講座

最近のベイズ理論の進展と応用 (III) ノンパラメトリックベイズ

持橋　大地[†a)]

An Introduction to Nonparametric Bayesian Models

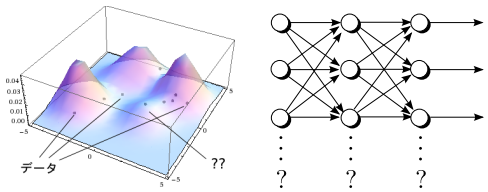
Daichi MOCHIHASHI[†a)]

1. はじめに

最近,「ノンパラメトリックベイズ」という統計モデルが (一部で) 流行っていると聞いたことのある方がいるかもしれない．または,「ディリクレ過程」などという言葉を耳にして，??と頭をひねった方もおられるのではないだろうか．

ノンパラメトリックベイズ法とはベイズ統計の新しい統計モデルであり，簡単に言うと「データに応じてモデル自体の複雑さも自動的に学習する」ことのできる統計モデルである．確率論での起源は 1973 年の論文 [1] に遡るが，実際の計算法が確立されて，統計学・統計的機械学習の分野で有望なモデルとなったのは 1990 年代後半から 2000 年代に入ってからである.「ノンパラメトリック」という言葉はパラメータがないことを意味するのではなく，単一の正規分布のような少数のパラメータで記述せずに，データにフィットする，柔軟な無限次元の離散分布を考えることを意味する．

われわれの扱う問題の中で，モデル自体の複雑さも学習したい, という場面は非常に多いと思われる．たとえば，音声処理では隠れマルコフモデル (Hidden Markov Model, HMM) やガウス混合モデル (Gaussian Mixture Model, GMM) が頻繁に使われるが，HMM の状態数や GMM の混合数はどうやって決めたらいいのだろうか (図 1)? また, 自然言語処理では「動詞」「助動詞」「人名」「地名」…のような，単語のもつ文法的・意味的カテゴリをテキストから自動的に



(a) GMM の混合数.　(b) HMM の状態数.

図 1　モデル選択問題.

導出することが大きな研究課題となっているが，このカテゴリの数は一体いくつにすればいいのだろうか．カテゴリが少なすぎれば言語の粗い記述しかできず，多すぎれば極端な場合，一単語一カテゴリとなって，カテゴリ化が無意味となってしまう．

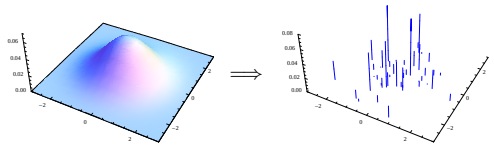
こうした問題は画像処理やデータマイニングなどを含め，多くの分野に共通する問題だと思われる．従来こうしたモデル選択には AIC や MDL が使われてきたが, ノンパラメトリックベイズ法はこれをパラメータ数だけに依存するのではなく(注1)，各モデルに適した, 非常に洗練された形でデータから学習することができる．これは，ディリクレ過程という魔法によって可能となる．

2. ディリクレ過程とは

ディリクレ過程 (Dirichlet process, DP) は，ノンパラメトリックベイズ法の最も基本となるモデルである．測度論的な定義は省略するが, これは図 2 のように，基底測度と呼ばれるある確率分布 G_0 を，それに

† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
a) E-mail: daichi@cslab.kecl.ntt.co.jp

(注1)：AIC や MDL の適用されるモデルの多くは特異モデルであり，この場合は汎化誤差最小という意味で適切でない，という理論的な指摘もある [2].



基底測度 G_0.　　$G \sim \mathrm{DP}(\alpha, G_0)$.

図 2　ディリクレ過程による無限離散分布 G の生成．見えない場所にも，指数的に小さい無限個の棒が立っている．

似た無限次元の離散分布によって「すかすか」にした分布 G を生成する確率過程であり，次のように書かれる．

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{1}$$

$\alpha > 0$ は G が平均的にどれくらい G_0 と似ているかを制御する，学習可能なパラメータである．G_0 が 1 次元上の連続分布の場合は G は無限次元の多項分布となり，ディリクレ過程は有限次元の多項分布を生成するディリクレ分布 (用語) の無限次元版といってよい．

2.1　Stick-Breaking Process

具体的には，G は次のような **Stick-breaking process** (SBP) とよばれる方法によってサンプルすることができる．図 3 にその様子を示した．

SBP では，確率の総和である長さ 1 の棒を左から切っていくことで G を生成する．まず，$[0,1]$ 上の確率分布であるベータ分布 $\mathrm{Be}(1,\alpha)$ からのサンプル $v_1 \sim \mathrm{Be}(1,\alpha)$ で棒を分割し，左側を $\pi_1 = v_1$ とする．次に，残った長さ $(1-v_1)$ の棒を $v_2 \sim \mathrm{Be}(1,\alpha)$ でまた分割し，左側を $\pi_2 = v_2(1-v_1)$ とする．次に，残った長さ $(1-v_2)(1-v_1)$ の棒を $v_3 \sim \mathrm{Be}(1,\alpha)$ で分割して…のように無限次元の多項分布 $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_\infty)$ を作り，高さ π_k のデルタ関数 $\delta(\theta_k)$ を G_0 からサンプルした場所 $\theta_k \sim G_0$ に立てていったものを G とする．式で書くと，次のようになる．

$$\begin{cases} v_k \sim \mathrm{Be}(1,\alpha) \\ \pi_k = v_k \prod_{i=1}^{k-1}(1-v_i) \quad (k = 1, \cdots, \infty) \\ \theta_k \sim G_0\,, \end{cases} \tag{2}$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)\,. \tag{3}$$

図 3(a) の $\boldsymbol{\pi}, \boldsymbol{\pi}', \boldsymbol{\pi}''$ からわかるように，こうして生成された G は乱数 $v_1, v_2, \cdots$ の値によって長い裾を持つことも，一部の要素に確率が集中することもあるが，その期待値は $E[G] = G_0$ に等しい．

2.2　ディリクレ過程混合モデル

この抽象的な確率過程が，なぜデータのモデル化に役立つのだろうか? 次式のような，データ x を生成するガウス混合モデルを考えてみよう．

$$p(x|\boldsymbol{\pi}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \pi_k \mathrm{N}(x|\mu_k, \sigma^2) \tag{4}$$

これは図 1(a) のように，K 個のガウス分布 $\mathrm{N}(\cdot|\mu_k, \sigma^2)$ を混合比 $\boldsymbol{\pi}$ で混ぜ合わせたモデルである．簡単のため，ガウス分布の分散は σ^2 ですべて等しいとしよう．本講座第 I 回の階層ベイズモデルの考え方を用いて，ガウス分布の中心 μ_k は事前分布

$$p(\mu) = \mathrm{N}(\mu_0, \sigma_0^2) \tag{5}$$

から，混合比 $\boldsymbol{\pi}$ は K 次元のディリクレ分布 [3]

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\alpha_1, \cdots, \alpha_K) \tag{6}$$

$$\propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{7}$$

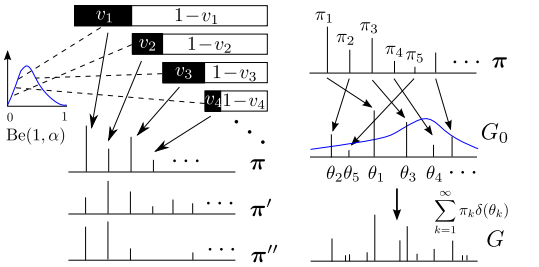
から生成されたとすれば，(4) 式の完全な生成モデルが得られるが，これは混合数が必ず K であるという制約がつきまとう．

ここで少し見方を変えて，無限個の $\mu_1, \mu_2, \ldots$ とその混合比 $\pi_1, \pi_2, \ldots$ を同時に生成することを考えてみよう．前節の議論を踏まえると，これは $p(\mu)$ を基底測度 G_0 とみて，ディリクレ過程によって離散化して

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\mu_k) \sim \mathrm{DP}(\alpha, p(\mu)) \tag{8}$$

とすれば可能なことがわかる (図 4(a))．(3) 式で離散化した位置 θ_k が，ここでは正規分布の中心 μ_k に対応している．こうして混合ガウス分布の中心と混合比が求まれば，分散は σ^2 で等しいとしたので(注2)，

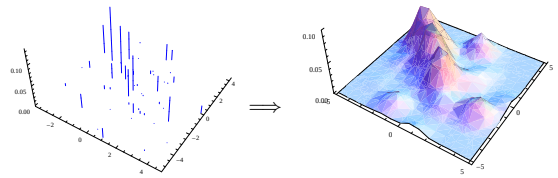
$$p(x|\alpha, p(\mu)) = \sum_{k=1}^{\infty} \pi_k \mathrm{N}(\mu_k, \sigma^2) \tag{9}$$



(a) 無限次元多項分布 $\boldsymbol{\pi}$ の生成.　(b) $\boldsymbol{\pi}$ と G_0 からの G の生成.

図 3　Stick-breaking process による G の生成.

(注2)：分散も同時に生成する場合は，$\mu \times \sigma$ の直積空間からのディリクレ過程によるサンプルを考える．



(a) $G=(\boldsymbol{\mu},\boldsymbol{\pi})\sim \mathrm{DP}(\alpha,p(\mu))$. (b) 無限ガウス混合モデル.

図 4 ディリクレ過程混合モデル (DPM) の構成.

という無限混合モデル, **ディリクレ過程混合モデル** (Dirichlet Process Mixtures, DPM) を得ることができる (図 4(b)).

このモデルの推定には SBP を直接用いて近似を行うことも可能であるが, 紙面の制約から次回の変分ベイズ法での解説に譲り, ここでは Chinese Restaurant Process とよばれる, 等価な, より簡単な推定法について見ていくことにする.

なお, ここでは DP を混合モデルの事前分布として間接的に用いたが, 自然言語処理などでは可能性として無限に存在する, 単語やカテゴリの確率分布のモデルとして直接用いることもできる. 詳しくは, [4] の記事を読まれたい.

2.3 Chinese Restaurant Process

ディリクレ過程混合モデルでは, 可能性として無限個ある混合要素 (例えば, ガウス分布) のどれかから一つ一つのデータが生成され, データ全体がクラスタリングされる. 隠れた分布 G を積分消去すると, このクラスタリングは, 次の **Chinese Restaurant Process** (CRP) と呼ばれる方法によって順番に生成されたと考えても等価なことが知られている. (注3)

CRP では, データ x_n の属するクラスタ z_n の事前確率は, これまでのデータ $x_1\cdots x_{n-1}$ に依存し,

$$p(z_n{=}k|z_1^{n-1}) = \begin{cases} \dfrac{n_k}{\alpha{+}n{-}1} & (k=1,\cdots,K) \\ \dfrac{\alpha}{\alpha{+}n{-}1} & (k=K+1) \end{cases} \tag{10}$$

で与えられる. ここで, n_k は $z_1^{n-1}=z_1\cdots z_{n-1}$ の中で k が現れた回数, K は現在までのクラスタ数を表す. また, この確率は $x_1\cdots x_{n-1}$ の順番によらない (交換可能性). (10) 式は, どのクラスタが選ばれるかの事前確率は

- そのクラスタの「人気度」 n_k に比例し, かつ
- α に比例する確率で新しいクラスタが生成

(注3) : 証明には, ディリクレ過程の測度論的な定義と, ディリクレ分布の簡単な性質を用いる [5].

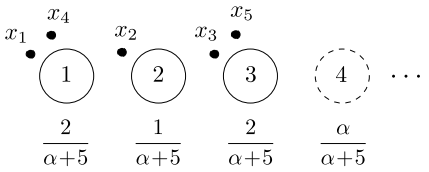


図 5 CRP に基づく x のクラスタの選択.

されることもあることを意味している.

いま, 丸テーブルが無限にある中華料理店に「客」$x_1\cdots x_n$ が順番に入り, テーブルに分かれて着席する状況を考えてみよう (図 5). 入った客 x は (10) 式に従い, 各テーブルの人数 n_k に比例した確率で座るテーブルを選び, α に比例した確率で誰もいない新しいテーブルに着席する. 最初は必ずテーブル 1 が選ばれるが, しだいに使われるテーブル数 K が増えていくことになる. このメタファーが, CRP という名前の語源になっている.

上で説明した CRP は x をこれから生成するモデルであり, 事前確率だけを与えていることに注意しよう. 逆に, x_n だけが与えられてその属するクラスタ z_n が未知のとき, その事後確率はベイズの定理から,

$$\begin{aligned} &p(z_n=k|x_n,z_1\cdots z_{n-1}) \\ &\propto p(x_n|z_n)\,p(z_n|z_1\cdots z_{n-1}) \end{aligned} \tag{11}$$

となり, この式の第 2 項は (10) 式の事前確率そのものである. 第 1 項はクラスタ k から x_n が生成される尤度で, これは k に属する他のデータで決まり, 結局 (11) 式は,

$$p(z_n=k|x_1\cdots x_{n-1},z_1\cdots z_{n-1}) \propto \begin{cases} p(x_n|k)\cdot\dfrac{n_k}{\alpha{+}n{-}1} & (k=1\cdots K) \\ p(x_n|k^{new})\cdot\dfrac{\alpha}{\alpha{+}n{-}1} & (k=K{+}1) \end{cases} \tag{12}$$

で求めることができる.

2.4 CRP に基づく DPM の学習

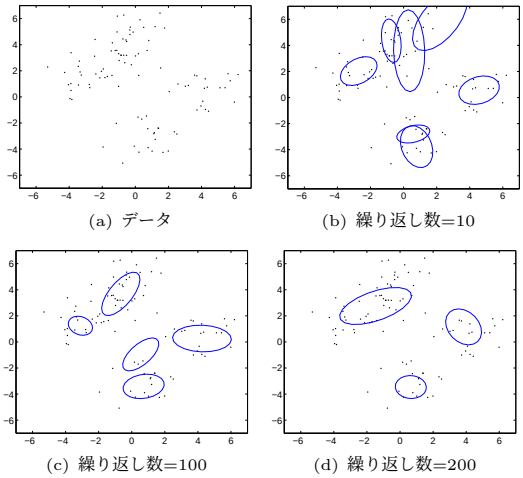
(12) 式を用いると, データ $X=x_1\cdots x_N$ が与えられたとき, その属するクラスタ番号 $Z=z_1\cdots z_N$ $(z_n\in 1\cdots N)$ をギブスサンプリングによって求めることができる. 本講座第 I 回でも登場した MCMC 法の最も簡単な場合であるギブスサンプリングとは, 隠れ変数 z_n を適当な初期値から始めて, それ以外を条件とした条件つき確率

$$z_n\sim p(z_n|X,Z_{-n}) \tag{13}$$

からサンプリングすることをランダムな順番ですべての n について繰り返せば, 真の分布 $p(Z|X)$ に従うサ

1: **while** not converged **do**
2: **for** n in randperm $(1, \cdots, N)$ **do**
3: x_n をクラスタ z_n から削除してパラメータを更新
4: $z_n \sim p(z_n|X, Z_{-i})$ をサンプル
5: x_n をクラスタ z_n に追加してパラメータを更新
6: **end for**
7: **end while**
8: $z_1, \cdots, z_N$ を出力

図 6　ディリクレ過程混合モデルのギブスサンプリング. randperm (X) は X のランダムな並び換えを表す.



(a) データ　(b) 繰り返し数=10　(c) 繰り返し数=100　(d) 繰り返し数=200

図 7　ディリクレ過程混合モデルの MCMC による学習.

ンプル Z が得られる, という方法のことである [6]. ここで, Z_{-n} は z_n を除いた Z を表している.

今の場合, これは「今の値 z_n をいったん忘れて」新しい z_n を (11) 式, すなわち (12) 式からサンプリングしていけば, 真の z_n が求まることを意味する. これは, (10) 式は X の順番によらないので, x_n をいつでも最後のデータとみなすことができる (交換可能性) からである.

アルゴリズムは図 6 のようになる. ハイパーパラメータ α も学習することができるが, やや複雑となるため割愛した. 図 7 に, 実際のデータに基づく DPM の学習過程を示した.(注4) (a) のようなデータに対し, MCMC 法によるサンプリングを繰り返すと, (b)–(c) のような中間状態を経て, (d) のようなクラスタリングに収束する. この場合は最終的に $K = 3$ となったが, 学習途中でクラスタ数 K が変化していく様子を図 8 に示す.

(注4)：この図の作成には, ディリクレ過程混合モデルの MATLAB ツールキット [7] を用いた. これは自由にダウンロードでき, 今回のような例を自分で試すことができる.

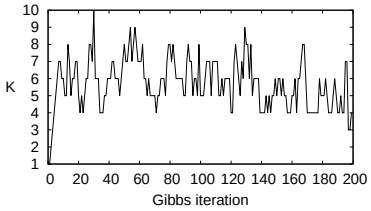


図 8　混合数 K の学習過程での変化.

3. ノンパラメトリックベイズ法の応用

ノンパラメトリックベイズ法は画像処理, バイオインフォマティクス, データマイニングなど様々な分野に適用されているが [8], その離散性から, 特にテキストデータに対する自然言語処理での発展が目覚ましい. ここではその中から, 音声など他の分野にも共通する基礎的なモデルとして, **無限隠れマルコフモデル** (Infinite HMM, 以下 IHMM) を考えてみよう.

HMM は図 9 に表したように, 観測値 $y_1 y_2 \cdots y_T$ の裏に隠れた状態系列 $s_1 s_2 \cdots s_T$ があり, そこから観測値が生成されたと考えるモデルであるが, 隠れ状態の総数は事前に決めておく必要があった. IHMM は, この隠れ状態の総数も自動的に決めることができる (画期的な) モデルである [9].

まず, 通常の HMM は本質的に, 混合モデルの時系列的な拡張であることに注意しよう. 時刻 t までの観測値 $y_1 \cdots y_t$ と状態 $s_1 \cdots s_t$ が与えられたとき, 次の観測値 y_{t+1} の確率は, s_{t+1} を隠れ状態とした $p(y_{t+1}|s_t) = \sum_{s_{t+1}} p(y_{t+1}|s_{t+1}) p(s_{t+1}|s_t)$ のような混合モデルであり, HMM ではこれが $t = 1 \cdots T$ まで時間的に連鎖している.

IHMM では, この状態遷移確率分布 $p(s_{t+1}|s_t)$ がディリクレ過程 $\mathrm{DP}(\alpha, G_0)$ に従っていると考える (図 10). ただし, 混合モデル間で状態を共有する必要があるため, 基底測度 G_0 自体がもう一つのディリクレ過程からのサンプル $G_0 \sim \mathrm{DP}(\gamma, H)$ である, とする. 直感的には, 「可能な遷移先とその事前確率を各状態間で共有する」ことだといってよい. ディリクレ過程が階層化されているため, これは階層ディリクレ過程

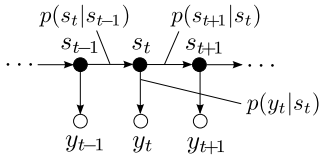


図 9　隠れマルコフモデルの構造. ○は観測値を表す.

(HDP) と呼ばれている [10].

HDP を用いた IHMM では, 同様に MCMC 法によって隠れ状態を推定する. 図 9 からわかるように, 隠れ状態 s_t の確率は, s_t を含む 3 つの確率の積

$$p(s_t = k|y_t) \propto p(y_t|s_t=k) \cdot p(s_t=k|s_{t-1}) \cdot p(s_{t+1}|s_t=k, s_{t-1})$$

に比例するから, (12) と同様にこれから $k = 1 \cdots K$ および $K+1$ の場合について確率を計算し, s_t の値と状態数 K を更新していく. 学習には階層化された CRP を用いるが, 手順がやや複雑になるため, 詳細については [10] を参照されたい.

“Alice” データの学習　図 11 に, “*Alice in Wonderland*” の最初の 20,000 語を使って学習した IHMM の状態数 K と, データの対数尤度の変化を示す. 対数尤度がほぼ一定となったところで, 状態数も $K = 7 \sim 8$ と学習されて安定していることがわかる. 表 1 に, 各状態に割り当てられた単語とその回数の上位を示す. かなり小さいデータだが, ほぼ状態 1 に主語, 状態 2 に修飾語, 状態 3 に動詞・助動詞, 状態 4 に前置詞, 状態 5 に名詞, 状態 6 に形容詞, という概念が学習されていることがわかる.

また, 上の状態の順番は任意ではなく, ディリクレ過程に従って若い順番ほど出やすい, 重要な状態であることに注意されたい. 隠れ状態数の学習も含め, こうした性質は通常の HMM の EM アルゴリズムによる最尤推定では得られず, 局所解に陥る心配もない.

4. まとめと今後の展望

本稿では, ノンパラメトリックベイズ法の基礎と実際の例について紹介した. 無限にモデルを伸縮できる, 柔軟な事前分布を用いることで, データの確率を最大化するだけで適切なモデルが完全に自動的に学習される. ノンパラメトリックベイズ法はディリクレ過程に限られるものではなく, ベータ過程や Pitman-Yor 過程, およびその階層化のようなより高度なモデル [8] も実際に用いられている. 今後は, 木やグラフといった, 複雑な組み合わせ論的対象を無限に生成する確率過程が理論的に重要になると考えられる.

また, 大量のデータを用いた実際の学習には高速な近似解法が不可欠であり, 次回の変分ベイズ (VB) 法はそのための有用な方法である. 一方ベイズ学習の並列化についても, 最近研究が進められている.

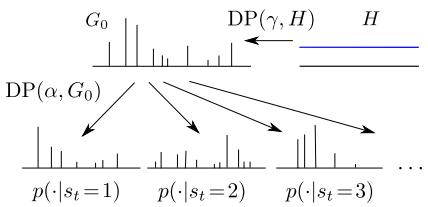
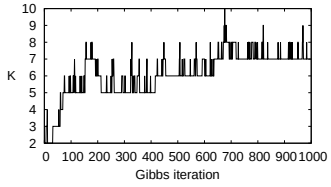
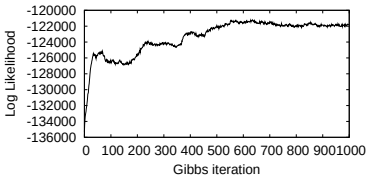


図 10　HDP による隠れ状態と状態遷移確率の生成.



(a) 隠れ状態数 K の学習



(b) データの対数尤度の変化

図 11　”Alice” データの IHMM による学習過程.

1		**2**		**3**	
she	432	the	1026	was	277
to	387	a	473	had	126
i	324	her	116	said	113
it	265	very	84	EOS	87
you	218	its	50	be	77
alice	166	my	46	is	73
and	147	no	44	went	58
they	76	his	44	were	56
there	61	this	39	see	52
he	55	EOS	39	could	52
that	39	an	37	know	50
who	37	your	36	thought	44
what	27	as	31	herself	42
i’ll	26	that	27	began	40
this	23	at	27	get	39

4		**5**		**6**	
EOS	845	way	45	little	92
and	466	mouse	41	great	23
of	343	thing	39	very	22
in	262	queen	37	long	22
said	174	head	36	large	22
to	163	cat	35	right	20
as	163	hatter	34	same	17
that	125	duchess	34	good	17
for	123	well	31	white	11
at	122	time	31	other	11
but	121	tone	28	poor	10
with	114	rabbit	28	first	10
on	83	door	28	best	9
so	77	march	26	own	8
when	63	dormouse	26	low	8

表 1　IHMM の各状態に割り当てられた単語とその回数. EOS は文末を示す特殊記号である.

この分野は急速に発展しているため, 現在のところまとまった教科書は存在していないが, 本稿が, 興味を持たれた方が [8] のようなサイトの参考文献でさらに学ばれる際のガイドとなればと願っている.

文　献

[1] T.S. Ferguson, “A Bayesian Analysis of Some Nonparametric Problems,” Annals of Statistics, vol.1, no.2, pp.209–230, 1973.

[2] 渡辺澄夫, “Model Selection in Singular Learning Machines”. http://watanabe-www.pi.titech.ac.jp/~swatanab/model-select.html.

[3] C.M. ビショップ, 元田, 栗田, 樋口, 松本, 村田 (監訳), パターン認識と機械学習 (上)(下) ベイズ理論による統計的予測, Springer, 2007,2008.

[4] 持橋大地, “生きた言葉をモデル化する—自然言語処理と数学の接点,” 『数学セミナー』2007 年 11 月号, pp.37–43, 2007.

[5] D. Blackwell and J.B. MacQueen, “Ferguson Distributions via Pólya Urn Schemes,” Annals of Statistics, vol.1, no.2, pp.353–355, 1973.

[6] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, Markov Chain Monte Carlo in Practice, Chapman & Hall / CRC, 1996.

[7] Y.W. Teh, “Bayesian Nonparametrics: DP Mixtures,” 2009. http://www.gatsby.ucl.ac.uk/~ywteh/teaching/npbayes/.

[8] NPBayes 2008. http://npbayes.wikidot.com/.

[9] M.J. Beal, Z. Ghahramani, and C.E. Rasmussen, “The Infinite Hidden Markov Model,” NIPS 2001. http://books.nips.cc/papers/files/nips14/AA01.pdf.

[10] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical Dirichlet Processes,” JASA, vol.101, no.476, pp.1566–1581, 2006.

（平成 xx 年 xx 月 xx 日受付）

持橋　大地

1998 東大・教養・基礎二卒. 2005 奈良先端大・情報・博士後期課程修了. ATR 音声言語コミュニケーション研究所を経て, 2007 より NTT コミュニケーション科学基礎研究所リサーチアソシエイト. 自然言語処理の研究に従事. 博士 (理学).