

# ノンパラメトリックベイズによるメディア処理

Media processing via Bayesian nonparametrics

中野允裕

Masahiro Nakano

NTT コミュニケーション科学基礎研究所

NTT communication science laboratories

nakano.masahiro@lab.ntt.co.jp

## Abstract

本稿では音、画像、動画、自然言語など多くのメディア処理における近年のベイズ（特にノンパラメトリックベイズ）的な手法の発展を分野横断的に紹介する。混合モデル、隠れマルコフモデル、確率文脈自由文法、 $n$ -gram、非負値行列因子分解、独立成分分析など多くの確率モデルに対するベイズ的な取扱いと、それらの無限モデルを構成する際に登場するディリクレ過程やレヴィ過程の特別な場合について紹介する。さらにモデルの拡張するための基本的な方法として階層化、相関の導入、入れ子構造の利用、木構造化などを示し、それらが各種メディア処理のどのような場面で活用されているかを紹介する。

## 1 はじめに

音、画像、動画、自然言語などのメディア処理全般において、近年確率的な生成モデルを用いて課題を解決しようとする研究が少なくない。特にここ最近 10 年ほどはノンパラメトリックベイズモデルの普及、推論アルゴリズムの発達、計算機能力の向上などと相まって、メディア処理における各種課題への有力な選択肢の一つとなってきた。

ベイズ的な手法を用いたメディア処理の基本的な戦略は、対象の確率的な生成モデルを描き、それを実際の観測データにフィッティングさせ、得られたモデルを介して課題を解くことにある。対象に対していかに適切なモデルを設計するかが重要になってくるため、対象とするメディアや扱う課題ごとに全く別の方針を考えなければならないように思われるが、実際のモデルの設計には共通した部分が非常に多く、また、あるメディアにおける標準的な戦略を別のメディアに輸入し成功した例も多い。本稿では各種メディアに対するベイズ的な手法を分野横断的に整理することでそれらの発展を概観したい (Fig. 1)。

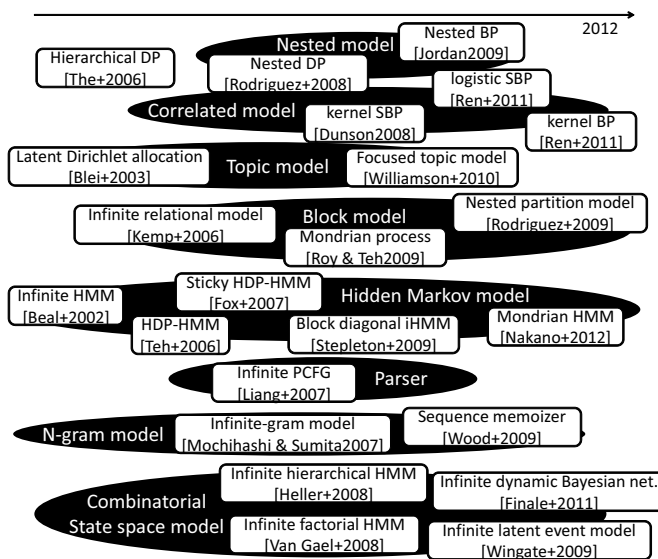


Figure 1: 本稿で扱うベイズモデルの一部。メディア処理におけるベイズモデルとして重要なものを全てを網羅することは出来ないが、その一部を対象とするメディアに関して分野横断的に概観する。

## 2 確率的生成モデルを用いたメディア処理

### 2.1 モデルの設計と推論の基本方針

ベイズ的な戦略の基本は、観測データを説明する確率的な生成モデルを設計することにある。観測データ（例えば 1 次元の実数） $x$  が平均  $\mu$ 、分散  $\sigma^2$  の正規分布  $\text{Normal}(\mu, \sigma^2)$  から生成されたとみるのが妥当であれば、 $x \sim \text{Normal}(\mu, \sigma^2)$  のように観測データの生成過程を記述する。もし 100 個の観測データ  $x_1, x_2, \dots, x_{100}$  がある一つの正規分布から独立に生成されたと見なすことが出来るのであれば  $x_n \sim \text{Normal}(\mu, \sigma^2)$  ( $n = 1, 2, \dots, 100$ ) のように表せばよい。ここで、正規分布の平均は 0 付近にあるはずだと分かっていたら、 $\mu$  を確率変数と見なし、 $\mu \sim \text{Normal}(0, \tau^2)$  のように  $\mu$  の生成過程を設計すればよい。

モデルが設計出来た後、我々の興味は観測データが与えられたときにそれらのパラメータがどうなっているか、またはパラメータの分布がどうなっているかにある。先の例であれば、 $x_1, x_2, \dots, x_{100}$  が観測されたときのパラメータ  $\mu, \sigma$  の分布  $p(\mu, \sigma | x_1, x_2, \dots, x_{100})$  が我々の興味である。このようなパラメータの分布を知るための推論には大きく二つの方針が用いられることが多く、一つはマルコフ連鎖モンテカルロ法と呼ばれるもので、もう一つは変分ベイズ法と呼ばれるものである。それぞれがどのようなものであるかについては膨大な教科書があるのでそれらを参照して頂きたい。

## 2.2 モデルの複雑度の設定

対象の確率的な生成モデルを設計する上で、データを説明するパラメータの個数をどの程度にすべきかは重要な問題である。パラメータを増やせば観測データをよく説明出来るようになるが、一方で過学習を起こし未知のデータを説明する能力に乏しくなる恐れがある。パラメータが少なすぎれば観測データを説明するための能力が足りなくなる恐れがある。このように、データを説明する際のパラメータ数（モデルの複雑度）をどのように設定すべきかは極めて重要な問題で、汎用の設計指針が求められてきた。

モデルの複雑度の設定に対する近年の標準的な取扱い方がノンパラメトリックベイズと呼ばれる枠組みで、これは無限のパラメータを用いた生成モデルを描こうとする考え方である。そもそもデータを説明するだけのパラメータの数などというものはそれ自身不確かなもので、本来観測データに説明させるべきものである。そこで、十分なパラメータを用意しておき、観測データにフィッティングさせた際に必要な分だけ説明に寄与するようなモデルの設計の仕方が考えられてきた。このとき、観測データは有限なため、無限のパラメータを使った生成モデルを考えることが出来ればデータの説明に寄与する分は観測データそれ自身の要請に従って必要な分だけ表出してくるように働くはずである。では無限のパラメータを使った生成モデルを描くにはどうすれば良いだろうか。

## 2.3 モデル化のための確率変数の設計

ノンパラメトリックベイズの基本的な戦略は無限のパラメータを用いた確率的な生成モデルを描くことにあったが、ではどのように無限のパラメータを取り扱うことが出来るだろうか。直観的には無限のパラメータを作った際にそれらの出現に対する確率が与えられているようなモデルと捉えることが出来る。つまり確率分布  $f(\theta)$  から確率変数  $x$  が生成ことを表す  $x \sim f(\theta)$  において、 $x$  として例えば無限実数列のようなものが作ればよい訳である。ノンパラメトリックベイズモデルは多くの場合確率過程を用いてこのような無限のパラメータから成るモデルを構成する。このことを以降の議論で直感的に了解するために、

確率空間、確率変数、分布とはどんなものであったかを簡単に整理し、先の無限数列のような色々なものが確率変数として扱うことが出来ることを確認しておく。

集合  $\Omega$  の要素  $\omega$  を標本点と呼び、 $\sigma$  加法族  $\mathcal{B}$ 、すなわち、1)  $\phi \in \mathcal{B}$ , 2)  $A_n \in \mathcal{B} (n = 1, 2, \dots) \Rightarrow \cup_n A_n \in \mathcal{B}$ , 3)  $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$ , によって決まる  $(\Omega, \mathcal{B})$  を可測空間と呼ぶ。 $\mathcal{B}$  を定義域とする関数  $P$  が、1)  $A \in \mathcal{B} \Rightarrow 0 \leq P(A) \leq 1$ , 2)  $P(\omega) = 1$ , 3)  $\{A_1, \dots, A_N\}$  が排反ならば  $P(\cup_n A_n) = \sum_n P(A_n)$ , のとき確率測度と呼び、 $P(A)$  を事象  $A$  の確率、 $(\Omega, \mathcal{B}, P)$  を確率空間と呼ぶ。今  $\Omega$  上に定義され実数値をとる関数  $X(\omega)$  で  $B(r) = \{\omega; X(\omega) \geq r\}$  に対して  $P(B(r))$  が定まるものを考えたとき、これを可測関数といい確率変数とみなすことが出来る。つまり確率変数とはとる値に対して必ずその確率を計算できるものであると捉えればよい。直観的な把握には色々な例があると思うが、例えばサイコロのようなものを連想しそのサイコロが相当に柔軟なものと考えてしまうのが簡単である。サイコロは偶然現象を表しており、サイコロを振ったときに一つの面が出るのが標本点  $\omega$  に対応している。偶然現象は「あの面  $\omega_i$  が出る」「この面  $\omega_i$  もしくはあの面  $\omega_j$  が出る」などの事象として表すことができ、これらの集合が  $\sigma$  加法族  $\mathcal{B}$  に対応している。各事象はサイコロの歪み方など  $P$  によって確率を計算することが出来る。いまサイコロの各面にはある実数が割り当てられていて、サイコロを振ってその目が出たとき、そこに描かれた実数を吐き出されるように見なすことで確率変数と捉えることが出来る。

確率変数としての可測関数は必ずしも実数値をとる必要はなく  $X(\omega)$  として多次元の実数や、無限次元の数列、関数（このような時に確率過程と呼ばれる）を考えることも出来る。確率変数として確率過程のようなもの考える場合においても、あくまでサイコロのような偶然現象を表すものがあり、出た面に割り当てられた出力が確率変数だという捉えることが直観的な理解を助けてくれる。

前述の通りノンパラメトリックベイズモデルにおいては無限のパラメータを扱う性質から、確率過程を導入することが多い。具体的なモデル化において登場するディリクレ過程、レヴィ過程（の一部であるベータ過程やガンマ過程）、などのそれぞれについては各論で触れていきたいが、それらと密接な関係があり以降の議論でも登場するポアソン過程を例としてここで簡単に紹介しておく [13]。

いま時間  $t$  にもなって変化する偶然現象  $X(t, \omega)$  を考えることにする。偶然を司る  $\omega$  が一つに決まれば（サイコロをようなものを振ったとして  $\omega$  が決まれば） $\omega$  を省略して  $X(t)$ 、すなわち時間  $t$  に伴って変化する関数が生成されたと捉えることが出来る。パラメータ  $\lambda > 0$  を持つポアソン過程  $X(t)$  とは次の性質を持つものを指す：

- $X(t)$  は非負整数値をとり、 $X(0)=0$ .
- $0 \leq t_1 \leq t_2 \rightarrow X(t_1) \leq X(t_2)$ .

- $0 \leq t_1 < t_2 < \dots < t_N$  について,  $X(t_1) - X(0), X(t_2) - X(t_1), \dots, X(t_N) - X(t_{N-1})$  は独立.
- $0 \leq t_1 \leq t_2, h > 0$  のとき,  $X(t_2) - X(t_1)$  と  $X(t_2 + h) - X(t_1 + h)$  の分布が同一.
- $P(X(h) = 1) = \lambda h + o(h), P(X(h) \geq 2) = o(h)$  ( $h \downarrow 0$ ). ただし,  $o(h)$  は  $P(X(h) \geq 2)/h \rightarrow 0$  ( $h \downarrow 0$ ).

$X(t)$  は時間  $t$  に伴って発生する何らかのイベントの発生回数を数えているもので, 発生しやすい程度が時間の経過に対して一律で  $\lambda$  に支配されていると考えたと直感的に理解しやすい. ポアソン過程には様々な良い性質があるがその一部だけ紹介すると:

- 時刻  $t_1$  から時刻  $t_2$  までに発生するイベントの回数は  $\text{Poisson}(\lambda(t_2 - t_1))$  に従う.
- あるイベントから次のイベントの発生までの時間は  $\text{Exp}(\lambda)$  に従う.
- $(0, t]$  において一回だけイベントが発生したとすれば, その発生時刻は  $\text{Uniform}(0, t)$  に従う.

ポアソン過程を実際に構成したい場合 (サイコロを振って一つの具体的なポアソン過程  $X(t)$  を生成したい場合) には上記の 2 番目の性質を使って, 指数分布に従うイベント発生時間間隔を次々に生成していけばよい. またイベント発生を表すパラメータ  $\lambda$  は時間の経過に対して一律である必要はなく, 時刻に依存するように  $\lambda(t)$  とすることも出来る. 例えば上記一つ目の性質は次のようになる:

- 時刻  $t_1$  から時刻  $t_2$  までに発生するイベントの回数は  $\text{Poisson}(\lambda \int_{t_1}^{t_2} \lambda(t))$  に従う.

1 次元のポアソン過程の説明として, 簡単のため  $t$  を時刻のように捉えてきたが, これは必要に応じて所望の対象を置き換えてよく, また多次元として考えることも出来る. ノンパラメトリックベイズモデルに頻出のディリクレ過程, ベータ過程, ガンマ過程, ベルヌーイ過程などはこのポアソン過程と密接な関係にある.

### 3 混合モデルと因子モデル

多くの確率的生成モデルは混合モデル (隠れマルコフモデル, 確率文脈自由文法, n-gram) と因子モデル (非負値行列因子分解, 独立成分分析) の二つに大別して捉えることが出来る. まずこの 2 つのモデルの基本的な考え方とそれぞれの無限モデル化について整理する.

#### 3.1 混合モデル

データの分類やクラスタリングは様々なメディアに現れる頻出の問題である. 例えば画像に「花」や「人」のラベルを付ける問題や, 音楽において各楽曲を「ジャズ」, 「ポップス」のように分類する問題がこの典型である. ベイズ的な方法によってこのような問題を扱う場合, 混合モデルという考え方が重要な役割を果たす.

$N$  個の観測データ  $y_1, y_2, \dots, y_N$  それぞれに対して,  $K$  種類のラベルの中の一つを割り当てることを考える. いま  $n$  番目のデータ  $y_n$  に割り当てられるラベルのインデックス ( $1, 2, \dots, K$  の中の一つ) を  $z_n$  と表すことにする. 典型的な分類やクラスタリングは  $y_1, y_2, \dots, y_N$  と分類したクラスタの数  $K$  が与えられたときに  $z_1, z_2, \dots, z_N$  を推定する問題だと捉えることが出来る ( $K$  が未知の場合の取扱い方は後述する). ベイズ的なアプローチにおける基本的な戦略は, 観測データの確率的な生成過程を描くことにある. 混合モデルは次のような生成過程によって観測データが確率的に生成されたと考えるものである.

1. 各ラベル  $k$  ごとにデータ生成用のパラメータ  $\theta_k$  をそれらの事前分布  $F$  から生成する:  $\theta_k \sim F$  ( $k = 1, 2, \dots, K$ ).
2. 各データに  $K$  種のラベルの中の一つを確率的に割り当てるための離散分布 (和が 1 となる  $K$  次元ベクトル) の重みを生成する:  $\pi = (\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K)$ .
3. 各データ  $y_n$  に離散分布  $\pi$  から一つのラベル  $z_n$  が割り当てられる:  $z_n \sim \text{Discrete}(\pi)$  ( $n = 1, \dots, N$ ).
4. 各データ  $y_n$  が  $z_n$  番目のラベルの持つパラメータ  $\theta_{z_n}$  からある分布  $f(\theta_{z_n})$  から生成される:  $y_n \sim f(\theta_{z_n})$  ( $n = 1, \dots, N$ ).

#### 3.1.1 ディリクレ過程と無限混合モデル

一般にベイズ的なモデリングにおいてモデルの複雑度をどのように設定するかは極めて重要な問題であり, 近年の標準的な取扱いとして, 無限混合モデルがよく用いられる (Rasmussen, 2000). 最もよく用いられるものの一つとして, ディリクレ過程 (Ferguson, 1973) を使ったモデルを紹介する.

ディリクレ過程は確率測度に対する確率分布であり, 集中度と呼ばれる正の実数  $\gamma$  と可測空間  $(\Theta, \mathcal{B})$  上の基底測度  $F$  が与えられたとき, ディリクレ過程から生成された確率測度  $G \sim \text{DP}(\gamma, F)$  は  $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \theta_k \sim F$  の形で表されることが知られている. 重み  $\beta$  を構成する方法としては棒折り過程 (Sethuraman, 1994) が有名である:  $\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l), \beta'_k \sim \text{Beta}(1, \gamma)$ . (この分布は簡単に  $\beta \sim \text{GEM}(\gamma)$  と表されることが多い)

ディリクレ過程の性質は前述の有限混合モデルと関連付けて捉えると分かりやすい. 有限混合モデルでは, 各コンポーネントの重みをディリクレ分布から生成し,  $k$  番目のコンポーネントに紐付いたラベルのもつパラメータ  $\theta_k$  を事前分布  $F$  から生成していた. ディリクレ過程から生成された確率測度は各コンポーネントの重み  $\pi_k$  と, そのラベルに紐付いたパラメータ  $\theta_k$  を同時に表現していると捉えると, ディリクレ過程は潜在的に無限のコンポーネン

トを持った混合モデルの構成に用いることが出来る:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \text{DP}(\gamma, F), \quad z_n \sim \pi, \quad y_n \sim f(\theta_{z_n}). \quad (1)$$

実際、前述の有限混合モデル (ディリクレ分布-離散分布) において  $K \rightarrow \infty$  とすればディリクレ過程混合モデルとなることが知られている (Ishwaran & Zarepour, 2002). 具体的な構成法や推論 (例えば Kenichi et al., 2007, Paisley et al., 2011, Walker, 2007, Papaspiliopoulos, 2008) については紙面の都合上省略させて頂く.

### 3.1.2 ディリクレ過程混合モデルを用いる際の注意: コンポーネント数は推定出来るのか?

ディリクレ過程はコンポーネントの数が未知の場合の混合モデルの構成に広く用いられるようになっているが、「コンポーネントの数を推定する」という文脈で使う場合には注意が必要であることが近年の研究で指摘されている. ディリクレ過程混合モデルは、混合分布の密度関数や混合の重みへのフィッティングの意味においては適しているが、コンポーネントの個数の推定には不向きであることが報告されている (Miller & Harrison [5]).

詳細は Miller & Harrison [5] を参照して頂きたいが、ここでは概略だけ紹介したい. ディリクレ過程は非常に少数のデータが割り当てられているようなコンポーネントが大量に発生することに寛容な性質を持っており、データ生成の尤度部分がこのディリクレ過程の事前分布における性質を上回ることが出来ずに無駄なコンポーネントの発生を抑えられない場合がある. コンポーネントの数自体を推定したい場合の代替案も併せて議論されている.

### 3.1.3 混合モデルの例: 隠れマルコフモデル

紙面の都合上無限隠れマルコフモデル (Beal et al., 2002) を例にとり、その生成モデルが混合モデルとして記述できることを確認していく. 隠れマルコフモデルは観測系列  $\mathbf{y} = (y_1, y_2, \dots, y_T)$  の背後に隠れ状態の系列  $\mathbf{z} = (z_1, z_2, \dots, z_T)$  を考えることにより系列データのモデル化を行う.  $z_t$  は時刻  $t$  における状態のインデックスを表している. 通常の隠れマルコフモデルでは 1 次のマルコフ性が仮定され、状態の遷移  $i \rightarrow j$  は状態遷移確率  $\pi_{i,j} = p(z_t = j \mid z_i = i)$  によって決まるものとして取り扱う. 現在の状態  $z_t$  は前の状態  $z_{t-1}$  から  $z_t \sim \pi_{z_{t-1}}$  と表現することが出来る. 各状態  $k$  はパラメータ  $\theta_k$  を持ち、観測系列は出力分布  $y_t \sim f(\theta_{z_t})$  から生成されたものと見なす. では無限モデルである無限隠れマルコフモデルはどのように構成できるだろうか. 状態遷移確率  $\pi_i$  ( $i = 1, 2, \dots$ ) は重みの総和が 1 となる多項分布であると考えられることから、それぞれの  $\pi_i$  をディリクレ過程に従う確率測度の重みと見なすことで無限次元多項分布のように振る舞うことが期待される. つまり、 $\pi_i \sim \text{GEM}(\gamma)$  ( $i = 1, 2, \dots$ )

とすることで無限状態を持つ隠れマルコフモデルが構成出来るかに思われる. しかしここで各重みに割り当てられたパラメータに注意を払わなければならない. 例えば  $\pi_3, \pi_5$  が独立なディリクレ過程の確率測度の重みだと考えたとき、 $\pi_{3,2}, \pi_{5,2}$  とともに“状態 2”の原子  $\theta_2$  に対応した重みである必要がある. しかし前述のような確率測度の重みだけの議論ではこのような原子の共有が陽に保証されていない. そこで次のような階層ディリクレ過程 (Teh et al., 2006) が用いられる.

階層ディリクレ過程とは、あるディリクレ過程から生成された測度を基底測度として共有するディリクレ過程の集合のことを指す. すなわち、確率測度の集合  $G_i \sim \text{DP}(\alpha, G_0)$  ( $i = 1, 2, \dots$ ) で、基底測度が  $G_0 \sim \text{DP}(\gamma, H)$  となるものを指す. ディリクレ過程の性質から、上の階層のディリクレ過程に従う  $G_0$  はアトミックになり、下の階層のディリクレ過程はアトミックになった基底測度を用いることで  $G_0$  においてアクティブな原子にだけ重みを持つようになる. 各確率測度は  $G_i = \sum_{k=1}^{\infty} \pi_{i,k} \delta_{\theta_k}$  ( $i = 1, 2, \dots$ ). と表せるようになり、各重みの受け持つ原子 (パラメータ) を陽に共有化することが出来る. そこで、 $\pi_{i,j}$  を状態  $i$  から状態  $j$  への遷移確率、 $\theta_k$  を状態  $k$  のもつ出力分布パラメータだと見なすと、階層ディリクレ過程に基づく隠れマルコフモデルは次のように表すことが出来る:

$$\begin{aligned} \beta &\sim \text{GEM}(\gamma), \quad \pi_i \sim \text{DP}(\alpha, \beta), \quad z_t \mid z_{t-1} \sim \pi_{z_{t-1}}, \\ \theta_k &\sim H, \quad y_t \mid z_t, \theta \sim f(\theta_{z_t}) \end{aligned} \quad (2)$$

## 3.2 因子モデル

混合モデルは各データが一つのラベルに割り当てられたパラメータから生成されるモデルになっていた. もう一つの典型的なモデルとして、各データがいくつかの特徴量の組み合わせから生成されたものと見なすものが因子モデルである. 各データが特徴量からの線形モデルとなる場合は非負値行列因子分解や独立成分分析などのメディア処理に頻出なモデルと捉えることが出来る. ノンパラメトリックベイズにおいては無限の因子とその結合の係数に対する事前分布を適切に設定することが重要になる.

### 3.2.1 ベータ過程, ガンマ過程, ベルヌーイ過程と無限因子モデル

潜在的に無限個の特徴量を持つような因子モデルを構成する方法として、特別な場合のレヴィ過程がよく用いられる. レヴィ過程  $X(\omega)$  は可測空間  $(\Phi, \mathcal{F})$  上に独立な跳ねを持つ確率過程で、無限因子モデルにはおいては離散的でそれらの跳ねが正となるようなベータ過程, ガンマ過程, ベルヌーイ過程がよく用いられる. 上記の特別な場合のレヴィ過程は  $R^+ \times \Omega$  の直積に対してレヴィ測度と呼ばれる  $\nu$  を持つポアソン点過程と見なすことが知られている (Sato, 1999, Wang & Carin, 2012).

まずレヴィ過程の特別な場合の例として、ベータ過程を紹介する。集中度を  $c > 0$ 、基底測度を  $\mu$  とするベータ過程  $B \sim \text{BP}(c, \nu)$  とは  $B(d\omega) \sim \text{Beta}(c\mu(d\omega), c(1-\mu(\omega)))$  となるようなレヴィ過程のことをいう。このとき、 $B$  はレヴィ測度を  $\nu(d\pi, d\omega) = c\pi^{-1}(1-\pi)^{c-1}d\pi\mu(d\omega)$  とする  $R^+ \times \Omega$  へのポアソン点過程とも見なすことが出来る。ここで  $c\pi^{-1}(1-\pi)^{c-1} = \text{Beta}(0, c)$  であり、 $(0, 1)$  において積分したときに無限になっていることに注意するとポアソン点過程の性質から、 $B$  には無限の点が発生していると思なすことができ、 $B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$  のようにアトム  $\omega_i$  において跳ね  $\pi_i$  を持つような形で表すことが出来ることが知られている。アトムを各特微量を表すパラメータ、跳ねを各特微量のデータへの貢献を表す重みとして用いることで無限因子モデルを構成することが出来る。

無限因子モデルの構成には必ずしもベータ過程を使う必要はなく、レヴィ測度はモデル化や推論の都合に合わせて適切なものを設定すればよい。以下に述べるガンマ過程も実用的によく用いられる。レヴィ測度を  $\nu(d\phi, d\theta) = c\theta^{-1}e^{-c\theta}d\theta F(d\phi)$ 、のように選ぶ。ただし、 $c > 0$  は集中度を表している。 $c\theta^{-1}e^{-c\theta}$  は  $\text{Gamma}(0, 1/c)$  となっており、 $(0, \infty)$  で積分した際に無限になるため、ガンマ過程  $G \sim \text{GP}(c, F)$  は無限のアトムをもち、次のように表すことが出来る： $G = \sum_{d=1}^{\infty} \theta_d \delta_{\phi_d}$ 。ベータ過程の跳ねは  $(0, 1)$  だったのに対し、ガンマ過程の跳ねは  $\mathbb{R}^+$  になっている。

無限因子モデルを利用する際に、無限の因子の各データに対する寄与の有無、すなわち 0 または 1 のバイナリな情報だけ用いたい場合も少なくない。そのような場合にはベータ過程-ベルヌーイ過程の階層モデルを用いるのが一般的である (Te et al., 2007, Thibaux & Jordan, 2007)。無限因子モデルの構成に用いられるインド料理過程はこの階層モデルの特別な場合に対応している (Thibaux & Jordan, 2007)。

### 3.2.2 因子モデルの例: 非負値行列因子分解

因子モデルの代表的のものとして非負値行列因子分解 (の確率的生成モデル) による音楽信号を分離する問題を例に、その構成法を紹介する。振幅スペクトログラムもしくはパワースペクトログラム  $Y = (Y_{\omega,t})_{\Omega \times T} \in R^{\geq 0, \Omega \times T}$  (ただし、 $\omega = 1, \dots, \Omega$  は周波数インデックス、 $t = 1, \dots, T$  は時間インデックスを表す) が基底  $H = (H_{\omega,d})_{\Omega \times D} \in R^{\geq 0, \Omega \times D}$  とアクティベーション  $U = (U_{d,t})_{D \times T} \in R^{\geq 0, D \times T}$  の積で表現できるという仮定に基づいている。これはすなわち、 $Y_{\omega,t} \approx \sum_d H_{\omega,d} U_{d,t}$  のように観測スペクトログラム  $Y$  を  $D$  個の頻出の基底スペクトル  $h_d = [H_{1,d}, \dots, H_{\Omega,d}]$  とそれぞれの音量変化を表すアクティベーションで近似しようとしていることに相当する。音響信号によく合った尤度関数の選び方はそれ自身大きな研究対象の一つではあるが、ここでは標準的なものの一つとして、指数分布を用いると次

のようになる： $Y_{\omega,t} \sim \text{Exp}(1/\sum_d H_{\omega,d} U_{d,t})$ 。これは板倉斎藤距離規準の非負値行列因子分解に対応している (Févotte et al., 2009)。 $H$  と  $U$  に対する事前分布には例えばガンマ分布を用いるのが標準的である： $H_{\omega,d} \sim \text{Gamma}(a_H, b_H)$ 、 $U_{d,t} \sim \text{Gamma}(a_U, b_U)$ 。一般にデータを説明するのに必要な因子の適切な数  $D$  を事前に決めるのは困難であり、無限非負値行列因子分解が有用になってくる。先のように尤度関数として指数分布を選んだときには推論の都合上ガンマ過程を用いるのが簡単である (Hoffman et al., 2010)。各コンポーネントごとのゲインの総量を表す  $\theta_d$  ( $d = 1, 2, \dots$ ) を導入し、 $Y_{\omega,t} \sim \text{Exp}(1/\sum_d \theta_d H_{\omega,d} U_{d,t})$ 、 $G = \sum_{d=1}^{\infty} \theta_d \delta_{\phi_d} \sim \text{GP}(c, F)$  のように、 $\theta_d$  ( $d = 1, 2, \dots$ ) をガンマ過程の跳ねだと考える方法が提案されている。 $\phi_d$  は  $d$  番目のコンポーネントにおけるパラメータ、すなわち  $(H_{1,d}, \dots, H_{\Omega,d})$ 、 $(U_{d,1}, \dots, U_{d,T})$  を表しており、 $F$  はそれらに対する事前分布 (ガンマ分布など) を考えればよい。

## 4 メディア処理のためのベイズモデル

各種メディアに対し数多のノンパラメトリックベイズモデルが提案されており、それらを網羅することは到底不可能であるが、紙面の許す範囲で一部だけでも紹介したい。

### 4.1 bag-of-words を表現したモデル

自然言語処理において文書中に登場する単語の生成過程を描く頻出のモデルとしてトピックモデルがある (Blei et al., 2003)。その汎用性の高さからコンピュータビジョンや音響信号処理分野でも盛んに応用されている。最も素朴なモデルは階層ディリクレ過程の例題としてもよく用いられる。各単語はある隠れたトピックから生成されたと仮定し、各文書における各トピックの出やすさと各単語における隠れたトピックの割り当てを同時に考えたモデルになっている。 $J$  個の文書が観測されたとき、 $j$  番目の文書における  $n$  番目の単語の生成過程は次のように描かれる：

1. 潜在的に無限のトピックの出やすさを表す確率測度を生成する： $G_0 = \sum_{t=1}^{\infty} \lambda_t \delta_{\phi_t} \sim \text{DP}(\gamma, F)$ 。 $\phi_t$  は  $t$  番目のトピックにおける各単語の出現頻度を表す離散分布であり、それらへの事前分布となるディリクレ過程の基底測度  $F$  には例えばディリクレ分布を使うのが一般的である。
2.  $j$  番目の文書中での各トピックの出現しやすさを表す確率測度を生成する： $G_j = \sum_{t=1}^{\infty} \theta_{j,t} \delta_{\phi_t} \sim \text{DP}(\alpha, G_0)$ 。各トピックがどの程度現れやすいかを表す  $G_0$  の傾向が階層ディリクレ過程によって  $G_j$  ( $j = 1, 2, \dots$ ) に反映されている。
3.  $j$  番目の文書の  $n$  番目の単語に対するトピックの割り当て  $z_{j,n}$  を生成する： $z_{j,n} \sim \theta_j$ 。
4.  $j$  番目の文書の  $n$  番目の単語に対して、 $z_{j,n}$  番目の

トピックにおける単語の出現頻度を表した離散分布  $\phi_{z_j,n}$  から単語  $x_{j,n}$  を生成する:  $x_{j,n} \sim \phi_{z_j,n}$ .

トピックモデルの改良として、混合モデルと因子モデルの両方を組み込んだ focused topic model と呼ばれるモデルが提案されている (Williamson et al., 2010). 先のトピックモデルは階層ディリクレ過程のおかげで、各文書においてあらゆるトピックの出現を 0 でない確率で考えることが出来た。一方、各文書中に現れるトピックは非常にスパース、つまり同一文書には数少ないいくつかのトピックしか出現しないものと仮定したモデルも考えられる。このような性質を陽を組み込んだ拡張として、先のトピックモデルにおける  $G_j$  ( $j = 1, 2, \dots$ ) を次のように修正することが考えられる:

$$B_{j,t} \sim \text{Bernoulli}(a_t), \quad a_t \sim \text{Beta}(a_0/T, b_0(T-1)/T), \quad (3)$$

$$\theta_j \sim \text{DP}(\alpha, B_j \odot \lambda) \quad . \quad (4)$$

$(a_1, \dots, a_T)$  はベータ過程 (簡単のため有限打ち切りで表記) の重み,  $(B_{j,1}, \dots, B_{j,T})$  はそれらを上の階層にしたベルヌーイ過程のバイナリな重みである。  $B$  はインド料理過程から生成されたと見なしても構わない。  $(B_{j,1}, \dots, B_{j,T})$  は  $j$  番目の文書における各トピックの出現有無をバイナリに表現しており,  $B_j \odot \lambda$  (ベクトルの要素ごとの積) によってトピックの候補がスパースになるように働く。

#### 4.2 状態空間を用いた系列データのモデル

隠れマルコフモデルのように観測系列をスパースな状態によって表現しようとするモデルはメディア処理の多くの場面で非常に役に立つ。観測データのそれぞれに一つの状態で表現されその状態遷移が一次のマルコフ性を持つという意味で隠れマルコフモデルを最も基本的な系列データのモデルと見なすことが出来るが、対象とするメディアや課題に合わせ様々なモデルが用いられている。いくつか主要な拡張として次のようなものを挙げる事が出来る。

- 状態遷移行列を所望の形へ誘導する: stick HDP-HMM (Fox et al., 2007), block diagonal infinite HMM (Stepleton et al., 2009)
- 系列の長い依存関係を表現すべく高次のマルコフ性を考える: n-gram (Mochihashi & Sumita, 2007), sequence memoizer (Wood et al., 2009)
- 系列の構文を解析すべく木構造の状態割り当てを考える: 確率文脈自由文法 (Liang et al., 2007)
- 状態空間を因子の組み合わせによって表現する: infinite factorial HMM (Van Gael et al., 2008), infinite latent event model (Wingate et al., 2009), infinite dynamic Bayesian net. (Doshi-Velez et al., 2011)

無限隠れマルコフモデルにおける状態遷移行列を所望の形に誘導しようとする研究として sticky HDP-HMM

や block diagonal infinite HMM などが知られている。Fox et al. (2008) はスティッキー階層ディリクレ過程隠れマルコフモデルと呼ばれるモデルを提案した。これは、系列をゆっくりと変化する隠れ状態の遷移によって表現しやすくなるように改良を加えたものである。何らかの系列を隠れマルコフモデルで解析する状況において、隠れ状態がゆっくりと変化してほしい (すなわち自己遷移が起こりやすくなってほしい) 場合は多々ある。このような性質を無限隠れマルコフモデルに導入する方法として、式 (2) において遷移確率を次のように変更する拡張が提案された:  $\pi_i \sim \text{DP}(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_i}{\alpha + \kappa})$ .  $\kappa > 0$  によって自己遷移への遷移確率にバイアスがかかることでゆっくりと変化するダイナミクスを捉えやすくなる。このように状態遷移行列を所望の形に誘導する拡張としてブロック対角の形 (Stepleton et al., 2009) がある。

系列の長い依存関係を表現するには高次のマルコフ性を考えた  $n$ -gram が有用である。  $n$ -gram は  $(n-1)$  次のマルコフ性をもった系列のモデルであり,  $(n-1)$  個の系列に依存して次が候補の中から確率的に選ばれるものと考えることで混合モデルの一種と見なすことが出来る。一般に依存関係の長さは系列の中でも可変として考えられる方が便利であり, 無限モデルとしての可変長  $n$ -gram が提案され, その構成には階層ピットマンヨー過程 (Pitman & Yor, 1997) が用いられている。ピットマンヨー過程もディリクレ過程同様それ自身が確率測度と見なすことが出来るため無限混合モデルの構成に用いることが出来る。また同種の系列データのために sequence memoizer と呼ばれるモデルも提案されている。これらは自然言語処理における単語系列のモデルとして使われることが多いが, 近年は音楽における旋律のモデル化などにも利用されている。さらに, 単語レベルでの  $\infty$ -gram に, 文字レベルの  $\infty$ -gram を入れ子にしたモデルを用いることで教師なしの単語分割に用いる研究も行われている。

自然言語処理における構文解析への応用に代表されるように確率文脈自由文法もメディア処理で度々登場するモデルで, これは一種の混合モデルとして構成することが出来る。簡単のためチョムスキー標準形 (親ノードが二つの子ノードに分岐) を考えるとすると, ある状態から二つの状態のペアへの確率的な分岐規則を司る確率測度を作ることが出来ればよく, これは隠れマルコフモデルの時と同様に階層ディリクレ過程によって表現することが出来る。確率文脈自由文法は自然言語処理における構文解析のみならず, 近年は音楽の多重音を 2 次元の確率文脈自由文法と因子モデルの組み合わせで表現しようという試みもある (Nakano et al., 2011, Kameoka et al., 2012)。

系列データを因子モデルによって表現しようとする研究もここ数年盛んに行われている。無限階乗隠れマルコフモデル (Van Gael et al., 2009) や無限階層隠れマルコフモデ

ル (Heller et al., 2009) では、観測系列を表現するために無限の (バイナリもしくは整数値をとる) 因子の組み合わせが導入された。また、無限イベントモデル (Wingate et al., 2009) や無限ダイナミックベイジアンネットワーク (Doshi-Velez et al., 2011) では因子の組み合わせおよびそれらの間のネットワークを同時に推定する手法が提案されてきた。

### 4.3 入れ子構造を用いたモデル

モデルの中に入れ子構造を導入することでしばしば有用な拡張を行うことが出来る (Jordan, 2009)。例えば画像のクラスタリングを階層的に行う問題、すなわち階層クラスタリングを考えてみる。パラの花が映ったある画像は「花」という大きな単位で分類され、その「花」の中でさらに細かい単位として「パラ」のように分類したい。いま  $n$  番目のデータ  $y_n$  に割り当てられる「花」のような大きなレベルのラベルのインデックス ( $1, 2, \dots, K$  の中の一つ) を  $z_n$  と表すことにすると混合モデルとして  $z_n \sim \pi$ ,  $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \text{DP}(\gamma, G_0)$  と表せる。ここで  $k$  番目のコンポーネントが持つ  $\phi_k$  およびその事前分布  $G_0$  をどのように扱えばよいだろうか。「花」のような大きな単位でのラベル  $z_n$  が割り当てられた後、データ  $y_n$  はさらに細かい単位のラベルとして「パラ」のような割り当てが行われることが期待されている。そこで細かい単位でのラベルのインデックスを  $z'_n$  とすると、これも混合モデルとみなすことが出来ることから

$$z'_n \sim \pi'_{z'_n}, G'_{z'_n} = \sum_{m=1}^{\infty} \pi'_{z'_n, m} \delta_{\phi_{z'_n, m}} \sim \text{DP}(\alpha, F), \quad (5)$$

$y_n \sim f(\phi_{z_n, z'_n})$  のように表現したい。このように考えると  $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$  のアトム  $\theta_k$  には細かいレベルでのラベル割り当てを行うための確率測度 (すなわちディリクレ過程)  $G'_k$  が対応していることが分かる。これはディリクレ過程が入れ子になっている様子を陽に表した形として

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{G'_k} \sim \text{DP}(\gamma, \text{DP}(\alpha, F)) \quad (6)$$

のようにも表される。 $\phi_{k, m}$  はデータを生成するためのパラメータが格納されており  $F$  がそれらの事前分布に対応していると考えればよい。

ここでは2段の入れ子構造を持ったディリクレ過程 (Rodriguez et al., 2008) のみを紹介したが、多段や無限の深さを考えることも出来る。また同種の考え方は因子モデルにおいても用いられており、nested beta process や nested gamma process などが考えられている (Jordan, 2009)。

### 4.4 共変量を介して依存関係を陽に表現したモデル

観測データのそれぞれには何らかの共変量があり、それらが似ているものについてはデータの生成過程も似た振る舞いになることが期待される場面がしばしばある。例え

ば、画像の領域分割はピクセルまたはスーパーピクセルのようにあらかじめ細分化された各領域に対してラベル付け (同一ラベルは同一領域) を行う問題と捉えることが出来るが、もともと画像上で隣り合った領域はそもそも同一ラベルが付きやすいはずである。したがって画像上での「位置」は細分化された領域に対するある種の共変量とみなすことが出来る。別の例として、音楽のような繰り返し構造を持った信号を各楽器音に分解する問題を考えたとき、これは信号が楽器音の線形結合で表せるとみなした因子モデルを適用するのが標準的な扱いであるが、音楽のようにAメロ、Bメロ、サビのような構造を持った信号の場合1回目のサビと2回目のサビでは似たような因子がアクティブになることが期待される。したがって、音楽の (隠れた) 繰り返し構造は各時刻の信号を分解する上での共変量となりうる。このようにデータ間の依存関係を導入しようとした研究は混合モデル、因子モデルともに数多く報告されている (Williamson et al., 2010, Ren et al., 2011)。

データ間の依存関係を導入した因子モデルの一つに kernel Beta process [42]を用いたものが挙げられる。これは各データと因子に共変量を導入しそれらが似ているものは共起しやすくなるようにベータ過程の跳ねを修正することで、共変量を介してデータ間の依存関係を表現しようとしたものである。データ  $Y_n$  に対応する共変量を  $x_n \in \mathcal{X}$ , factor  $\omega_i$  に対応する共変量を  $x_i^* \in \mathcal{X}$  と表すことにする。各因子は共変量空間において、それぞれの局所的な性質を表すためのパラメータ  $\psi^* \in \Psi$  を用いると、共変量空間においてデータと因子の間の類似度  $K(x, x^*; \psi^*)$  (例えば  $K(x, x^*; \psi^*) = \exp(-\psi^* \|x - x^*\|_2)$ ) を考えることが出来る。ここで、共変量  $x$  の影響を受けた  $B_x$ :

$$B_x = \sum_{i=1}^{\infty} \pi_i K(x, x_i^*; \psi^*) \delta_{\omega_i} \quad (7)$$

を考えると、これはベータ過程の重みが共変量の近さによって修正されたものと見なすことが出来る。つまり、共変量を介してデータ間の依存関係を陽に導入することが出来る。 $x^*, \psi^*$  がそれぞれ確率測度  $S, Q$  から生成されたとすると、 $K(x_n, x_i^*; \phi_i^*) \in (0, 1]$ ,  $K(x_n, x_i^*; \phi_i^*) \rightarrow 0$  ( $\|x_n - x_i^*\|_2 \rightarrow \infty$ ),  $K(x^*, x^*; \phi^*) = 1$  を満たすとき、これは  $\nu_{\mathcal{X}} = S(dx^*)Q(d\psi^*)\nu(d\pi, d\omega)$  をレヴィ測度 ( $\nu(d\pi, d\omega)$  は通常のベータ過程と同様) とするレヴィ過程となることが知られている (Ren et al., 2011)。

### 4.5 互いに似ていないものの同時出現を表現したモデル

時々刻々と更新されていくテキストの中から、(時間に伴って変化する) ニュースの見出しを作る問題を考えてみる。その時々でいくつかのニュースの見出しを作る際、それぞれ一つ一つが重要なものを指しているのと同時に、お互いが違うことを指しているのが望ましい。例えば、ある時刻に同一のことを指して「火事」と「火災」の二つの見出し



が出現するのは望ましくない。このように似ていないものが同時に出現しやすくなるような機構をモデル化したい場面は少なくない。近年このような互いに似ていないものを表現するために行列式点過程を用いる方法が提案されている (Kulesza& Taskar, 2010, Affandi et al., 2012)。

## 5 おわりに

各種メディア処理においてベイズ的な手法は有力の選択肢の一つとなってきた。近年の発展を把握する上で本稿および発表がその概観に役立てば幸いである。

## 参考文献

- [1] T. S. Ferguson, “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, 1(2): pp. 209-230, 1973.
- [2] J. Pitman and M. Yor, “The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator,” *Annals of Probability*, 25: pp. 855-900, 1997.
- [3] J. Sethuraman, “A constructive definition of Dirichlet priors,” *Statistica Sinica*: 4, pp. 639-650, 1994.
- [4] C. E. Rasmussen, “The infinite Gaussian mixture model,” In *Advances in Neural Information Processing Systems*, vol. 12, 2000.
- [5] J. W. Miller and M. T. Harrison, “Dirichlet process mixtures are inconsistent for the number of components in a finite mixture,” in *ICERM*, 2012.
- [6] K. Kenichi, M. Welling, and Y. Whye Teh, “Collapsed variational dirichlet process mixture models” in *Proc. IJCAI*, 2007.
- [7] D. J. Aldous, “Representations for Partially Exchangeable Arrays of Random Variables,” *Journal of Multivariate Analysis*, 11: pp. 581-598, 1981.
- [8] J. Paisley, C. Wang and D. Blei, “The discrete infinite logistic normal distribution for mixed-membership modeling”, in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2011.
- [9] S. G. Walker, “Sampling the Dirichlet mixture model with slices,” *Communications in Statistics - Simulation and Computation*, 36:45, 2007.
- [10] O. Papaspiliopoulos and G. O. Roberts, “Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models,” *Biometrika*, 95(1): pp. 169-186, 2008.
- [11] K. Sato, “Lévy processes and infinitely divisible distributions,” Cambridge University Press, 1999.
- [12] Y. Wang and L. Carin, “Lévy Measure Decompositions for the Beta and Gamma Processes,” in *Proc. of ICML*, 2012.
- [13] J. F. C. Kingman, “Completely random measure,” *Pacific Journal of Mathematics*, vol. 21(1): pp. 59-78, 1967.
- [14] M. I. Jordan, “Hierarchical models, nested models and completely random measures,” *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. New York: Springer, 2009.
- [15] C. Févotte, N. Bertin and J. L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Computation*, 21 (3): pp. 793-830, 2009.
- [16] M. Hoffman, D. Blei and P. Cook, “Bayesian nonparametric matrix factorization for recorded music” in *Proc. ICML*, pp. 641-648, 2010.
- [17] S. Williamson, C. Wang, K. Heller and D. Blei, “The IBP compound Dirichlet process and its application to focused topic modeling,” in *Proc. ICML*, 2010.
- [18] T. Stepleton, Z. Ghahramani, G. Gordon and T. S. Lee, “The block diagonal infinite hidden Markov model,” in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2009.
- [19] J. Van Gael, Y. Saatchi, Y. W. Teh and Z. Ghahramani, “Beam sampling for the infinite hidden Markov model,” in *Proc. of the International Conference on Machine Learning*, 2008.
- [20] R. Thibaux, and M. I. Jordan, “Hierarchical beta processes and the indian buffet process,” in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2007.
- [21] K. A. Heller, Y. W. Teh and D. Görür, “Infinite hierarchical hidden Markov models,” in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2009.
- [22] J. Van Gael, Y. W. Teh and Z. Ghahramani, “The infinite factorial hidden Markov model,” in *Advances in Neural Information Processing Systems*, 2009.
- [23] D. Wingate, N. D. Goodman, D. M. Roy, D and J. B. Tenenbaum, “The infinite latent events model,” in *Proc. of the International Conference on Uncertainty in Artificial Intelligence*, 2009.
- [24] F. Doshi-Velez, D. Wingate, N. Roy and J. Tenenbaum, “Infinite dynamic Bayesian networks,” in *Proc. of International Conference in Machine Learning*, 2011.
- [25] H. Ishwaran and M. Zarepour, “Exact and approximate sum-representations for the Dirichlet process,” *Canadian Journal of Statistics*, 30, 269-283, 2002.
- [26] Y. W. Teh, M. I. Jordan, M. Beal and D. Blei, “Hierarchical Dirichlet processes,” *Journal of the American Statistical Association*, 101, 1566-1581, 2006.
- [27] M. Beal, Z. Ghahramani and C. Rasmussen, “The infinite hidden Markov model,” in *Advances in Neural Information Processing Systems*, 2002.
- [28] Y. W. Teh, D. Görür and Z. Ghahramani, “Stick-breaking construction for the Indian buffet process,” in *Proc. of the International Conference on Artificial Intelligence and Statistics*, vol. 11, 2007.
- [29] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, 3:993-1022, 2003.
- [30] P. Liang, S. Petrov, M. I. Jordan, and D. Klein, “The infinite PCFG using hierarchical Dirichlet processes,” in *Proc. of EMNLP*, pp. 688-697, 2007.
- [31] H. Kameoka, K. Ochiai, M. Nakano, M. Tsuchiya, S. Sagayama, “Context-free 2D tree structure model of musical notes for Bayesian modeling of polyphonic spectrograms,” in *Proc. of ISMIR*, 2012.
- [32] M. Nakano, Y. Ohishi, H. Kameoka, R. Mukai, K. Kashino, “Bayesian nonparametric music parser,” in *Proc. of ICASSP*, 2011.
- [33] D. Mochihashi and E. Sumita, “The infinite Markov model,” in *Advances in Neural Information Processing Systems*, 2007.
- [34] F. Wood, C. Archambeau, J. Gasthaus, L. F. James and Y.W. Teh, “A Stochastic Memoizer for Sequence Data,” in *Proc. of ICML*, 2009.
- [35] A. Rodriguez, D. B. Dunson and A. E. Gelfand, “The nested Dirichlet processThe nested Dirichlet process,” *Journal of American Statistics Association* 103, 1131-1154, 2008.
- [36] A. Rodriguez and K. Ghosh, “Nested partition models,” *Jack Baskin School of Engineering*, Technical report, 2009.
- [37] E. B. Fox, E. B. Sudderth, M.I. Jordan, A.S. Willsky, “A Sticky HDP-HMM with Application to Speaker Diarization,” *Annals of Applied Statistics*, 2011.
- [38] A. Kulesza and B. Taskar, “Structured determinantal point processes,” in *Proc. of NIPS*, 2010.
- [39] R. H. Affandi, A. Kulesza and E. B. Fox, “Markov determinantal point process,” in *Proc. of UAI*, 2012.
- [40] S. Williamson, P. Orbanz, and Z. Ghahramani, “Dependent Indian buffet processes,” in *Proc. AISTATS*, 2010.
- [41] M. Zhou, H. Yang, G. Sapiro, D. Dunson, and L. Carin, “Dependent hierarchical beta process for image interpolation and denoising,” in *Proc. AISTATS*, 2011.
- [42] Lu Ren, Y. Wang, D. Dunson, and L. Carin, “The Kernel Beta process,” in *Proc. NIPS*, 2011.