ノンパラメトリックベイズ

モデルの複雑さが不明な場合

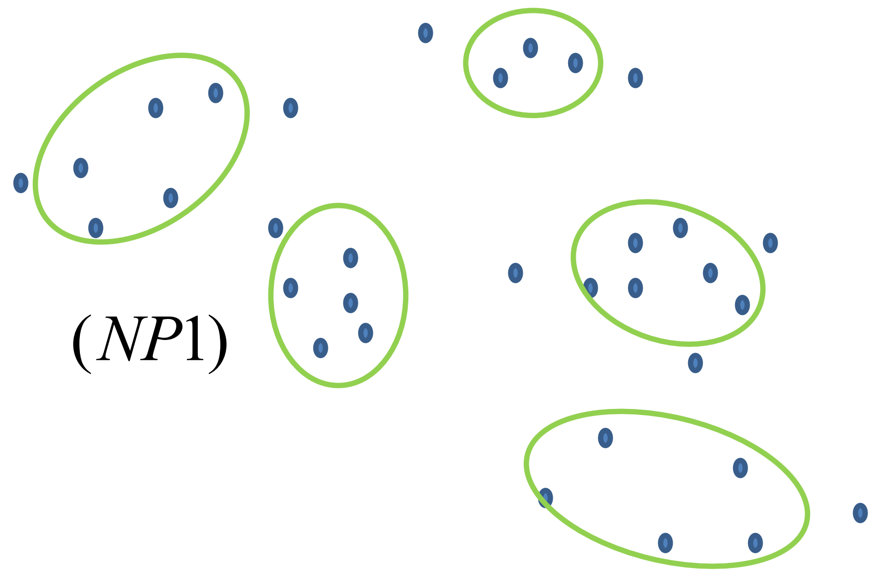
- これまでに説明してきたK-means、EM、変分ベイズなどは、モデルの複雑さ、たとえばクラスタリングにおけるクラスタ数は予め分かっているとしてモデル推定した。
- しかし、実際はクラスタ数が不明の場合が多い。

ノンパラメトリックとは

- 観測データに応じてモデル自体の複雑さも学習する
  - クラスタリングの場合は、クラスタ数が予め分かっていない場合。観測データに適したクラスタ数も推定

データから学習するということの直観

$$p(x \mid \theta, \pi) = \sum_{k=1}^{K} \pi_k N(x \mid \theta_k) \qquad (NP1)$$



- 有限個の正規分布を配分比π_kでの混合モデル:式($NP1$)
- クラスタ数Kの値を観測データから最適化することによって推定する。
- 基本的アイデア：無限次元の連続分布から観測データに適応した有限次元の離散分布を学習する。

無限次元からサンプリングして有限次元へ

$$G_0 \quad \Rightarrow \quad p(x \mid \theta, \pi) = \sum_{k=1}^{\boxed{\infty}} \pi_k N(x \mid \theta_k) \qquad (NP2)$$

$$\Rightarrow \quad p(x \mid \theta, \pi) = \sum_{k=1}^{\boxed{K}} \pi_k N(x \mid \theta_k) \qquad (NP1)$$

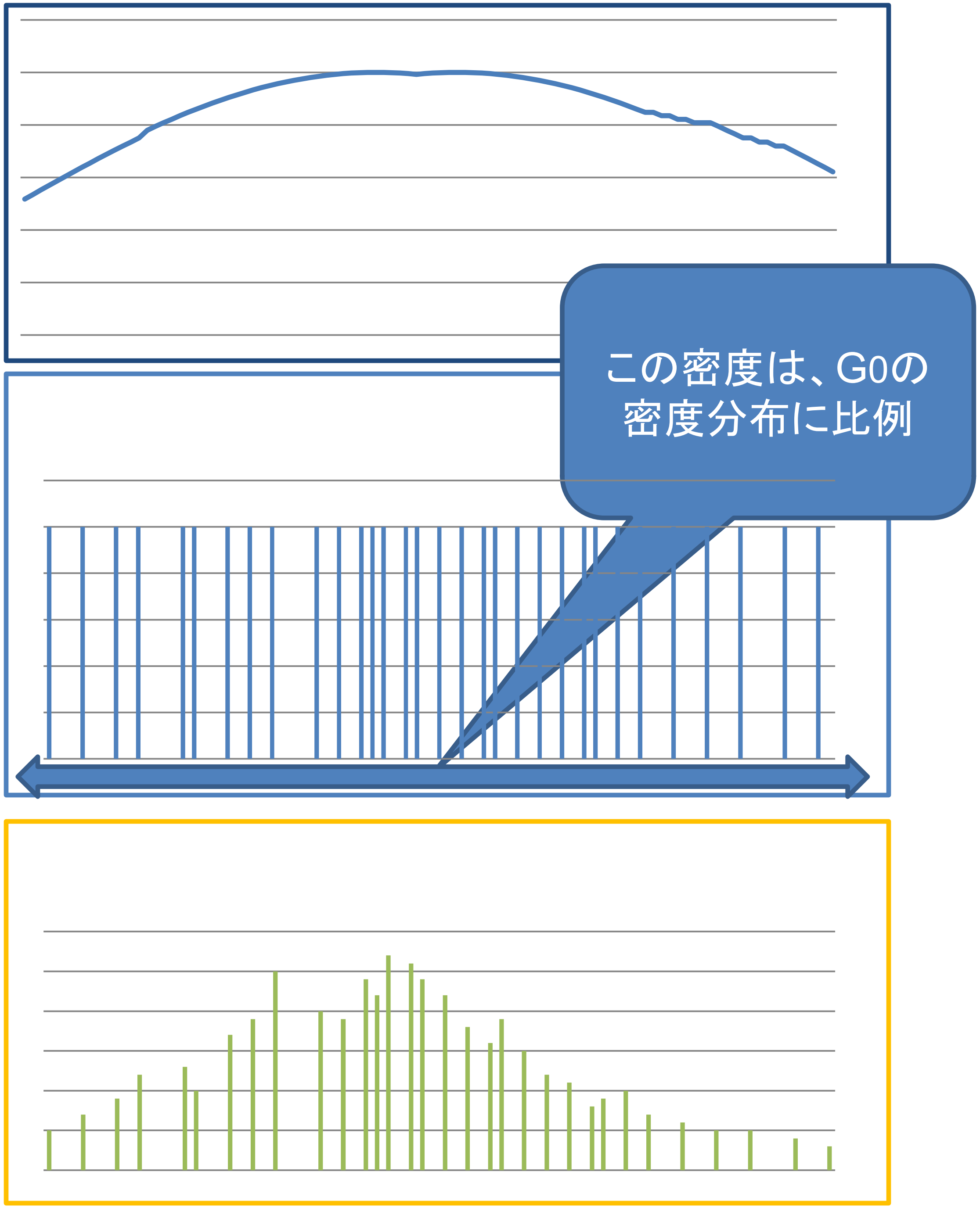
- 基底測度という連続の確率分布G_0から無限次元の離散基底分布($NP2$)で近似し、そこから有限次元の離散分布G($NP1$) を推定する。

連続ドメインの基底測度 G_0

連続ドメインの基底測度を
- 可算無限次元の離散分布 (NP2)

で近似して事前分布とする。

有限次元離散分布 (NP1)
- 観測データから推定



Chinese Restaurant Process

観測データが到着するごとに無限次分布*NP2*からクラスタを生成していくにあたって、以下の2者択一

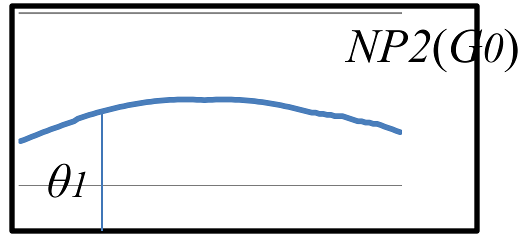
- 新しく生成　→ G_0 に近い
- 既にあるクラスタに追加 →観測データへの追従性

これによって、 G_0 に観測データを組み合わせた事後分布を得る
→Chinese Restaurant Process

G_0への近さを表すパラメターを集中度パラメター

(concentration papameter) α_0 とする。

イメージは次のスライドを参照



Z_1

θ_1

θ_2

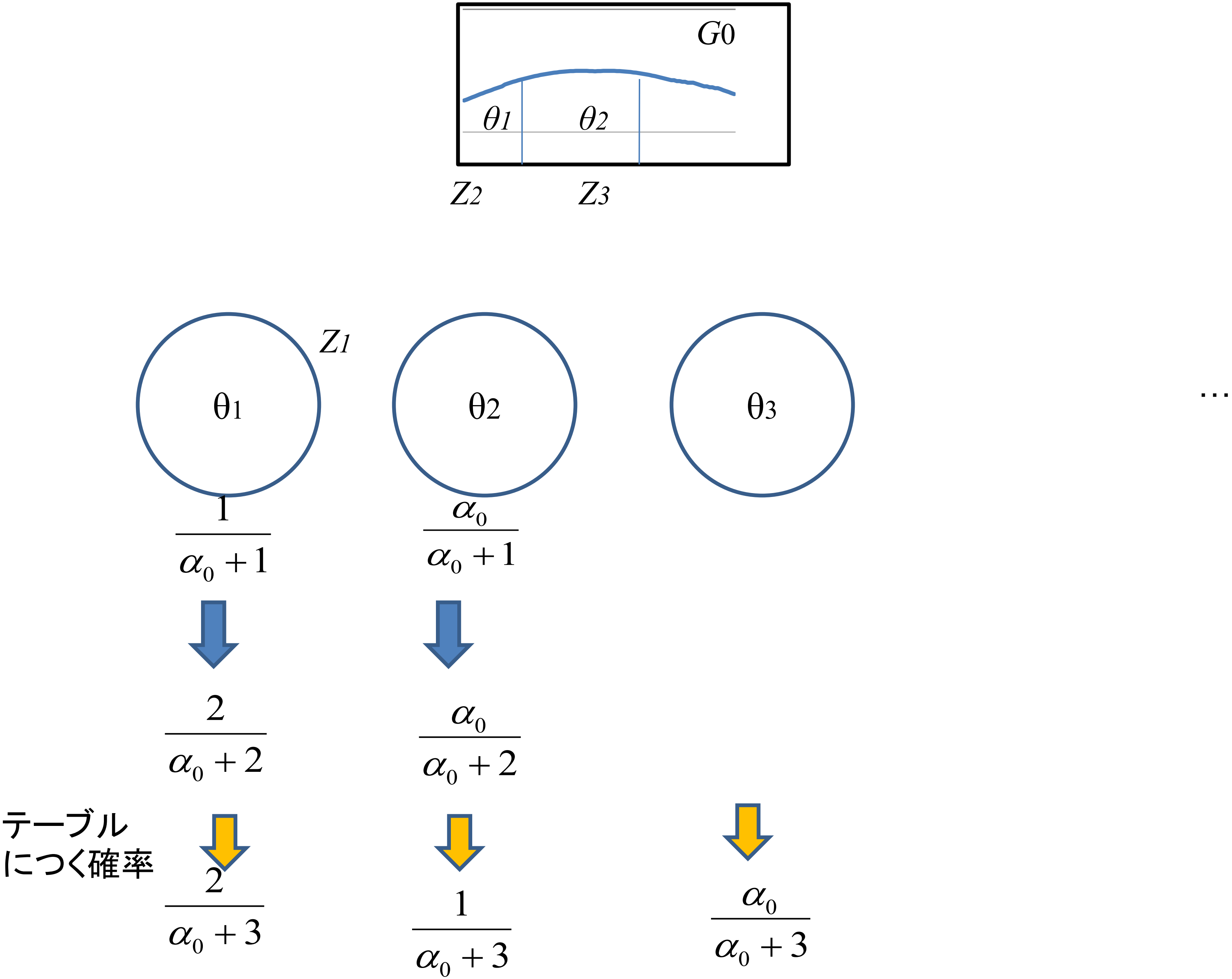
….

$$\frac{1}{\alpha_0+1}$$

1のテーブル
につく確率

$$\frac{\alpha_0}{\alpha_0+1}$$

残ったテーブルのうちの一つにつく確率（ただし、次の候補のテーブルはθ2に決まっているとする。）



$G0$
$\theta 1$
$\theta 2$
$Z2$
$Z3$
$Z1$
θ_1
θ_2
θ_3
....
$\frac{1}{\alpha_0+1}$
$\frac{\alpha_0}{\alpha_0+1}$
$\frac{2}{\alpha_0+2}$
$\frac{\alpha_0}{\alpha_0+2}$
テーブル
につく確率
$\frac{2}{\alpha_0+3}$
$\frac{1}{\alpha_0+3}$
$\frac{\alpha_0}{\alpha_0+3}$

N+1人目が k のテーブルに着く確率$p(Z_{N+1}= k|Z_1, \ldots, Z_N)$

N人目までに着席したテーブル数 $= K$

テーブルkに着席した人数をn_kとすると

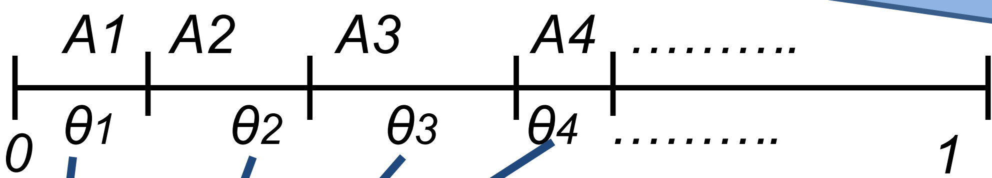
$$p(Z_{N+1} = k \mid Z_1, \ldots, Z_N) = \begin{cases} \dfrac{n_k}{\alpha_0 + N} & (k = 1, \ldots, K) \\ \dfrac{\alpha_0}{\alpha_0 + N} & (k = K + 1) \end{cases}$$

- α_0が大きいほど、無限次元の分布G_0に近い
- α_0が小さいほど、観測データに追従した事後分布となる
- テーブルをクラスタとみなすと、着席しているテーブルの数が観測データから推定されたクラスタ数

[Ferguson,1973]

- 基底測度G_0においてドメインΩ=[0,1]上の分割$\{A_i\}_{i=1}^{K}$の確率測度:$G(A_i)$

分割Aiをどのように作るかはまだ分かっていない



$$(G(A_1),G(A_2),\cdots,G(A_K)) \sim Dir(\alpha_0 G_0(A_1),\alpha_0 G_0(A_2),\cdots,\alpha_0 G_0(A_K))$$

α0G0(Ai)に比例する回数のsamplingによって離散分布化しxiを得る

$$G \sim DP(\alpha_0, G_0) \qquad G \propto \prod_{i=1}^{K} \theta_i^{x_i-1}, \qquad x_i \propto \alpha_0 G_0(A_i)$$

$$\theta \sim G \qquad \theta = (\theta_1,..,\theta_K)$$

xiを用いて、ディリクレ分布の確率変数θiの値を得る

$A \sim B(\phi)$は、Bという生成過程でパラメターϕを用いてAという分布が生成されるという意味

DPの構成定理

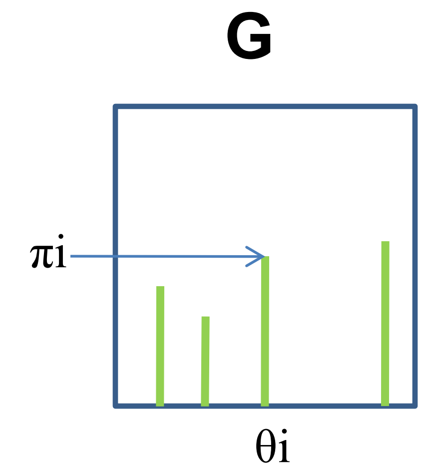
[Sethuraman,1994]

$G \sim DP(\alpha_0, G_0)$ は

分割Aiに対応する配分比πi

$\theta \sim G_0, \pi_k > 0, \sum_{k=1}^{\infty} \pi_k = 1$ を用いて

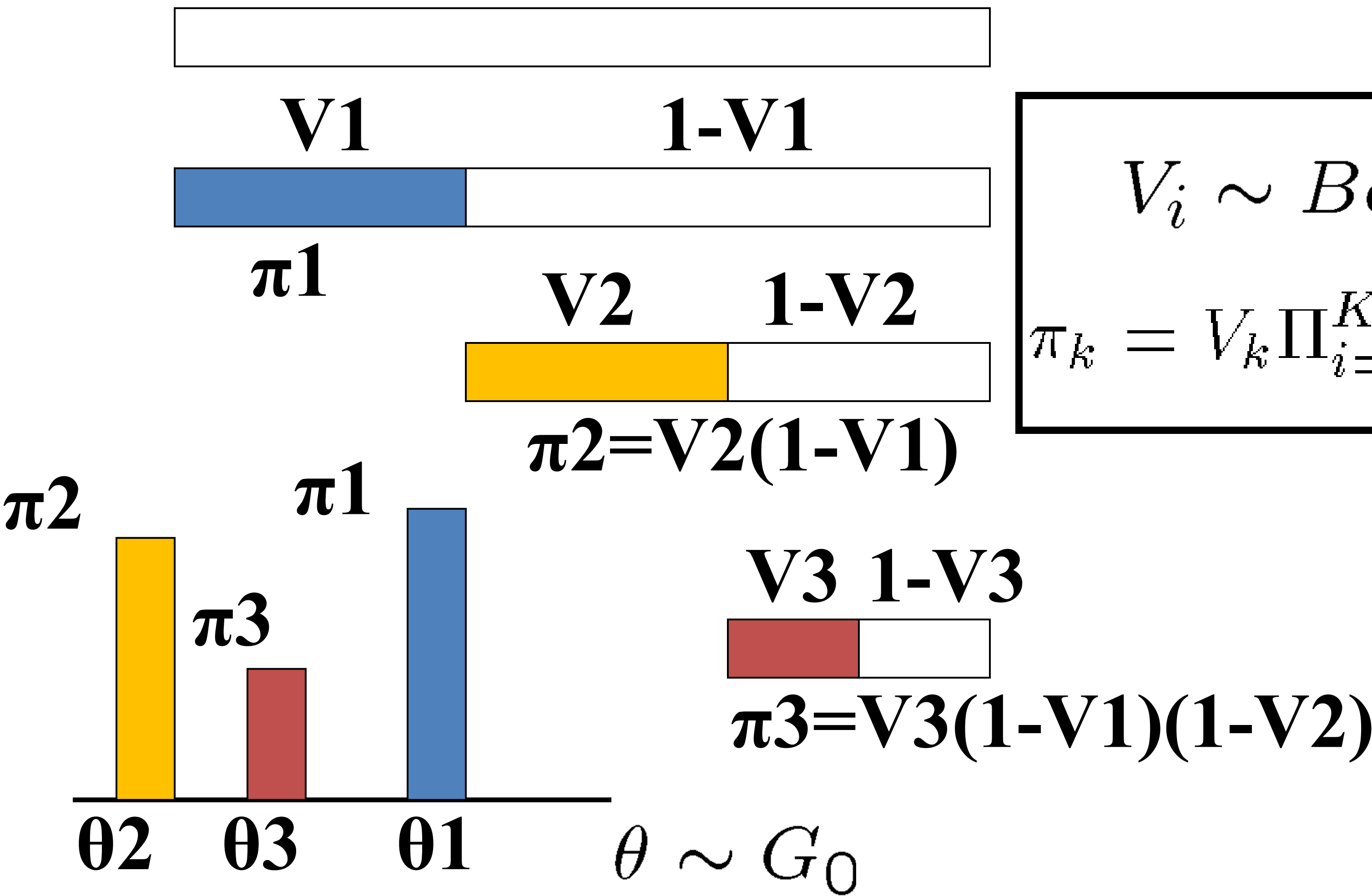
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^k}(\theta)$$ と構成可能



配分比πi は次に述べる
Stick-breaking Processで実現可能

Stick-breaking Process

- 長さ1の棒(stick)の切断(breaking)



$$V_i \sim Beta(1, \alpha)$$

$$\pi_k = V_k \Pi_{i=1}^{K-1}(1 - V_i)$$

Beta(1, α)

$$Beta(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

$$= Dirichlet(\mu_1,\mu_2 \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu_1^{\,a-1}\mu_2^{\,b-1}$$

$$\mu_1 + \mu_2 = 1$$

$$Beta(\mu \mid 1,\alpha) = \frac{\Gamma(1+\alpha)}{\Gamma(1)\Gamma(\alpha)}(1-\mu)^{\alpha-1}$$

連続ドメインの基底測度 G_0

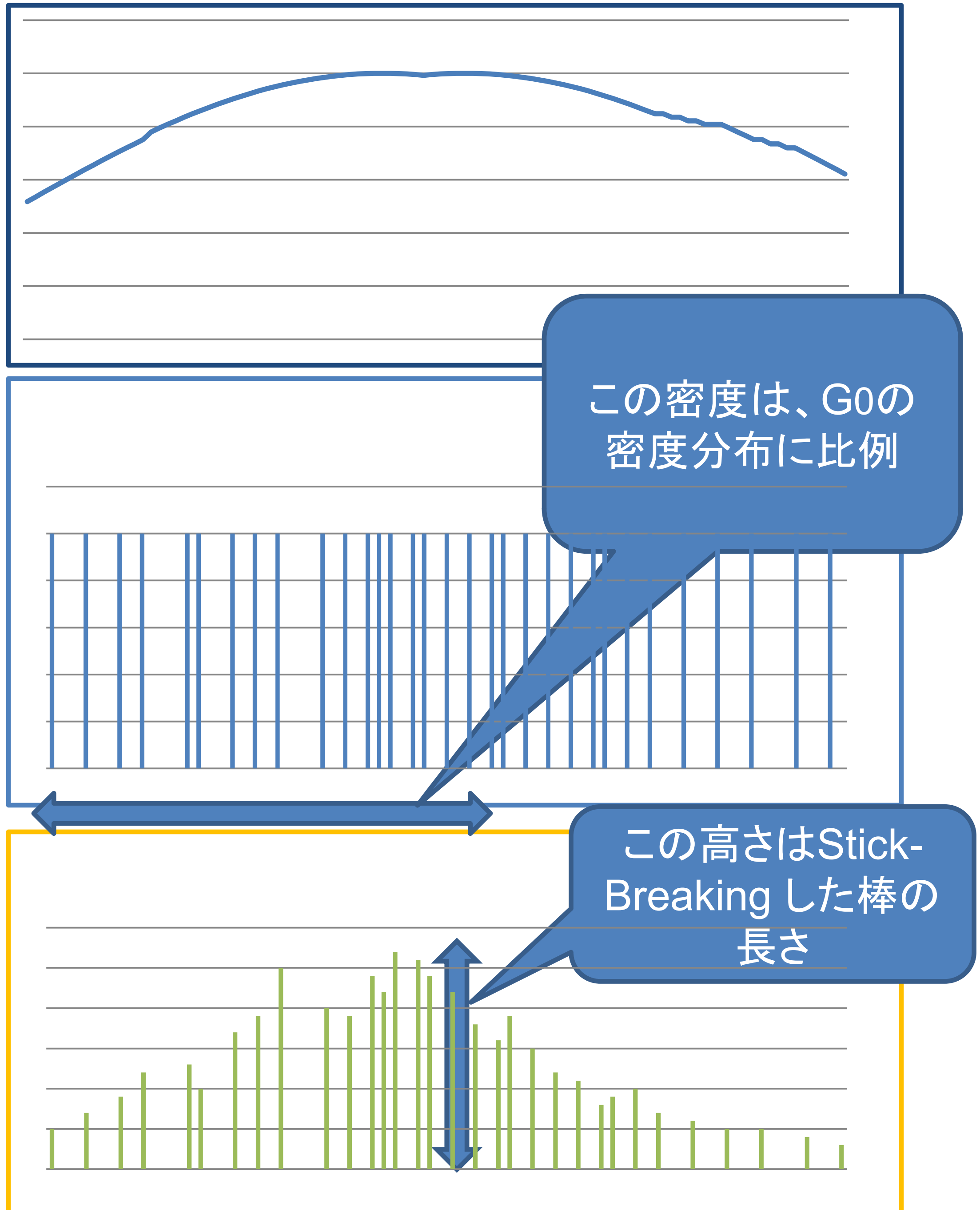
連続ドメインの基底測度を

- 可算無限次元の離散分布 G

で近似して事前分布とする。

有限次元離散分布

- 観測データから推定



while 収束するまで以下を繰り返す do

for n in randperm (1,・・・,N) do

Xn をクラスタ(古い）Zn から削除してパラメータを更新

Zn ~ p(Zn|X, Z−i) をサンプル(新しいZn)

Xn をクラスタZn に追加してパラメータを更新

end for

end while

Z1,・・・, ZN を出力

ギブスサンプリングによるznの推定.

randperm (x) はx のランダムな並び換えを表す.

Dirichlet 分布の事後分布

$$\boxed{P(\boldsymbol{\pi}|\{x_i\}_{i=1}^N) \propto P(\{x_i\}_{i=1}^N|\boldsymbol{\pi})P(\boldsymbol{\pi})}$$

Multinomial分布

πの事前分布はDirichlet分布

$$\pi \mid \{x_i\}_{i=1}^{N} \sim Dir(\mu'_1,\ldots,\mu'_K)$$

$$\mu'_i = \frac{\alpha_0\mu_i + n_i}{\alpha_0 + N}$$ n_iは観測データにおけるiの出現回数

Dirichlet 分布の事後予測分布

$$\boxed{P(x_{N+1}|\{x_i\}_{i=1}^N) = \int P(x_{N+1}|\boldsymbol{\pi})P(\boldsymbol{\pi}|\{x_i\}_{i=1}^N)d\boldsymbol{\pi}}$$

$$\begin{aligned}
P(x_{N+1} = x^j|\{x_i\}_{i=1}^N) &= \int P(x_{N+1} = x^j|\boldsymbol{\pi})P(\boldsymbol{\pi}|\{x_i\}_{i=1}^N)d\boldsymbol{\pi} \\
&= \int \pi_j P(\boldsymbol{\pi}|\{x_i\}_{i=1}^N)d\boldsymbol{\pi} \\
&= \mu_j^{'} = \frac{\alpha_0\mu_i + n_i}{\alpha_0 + N} \\
&= \boxed{\frac{\alpha_0}{\alpha_0 + N}\mu_j + \frac{N}{\alpha_0 + N}\sum_{i=1}^N \frac{\delta_j(x_i)}{N}}
\end{aligned}$$

Dirichlet Processの事後予測分布

[Blackwell and MacQueen,1973]

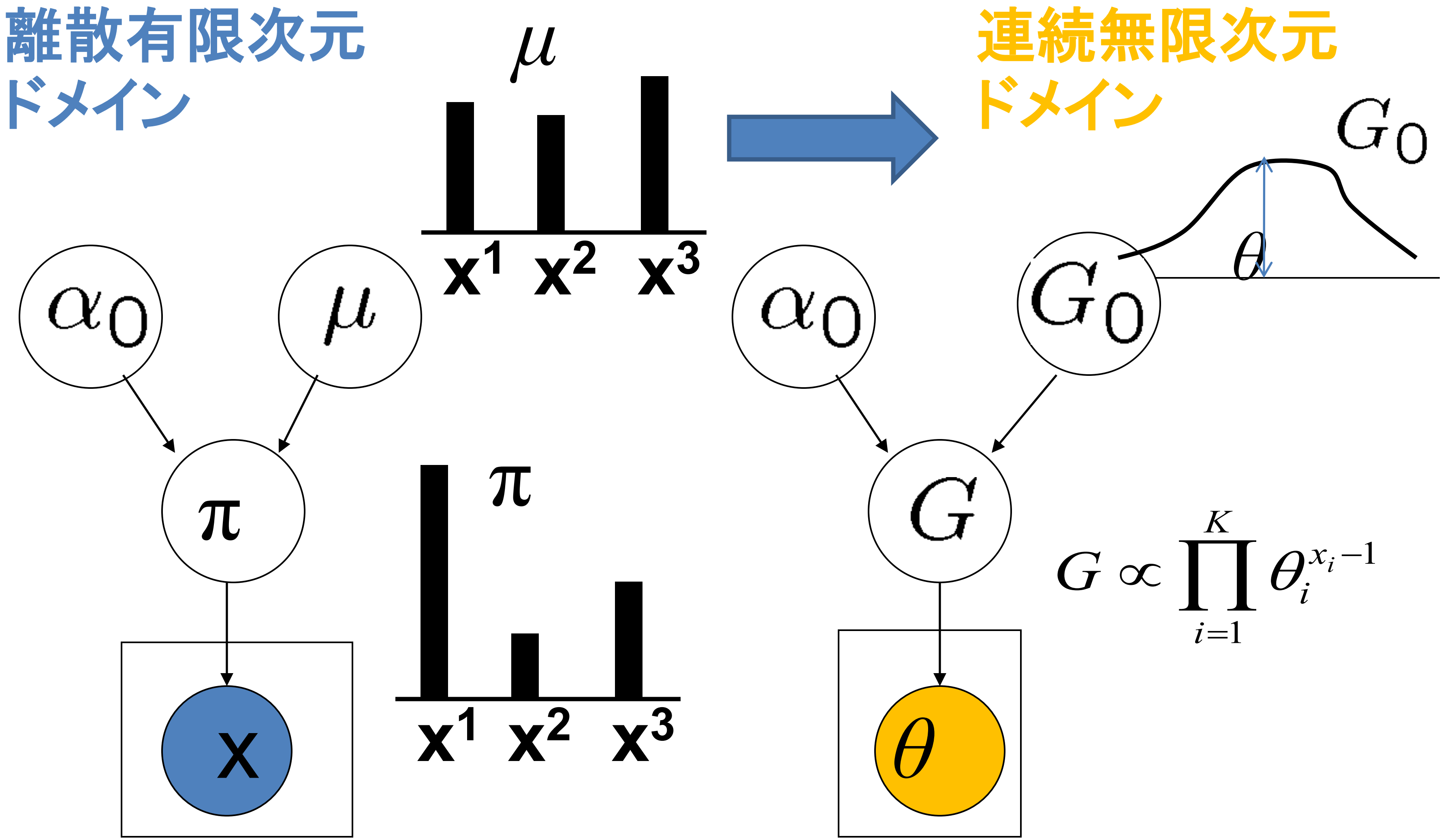
$$P(\theta_{N+1}|\{\theta_i\}_{i=1}^N) = \int P(\theta_{N+1}|G)P(G|\{\theta_i\}_{i=1}^N)dG$$

$$= G' = \frac{\alpha_0}{\alpha_0 + N}G_0 + \frac{N}{\alpha_0 + N}\sum_{i=1}^{N}\frac{\delta_{\theta j}(\theta_i)}{N}$$

新たにサンプリングされる確率

$$\theta_{N+1}|\{\theta\}_{i=1}^N \sim \begin{cases} G_0 & (\frac{\alpha_0}{\alpha_0+N}) \\ \sum_{i=1}^N \frac{\delta_{\theta j}(\theta_i)}{N} & (\frac{N}{\alpha_0+N}) \end{cases}$$

既存のθが選ばれる確率



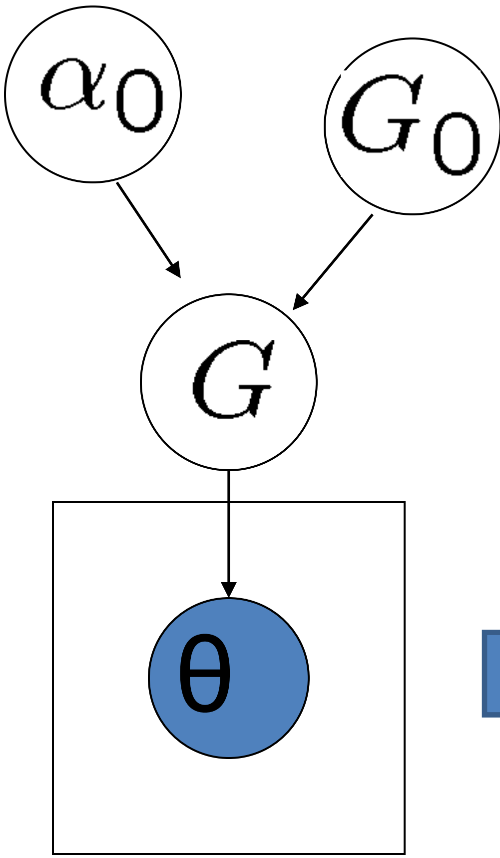
$$G \propto \prod_{i=1}^{K} \theta_i^{x_i - 1}$$

$$\pi \sim Dir(\alpha_0, \mu) \qquad \mu = (\mu_1, \ldots, \mu_K)$$

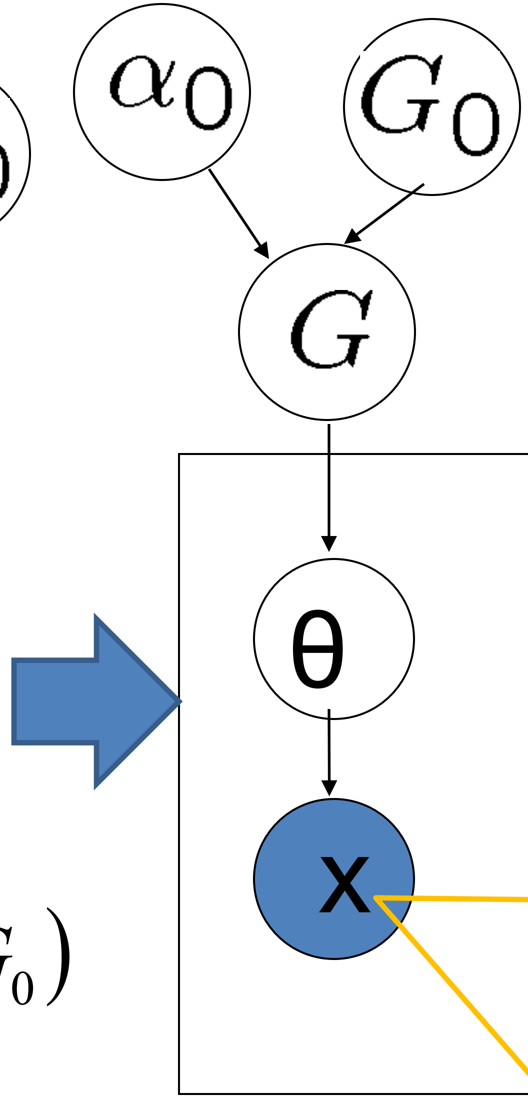
$$x \sim Multi(\pi) \propto \prod_{i=1}^{K} \pi_i^{x_i} \qquad x = (x_1, \ldots, x_K)$$

$$G \sim DP(\alpha_0, G_0)$$

$$\theta \sim G$$



$$G \sim DP(\alpha_0, G_0)$$
$$\theta \sim G$$

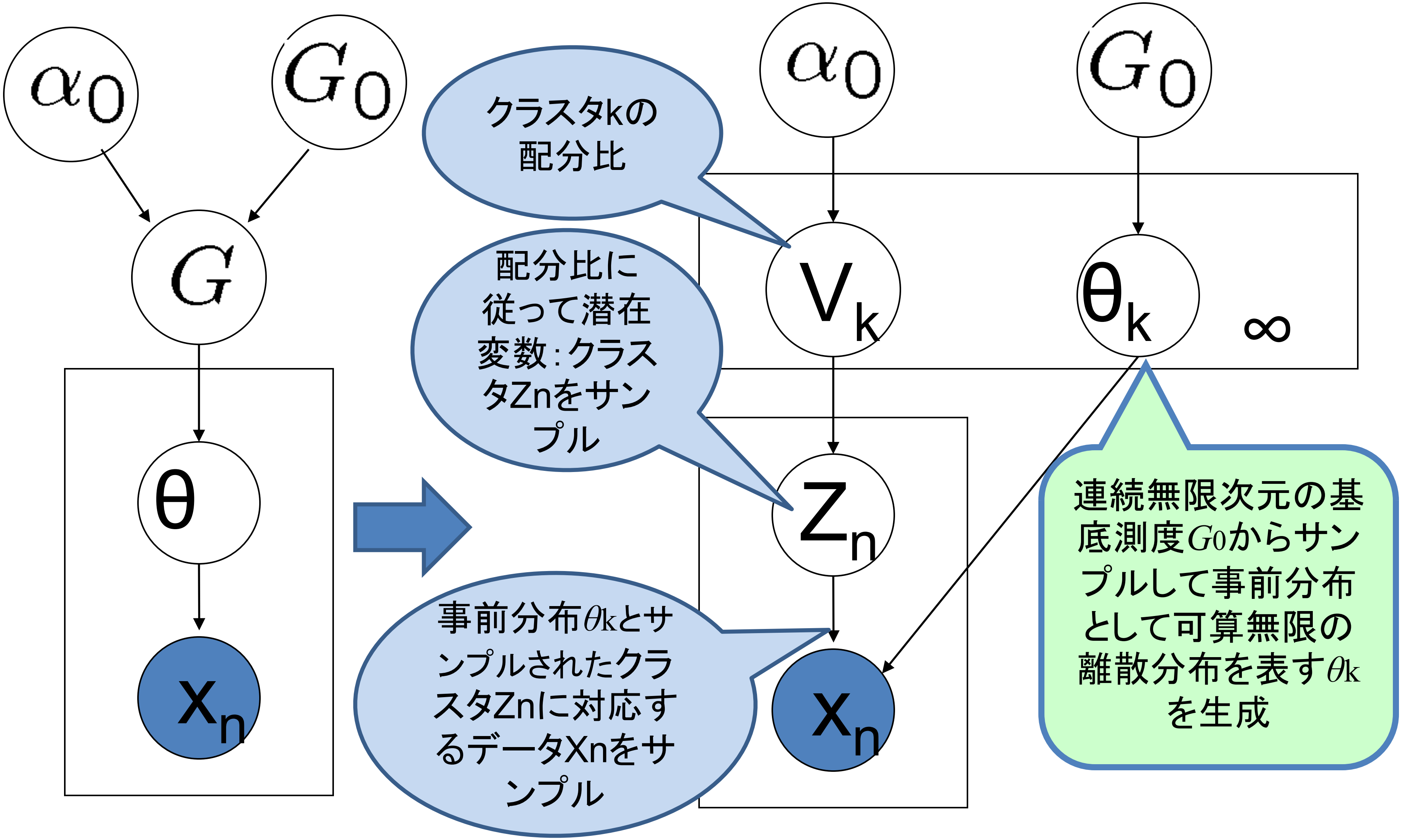


$$G \sim DP(\alpha_0, G_0)$$
$$\theta \sim G \qquad \propto \prod \theta_i^{x_i - 1}$$
$$x \sim P(x \mid \theta) \qquad \propto \prod \theta_i^{x_i}$$

- ●このようにして得たθは事前分布。
- ●Xは観測データからサンプルして、θなどのパラメターの推定を収束するまで繰り返す（ex. Gibbs Sampling）
- ●しかし、観測データを考慮して事後分布にする推論方法はまだ分からない

→そこでStick Breaking

Dirichlet 混合過程



$$G \sim DP(\alpha_0, G_0)$$
$$\theta \sim G$$
$$x_n \sim P(x|\theta)$$

$$V \sim Beta(1, \alpha_0)$$
$$\theta \sim G_0$$
$$x_n \sim p(x|Z_n, \theta)$$

Dirichlet 過程

- Chinese Restaurant Process や混合正規分布の混合比：πi (i=1,…∞)は、離散分布
- 離散分布で一般的なMultinomial あるいはその事前分布のDirichlet分布で考えていく。

$$Mult(m_1, m_2, \ldots, m_K \mid \boldsymbol{\mu}, N)$$

$$= \binom{N}{m_1 m_2 .. m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\mathrm{E}[m_k] = N\mu_k$$

$$\mathrm{var}[m_k] = N\mu_k(1-\mu_k)$$

$$\mathrm{cov}[m_j m_k] = -N\mu_j\mu_k$$

ただし、 $\sum_{i=1}^{K} \mu_k = 1$

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{pmatrix}$$

$$\sum_{k=1}^{K} \mu_k = 1 \qquad 0 \le \mu_k \le 1$$

$$Dir(\mu \mid \alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\mathrm{E}[\mu_k] = \frac{\alpha_k}{\alpha_1 + \cdots + \alpha_K}$$

参考文献

[1]Sethuraman. A Constrauctive Definition of Dirichlet Prior. Statistica Sinica (4), 639-650. 1994

[2]持橋大地. 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. 電子情報通信学会論文誌 (to be published)