語順が NA 型) の関係を捉えたいからである。多値特徴は $(1,1),(1,2),\cdots,(1,F_1),(2,1),\cdots,(i,j),\cdots,(N,F_N)$ のように索引づけされる。これに対応する 2 値化特徴の索引を $1,\cdots,m,\cdots,M$ とする ($M = \sum_{i=1}^{N} F_i$)。この 2 種類の索引は関数 $f(m) = (i,j)$ と $f^{-1}(i,j) = m$ により対応づけられる。

提案手法 IBP-BP は、潜在的パラメータをノンパラメトリック・ベイズのモデルであるインディアンビュッフェ過程 (IBP) によって生成する。図 1 に生成過程を示す。IBP は L 個の行と、加算無限個 (K 個) の列からなる 2 値行列 $Z \in \{0,1\}^{L\times K}$ を生成する。ここで、Z の要素 $z_{l,k}$ は言語 l の k 番目のパラメータである。IBP を実現する手段はいくつかあるが、ここでは棒折り過程 [9] を採用する。$z_{l,k}$ は Bernoulli(μ_k) に従う。いま、これらの μ_k を $\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$ となるように並べ替えると、$\mu_{(k)}$ は以下の手続きにより得られる。

$$\nu_{(k)} \sim \mathrm{Beta}(\alpha,1) \quad \mu_{(k)} = \prod_{l=1}^{k} \nu_{(l)}$$

L 個の言語のうち 1 個以上が $z_{l,k} = 1$ をとるパラメータを有効パラメータとよぶ。$\mu_{(k)}$ は単調に減少し、有効パラメータは有限個 (K^+) にすぎない。この K^+ 個の有効パラメータに対して順序を陽に保持する必要がないことが知られているため、ここでは半順序つき棒折り表現 [9] を採用する。なお、IBP のハイパーパラメータは $\alpha \sim \mathrm{Gamma}(1,1)$ とする。

次に、Z に対応する重み行列 $W \in \mathbb{R}^{K\times M}$ を生成し、特徴スコア $\tilde{\Theta} = ZW$ を得る。$\tilde{\Theta}$ の要素に着目すると、$\tilde{\theta}_{l,m} = \sum_{k=1}^{\infty} z_{l,k} w_{k,m}$ であり、$z_{l,k} = 1$ であるパラメータだけが特徴スコアに影響を与える。各 $w_{k,m}$ はスチューデントの t 分布 (自由度 $\nu = 1$) から生成する。この分布を選んだ理由は、(1) 推論の都合上、対数確率関数が微分可能でなければならないこと、(2) 正規分布とくらべて裾が広く、一部の要素が比較的大きな値を取りやすいことである。

Algorithm 1 IBP-BP の推論

for 反復 $t \leftarrow 1, T$ **do**
  s を更新
  $\mu^{\circ}_{(K^{\circ}+1)} < s$ となるまでパラメータを追加
  追加パラメータに対して $w_{k,m}$ を t 分布から生成
  $z_{l,k}$, $w_{k,*}$, 欠損値 $x_{l,i}$ をランダムな順序で更新
  非有効パラメータと対応する $w_{k,m}$ を削除
  μ_k を更新
  α を更新
end for

$\tilde{\Theta}$ に対し局所的に多項ロジスティック回帰を適用し、言語 l の i 番目の表層的特徴の値が j である確率 $\theta_{l,i,j}$ を得る。この確率にしたがって $x_{l,i}$ を生成する。

$$\theta_{l,i,j} = \frac{\exp(\tilde{\theta}_{l,f^{-1}(i,j)})}{\sum_{j'} \exp(\tilde{\theta}_{l,f^{-1}(i,j')})}$$
$$P(x_{l,i}|-) = \theta_{l,i,x_{l,i}} \quad (1)$$

$\theta_{l,i,j}$ を展開すると

$$\begin{aligned}\theta_{l,i,j} &\propto \exp(\sum_{k=1}^{\infty} z_{l,k} w_{k,f^{-1}(i,j)}) \\ &= \prod_{k=1}^{\infty} \exp(z_{l,k} w_{k,f^{-1}(i,j)})\end{aligned}$$

となることからわかるように、このモデルは専門家の積 [5] である。$z_{l,k} = 0$ のとき、$\exp(z_{l,k} w_{k,f^{-1}(i,j)}) = 1$ であり、$\theta_{l,i,j}$ に影響を及ぼさない。$z_{l,k} = 1$ のとき、$w_{k,f^{-1}(i,j)} > 0$ であれば $\theta_{l,i,j}$ を大きくし、$w_{k,f^{-1}(i,j)} < 0$ であれば $\theta_{l,i,j}$ を小さくする。ある一群の言語について $z_{l,k} = 1$ となるようなパラメータ k に着目すると、それらの言語において多値特徴の値 (i_1, j_1) と (i_2, j_2) が同時に現れやすいことは、$w_{k,f^{-1}(i_1,j_1)}$ と $w_{k,f^{-1}(i_2,j_2)}$ の両方を正の値にすることで表現できる。反対に、同時に現れにくい場合は、一方の重みを正に、他方を負にすればよい。

3.2 推論

IBP-BP の推論は Gibbs サンプリングにより行う。擬似コードをアルゴリズム 1 に示す。

パラメータは加算無限個存在するが、スライスサンプリングを用いることで、近似を行うことなく、適当な回数で棒折りを打ち切れる [9]。半順序つき棒折り表現において、K^+ 個の有効パラメータのうち、μ_k の最小値を μ_* とすると、まず補助変数 $s \sim \mathrm{Uniform}[0, \mu_*]$ を得る。次に、順序つき非有効パラメータを $\mu^{\circ}_{(K^{\circ}+1)} < s$ となるまで 1 個ずつ生成する (合計 K° 個)。非有効パラメータの生成確率は

$$p(\mu^{\circ}_{(k)}|\mu^{\circ}_{(k-1)}, z_{*,>k} = 0) \propto \exp(\alpha \sum_{l=1}^{L} \frac{1}{l}(1-\mu^{\circ}_{(k)})^l)$$
$$(\mu^{\circ}_{(k)})^{\alpha-1}(1-\mu^{\circ}_{(k)})^L \mathbb{I}(0 \le \mu^{\circ}_{(k)} \le \mu^{\circ}_{(k-1)})$$

手法	精度
FREQ	60.9%
MCR	69.9%
IBP-BP ($K_0 = 50$)	**71.8%**
IBP-BP ($K_0 = 100$)	70.4%
IBP-BP ($K_0 = 250$)	68.3%

表 2: 欠損値補完の精度

である。この分布は複雑だが、$\log \mu^{\circ}_{(k)}$ に対して対数凹であることから、適応的棄却サンプリングが利用できる [9]。こうして追加された K° 個のパラメータについて、対応する $w_{k,m}$ を t 分布から生成する。対応する $z_{l,k}$ はすべて 0 で初期化する。

$z_{l,k}$, $w_{k,m}$, 欠損値 $x_{l,i}$ を更新 (後述) したのち、非有効パラメータ k と、それに対応する $w_{k,m}$ を削除する。残った有効パラメータについては、μ_k を

$$\mu_k|z_{*,k} \sim \mathrm{Beta}(\sum_{l=1}^{L} z_{l,k}, 1 + L - \sum_{l=1}^{L} z_{l,k})$$

に従って更新する。α は事後分布

$$\alpha|Z \sim \mathrm{Gamma}(1 + K^+, 1 + \sum_{l=1}^{L} \frac{1}{l})$$

に従って更新する [2]。

$x_{l,i}$ は式 (1) により更新する。$P(z_{l,k}|-)$ は μ_k と言語 l のすべての i についての式 (1) の積に比例する。問題は残る $w_{k,m}$ の更新である。$w_{k,m}$ の事前分布は尤度関数と共役ではないため、事後分布が十分統計量からは求まらない。しかも、数が $(K^+ + K^{\circ}) \times M$ 個と非常に多く、単純な Metropolis-Hastings アルゴリズムによる酔歩 [2] では現実的な速度で収束しない。

そこで、Hamiltonian Monte Carlo (HMC) [8] を用いる。HMC の利点は、勾配を用いた数値計算により、酔歩を避けて効率的に大きく値を動かせることである。HMC では、サンプリングしたい変数 $q \in \mathbb{R}^M$ を力学系の一般化座標とみなし、位置エネルギー関数 $U(q) = -\log P(q)$ を定義する。さらに補助変数として運動量変数 $p \in \mathbb{R}^M$ を導入し、運動関数 $K(p)$ を定義する。このとき、力学系 $H(q,p) = U(q) + K(p)$ が時間に対して一定であることを利用する。初期値、つまり時刻 $t = 0$ における q, p が与えられたとき、q, p の時間発展が定まる。適当な時刻 $t = \tau$ における q, p が求まり、この q を新たなサンプルとする。実際には離散的な数値計算が誤差を生じさせるため、最後に Metropolis-Hastings アルゴリズムにより補正する。

HMC を用いて、$w_{k,*}$ (M 個の $w_{k,m}$) をブロックサンプリングする。HMC の実装に必要なのは、$U(w_{k,*})$ とその勾配 $\nabla U(w_{k,*})$ である。$U(w_{k,*})$ は (1) 各 $w_{k,m}$ の t 分布からの生成確率の積と (2) 式 (1) の尤度の積という 2 種類の確率の積に対して負の対数をとったものである。いずれの対数確率も微分可能である。

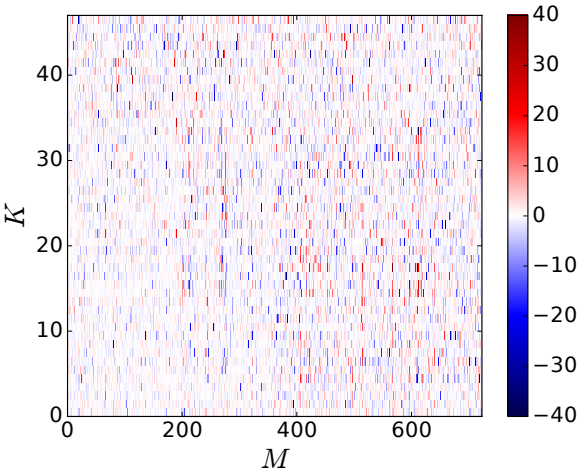

図 2: IBP-BP ($K_0 = 50$) において推定された W

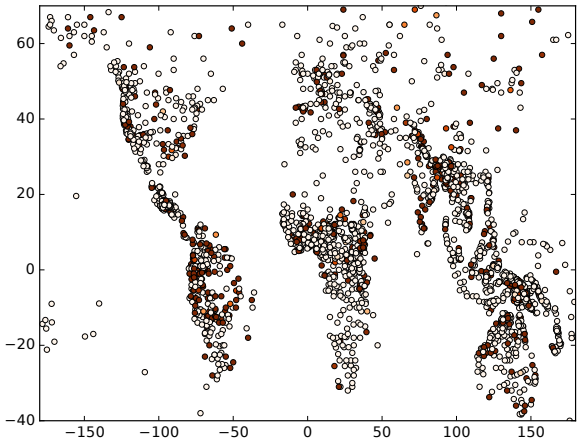

図 3: IBP-BP ($K_0 = 50$) におけるパラメータ $k = 16$ の地理的分布 (色が濃いほど $z_{l,k} = 1$ の割合が大きい)

4 結果と議論

まず、IBP-BP の定量的評価に欠損値補完を用いた。具体的には、既知の特徴の一部を隠して他の欠損値と同様に推定し、それらの値がどの程度正確に復元されるかを 10 分割交差確認により評価した。比較手法として、最頻値 (FREQ)、AE-CP が前処理として採用した多重対応分析に基づく欠損値補完 (MCR) [6] を用いた。IBP-BP は 500 反復ののち、言語 l ごとに $z_{l,k}$, $x_{l,i}$ のサンプリングを 100 反復行い、得られた 100 サンプルのなかから各欠損値 $x_{l,i}$ の最頻値を採用した。

Gibbs サンプリングは、理論上は無限回の反復により定常分布に収束する。しかし、現在の推論手続きではパラメータの混合が遅い。また、新しく追加されたパラメータが適切な重みを持つように育つことは稀で、ほとんどの場合、非有効パラメータとしてすぐに削除される。結果として、初期有効パラメータ数 K_0 ごとに異なる有効パラメータ数に収束する傾向が見られる。そこで、K_0 として $50, 100, 250$ を試した。

欠損値補完の結果を表 2 に示す。IBP-BP はすべての設定で FREQ を上回り、$K_0 = 50, 100$ で MCR を上回った。したがって、IBP-BP は表層的特徴列に潜む規則性を捉えることに成功していると判断できる。K_0 が小さいほど精度が高い傾向が見られる。K_0 が大きいほど $P(X|-)$ の尤度も大きくなるが、欠損値の多さから過学習が起きている可能性がある。

次に、すべての観測値を用いて IBP-BP の推論を行った。$K_0 = 50$ とし、1,000 反復を実行した。重み W の推定結果を図 2 に示す。ここでは、$K^+ = 47$ 個の有効パラメータを μ_k に従って降順に並べている (つまり、k が小さいほど $z_{l,k} = 1$ となる確率が高い)。同じ行で正負の値がモザイク状に分布しているのは、一般に同じ特徴の別の値は競合するからである。

図 3 にパラメータの地理的分布の例を示す。欠損値推定の場合と同様に、$z_{l,k}$, $x_{l,i}$ のサンプリングを 200 反復行い、得られた 200 サンプルから言語 l ごとに $z_{l,k}$ の頻度を計算したものである。こうした分布が持つ意味の解明は今後の課題である。

5 おわりに

本稿では、生成文法における原理とパラメータの枠組みに触発され、類型論的特徴から潜在的 2 値パラメータを獲得するための確率モデルを提案した。パラメータがインディアンビュッフェ過程により表現できること、非共役性ゆえに多数の潜在変数の推論が必要となるものの、Hamiltonian Monte Carlo により効率的にブロックサンプリングできることを示した。

今後は得られたパラメータの定性的、定量的分析を進めたい。定量的分析については、2 値パラメータには、言語や遺伝子の特徴列を分析する従来手法 (例えば [10]) がほぼそのまま利用できる。

謝辞 本研究は一部 JSPS 科研費 26730122 の助成を受けた。

参考文献

[1] Mark C. Baker. *The Atoms of Language: The Mind's Hidden Rules of Grammar*. Basic Books, 2002.
[2] Dilan Görür, Frank Jäkel, and Carl Edward Rasmussen. A choice model with infinitely many latent features. In *Proc. of ICML*, pp. 361–368, 2006.
[3] Joseph H. Greenberg, editor. *Universals of language*. MIT Press, 1963.
[4] Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. *The World Atlas of Language Structures*. Oxford University Press, 2005.
[5] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800, 2002.
[6] Julie Josse, Marie Chavent, Benot Liquet, and François Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, Vol. 29, No. 1, pp. 91–116, 2012.
[7] Yugo Murawaki. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proc. of NAACL-HLT*, pp. 324–334, 2015.
[8] Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pp. 113–162. CRC Press, 2011.
[9] Yee Whye Teh, Dilan Görür, and Zoubin Ghahramani. Stick-breaking construction for the Indian buffet process. In *AISTATS*, pp. 556–563, 2007.
[10] Kenji Yamauchi and Yugo Murawaki. Contrasting vertical and horizontal transmission of typological features. In *Proc. of COLING*, pp. 836–846, 2016.