ノンパラメトリックベイズ

産業技術総合研究所
情報技術研究部門
吉井 和佳
k.yoshii@aist.go.jp

統計的学習の基礎

観測データと確率モデル

- 観測データ
  - 我々が実際に観測できる量で、何らかの確率分布に従うことを仮定する (確率変数とみなす)
    - 例：サイコロをN回振って出た目の系列

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$$

- 確率モデル
  - 観測データに対してそれらがどのような過程を経て確率的に生成されたかを記述するもの
    - 確率モデルから生成されるデータは無数に考えられ観測データとはその１つの実現値
    - 実現値について確率(密度)を与えることができる

$$p(X \mid \Theta)$$

モデルパラメータ Θ

例: 各面の出る確率を示す6個のパラメータ

確率分布と確率モデル

- 確率分布
  - ある確率変数に対して確率(密度)を与える $p(x \mid \Theta)$
- 確率モデル
  - 観測データに対して確率(密度)を与える $p(X \mid \Theta)$

例：サイコロを振って出た目の系列に対する確率モデル

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$$

$$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\}$$

6次元ベクトル $\boldsymbol{x}_n$

出た目に対応する要素が1で他は0

離散分布

$$p(\boldsymbol{x} \mid \Theta) = \prod_{k=1}^{6} \theta_k^{x_k} \qquad \int p(\boldsymbol{x} \mid \Theta) d\boldsymbol{x} = \sum_{k=1}^{6} \theta_k = 1$$

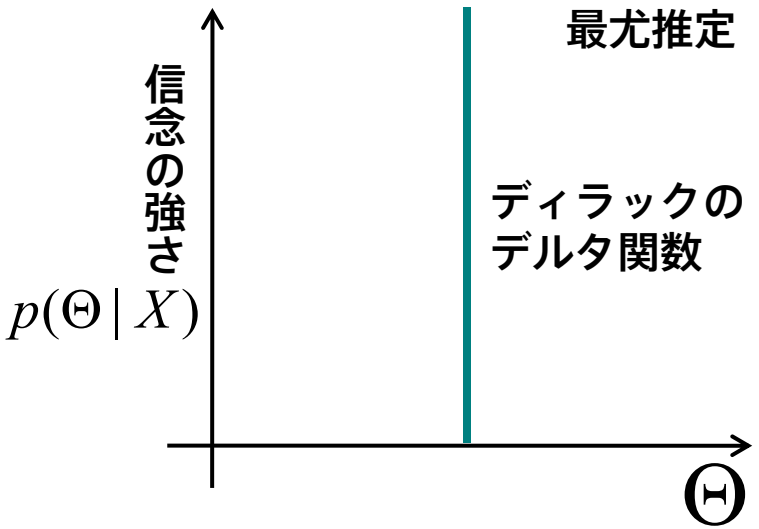
多項分布

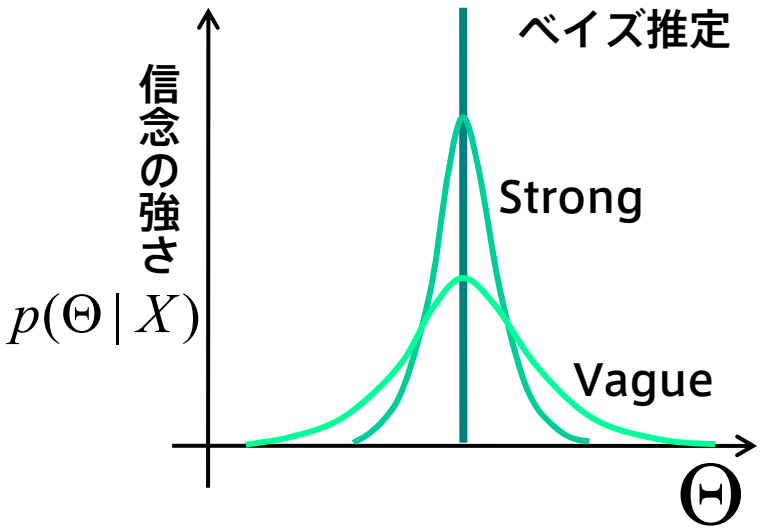
$$p(X \mid \Theta) = \prod_{n=1}^{N} \prod_{k=1}^{6} \theta_k^{x_{n,k}}$$

確率モデルの学習

- 観測データの生成確率が最大になるような確率モデルのパラメータを推定すること
  - 最尤推定
    - 最適なパラメータを点推定する (一意に決める)
  - ベイズ推定
    - パラメータの事後分布を推定する (信念の強さを反映する)

データが無限にあれば両者は一致





最尤推定

- 観測データだけから最適なパラメータを点推定

例：サイコロの各面が出る確率を最尤推定する

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$$

$$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\}$$

$$p(X \mid \Theta) = \prod_{n=1}^{N} \prod_{k=1}^{6} \theta_k^{x_{n,k}}$$

拘束条件 $\sum_{k=1}^{6} \theta_k = 1$

1. 対数をとって凸関数化：これを最大化

$$\log p(X \mid \Theta) = \sum_{n=1}^{N} \sum_{k=1}^{6} x_{n,k} \log \theta_k = \sum_{k=1}^{6} \left(\sum_{n=1}^{N} x_{n,k} \right) \log \theta_k = \sum_{k=1}^{6} n_k \log \theta_k$$

2. 拘束条件付きの最適化：ラグランジュの未定乗数法

$$F = \log p(X \mid \Theta) + \lambda \left(1 - \sum_{k=1}^{6} \theta_k \right)$$

4. 拘束条件に代入

3. 偏微分して０とおく

$$\frac{\partial F}{\partial \theta_k} = \frac{n_k}{\theta_k} - \lambda \equiv 0 \Rightarrow \theta_k = \frac{n_k}{\lambda} \longrightarrow \lambda = N$$

$$\theta_k = \frac{n_k}{N}$$

過学習

- 確率モデルが観測データにフィットしすぎて汎化能力(未知データの予測能力)が失われる

例：サイコロの各面が出る確率を最尤推定する

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$$

$$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\}$$

$$p(X \mid \Theta) = \prod_{n=1}^{N} \prod_{k=1}^{6} \theta_k^{x_{n,k}}$$

$$\theta_k = \frac{n_k}{N}$$

N=10の観測データ

目 k	回数 n_k
1	4
2	1
3	0
4	0
5	2
6	3

最尤推定値 θ_k
0.4
0.1
0.0
0.0
0.2
0.3

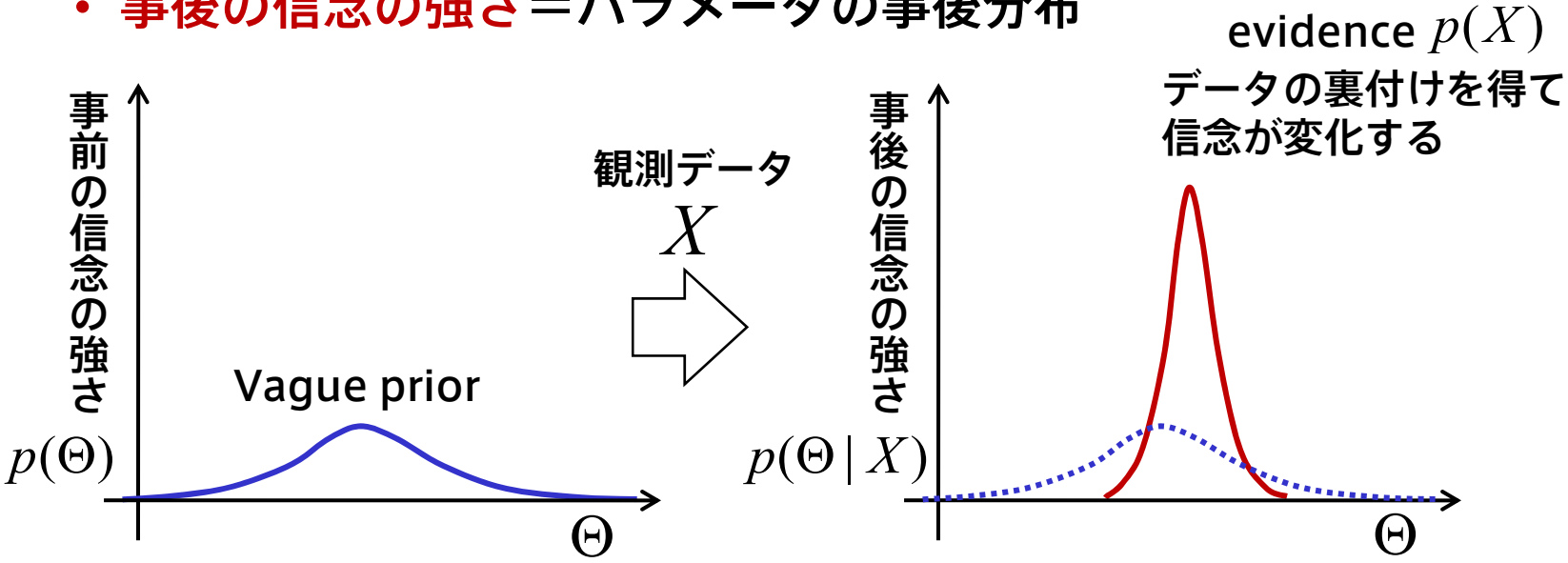
各面の出る確率が
偏りすぎではないか？

$\boldsymbol{x}_{N+1}$ はどうなるのだろうか？

3,4が出る確率はゼロ！

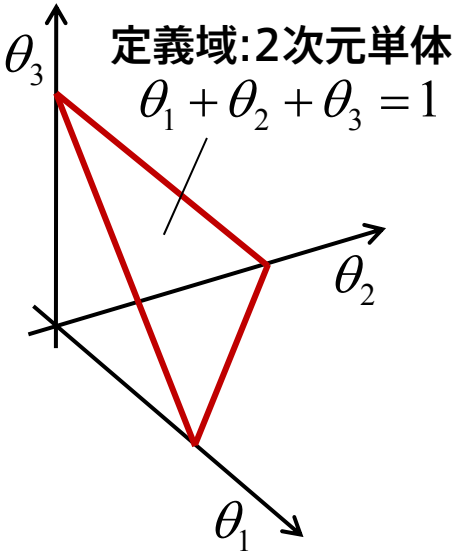
ベイズ推定

- パラメータを一意に決めずに
  パラメータ上の事後確率分布を推定する
  - 事前分布を導入してパラメータは「こうあるべき」という予断を与える
    - 事前分布のパラメータ(ハイパーパラメータ)を変えることで事前の信念の強さを反映することができる
  - 観測データが与えられると信念の強さが変化する
    - 事後の信念の強さ＝パラメータの事後分布

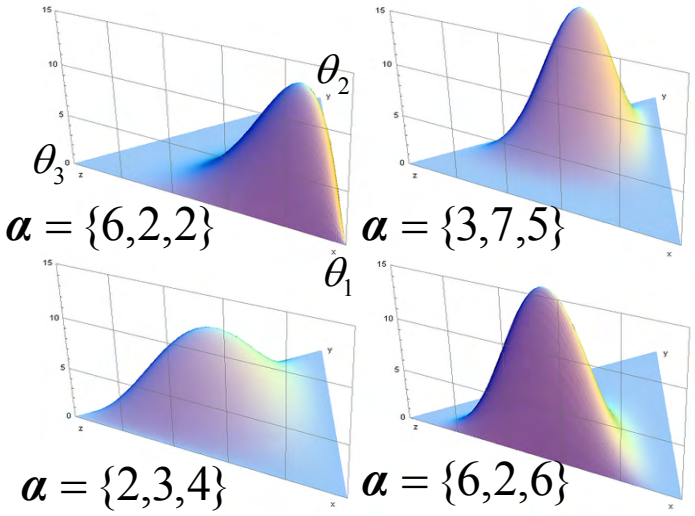


事前分布

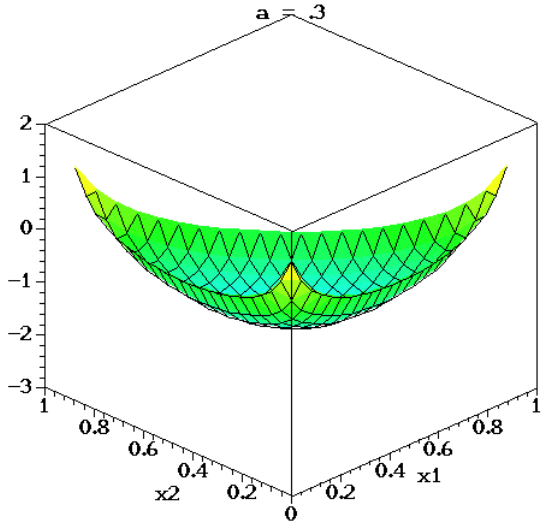
- 事前分布とは確率分布上の確率分布
  - Distribution over distributions
    - 確率分布を記述するパラメータが従う確率分布とも言える
  - 良く使われるのは**共役事前分布**
    - 事前分布 $p(\Theta)$ と事後分布 $p(\Theta \mid X)$ が同じ形になる
      - 離散分布上の分布：ディリクレ分布
      - ガウス分布上の分布：ガウス・ウィシャート分布



$\Theta = \{\theta_1, \theta_2, \theta_3\}$ の分布 $\mathrm{Dir}(\Theta \mid \boldsymbol{\alpha})$



$0 < \alpha < 2$ における変化



ベイズ推定

- **共役事前分布を用いて事後分布を計算**

例：サイコロの各面が出る確率をベイズ推定する

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\} \quad \Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\}$$

$$p(\Theta) = \mathrm{Dir}(\Theta \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{6} \alpha_k\right)}{\prod_{k=1}^{6} \Gamma(\alpha_k)} \prod_{k=1}^{6} \theta_k^{\alpha_k - 1} = C(\boldsymbol{\alpha}) \prod_{k=1}^{6} \theta_k^{\alpha_k - 1}$$

$a_k - 1$：**事前の観測回数に相当 (事前の信念の強さ)**

真っ当なサイコロだと例えば $a_k = 101$　$a_k = 1$ **だと**

無情報事前分布

Noninformative prior

信念が変化

$$p(X \mid \Theta) = \prod_{n=1}^{N} \prod_{k=1}^{6} \theta_k^{x_{n,k}} = \prod_{k=1}^{6} \theta_k^{n_k}$$

$$p(\Theta \mid X) = \frac{p(X \mid \Theta) p(\Theta)}{p(X)} \propto p(X \mid \Theta) p(\Theta) = C(\boldsymbol{\alpha}) \prod_{k=1}^{6} \theta_k^{\alpha_k + n_k - 1}$$

$$p(\Theta \mid X) = C(\boldsymbol{\alpha} + \boldsymbol{n}) \prod_{k=1}^{6} \theta_k^{\alpha_k + n_k - 1} = \mathrm{Dir}(\Theta \mid \boldsymbol{\alpha} + \boldsymbol{n})$$

正規化係数は分布が正しく正規化されるようにあとで計算すればよい

汎化能力の改善

- 適切な事前分布を与えることで過学習を抑制し、汎化能力を向上させることができる

事前の観測回数 $\alpha_k - 1$	実際の観測回数 n_k	事後の観測回数 $a_k + n_k - 1$
100	4	104
100	1	101
100	0	100
100	0	100
100	2	102
100	3	103

ベイズ推定による事後分布の保持

$$\mathrm{Dir}(\Theta \mid \boldsymbol{\alpha} + \boldsymbol{n})$$

Maximum a posterior: 事後確率最大化

	最尤推定値	MAP推定値
点推定	$\Theta = \{0.4, 0.1, 0.0, 0.0, 0.2, 0.3\}$	$\hat{\Theta} = \{0.170, 0.166, 0.164, 0.164, 0.167, 0.169\}$

潜在変数モデル

- 観測データだけでなく非観測データも考える
  - 例：混合モデル
    - 性別不明の身長データに対する混合ガウスモデル
      - 観測変数X: 身長 (男 or 女 のガウス分布)
      - 潜在変数Z: 男 or 女 (離散分布)
      - パラメータΘ: ガウス分布の平均と精度・混合比

最尤推定 パラメータ Θ　事後分布 $p(Z \mid X;\Theta)$ を求めたい

確率変数ではない　あらゆるZの可能性をその確率を重みとして考慮

$$p(X;\Theta) = \int p(X \mid Z;\Theta)\, p(Z;\Theta) dZ$$

ベイズ推定 事後分布 $p(\Theta, Z \mid X)$ を求めたい

$$p(X) = \int p(X \mid Z,\Theta)\, p(Z \mid \Theta)\, p(\Theta) d\Theta dZ$$

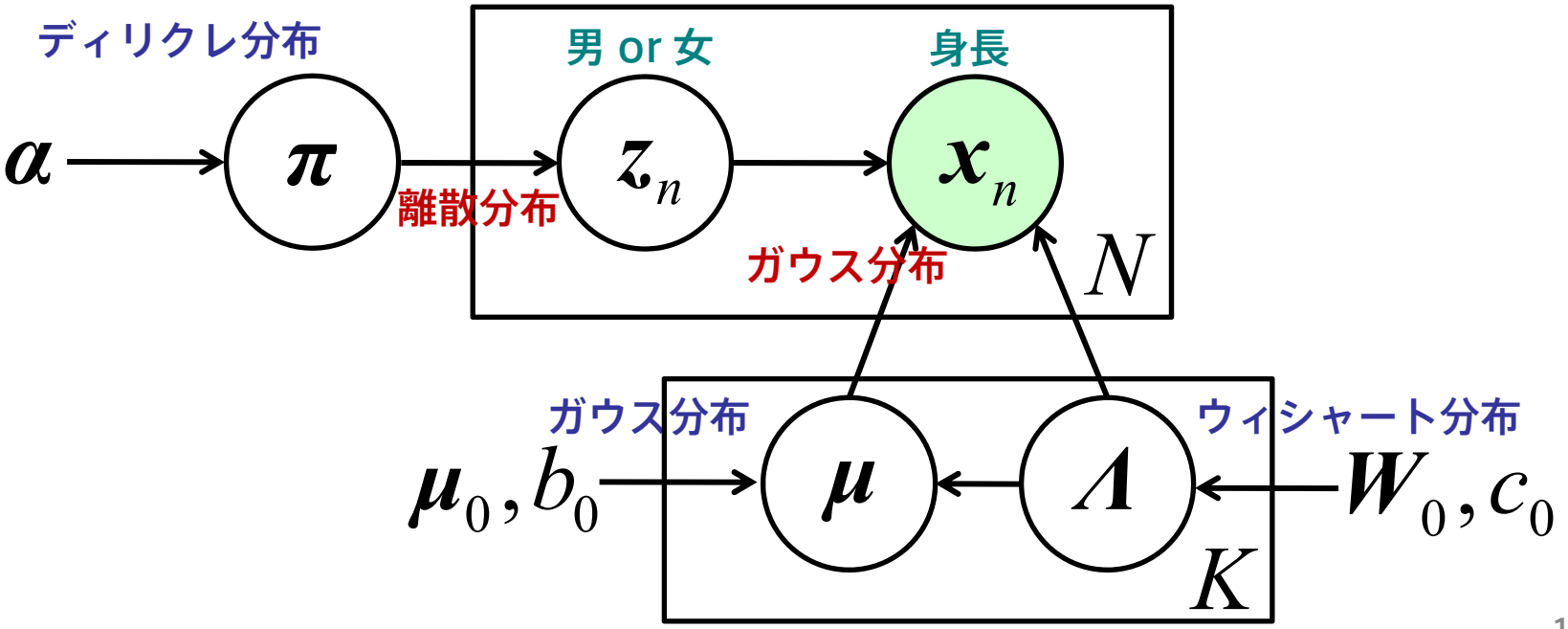
事後分布の計算にはこれらの正規化係数の計算が必要だが解析的には解けない
→ 変分近似で直接計算を可能にする (VB)
→ 陽に求めなることなく事後分布を推定する (MCMC)

グラフィカル表現

- 確率変数間の依存関係を図で表現すると便利
  - 階層ベイズのような複雑な確率モデルになると依存関係を可視化すると理解しやすい

例：性別不明の身長データに対するベイズ混合ガウスモデル

$$\Theta = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\pi}\} \quad K = 2$$



ノンパラメトリックベイズ

ノンパラメトリックベイズ

- ベイズ推論のための新しいパラダイム
  - 確率モデルの関数形を仮定しない
  - 確率モデルの複雑さを仮定しない
    - モデルパラメータに適切な事前分布を与えれば陽に考える必要がなくなる
    - モデルパラメータを積分消去する

integrate out/marginalize out/collapse

例：混合数

$$\underset{\text{エビデンス}}{p(X;K)} = \int \underset{\text{(狭義の)モデル}}{p(X \mid \Theta;K)}\, \underset{\text{事前分布}}{p(\Theta;K)} d\Theta$$

Kも確率変数と思ってΘに含ませる

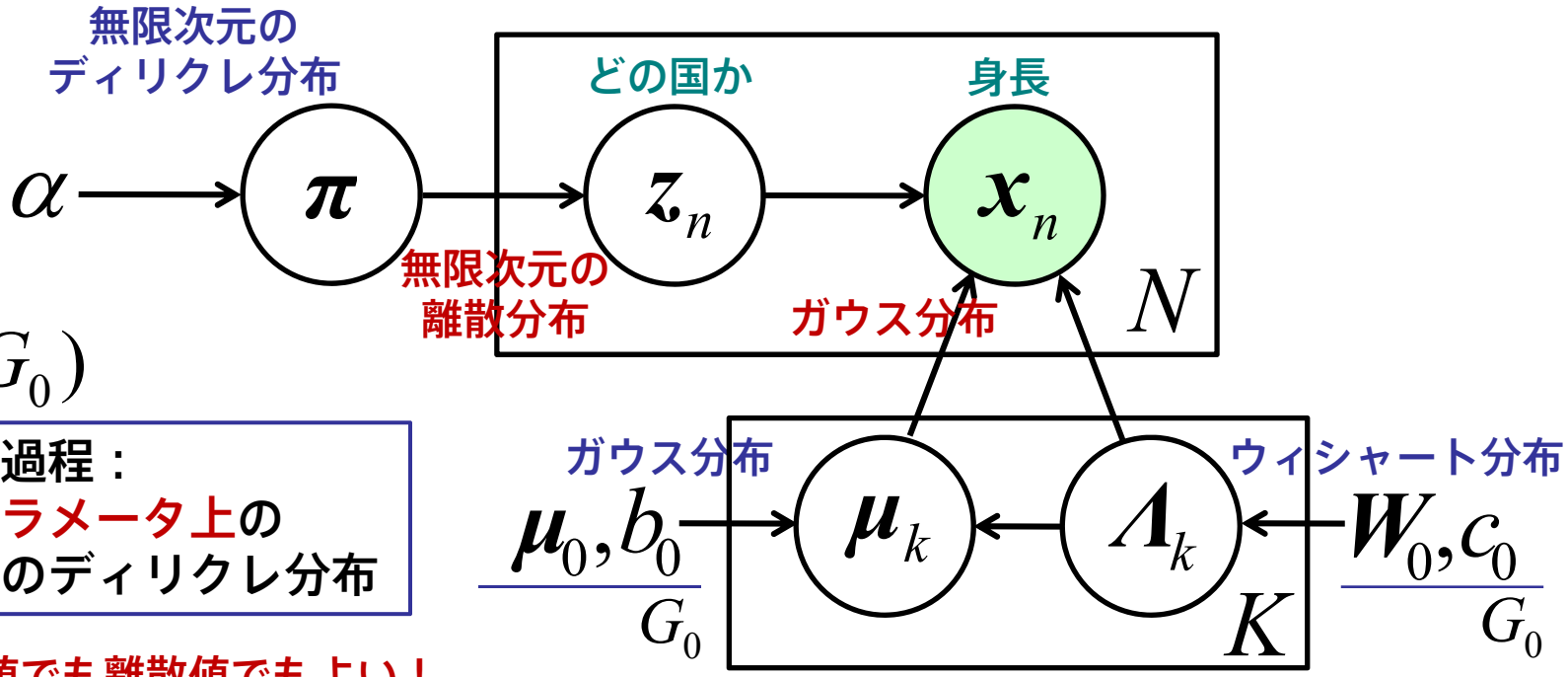
$$p(X) = \int p(X \mid \Theta)\, p(\Theta) d\Theta$$

無限潜在変数モデル

- 可算無限個の潜在クラスを許容する確率モデル
  - 例：N人の外国人を身長でクラスタリング
    - 従来の興味：国の数 K　各国の平均と精度 μ, Λ　混合比 π

無限個存在　　無限次元

$$\Theta = \{\mu_1, \mu_2, \cdots, \Lambda_1, \Lambda_2, \cdots \pi\} \qquad K \to \infty$$



$\mathrm{DP}(\alpha, G_0)$

ディリクレ過程：
モデルパラメータ上の
無限次元のディリクレ分布

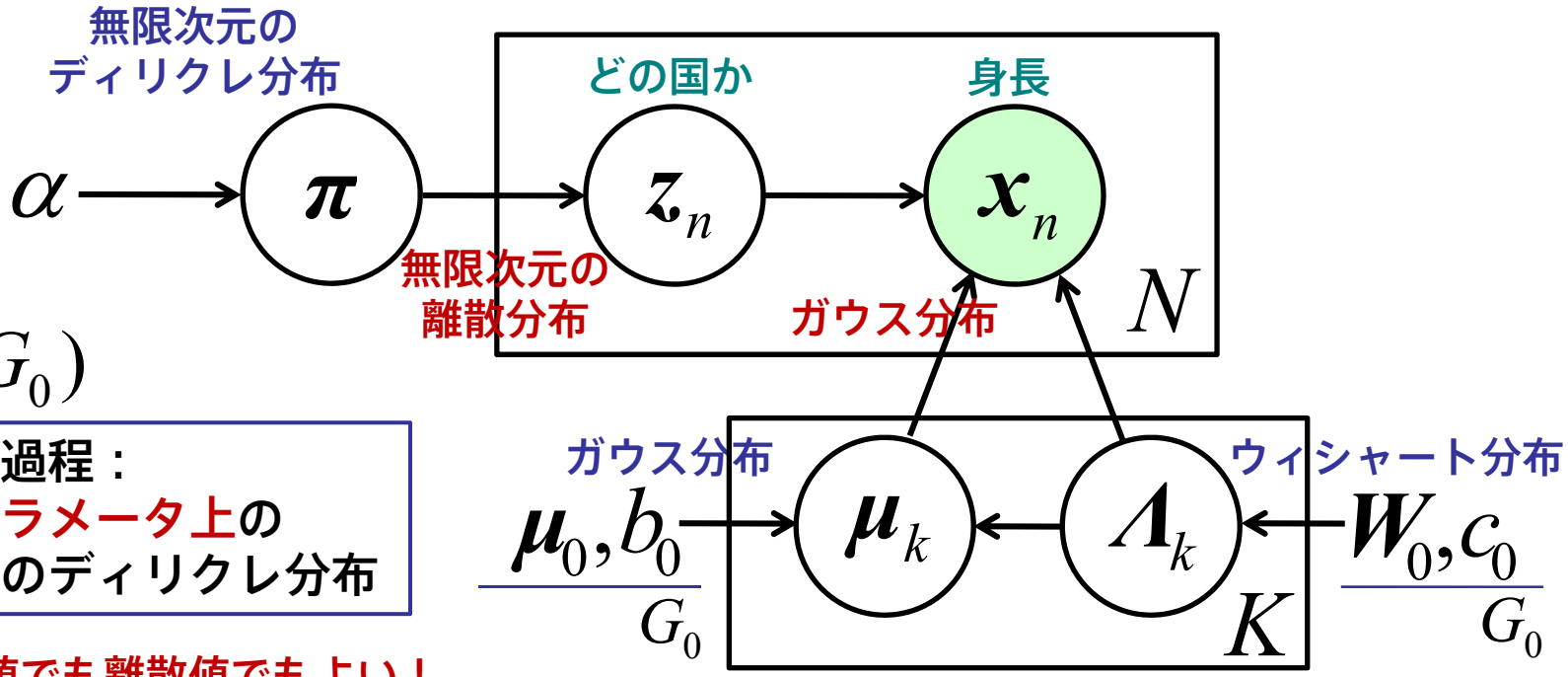
連続値でも離散値でもよい！

無限潜在変数モデル

- 可算無限個の潜在クラスを許容する確率モデル
  - 例：N人の外国人を身長でクラスタリング
    - 従来の興味：国の数 K　各国の平均と精度 μ, Λ　混合比 π

無限個存在　　無限次元

$$\Theta = \{\mu_1, \mu_2, \cdots, \Lambda_1, \Lambda_2, \cdots \pi\} \qquad K \to \infty$$



$\mathrm{DP}(\alpha, G_0)$

ディリクレ過程：
モデルパラメータ上の
無限次元のディリクレ分布

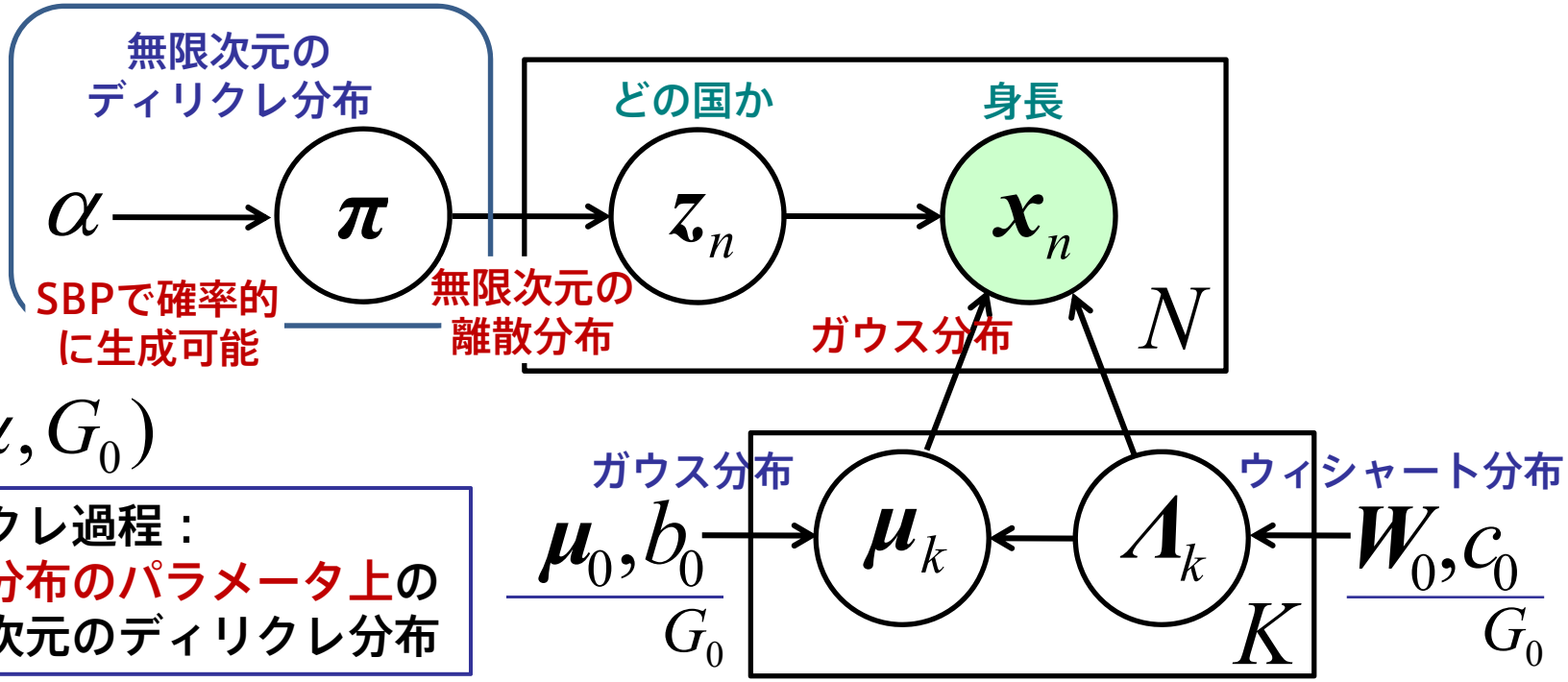
連続値でも離散値でもよい！

棒折り過程 (SBP)

- 無限個のパラメータを陽に生成
  - 例：N人の外国人を身長でクラスタリング
    - 従来の興味：国の数K　各国の平均と精度μ, Λ　混合比π

無限個存在　　無限次元

$$\Theta = \{\mu_1, \mu_2, \cdots, \Lambda_1, \Lambda_2, \cdots \pi\} \qquad K \to \infty$$



$\mathrm{DP}(\alpha, G_0)$

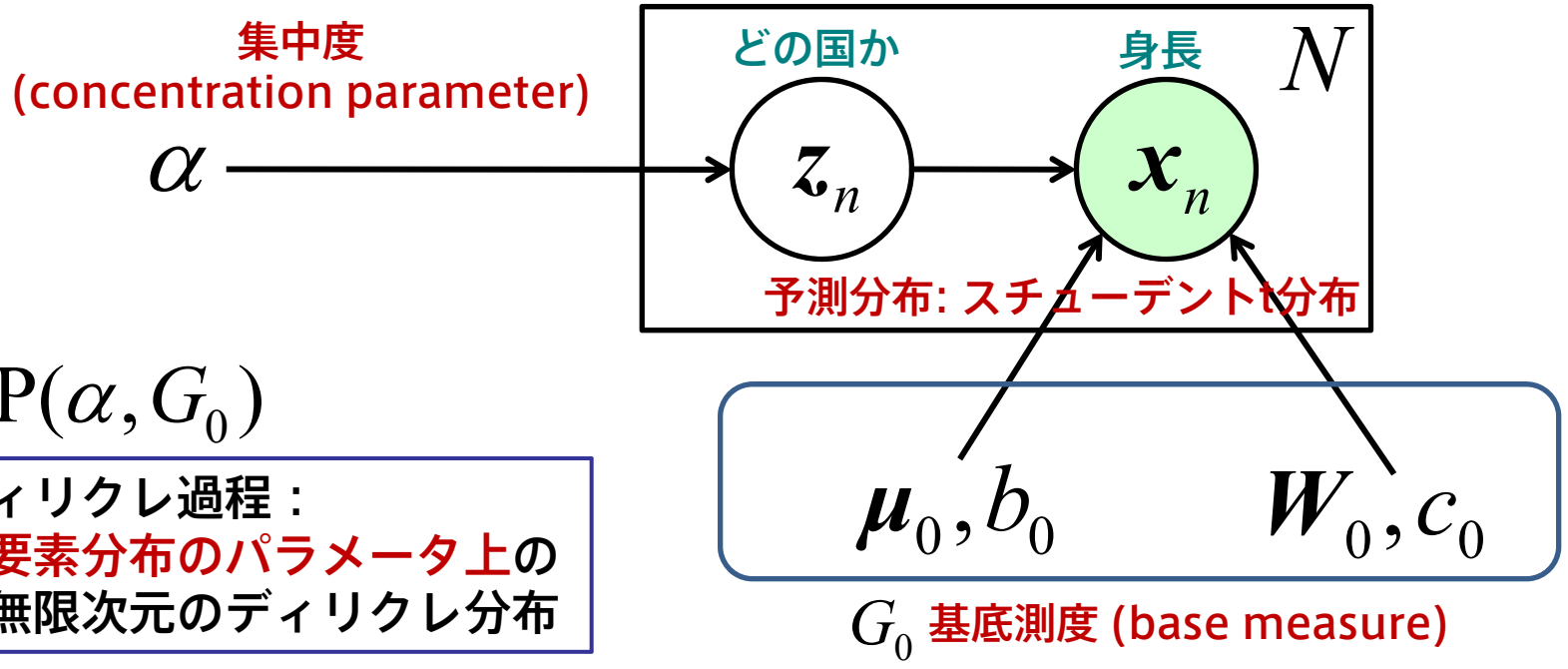
ディリクレ過程：
要素分布のパラメータ上の
無限次元のディリクレ分布

中華料理店過程 (CRP)

- パラメータを積分消去して直接サンプルを生成
  - 例：N人の外国人を身長でクラスタリング
    - 従来の興味：国の数 K　各国の平均と精度 $\boldsymbol{\mu}, \boldsymbol{\Lambda}$　混合比 $\boldsymbol{\pi}$

      → 無限個のパラメータを陽に考えなくてもよくなる！

$$\Theta = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \cdots \boldsymbol{\pi}\} \qquad K \to \infty$$



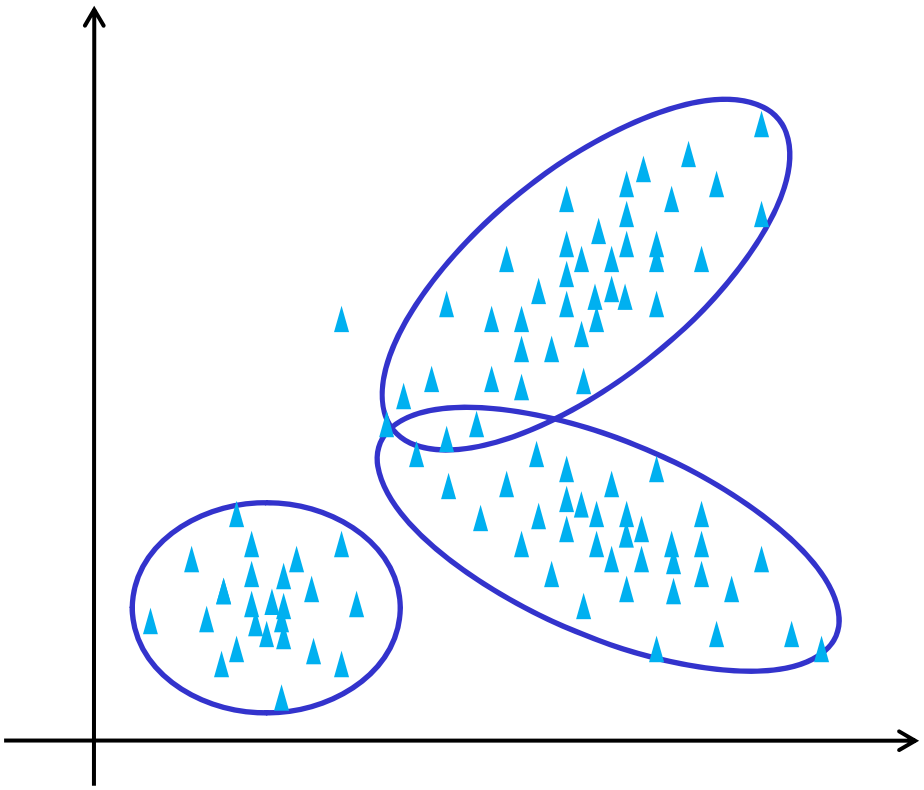
$\mathrm{DP}(\alpha, G_0)$

ディリクレ過程：
要素分布のパラメータ上の
無限次元のディリクレ分布

有限モデルから
無限モデルへの拡張

Infinite GMM

- 可算無限個のクラス (潜在変数の種類数) を許容するベイズ混合ガウスモデル
  - ディリクレ過程 (DP) を利用



有限混合GMM:
1. クラス数Kを変化させて大量のモデルを学習
2. AICやBICで適切なものをあとから選択

無限混合GMM:
一度の学習でクラス数の事後分布が求まる

Yes or No?

Q：ノンパラメトリックベイズモデルとは
その複雑さを観測データに合わせて
変化させることができるモデルである。

A: No!

ノンパラメトリックベイズモデルの複雑さは常に無限

- モデル選択問題（モデルの「真の」複雑さを決める問題）は
「解決した」のではなく「消滅した」
- 変化するのは「モデル自身の複雑さ」ではなく
「観測データに対する説明の複雑さ」
  - モデルの「実効的な」複雑さが変化するとは言える

無限混合の意味

- この世における森羅万象は無限
  - 無限個のクラスが存在すると考える

$$p(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k N(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$$

無限混合ガウス分布

- 我々が観測できるデータは有限
  - 高々有限個のクラスから構成されている
    - ほとんどのクラスの混合比πは極めて小さく観測データに出現しない
    - 潜在変数Zは、有限個のクラスにクラスタリングされている

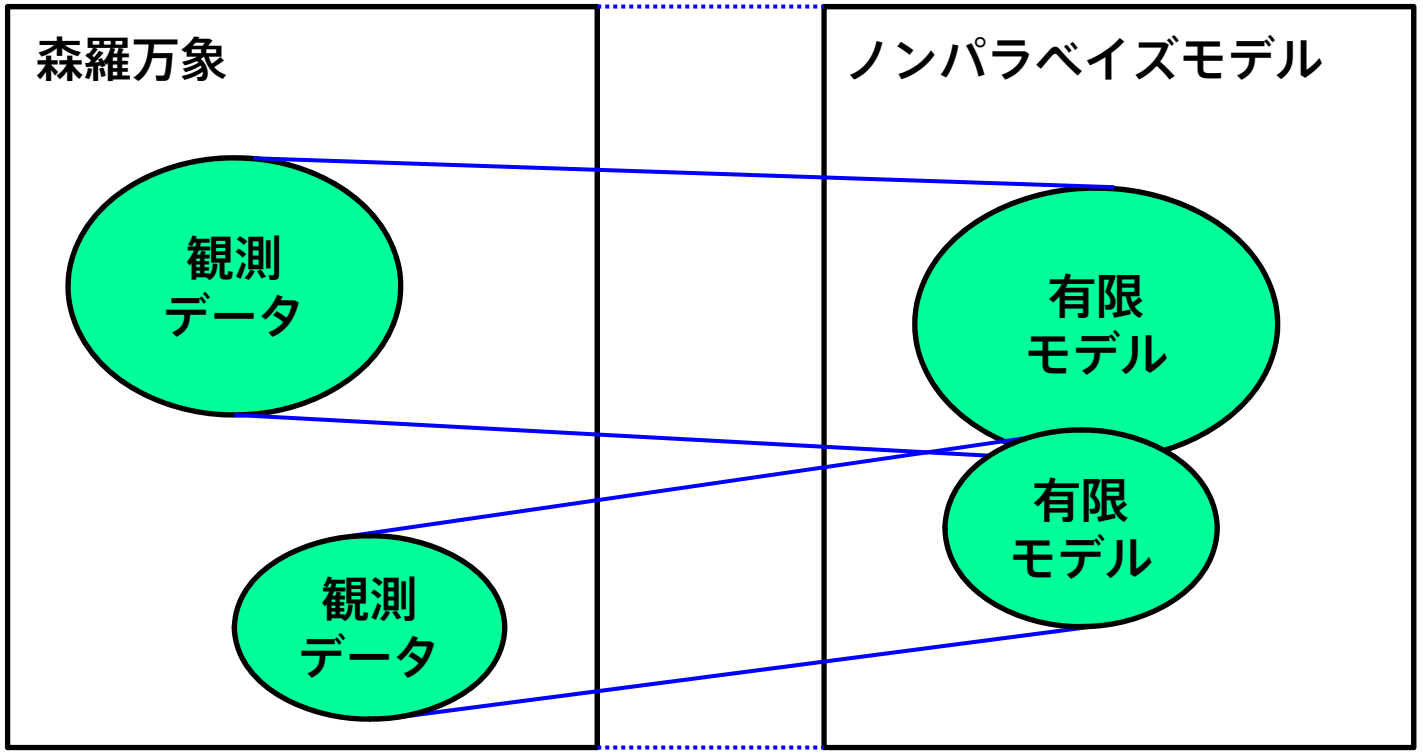
→ 有限混合分布を考えているのではない！

infinite と unbounded の違いを理解する

森羅万象の分割　　観測データの分割

無限混合の意味

- ノンパラメトリックベイズモデルはあらゆる複雑さの古典的なベイズモデルをすべて内包
  - 観測データを表現するために一部のみが利用される



事前分布の設計

- 有限混合ガウスモデルの場合
  - $\boldsymbol{\pi}$ の事前分布：ディリクレ分布
  - $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ の事前分布：ガウス・ウィシャート分布
- 無限混合ガウスモデルの場合
  - 無限次元の $\boldsymbol{\pi}$ の事前分布：ディリクレ過程
  - 無限個の $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ の事前分布：ガウス・ウィシャート分布

無限個の要素分布 $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k$ を確率的に生成できる
無限個の確率値 π_k は総和が1に正規化されている

$$p(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k N(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$$

有限面のサイコロ → 無限面のサイコロ
という置き換えで考えると理解しやすい

無限混合分布の生成

- ディリクレ過程 (DP)
  - 可算無限個の要素分布 θ を確率的に生成
  - 各要素分布に対する可算無限個の重み π も生成

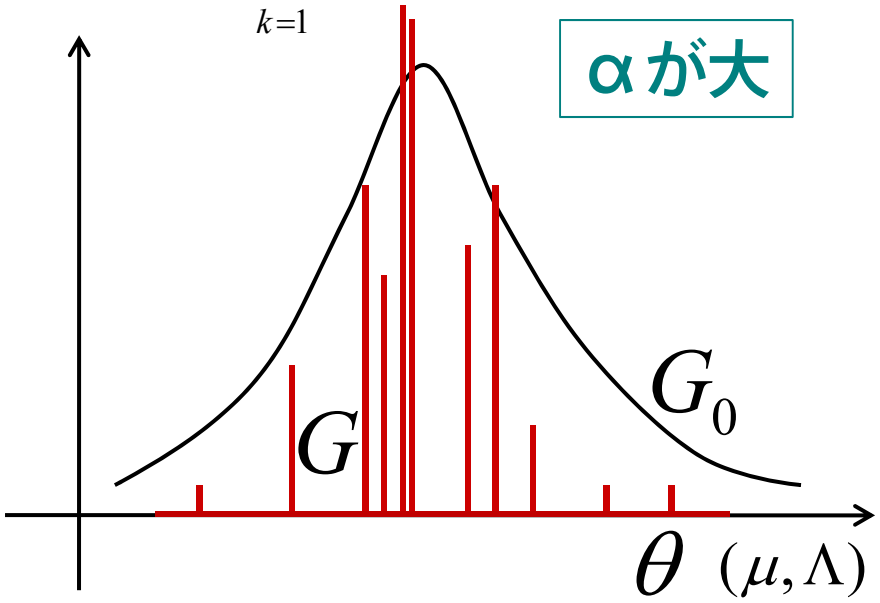
$$G \sim \mathrm{DP}(\alpha, G_0)$$

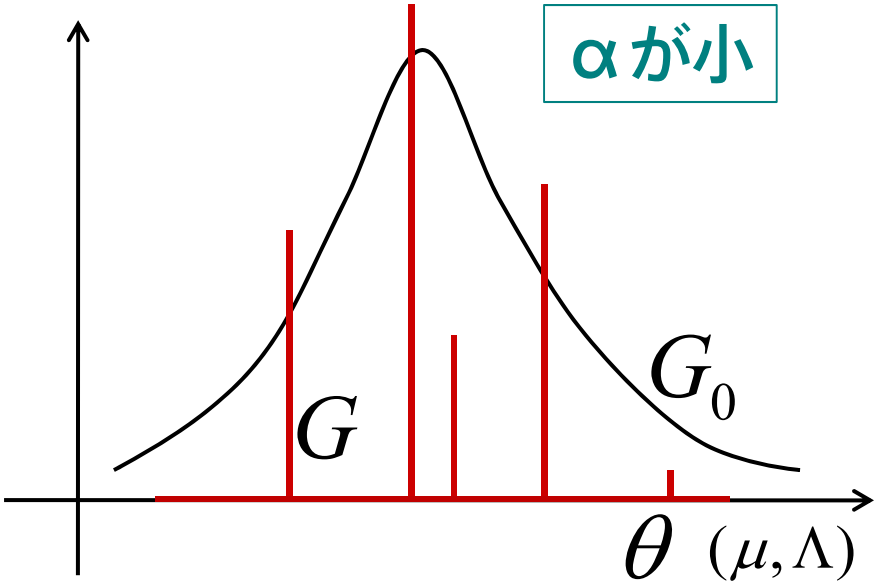
集中度　基底測度

G_0 が連続分布でも
G は離散分布になる

無限個のデルタ関数の和

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta - \theta_k)$$





具体的な実現方法

- Stick-Breaking Process (SBP)
  - 無限個の要素分布(森羅万象)をあらかじめ考えておき各潜在変数(観測データ)はそのうちの有限個に対応
    - 利点
      - 要素分布の確率分布が具体的でイメージしやすい
    - 欠点
      - 「無限」を明示的に考える必要がある
      - 計算機で扱うには打ち切り近似が必要 (VB)
- Chinese Restaurant Process (CRP)
  - 観測データに合わせて考えるべき要素分布を増減
    - 利点
      - 無限個の要素分布(パラメータ)を積分消去することで潜在変数上のクラスタリング結果のみを考えればよい
      - 近似は必要ない (MCMC/Gibbs Sampling)

無限個の混合比の生成

- 棒折り過程 (Stick-Breaking Process)
  - 長さ1の棒を無限回折りとっていく
  - どこで折りとるかは確率的に決まる

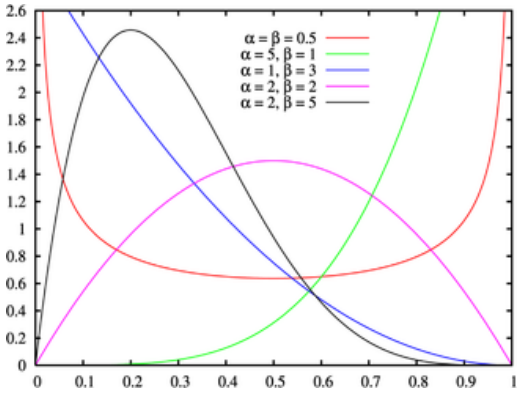
$$\tilde{\pi}_k \sim \mathrm{Beta}(1, \alpha)$$

平均的には
1:αで折る

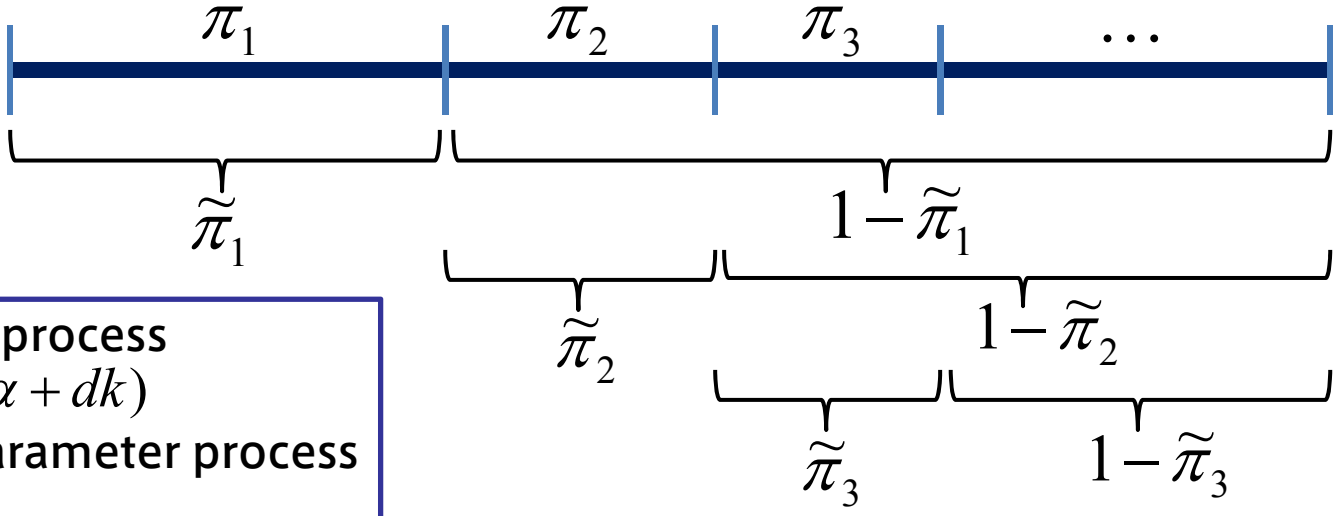
$$\pi_k = \tilde{\pi}_k \prod_{k'=1}^{k-1} (1 - \tilde{\pi}_{k'})$$

GEM分布

$\mathrm{Beta}(\alpha, \beta)$



[0 1]上の確率分布



DPの一般化

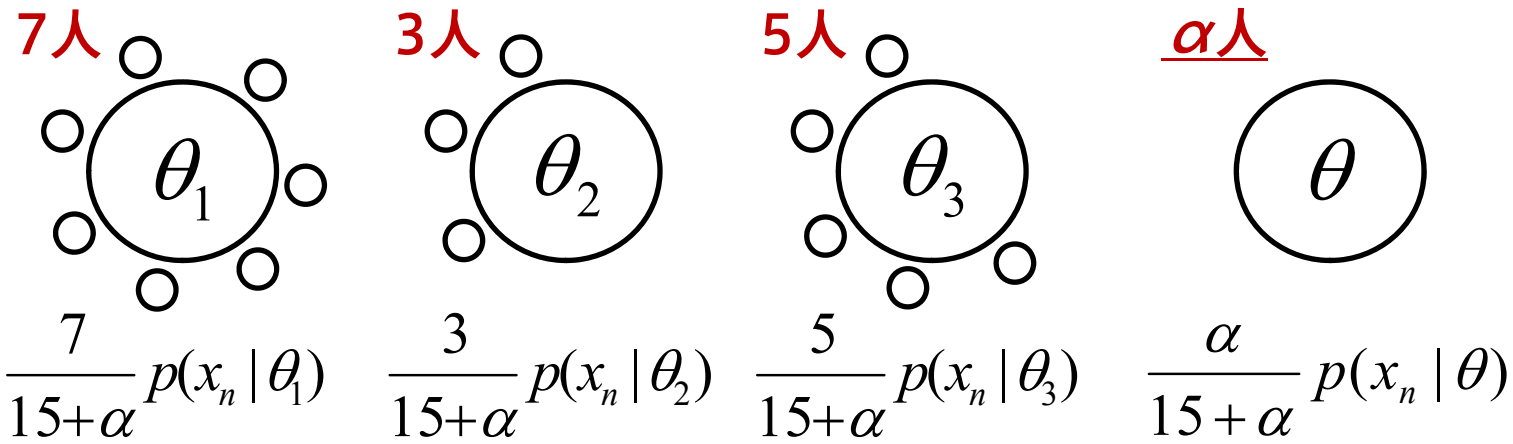
Pitman-Yor process
$\mathrm{Beta}(1-d, \alpha+dk)$

Beta two-parameter process
$\mathrm{Beta}(\alpha, \beta)$

サンプルの逐次生成

- 中華料理店過程 (CRP: Chinese Restaurant Process)
  - 客z_nより以前の客$z_{1:n-1}$がテーブルに着席しているときに客z_nがどのテーブルに座るか
    - 既存テーブルについて、同じ料理を食べる
    - 新規テーブルについて、料理を注文
  - テーブルが決まったらそこの料理θ_kを使って観測値x_nを出す

$$G \sim \mathrm{DP}(\alpha, G_0)$$
$$x, z \sim G$$

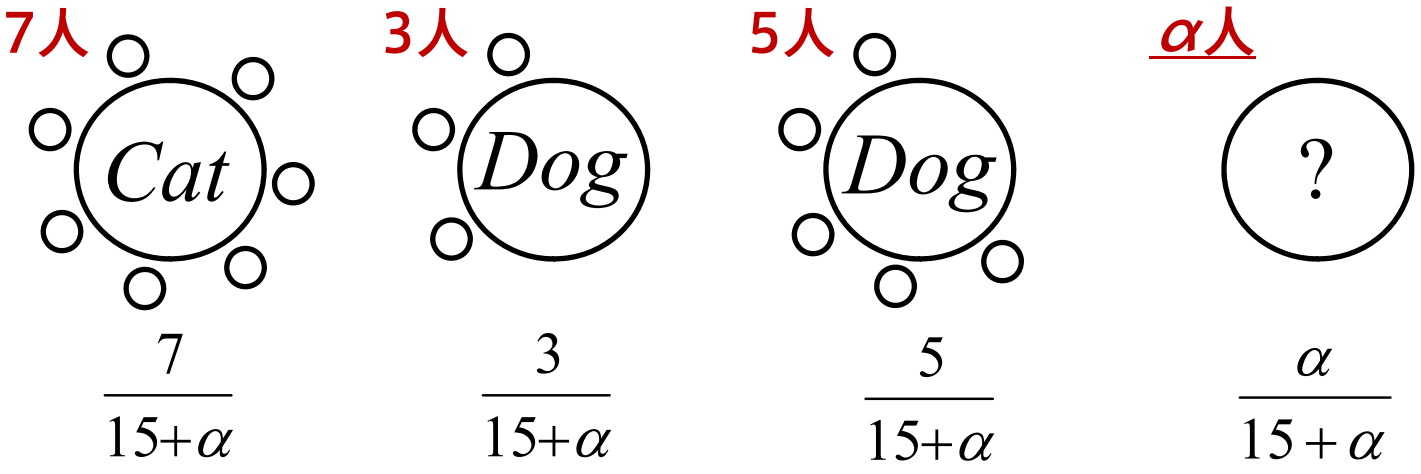


次の観測値x_nは上記の確率に従って決まる

料理が観測可能な場合

- N-gramモデルは混合モデルの一種とみなせる
  - 客z_nが食べた料理 (単語) は観測可能 $p(x_n \mid \theta_k) = \begin{cases} 1 & (x_n = \theta_k) \\ 0 & (\text{otherwise}) \end{cases}$
  - 客z_nが着いたテーブルは観測不可能
    - 同じ料理が複数のテーブルで出されていることあり
      - 基底測度が離散分布だと同じ料理が生成され得る
      - 参考：HDP, HPY

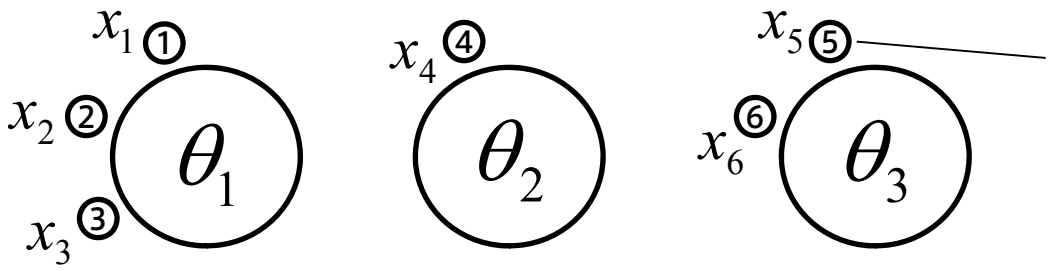
$p(z_n \mid Z_{-n})$ のみを考えればよいので単純



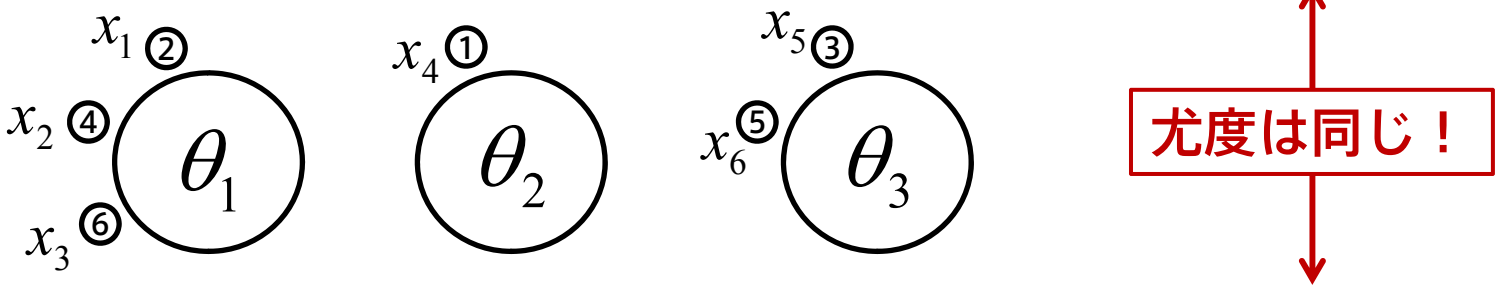
次の単語z_nは上記の確率に従って決まる

交換可能性

- サンプル群の尤度は生成順序に関係なし
  - 最終的な席の配置のみが重要
  - 混合分布で一般的な仮定 (サンプル生成はi.i.d.) に対応



$$p(X \mid \Theta) = \frac{\alpha}{\alpha} p(x_1 \mid \theta_1) \cdot \frac{1}{1+\alpha} p(x_2 \mid \theta_1) \cdot \frac{2}{2+\alpha} p(x_3 \mid \theta_1) \cdot \frac{\alpha}{3+\alpha} p(x_4 \mid \theta_2) \cdot \frac{\alpha}{4+\alpha} p(x_5 \mid \theta_3) \cdot \frac{1}{5+\alpha} p(x_6 \mid \theta_3)$$



$$p(X \mid \Theta) = \frac{\alpha}{\alpha} p(x_4 \mid \theta_2) \cdot \frac{\alpha}{1+\alpha} p(x_1 \mid \theta_1) \cdot \frac{\alpha}{2+\alpha} p(x_5 \mid \theta_3) \cdot \frac{1}{3+\alpha} p(x_2 \mid \theta_1) \cdot \frac{1}{4+\alpha} p(x_6 \mid \theta_3) \cdot \frac{2}{5+\alpha} p(x_3 \mid \theta_1)$$

データ生成とベイズ推論

- これまでは「データ生成」に関する説明
  - 観測データがどのような過程を経て生成されたか
    - 考え方1
      - 棒折り過程 (SBP) で無限個の混合比を生成
      - 無限個の混合比を用いて各サンプルを生成
    - 考え方2
      - 中華料理店過程 (CRP) でいきなりサンプルを生成
  - しかし最終結果である観測データ以外は未知
    - 確率的に生成されたパラメータの値やクラス割り当て

- したがって「ベイズ推論」したい
  - 観測データの生成後に未知の過程を推定したい
    - 後から振り返ってみてパラメータやクラスの割り当ては
      こうではないかと予想できるのではないか

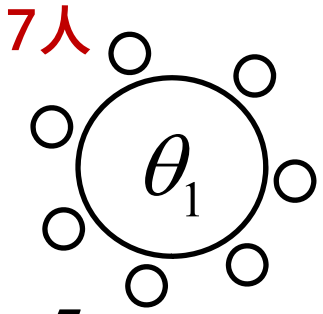
「事後的な」サンプルの逐次生成

- 中華料理店過程 (CRP: Chinese Restaurant Process)
  - 観測値x_nはすでに決定済み
  - 客z_n以外の着席がすでに判明している場合に客z_nがどのテーブルに座っていそうか
    - じつは、既存テーブルで、同じ料理を食べていた
    - じつは、新規テーブルで、料理を注文していた

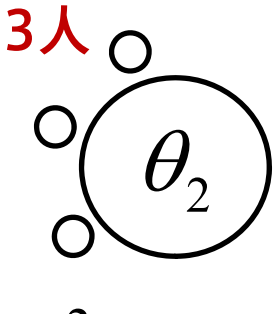
$$G \sim \mathrm{DP}(\alpha, G_0)$$
$$x, z \sim G$$

$$p(z_n \mid Z_{-n}, \theta) \propto p(x_n, z_n \mid Z_{-n}, \theta) = p(x_n \mid z_n, \theta) p(z_n \mid Z_{-n})$$

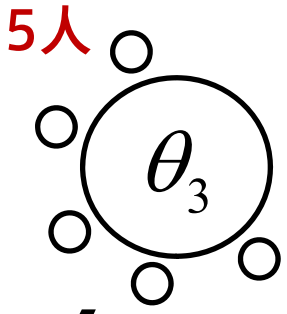
に比例する確率でテーブルに着席していたと予測



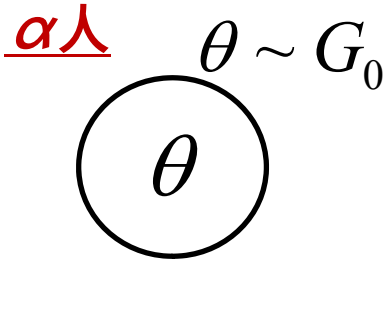
$$\frac{7}{15+\alpha} p(x_n \mid \theta_1)$$



$$\frac{3}{15+\alpha} p(x_n \mid \theta_2)$$



$$\frac{5}{15+\alpha} p(x_n \mid \theta_3)$$



$$\frac{\alpha}{15+\alpha} p(x_n \mid \theta)$$

ギブスサンプリング

- EステップとMステップを交互に実行
  - ある潜在変数の値を**自分以外の潜在変数およびパラメータの値を固定**した条件付き分布に従ってサンプル
    - 例：ガウス分布 x 離散分布に従ってサンプル

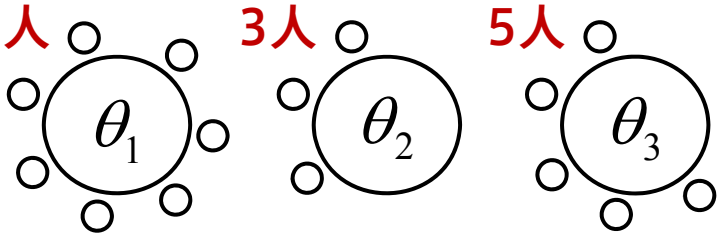
$$p(z_n \mid Z_{-n}, \theta) \propto p(x_n, z_n \mid Z_{-n}, \theta) = p(x_n \mid z_n, \theta) p(z_n \mid Z_{-n})$$

  - パラメータの値を**全ての潜在変数の値を固定**した条件付き分布に従ってサンプル
    - 例：ガウス・ウィシャート事後分布からサンプル

$$p(\theta_k \mid X, Z) \propto p(X \mid Z, \theta_k) p(\theta_k)$$

共役事前分布であれば解析的に求まる

全ての客のシーティングが既知であれば



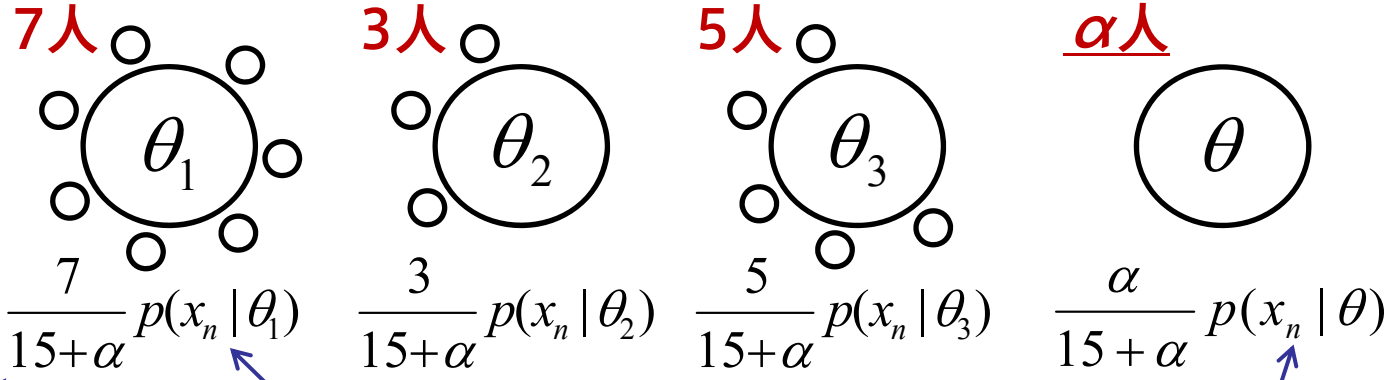
各Θ_kの事後分布を求めるのはたやすい(単純なベイズ推定)

「事後的な」サンプルの逐次生成

- 中華料理店過程 (CRP: Chinese Restaurant Process)
  - 観測値x_nはすでに決定済み
  - 客z_n以外の着席がすでに判明している場合に客z_nがどのテーブルに座っていそうか
    - じつは、既存テーブルで、同じ料理を食べていた
    - じつは、新規テーブルで、料理を注文していた

$$G \sim \mathrm{DP}(\alpha, G_0)$$
$$x, z \sim G$$

$$p(z_n \mid X, Z_{-n}) \propto p(x_n, z_n \mid X_{-n}, Z_{-n}) = p(x_n \mid Z, X_{-n}) p(z_n \mid Z_{-n})$$



$$\frac{7}{15+\alpha} p(x_n \mid \theta_1) \qquad \frac{3}{15+\alpha} p(x_n \mid \theta_2) \qquad \frac{5}{15+\alpha} p(x_n \mid \theta_3) \qquad \frac{\alpha}{15+\alpha} p(x_n \mid \theta)$$

他の客Z_{-n}のテーブルが決まっているときにどのテーブルに着くか

$$p(z_n = k \mid Z_{-n})$$

Θを事後分布で積分消去して予測分布に置き換え可能

$$p(x_n \mid z_n = k, Z_{-n}, X_{-n}) = \int p(x_n \mid \theta_k) p(\theta_k \mid Z_{-n}, X_{-n}) d\theta_k$$

客を観測していないテーブルではΘを事前分布で積分消去

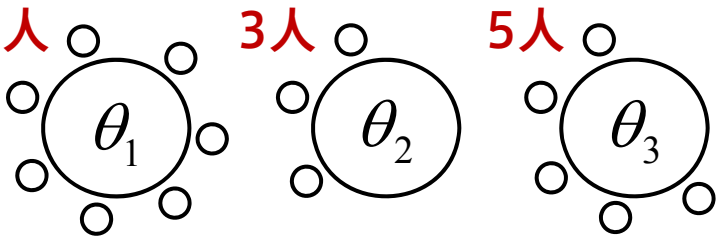
$$p(x_n \mid z_n = \mathrm{new}, Z_{-n}, X_{-n}) = \int p(x_n \mid \theta) G_0(\theta) d\theta$$

周辺化ギブスサンプリング

- Eステップのみを実行
  - ある潜在変数の値を**自分以外の潜在変数の値を固定**した条件付き分布に従ってサンプル

$$p(z_n \mid X, Z_{-n}) \propto p(x_n, z_n \mid X_{-n}, Z_{-n}) = p(x_n \mid Z, X_{-n}) p(z_n \mid Z_{-n})$$

全ての客のシーティング

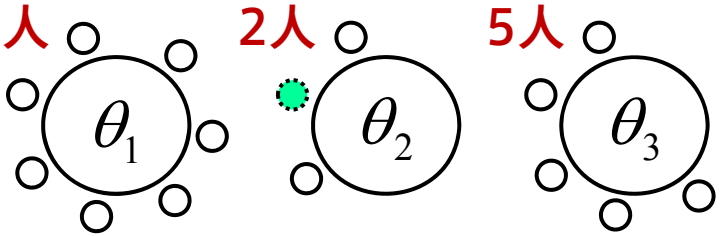


Exchangeability

交換可能性：客順は無関係！

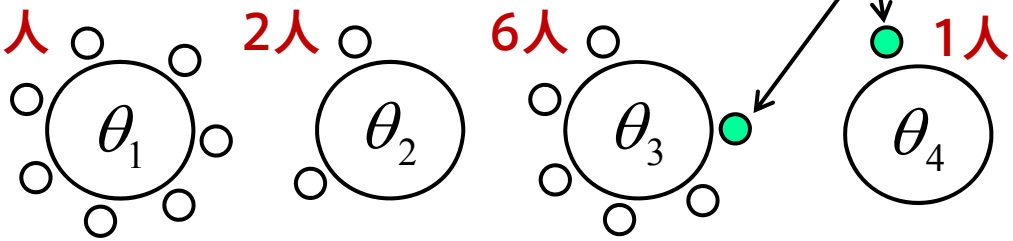
サンプルしたい客が一番最後に到着したと考えてよい

Remove customer



既存テーブルのいずれかに着くか新規テーブルに着くかは確率的に決まる

Add customer



ランダムな客順で収束するまで反復

ベイズ推論手法

- 確率変数の(同時)事後分布は通常、解析的に求めることは困難なので反復解法を用いる
  - 変分ベイズ法 (VB)
    - 収束は高速だが近似解しか得られない
      - 事後分布の関数形を制限する
      - 打ち切り近似を入れる
        - » 無限個の要素分布を考える必要があるため
    - 共役なモデルでないと事後分布が導出できない
      - 階層モデルへの適用は困難
  - マルコフ連鎖モンテカルロ (MCMC)
    - 無限の時間があれば真の事後分布に収束する
      - 複雑なモデルでも局所解に陥りにくい
      - 無限個の要素分布を考える必要はない

モデルの複雑さが上がるにつれMCMCが好まれる傾向

Infinite HMM

- 可算無限個の状態(と出力シンボル)を許容するベイズ隠れマルコフモデル [Beal2001]
  - 階層ディリクレ過程 (HDP) を利用
  - ギブスサンプリング
    - 遷移確率も出力確率も積分消去
    - 各時刻の潜在変数を反復的にサンプル
      - 十分時間をかければ真の事後分布に収束
      - HMMは隣同士の相関が強くて収束が非常に遅い
  - ビームサンプリング [VanGael2007]
    - 問題：状態数が無限なので動的計画法が実行できない
      - かといって打ち切り近似はしたくない・・・
    - 解決法：スライスサンプリングを組み合わせる
      - ある閾値以下の確率の遷移を無視する
      - 閾値も確率的にサンプルすることで真の事後分布に収束

実験結果

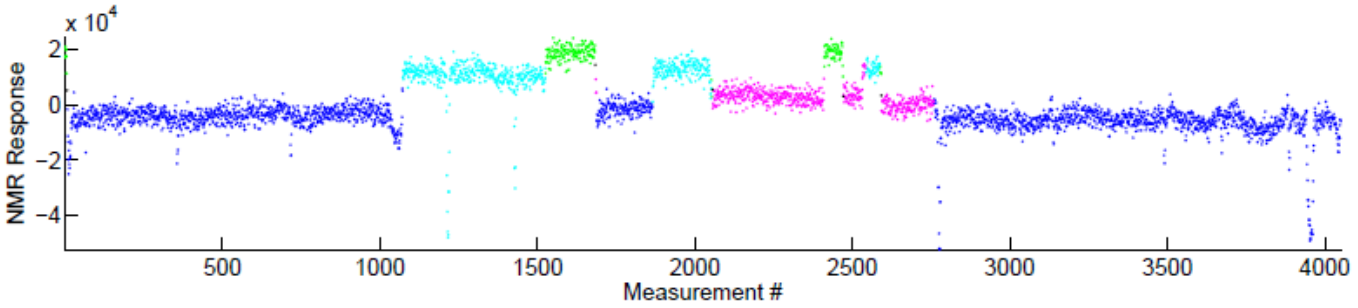


Figure 5. The 40'th sample of the beam sampler with every state represented by a different color on the well-log dataset.

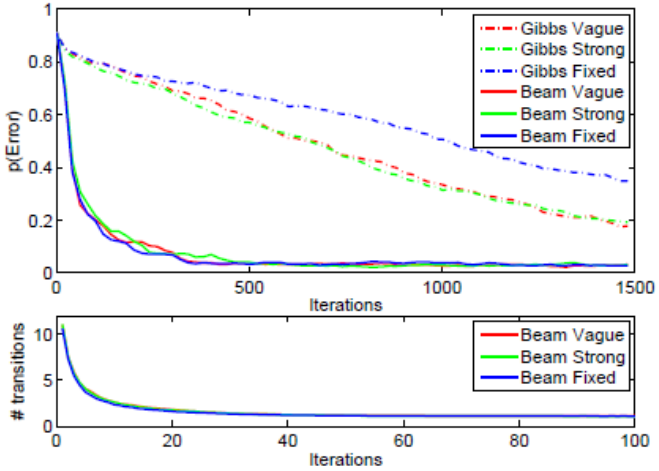


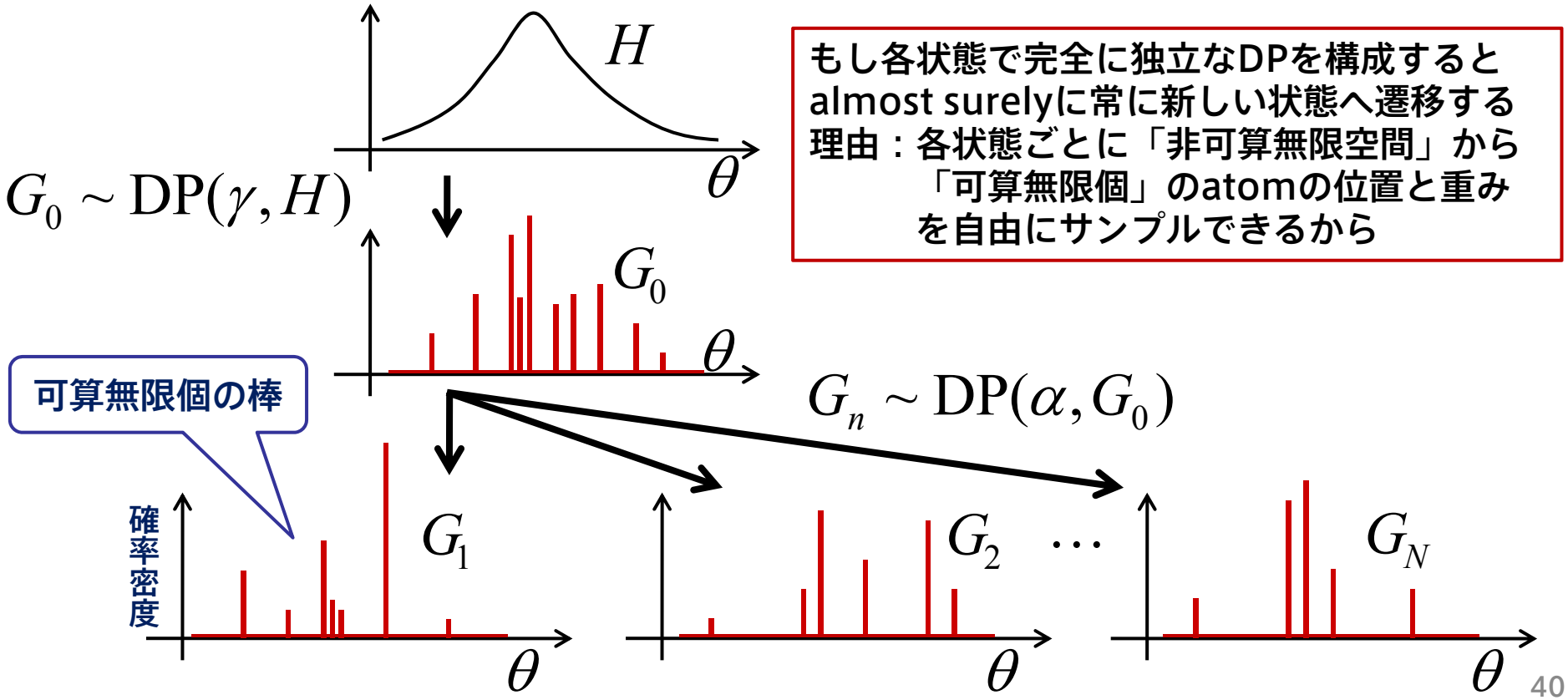
Figure 3. iHMM performance on strong negatively correlated data. The top plot shows the error of the Gibbs and beam sampler for the first 1500 iterations averaged over 20 runs. The bottom plot shows the average number of previous states considered in equation (4) for the first 100 iterations of the beam sampler.

ギブスサンプラーより
ビームサンプラーの方が
はるかに収束が速い

個人的な予備実験では
ビームサンプラーでも
収束が遅いことも多々ある
↓
現実的には有限打ち切りで
近似した方が効率的

階層ディリクレ過程

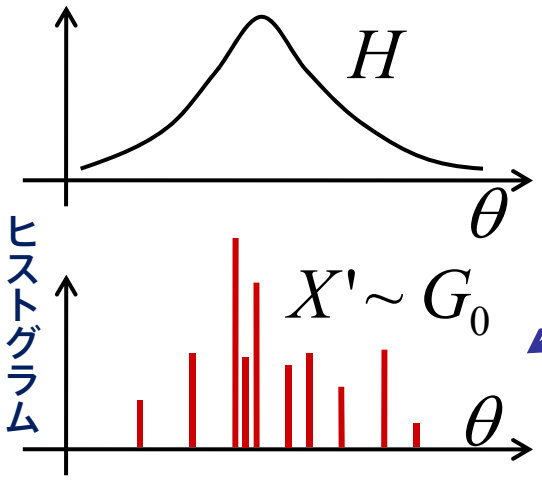
- 各状態が遷移先の状態を共有するためには
  ディリクレ過程を階層化すればよい [Teh2006]
  - 各状態kのディリクレ過程の基底測度G_0を
    グローバルなディリクレ過程とする



もし各状態で完全に独立なDPを構成すると
almost surelyに常に新しい状態へ遷移する
理由：各状態ごとに「非可算無限空間」から
「可算無限個」のatomの位置と重み
を自由にサンプルできるから

階層ディリクレ過程からのサンプル

- 中華料理店フランチャイズ (CRF) [Teh2004]
  - 客(サンプル)はまず下位のレストラン G_n に入る
    - 既存テーブルについて料理 θ を食べる
    - 新規テーブルについて代理客を上位レストラン G_0 に送り込む
      - 既存テーブルについて料理を下位レストランへ持ち帰る
      - 新規テーブルについて新しい料理を H から開発

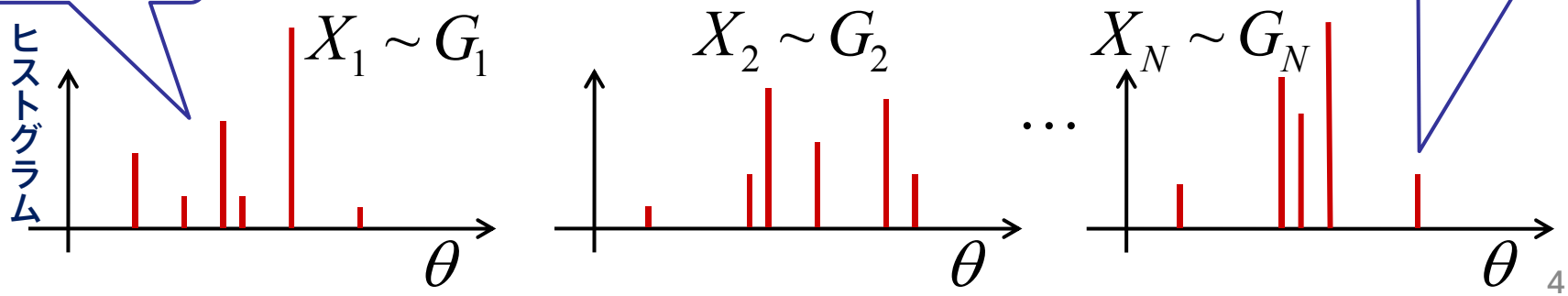


上位レストランの客
＝下位レストランのテーブル

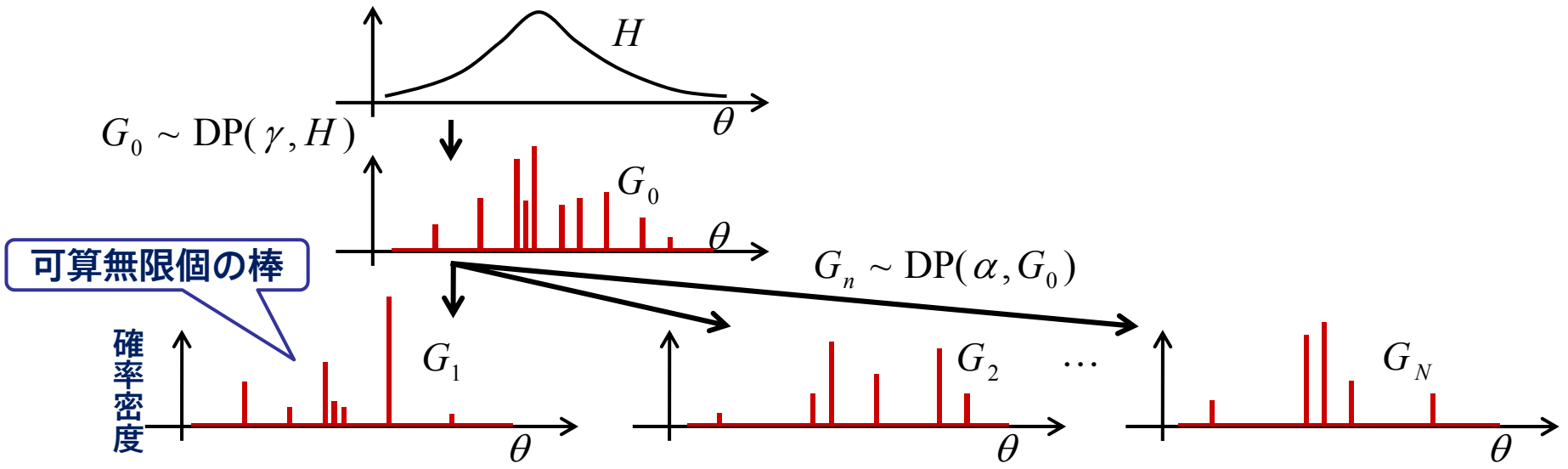
下位のヒストグラムで棒が立っている位置は
上位のヒストグラムで必ず棒が立っている

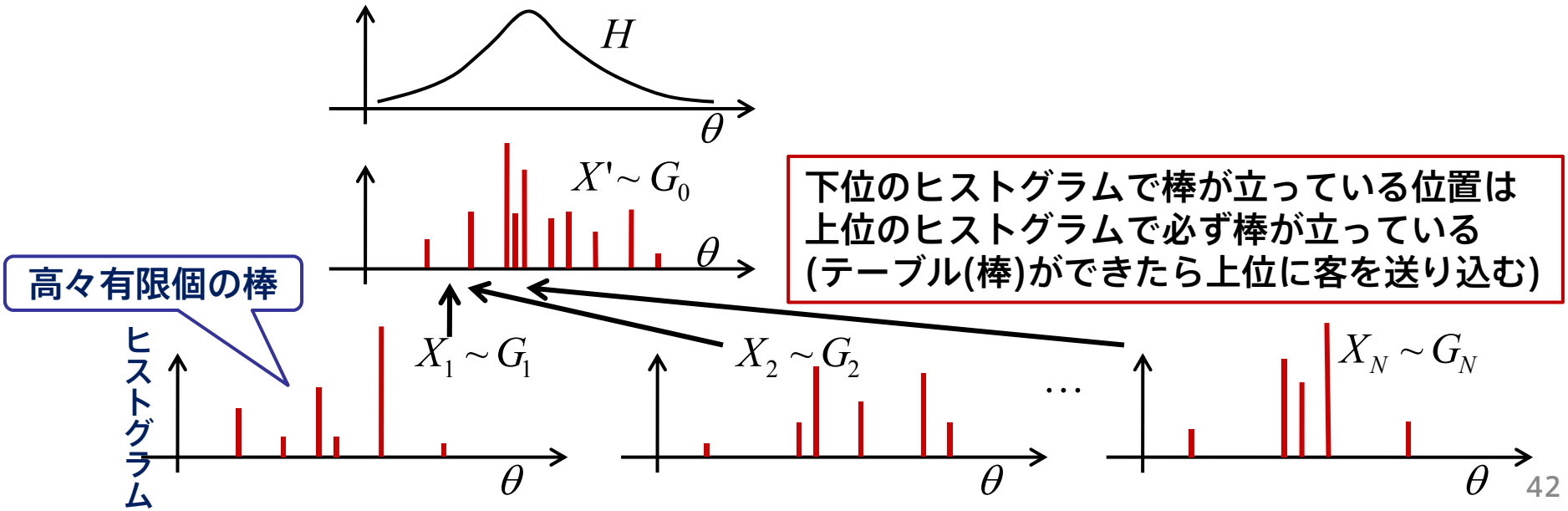
高々有限個の棒

同じ料理θを食べている客が
複数の異なるテーブルに
ついていることがある



確率過程とサンプル


H
θ
$G_0 \sim \mathrm{DP}(\gamma, H)$
G_0
θ
可算無限個の棒
$G_n \sim \mathrm{DP}(\alpha, G_0)$
確率密度
G_1
G_2
…
G_N
θ
θ
θ
θ


H
θ
$X' \sim G_0$
θ
下位のヒストグラムで棒が立っている位置は
上位のヒストグラムで必ず棒が立っている
(テーブル(棒)ができたら上位に客を送り込む)
高々有限個の棒
ヒストグラム
$X_1 \sim G_1$
$X_2 \sim G_2$
…
$X_N \sim G_N$
θ
θ
θ
θ

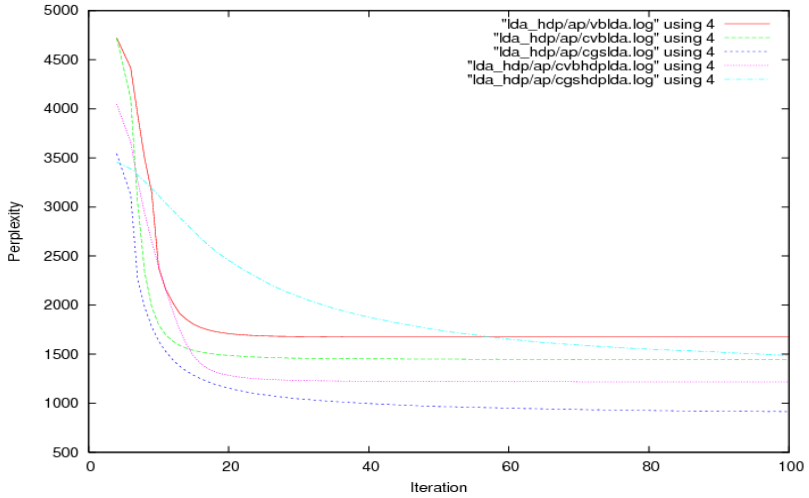
潜在的ディリクレ配分法

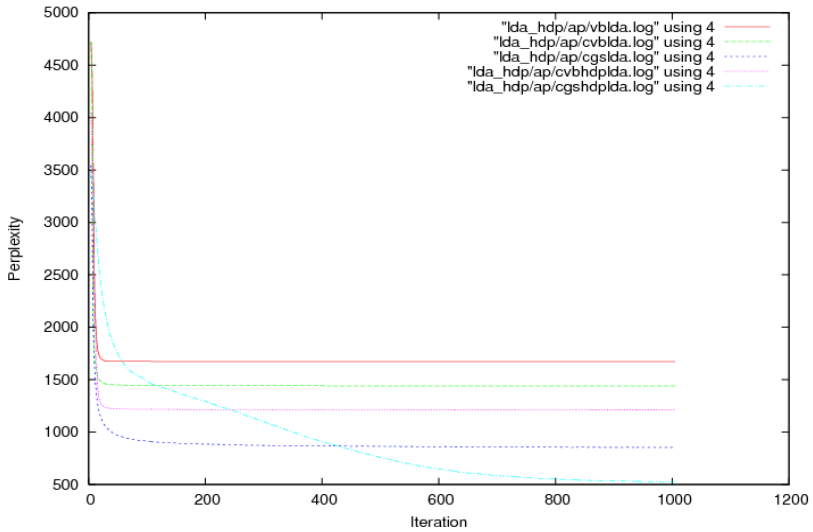
- 最もポピュラーなトピックモデル
  - PLSI: 尤度を最大化するパラメータを最尤推定
    - 各文書におけるトピック分布
    - 各トピックにおける単語分布
  - LDA: ディリクレ事前分布を導入してベイズ推定
    - 各文書におけるトピック分布の事後分布
    - 各トピックにおける単語分布の事後分布
  - HDP-LDA: トピック数を無限化
    - 文書間で無限個のトピック分布を共有する必要あり
    - 階層ディリクレ過程が必要
- 多様な推論手法
  - 変分ベイズ法・周辺化変分ベイズ法・Power EP
  - 周辺化ギブスサンプラ

HDP-LDAを実装してみた

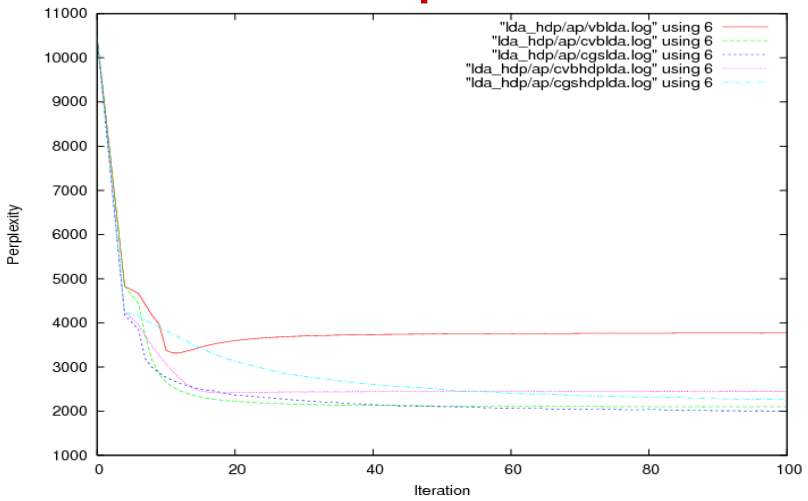
- D. Bleiのページの APデータ: 2246文書 10473単語
- 90%(のべ392231単語)で学習・10%(のべ43607単語)を評価

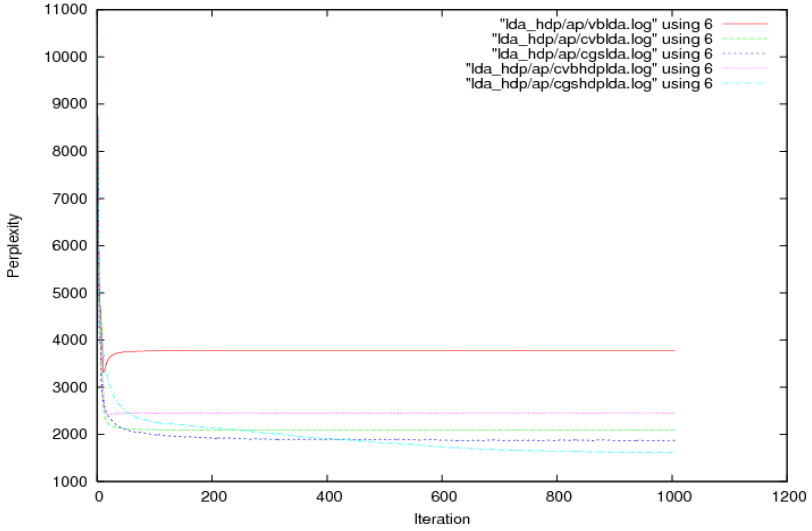
closed





open





確率的文脈自由文法

- シンボルの導出規則に確率が付与されたもの
  - 予めチョムスキー標準系に変換しておく

S → NP VP	1.0
PP → P NP	1.0
VP → V NP	0.7
VP → VP PP	0.3
P → *with*	1.0
V → *saw*	1.0

NP → NP PP	0.4
NP → *astronomers*	0.1
NP → *ears*	0.18
NP → *saw*	0.04
NP → *stars*	0.18
NP → *telescopes*	0.1

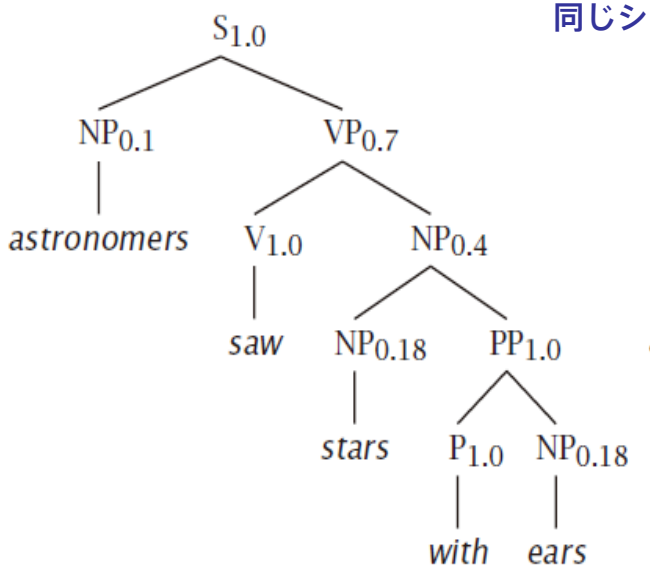
同じシンボルで開始する規則の確率の総和は1

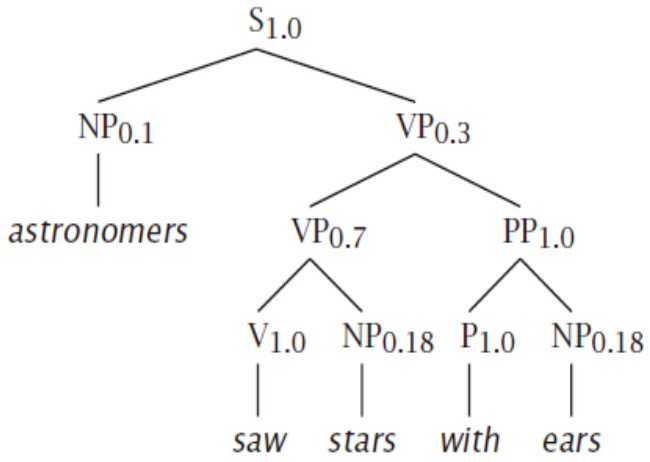
PCFGの最尤学習：
単語列が与えられた時に
各導出規則の確率を求める

PCFGのベイズ学習：
各シンボルごとに導出確率の
ディリクレ事後分布を求める

[栗原2004]

t_1:





$$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$$

$$P(t_2) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$$

$$P(w_{15}) = P(t_1) + P(t_2) = 0.0015876$$

Infinite PCFG

- 無限個のシンボルおよび導出規則を許容するベイズ文脈自由文法 [Liang2007]
  - 階層ディリクレ過程 (HDP) を利用
  - 手動で導出規則を与えなくてよい
    - 必要なシンボル・必要な導出規則が自動的に生成

有限モデル

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → *astronomers*	0.1
VP → V NP	0.7	NP → *ears*	0.18
VP → VP PP	0.3	NP → *saw*	0.04
P → *with*	1.0	NP → *stars*	0.18
V → *saw*	1.0	NP → *telescopes*	0.1

無限モデル

S1 → S1 S2

S1 → S1 S5

S2 → S2 S3

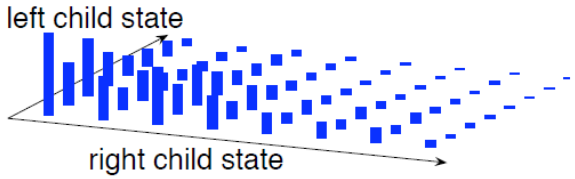
⋮

1. シンボルを無限に生成

$\beta \sim \mathrm{GEM}(\alpha)$

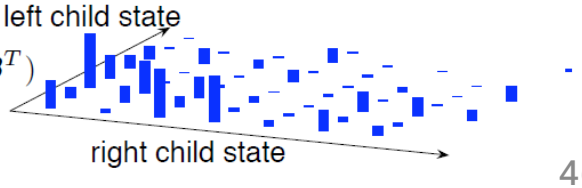
2. 導出規則の右側を無限に生成

$\beta\beta^T$



3. 導出規則をサンプル

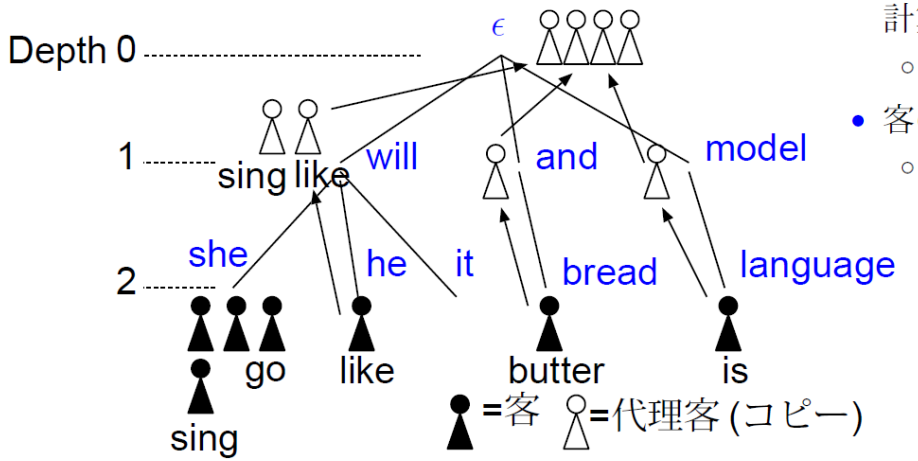
$\phi_z^B \sim \mathrm{DP}(\beta\beta^T)$



HPYLM: 固定長n-gram

- 高精度な階層的ベイズスムージングが施されたn単語連鎖の言語モデル [Teh2006]
  - 実験的に最高性能と言われるInterpolated Kneser-Ney (IKN)スムージングはHPYLMの近似
  - 階層Pitman-Yor過程 (HPY) を利用

- n グラム分布は, 深さ $(n-1)$ の Suffix Tree で表せる
- 例として, トライグラムを考える

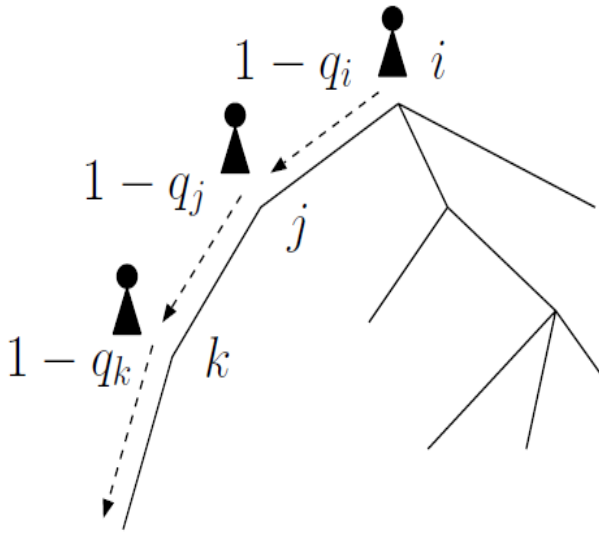


- ‘she will’→‘sing’ を予測…木を $\epsilon \rightarrow$ will $\rightarrow$ she の順にたどる
- 止まった, 深さ 2 のノード (トライグラム) から, sing の確率を計算

- ノードの持つ客 (単語カウント) の分布から, p(sing|she will) を計算 $\rightarrow$ p(like|she will) はどうする?
  - ‘like’ のカウントがない
- 客のコピー (代理客) を上のノードに確率的に送る
  - ‘he will like’ から送られた上のノードの客 ‘like’ を使って, バイグラムと補間して確率を計算

VPYLM: 可変長n-gram

- 高精度な階層的ベイズスムージングが施された可変長単語連鎖の言語モデル [持橋2007]
  - 完全なベイズ生成モデル
    - あらかじめ高次n-gramを学習して枝刈りする従来の方法は可変長モデルの構成意図と矛盾
  - N-gramのNが積分消去されて∞-gram
    - 頻度の高い(低い)単語ほど確率的に深く(浅く)



- 客 (カウント) を, 木のルートから確率的にたどって追加
- ノード i に, そこで止まる確率 q_i ($1-q_i$:「通過確率」) がある
  - q_i は, ランダムにベータ事前分布から生成される

$$q_i \sim \mathrm{Be}(\alpha, \beta)$$

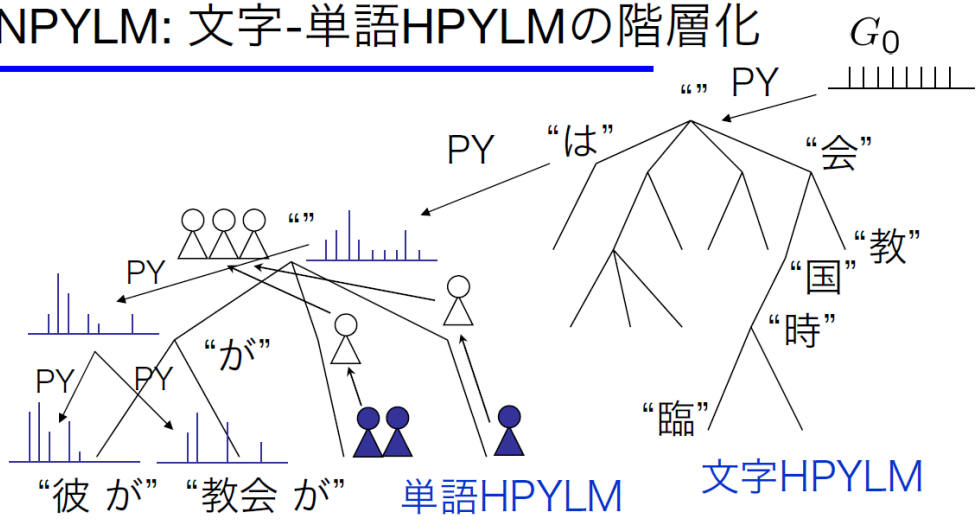
  - ゆえに, 深さ n のノードで止まる確率は

$$p(n|h) = q_n \prod_{i=0}^{n-1} (1-q_i)\,.$$

教師なし形態素解析

- 学習データなしで文を形態素に分割 [持橋2009]
  - 単語系列に内在する規則性を自己組織的に学習
    - 言語として最も自然な単語分割を見つける
    - 未知の言語でも適用可能
  - 単語gramと文字gramを階層ベイズ的に統合
    - 単語3-gram＋文字∞-gram
    - あらゆる単語に確率を与えるので未知語が存在しない

NPYLM: 文字-単語HPYLMの階層化



Gibbs Samplingと単語分割

1 神戸では異人館 街の 二十棟 が破損した 。
2 神戸 では 異人館 街の 二十棟 が破損した 。
10 神戸 では 異人館 街の 二十棟 が破損した 。
50 神戸 で は異人 館 街 の 二 十 棟 が 破損 し た 。
100 神戸 で は 異 人館 街 の 二 十 棟 が 破損 し た 。
200 神戸 で は 異人館 街 の 二 十 棟 が 破損 し た 。

- ギブスサンプリングを繰り返すごとに、単語分割とそれに基づく言語モデルを交互に改善していく。

VPYLM/NPYLMの実装

- テンプレートを使うとエレガントに実装可能

```
#ifdef ENABLE_NPYLM
  // NPYLM言語モデルの生成　(二階層)
  UNILM<char> *lm0;
  VPYLM<char UNILM<char> > *lm1;
  VPYLM<std::string, VPYLM<char, UNILM<char> > > *lm2;
#endif

#ifdef ENABLE_VPYLM
  // VPYLM言語モデルの生成 (一階層)
  UNILM<std::string> *lm1;
  VPYLM<std::string, UNILM<std::string> > *lm2;
#endif

#ifdef ENABLE_HPYLM
  // HPYLM言語モデルの生成 (一階層)
  UNILM<std::string> *lm1;
  VPYLM<std::string, UNILM<std::string> > *lm2;
#endif
```

VPYLM/NPYLMのテスト

- C++のテンプレート機能を活用
  - boost::random
  - boost::serialization
  - boost::flyweight
- 実験結果
  - データ
    - Alice in Wonderland
    - 語彙数：2598 (34^6=1544804416) 文字数34
  - VPYLM (2598単語)
    - 学習セットPPL：18.7241 評価セットPPL：99.3449
  - VPYLM (1544804416単語)
    - 学習セットPPL：18.6393 評価セットPPL：184.288
  - NPYLM (無限語彙 & 34文字)
    - 学習セットPPL：18.8436 評価セットPPL：149.166

NPYLMのサンプル

iteration = 100
char_customers = 15593
char_restaurants = 1955
gamma_a = 15601, gamma_b = 2599, lambda = 5.983822
discount[0] = 0.333540, strength[0] = 0.298315, order[0] = 2609
discount[1] = 0.382044, strength[1] = 1.670189, order[1] = 3386
discount[2] = 0.553658, strength[2] = 1.834798, order[2] = 5855
discount[3] = 0.024605, strength[3] = 2.064980, order[3] = 3201
discount[4] = 0.122095, strength[4] = 1.251715, order[4] = 481
discount[5] = 0.490955, strength[5] = 0.305375, order[5] = 53
discount[6] = 0.712990, strength[6] = 0.644234, order[6] = 8
discount[7] = 0.635828, strength[7] = 1.259089, order[7] = 0
discount[8] = 0.785585, strength[8] = 0.328446, order[8] = 0
word_customers = 28116
word_restaurants = 8994
gamma_a = 28124, gamma_b = 1848, lambda = 15.154497
discount[0] = 0.591050, strength[0] = 0.243706, order[0] = 2023
discount[1] = 0.729518, strength[1] = 1.683618, order[1] = 16422
discount[2] = 0.048505, strength[2] = 5.294701, order[2] = 7995
discount[3] = 0.095503, strength[3] = 2.323097, order[3] = 1391
discount[4] = 0.877284, strength[4] = 0.835016, order[4] = 238
discount[5] = 0.638015, strength[5] = 0.341623, order[5] = 35
discount[6] = 0.558890, strength[6] = 0.194527, order[6] = 11
discount[7] = 0.907094, strength[7] = 0.049327, order[7] = 1
discount[8] = 0.681959, strength[8] = 0.313109, order[8] = 0

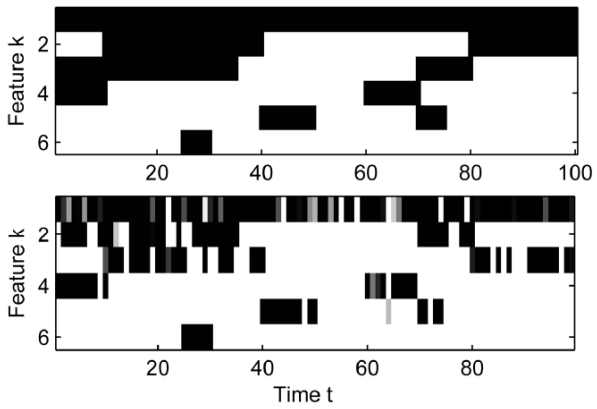
ベイズモデル実装のポイント

- C++ & Intel C++ Compiler を選択
- テンプレート機能をフル活用
  - boost::random
    - 様々な確率分布からのサンプル
  - boost::serialization
    - バイナリ形式・xml形式で書き出し
    - ポインタ・木構造・循環参照 すべて問題なし
  - boost::mpi
    - EMやVBと非常に相性が良い (MapReduce型推論)
    - 「周辺化」ギブスサンプラとはあまり相性が良くない
    - boost::serializationと組み合わせると神
  - boost::flyweight
    - インスタンスの共有を自動的にやってくれる
    - HPYLMにおいてメモリの大幅削減が可能

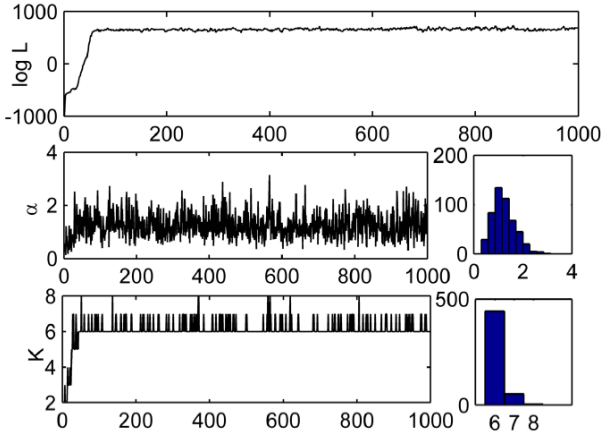
混合モデルから
因子モデルへ

Infinite ICA/SFA

- 可算無限個の素性を許容する独立成分分析およびスパース因子分析 [Knowles2007]
  - インド料理過程 (IBP) を利用
    - 各サンプルは**複数のクラス**に属してよい
    - 実現方法
      - ある客はそれまでの人気に比例して料理を複数選ぶ
      - さらに新しい料理にもチャレンジする



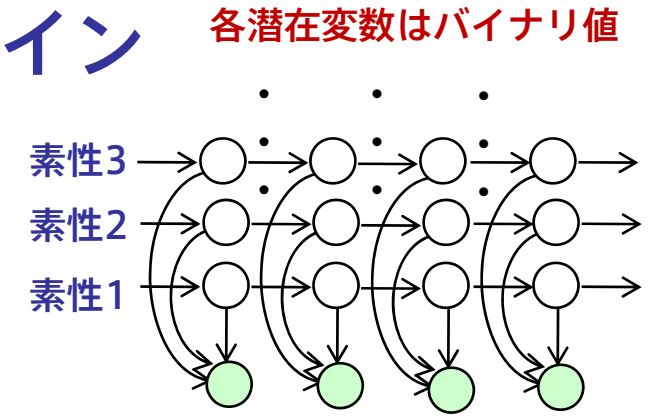
(a) *Top:* True $\mathbf{Z}$. *Bottom:* Inferred $\mathbf{Z}$

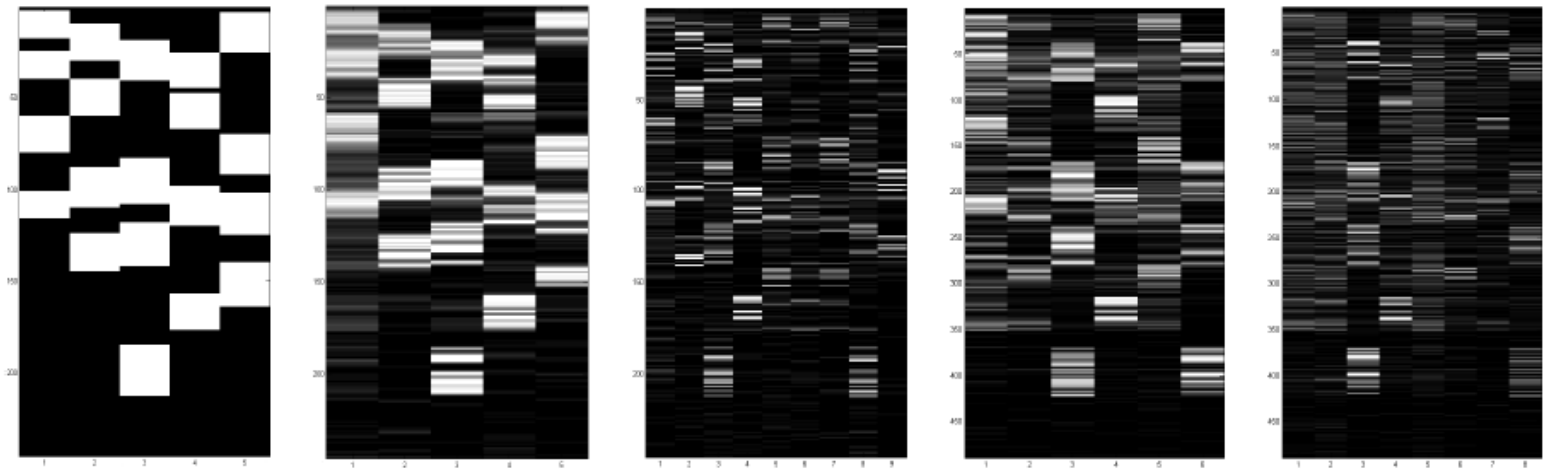


(b) Log likelihood, α and K^+ for duration of 1000 iteration run

Infinite Factorial HMM

- 可算無限本の隠れマルコフチェインを許容するHMM [VanGael2008]
  - インド料理過程 (IBP) を利用
  - ICAと組み合わせることが可能





(a) Ground Truth (b) ICA iFHMM (c) iICA (d) ICA iFHMM (e) iICA

Figure 4: Blind speech separation experiment; figures represent which speaker is speaking at a certain point in time: columns are speakers, rows are white if the speaker is talking and black otherwise. The left figure is ground truth, the next two figures in are for the 10 microphone experiment, the right two figures are for the 3 microphone experiment.

参考文献

- HDP
  - Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei: Hierarchical Dirichlet processes, JASA, *101(476):1566–1581, 2006.*
- HDP-LDAに対する周辺化VB
  - Y. W. Teh, K. Kurihara, M. Welling: Collapsed Variational Inference for HDP, NIPS 2007.
- HDP-LDAに対するVB（オンライン版含む）
  - C. Wang, J. Paisley, and D. M. Blei: Online Variational Inference for the Hierarchical Dirichlet Process, AISTATS 2011.
- iHMM
  - M. J. Beal, Z. Ghahramani, and C. E. Rasmussen: The Infinite Hidden Markov Model, NIPS 2002.
  - Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei: Hierarchical Dirichlet processes, JASA, *101(476):1566–1581, 2006.*
- iHMMに対するビームサンプラー
  - J. Van Gael, Y. Saatci, Y. W. Teh, and Z. Ghahramani: Beam Sampling for the Infinite Hidden Markov Model, ICML 2008.
- iPCFG
  - P. Liang , S. Petrov , M. I. Jordan , and D. Klein: The infinite PCFG using hierarchical Dirichlet processes, EMNLP 2007.
- iICA/iFA
  - D. Knowles, and Z. Ghahramani: Infinite Sparse Factor Analysis and Infinite Independent Components Analysis, ICA 2007.
  - J. Paisley and L. Carin: Nonparametric Factor Analysis with Beta Process Priors, ICML 2009.
- iFHMM
  - J. Van Gael , Y. W. Teh, and Z. Ghahramani: The Infinite Factorial Hidden Markov Model, NIPS 2008.