Nonparametric Bayesian Model

Chikahara Yoichi

Model-based Clustering

- 生成的アプローチ：
  - $\pi=(\pi_1, \ldots, \pi_k)$で条件付けられるK個のクラスターのうち１つを選択
  - cluster-specificな確率分布からデータ点を生成

- Bayesian finite mixture model:

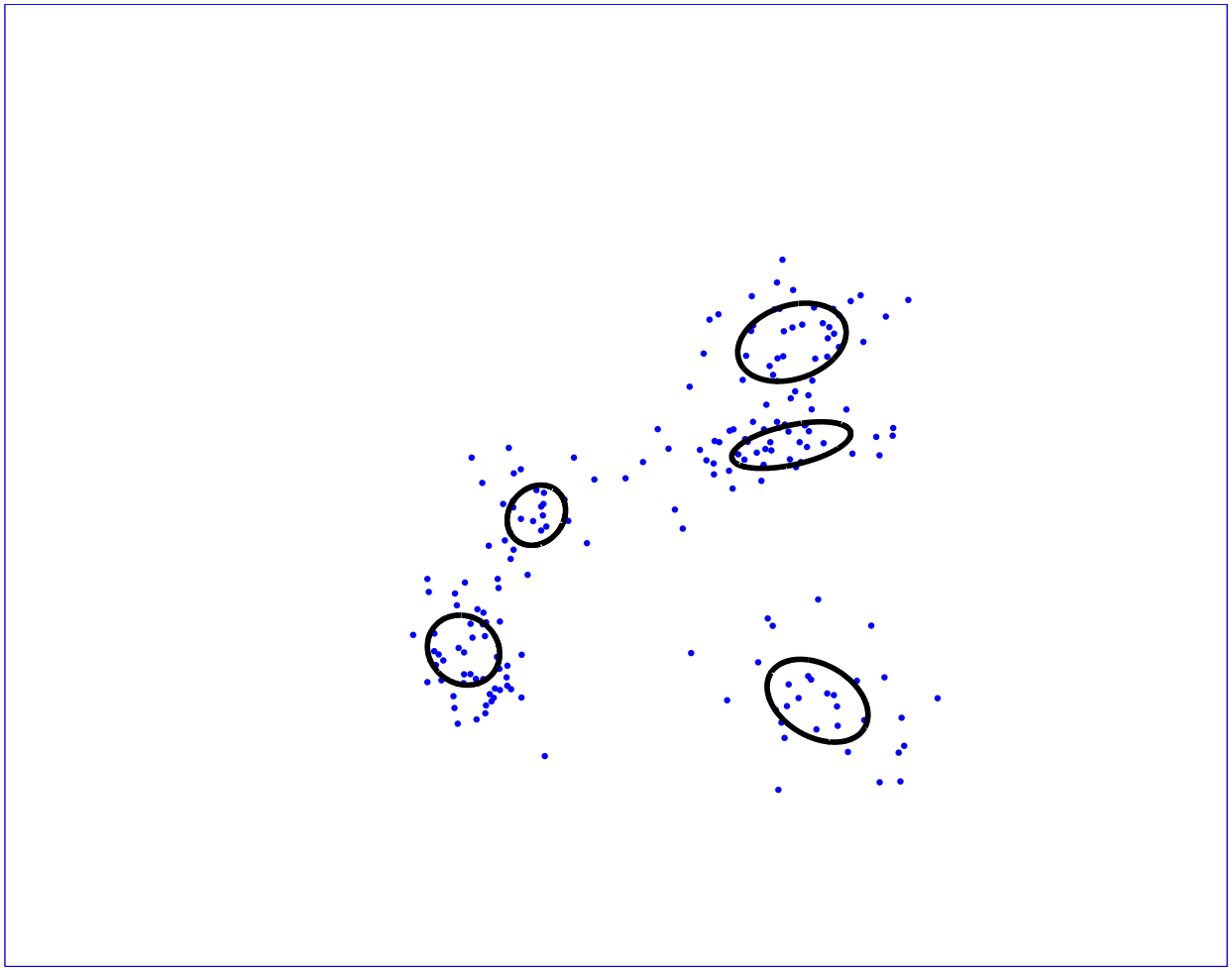
$$P(x|\theta,\pi) = \sum_{k=1}^{K} \pi_k P(x|\theta_k)$$

$\pi = (\pi_1, \cdots, \pi_K)$: クラス確率
$\theta = (\theta_1, \cdots, \theta_K)$: 各クラスに対応したパラメータ

- 各データ点$\{x_i\}_{i=1}^{n}$は上記混合モデルよりi.i.d.に生成される
- k番目のクラスターはk番目のmixture componentから生成されたデータ点集合

Finite Mixture Models

- E.g., for Gaussian mixtures, $\phi_k = (\mu_k, \Sigma_k)$ and $p(x \mid \phi_k)$ is a Gaussian density with mean μ_k and covariance matrix Σ_k



http://www.cs.berkeley.edu/~jordan

Finite Mixture Model 1

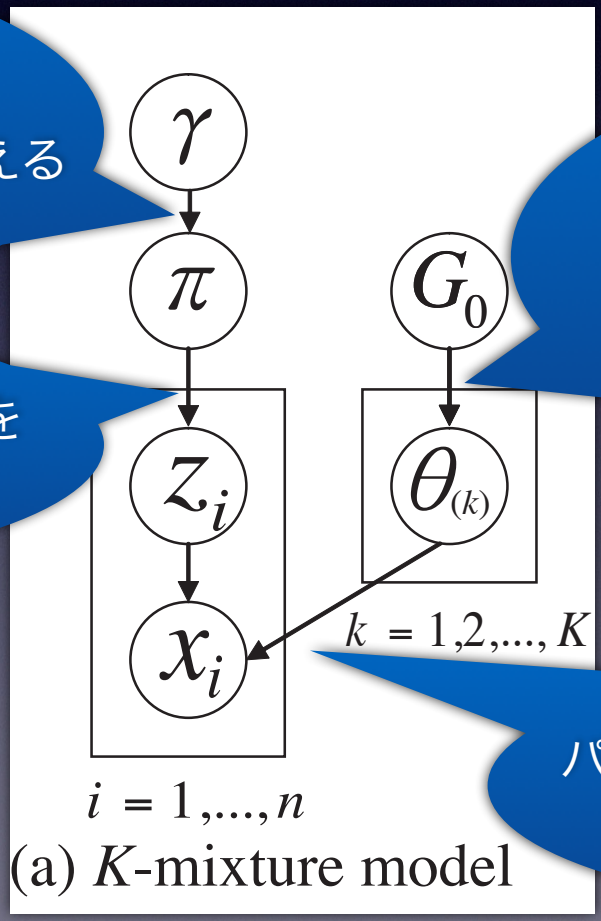
(1) $\theta_{(k)} \sim G_0$, for $k = 1, \ldots, K$

(2) $\pi = (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\pi;\ \gamma_1, \ldots, \gamma_K)$

(3) $z_i \,|\, \pi \sim \text{Multinomial}(z;\ \pi)$, and $x_i \,|\, \theta_{(z_i)} \sim p(x \,|\, \theta_{(z_i)})$, for $i = 1, \ldots, n$

(2)
K個の要素分布の選択確率
$\pi_1, \ldots, \pi_K$をDirichlet分布で与える

(3)-1
K個の要素分布から1個の分布を
選ぶ (多項分布)



(a) K-mixture model

(1)
共通の事前分布G_0から
K個の要素分布(e.g. Gaussian)
のパラメータ$\theta_{(k)}$を生成

(3)-2
パラメータ$\theta_{(z_i)}$で規定される要素
分布に従う形でx_iが生成

Finite Mixture Model 2

- (3)–1式について、これはΘが分布

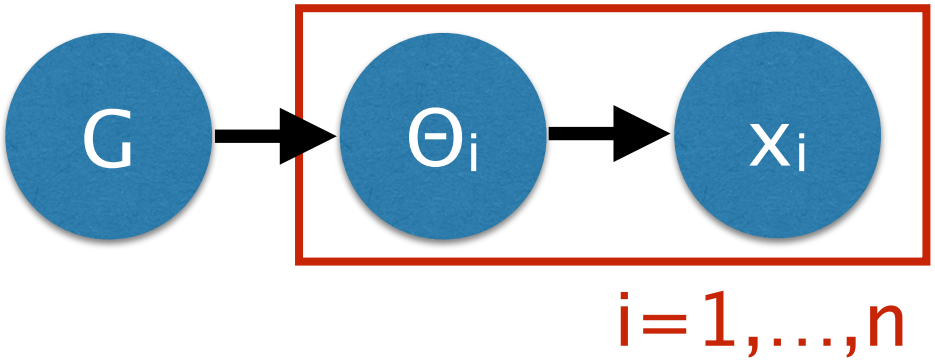
$$G(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$$

$$\delta_{\theta_k} = \begin{cases} 1 & (\theta = \theta_k) \\ 0 & (otherwise) \end{cases}$$

に従って生成されていると考えることもできる

- これよりデータ点x_i (i=1,…,n)を得るプロセスを

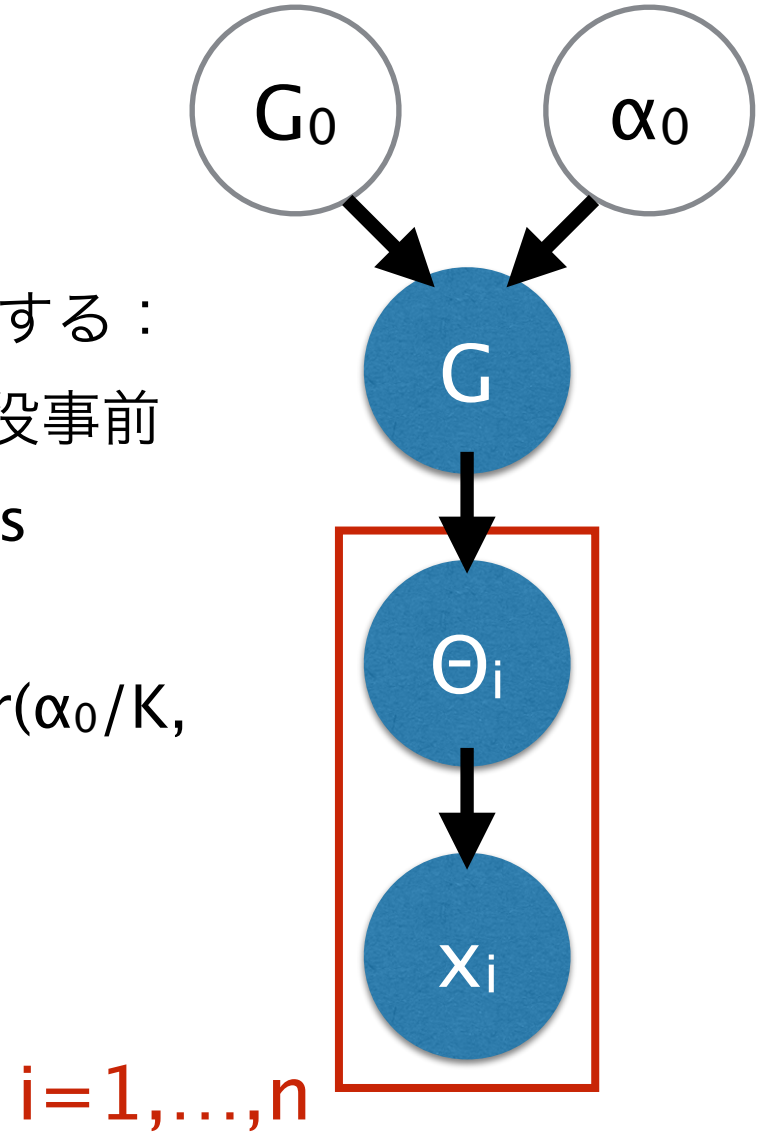
$$\theta_i \sim G$$
$$x_i \sim P(\cdot \,|\theta_i)$$



と表すこともできる

Finite Mixture Models

- Θとπに関するpriorを設定する必要がある
- Θに関するprior G_0はmodel-specificに選択する：(e.g.) Gaussian mixture modelならその共役事前分布であるnormal / inverse-gamma priors
- πに関するpriorは対称なDirichlet prior, Dir(α_0/K, …, α_0/K)で与えると良い
  このときΘ_kのprior meanは1/Kに等しい



Finite Mixture Models

$$\theta_k \sim G_0$$

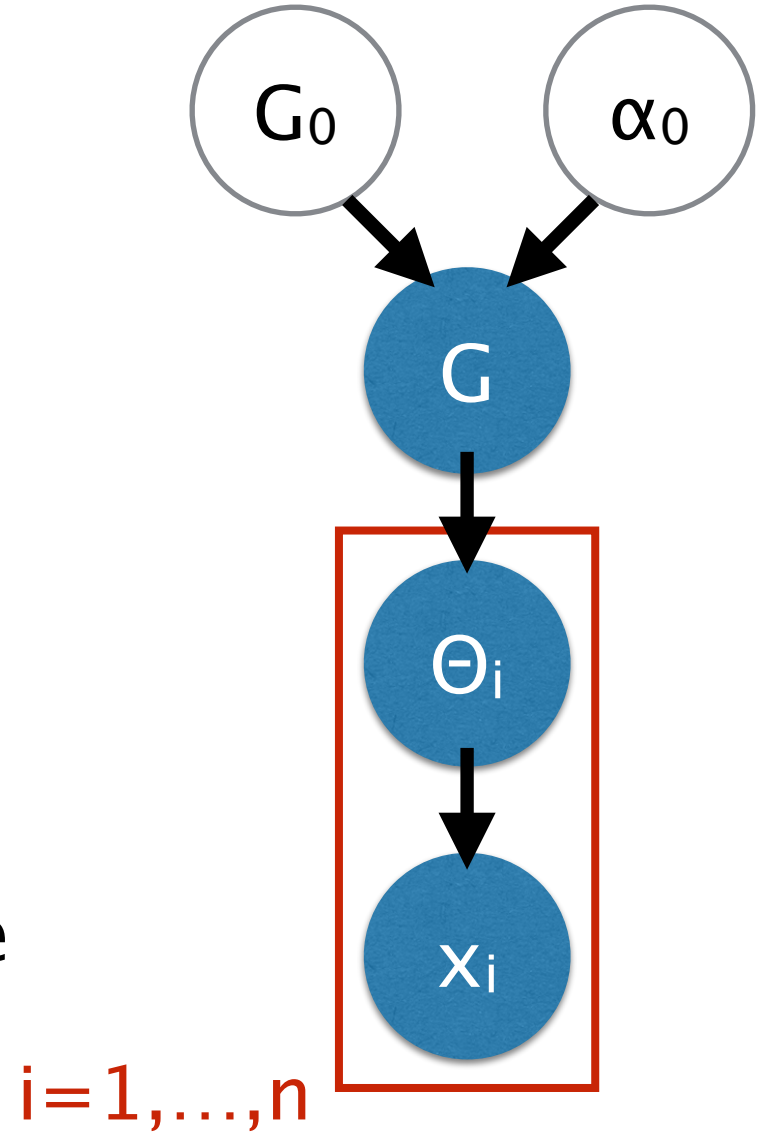
$$\pi_k \sim Dir(\alpha_0/K, \cdots, \alpha_0/K)$$

$$G(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$$

$$\theta_i \sim G$$

$$x_i \sim P(\cdot \,|\theta_i)$$

※ G: random measure



Model Selection

- mixture componentの数 Kを如何にして決めるか

- 様々な手法が考えられる：
  - cross-validation
  - information criterion – AIC, BIC, MDL, …

Dirichlet過程、Chinese Restaurant Processを用いた
ノンパラメトリックベイズによる代替手法を考える

Parametric Model Selection

(1) $\theta_{(k)} \sim G_0$, for $k = 1, \ldots, K$

(2) $\pi = (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\pi;\ \gamma_1, \ldots, \gamma_K)$

(3) $z_i \,|\, \pi \sim \text{Multinomial}(z;\ \pi)$, and $x_i \,|\, \theta_{(z_i)} \sim p(x \,|\, \theta_{(z_i)})$, for $i = 1, \ldots, n$

- (パラメトリックモデル補足) Kの値は経験ベイズで与えられることが多い

左のような階層ベイズモデルで事後確率

$$p(K, \Theta | D) \propto p(D|\Theta)p(\Theta|K)p(K)$$

を最大化するKを考えるべきだが、

Θに関して周辺化できる場合、p(K)を一様分布で与えて

$$\hat{K} = \arg\max_K \int p(D|\Theta)p(\Theta|K)d\Theta$$
$$= \arg\max_K p(D|K)$$

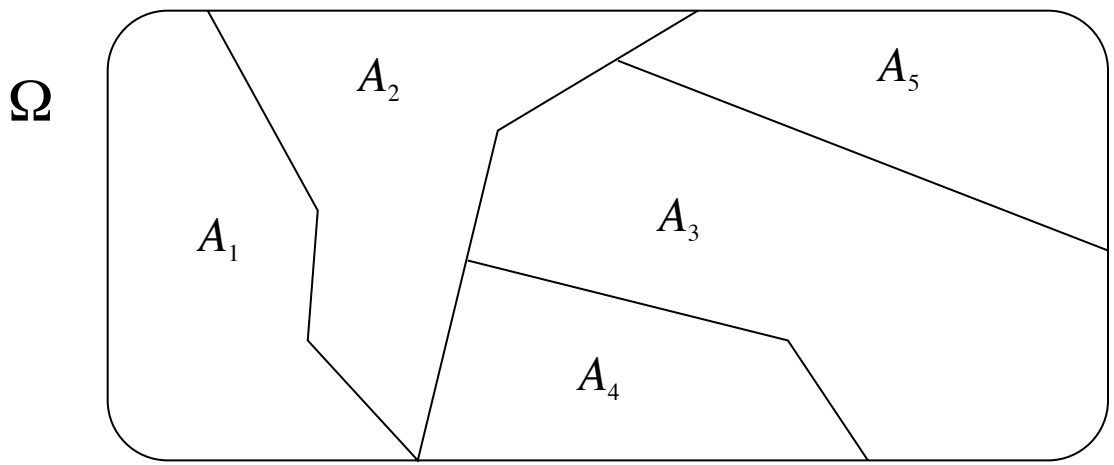
と点推定で与える方針が考えられる

Dirichlet Process

Definition 1. *Let $(\Omega, \mathcal{B})$ by a measurable space, with G_0 a probability measure on the space, and let α_0 be a positive real number. A Dirichlet process is the distribution of a random probability measure G over $(\Omega, \mathcal{B})$ such that, for any finite partition $(A_1, \ldots, A_r)$ of Ω, the random vector $(G(A_1), \ldots, G(A_r))$ is distributed as a finite-dimensional Dirichlet distribution:*

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}\big(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)\big) \tag{2}$$

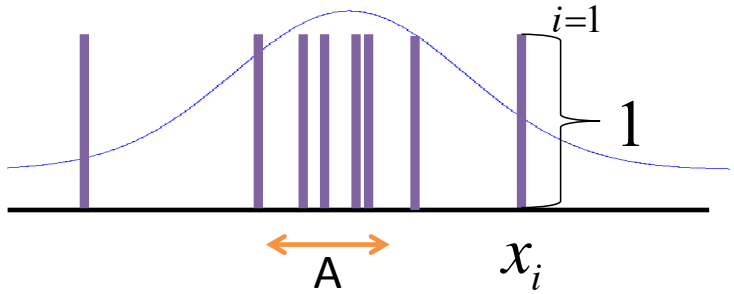
We write $G \sim \mathrm{DP}(\alpha_0 G_0)$ if G is a random probability measure distributed according to the Dirichlet process. Call G_0 the base measure of G and call α_0 the concentration parameter.



http://www.cs.berkeley.edu/~jordan

ディリクレ過程とは
(例)　Poisson過程

Poisson過程：$N(A) = \sum_{i=1}^{\infty} 1\delta(x_i \in A)$



棒の従うルール
(合計は整数値)

任意のA(⊂X)に関して、N(A)が

$$N(A) \sim \mathrm{Poison}(\alpha_0 H(A))$$

点の従うルール
$x_i \sim H(x)$

$\alpha_0 H: X \rightarrow R^+$
α_0:非負スカラー値
H:基底(確率)測度

http://www.slideshare.net/issei_sato/ppml

ディリクレ過程とは

Dirichlet 過程

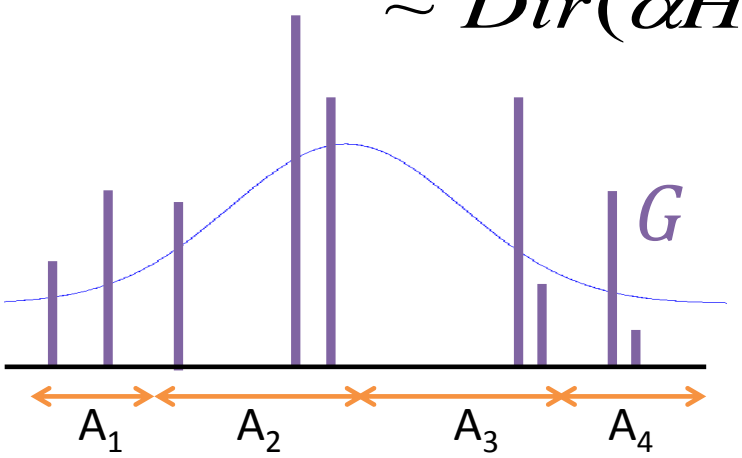
[Ferguson,1973]

Xの任意の分割$A_1,\ldots A_m$に対して

$$(G(A_1),\cdots,G(A_m))$$

棒の従うルール
（全棒の合計が1）

$$\sim Dir(\alpha H(A_1),\cdots,\alpha H(A_m))$$



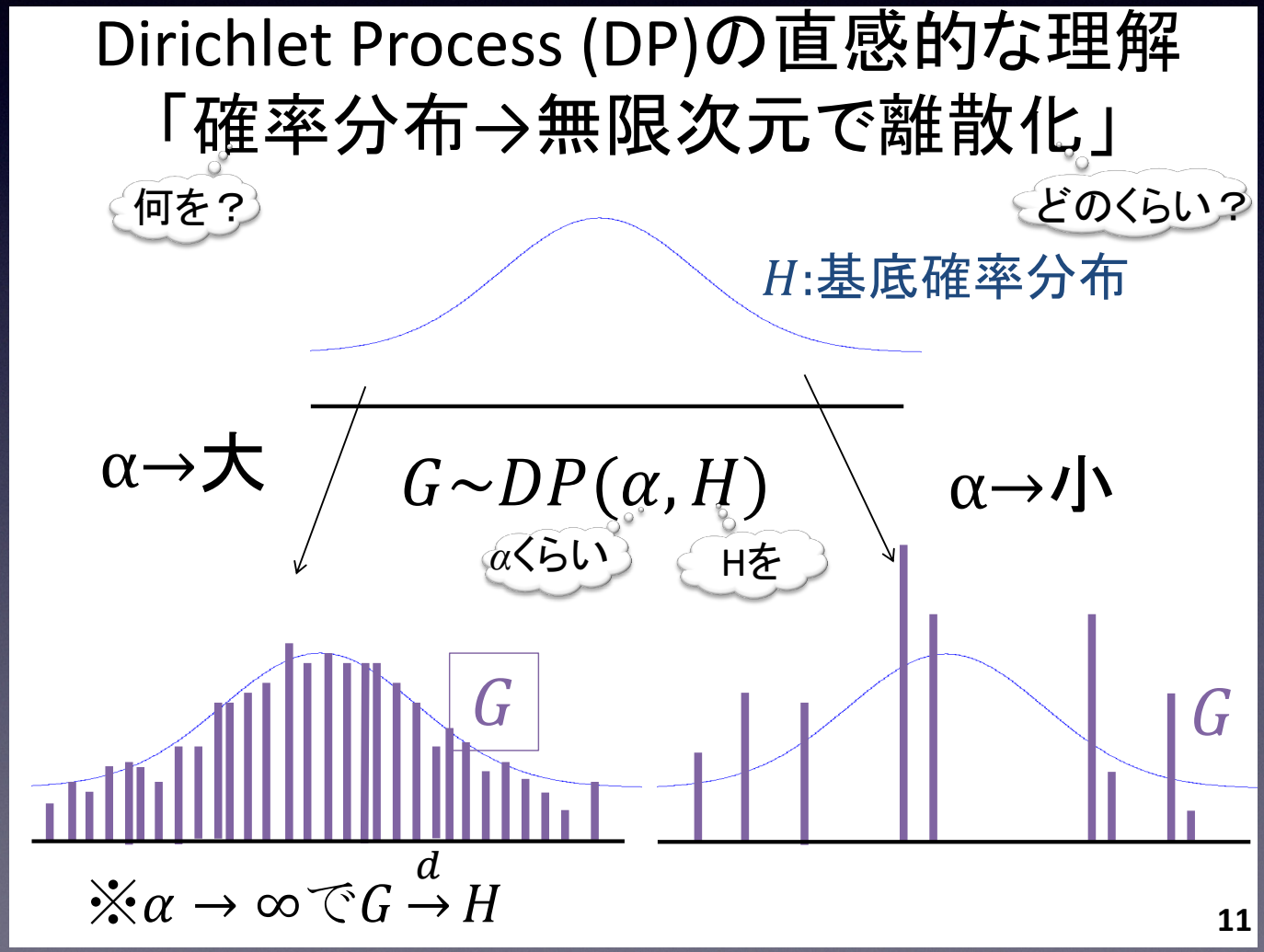
$$\mathrm{G}(A)=\sum_{i=1}^{\infty} w_i \delta\left(x_i \in A\right)$$

$$\sum_{i=1}^{\infty} w_i = 1$$

http://www.slideshare.net/issei_sato/ppml

ディリクレ過程とは

αが大きいほど事前分布Hを信頼することになる

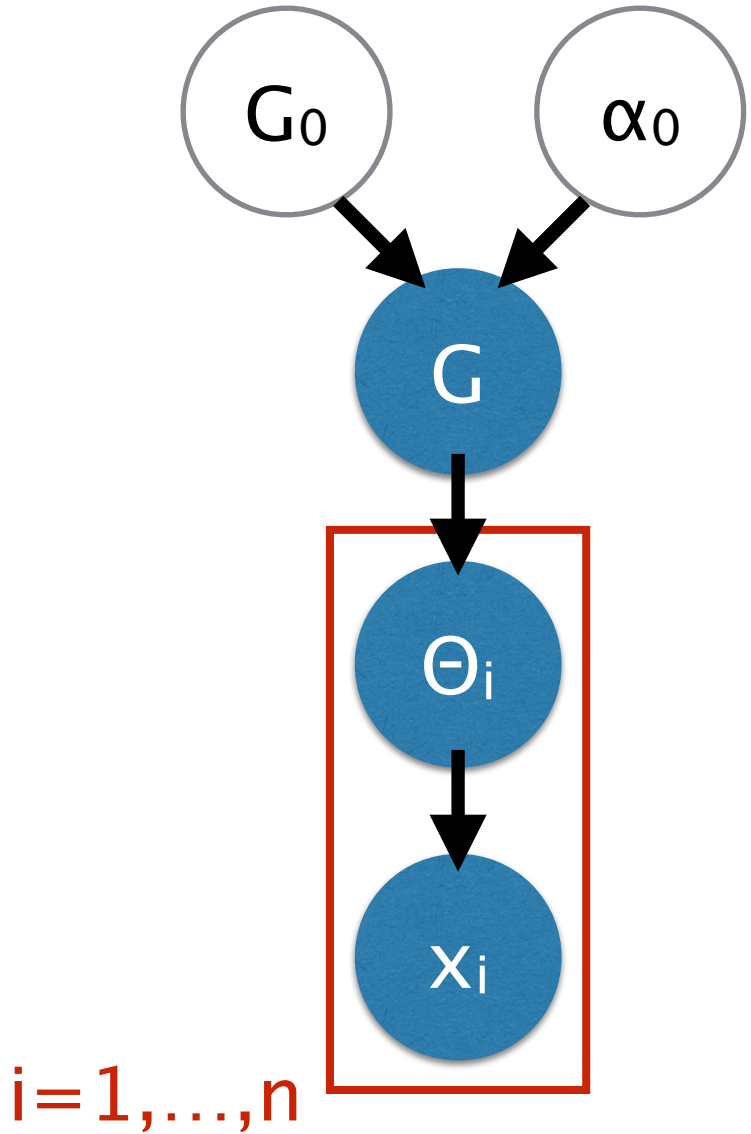


http://www.slideshare.net/issei_sato/ppml

Dirichlet Process Mixture Models

- Θ_iに関するpriorを得るのにDirichlet過程を導入する

$$
\begin{cases}
G \sim DP(\alpha_0 G_0) \\
\theta_i | G \sim G \\
x_i | \theta_i \sim F(x_i | \theta_i)
\end{cases}
$$



Posterior Dirichlet Process

- Dirichlet過程によりGをサンプリングし、GよりΘ$_1$をサンプリングしたとするとき、posterior processは？

- このとき、一定の分割A=(A$_1$,…, A$_m$)に対して、Θ$_1$から生成されたデータ点x$_1$を含む領域A$_k$に関するDirichlet分布のパラメータのみ１増える

- これより次を得る

$$G|\theta_1 \sim DP(\alpha_0 G_0 + \delta_{\theta_1})$$

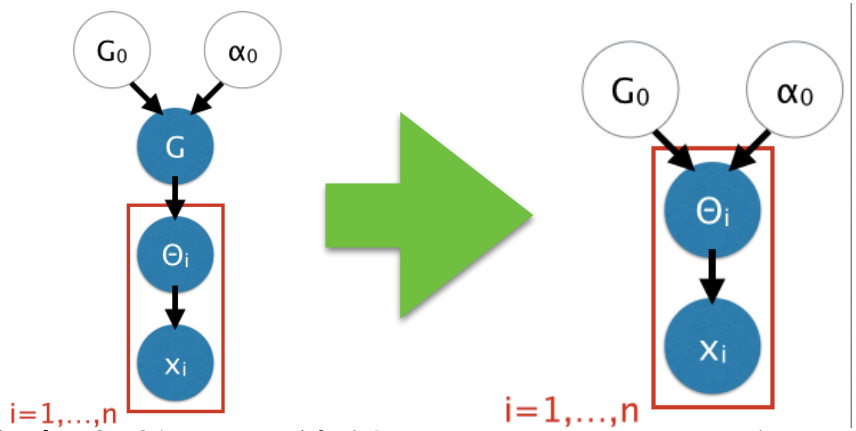
$$\delta_{\theta_i} = \begin{cases} 1 & (x_i \in A_k) \\ 0 & (otherwise) \end{cases}$$
$$k = 1, \cdots, K$$

- このようなposterior updateを繰り返すと、次を得る

$$G|\theta_1, \cdots, \theta_n \sim DP(\alpha_0 G_0 + \sum_{i=1}^{n} \delta_{\theta_i})$$

Polya Urn Process
Chinese Restaurant Process

- パラメータ$\Theta_1,\ldots,\Theta_n$についての周辺確率を求めるには、Gをintegrate outする必要がある



- Dirichlet分布に従う確率変数の期待値の公式から、どのような集合A (⊆Ω)に対しても、

$$E[G(A)|\theta_1,\cdots,\theta_n] = \frac{\alpha_0 G_0(A) + \sum_{k=1}^{K} n_k \delta_{\theta_k}(A)}{\alpha_0 + n}$$

(ア)AがΘ_k(k=1,...,K)により生成される集合とするとき、Θ_kによる生成を再度観測する確率はn_kに比例する

(イ)Aが新たなクラスタにより生成される集合であるとするとき、その確率はα_0に比例する

MCMC for Dirichlet Process Mixtures

- Chinese Restaurant priorに従うcustomer iの座る席は、他のcustomer達の席の選択を条件としてどう記述できるか？

CRPはde Finettiの表現定理(1935)に従うため、
customer i (i=1,…,n)はexchangeableである
→　任意のiに対して最後のcustomerとして扱える

$$P(x_i \in A_k | T_{m-1}) \propto n_k \qquad (k = 1, \cdots, K)$$
$$P(x_i \notin A_k | T_{m-1}) \propto \alpha_0 \qquad (k = 1, \cdots, K)$$

$$for \quad all \quad i$$

Collapsed Gibbs Sampling
for Dirichlet Process Gaussian Mixture Models

1) DPGMM: the Model

$$P(x_{1:N}) = \prod_{i=1}^{N} \sum_{k=1}^{K} P(x_i|\theta_{z_i}) P(z_i = k)$$

$$\left\{ \begin{array}{ll} i \in \{1, \cdots, N\} & \text{：観測データインデックス} \\ k \in \{1, \cdots, K\} & \text{：クラスターインデックス} \\ z_i & \text{：クラスター割当て} \\ \theta_k & \text{：クラスターkに関するパラメータ} \end{array} \right.$$

Collapsed Gibbs Sampling for Dirichlet Process Gaussian Mixture Models

2) Fitting Data

$$\begin{cases} \pi \sim CRP(\alpha) & \text{:mixing rates} \\ z_i \sim \pi & \text{:cluster assignment} \\ \theta_k \sim H(\lambda) & \text{:parameters} \\ x_i \sim F(\theta_{z_i}) & \text{:values} \end{cases}$$

$$\begin{cases} z_{-i} = \{z_j | j \neq i\} \\ x_{-i,k} = \{x_j | z_j = k, j \neq i\} \\ k^* \quad \text{: new cluster} \\ N_{k,-i} : x_{-i,k}\text{の要素数} \end{cases}$$

として、i番目の観測点がcluster kに割り当てられる確率は

$$P(z_i = k | z_{-i}, x, \alpha, \lambda) \propto \underbrace{P(x_i | x_{-i}, z_i = k, z_{-i}, \lambda)}_{\text{(ア)}} \underbrace{P(z_i = k | z_{-i}, \alpha)}_{\text{(イ)}}$$

と分解できる。

Collapsed Gibbs Sampling
for Dirichlet Process Gaussian Mixture Models

2) Fitting Data

Chinese Restaurant Process

(イ－1)

$$\underbrace{P(z_i = k | z_{-i}, \alpha)}_{\text{(イ)}} = \begin{cases} \dfrac{N_{k,-i}}{\alpha + N - 1} & \text{: if k is an existed cluster} \\ \dfrac{\alpha}{\alpha + N - 1} & \text{: if k is a new cluster} \end{cases}$$

(イ－2)

αはCRPの制御パラメータ

αが大きいほどクラスター数は増えやすい

Collapsed Gibbs Sampling for Dirichlet Process Gaussian Mixture Models

2) Fitting Data

Marginal Likelihood of all the data assigned to cluster k, including i

$$\underbrace{P(x_i|x_{-i}, z_i = k, z_{-i}, \lambda)}_{(ア)} = P(x_i|x_{-i,k}, z_i = k, \lambda)$$

$$\propto P(x_i, x_{-i,k}, \lambda|z_i = k)$$

$$= \int \underbrace{P(x_i|\theta_k)[\prod_{j \neq i, z_j = k} P(x_j|\theta_k)]}_{\text{Likelihood}} \underbrace{H(\theta_k|\lambda)}_{\text{Prior}} d\theta_k$$

$z_i = k^*$ (i.e. new cluster) ならば

$$P(x_i|x_{-i}, z_i = k^*, z_{-i}, \lambda) = \int \underbrace{P(x_i|\theta)}_{\text{Likelihood}} \underbrace{H(\theta|\lambda)}_{\text{Prior}} d\theta$$

Collapsed Gibbs Sampling
for Dirichlet Process Gaussian Mixture Models

3) Conjugate Priors

likelihood Fをmultivariate normalで与えるとき、

μ, Σを与えるconjugate prior Hはnormal-inverse-Wishart

$$f(\mu, \Sigma | \mu_0, \kappa_0, \Psi_0, \nu_0) = N(\mu | \mu_0, \frac{1}{\kappa_0}\Sigma) IW(\Sigma | \Psi_0, \nu_0)$$

$$\left(\Leftarrow \begin{cases} \mu | \mu_0, \kappa_0, \Sigma \sim N(\mu | \mu_0, \frac{1}{\kappa_0}\Sigma) \\ \Sigma | \Psi_0, \nu_0 \sim IW(\Sigma | \Psi_0, \nu_0) \end{cases} \right)$$

で与えるとよい

※　逆Wishart分布の自由度パラメータν_0が大きいほど

hyper-parameter ψ_0を信用することになる

Collapsed Gibbs Sampling
for Dirichlet Process Gaussian Mixture Models

3) Conjugate Priors

このとき、MAP estimatesはそれぞれ、

$$\begin{cases} \mu_n = \dfrac{\kappa_0 m u_0 + n\tilde{x}}{\kappa_0 + n} = \mu_{k,-i} \\ \kappa_n = \kappa_0 + n \\ \nu_n = \nu_0 + n \\ \Psi_n = \Psi_0 + C + \dfrac{\kappa_0 n}{\kappa_0 + n}(\tilde{x} - \mu_0)(\tilde{x} - \mu_0)^T \\ \Sigma_{k,-i} = \dfrac{\kappa_n + 1}{\kappa_n(\nu_n - d + 1)}\Psi_n \end{cases}$$

$$n = N_{k,-i}$$

$$\tilde{x} = \frac{\sum x_{-i,k}}{n}$$

$$C = \sum_{p=1}^{n}(x_p - \tilde{x})(x_p - \tilde{x})^T$$

で与えられ、これより

$$P(x_i | x_{-i}, z_i = k, z_{-i}, \lambda) \propto N(\mu_{k,-i}, \Sigma_{k,-i}) \quad \text{(ア)}$$

を得る

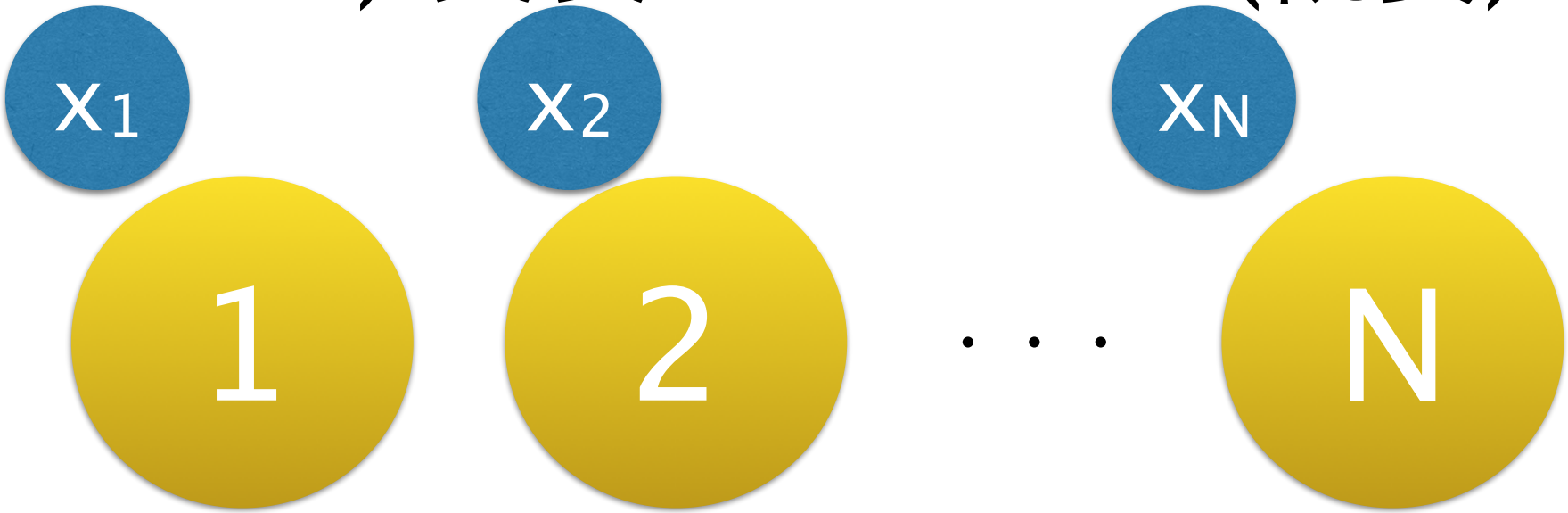
Collapsed Gibbs Sampling for Dirichlet Process Gaussian Mixture Models

4) 実装のイメージ (概要)

```
while (the cluster labels does not converge)
  for (i in 1:N)
    remove x[i] from cluster z[i]
    if (any cluster is empty)
      delete the cluster and decrease K
    for (k in 1:K)
      compute tmp[k] = (ア) × (イ－1)
    tmp[K+1] = compute (ア) × (イ－2)
    normalize tmp[k], and choose cluster
```

Collapsed Gibbs Sampling for Dirichlet Process Gaussian Mixture Models

4) 実装のイメージ (概要)



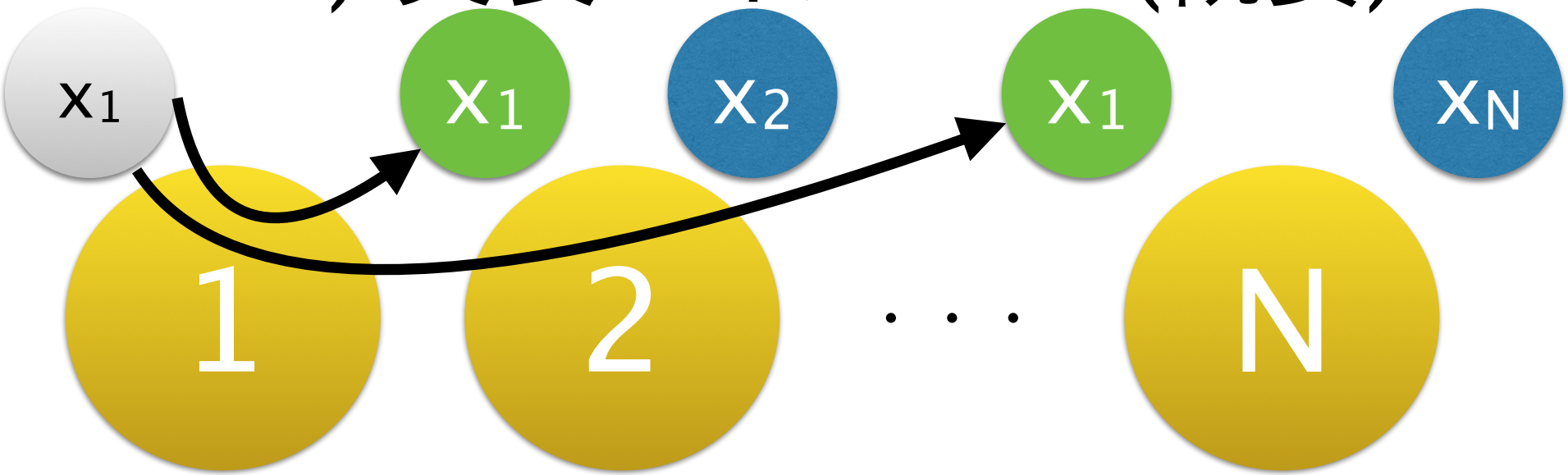
最初にN個の観測点をクラスタに割り当てる

1) N個のクラスタを用意し、１点ずつ割り当てる (上図)

2) K (< N)個のクラスタを用意し、ランダムに割り当てる

Collapsed Gibbs Sampling for Dirichlet Process Gaussian Mixture Models

4) 実装のイメージ (概要)



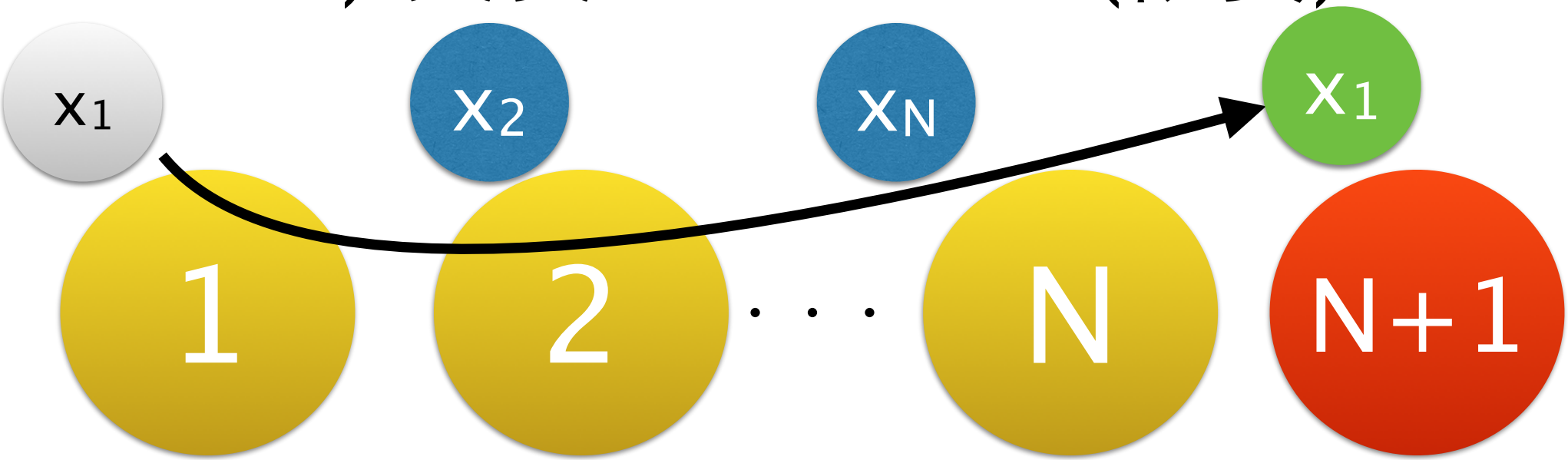
クラスタZ[i]からデータx[i]を取り除き、

各kについて割当て確率 (ア) × (イ−1)を計算し、結果をtmp[k]に格納する

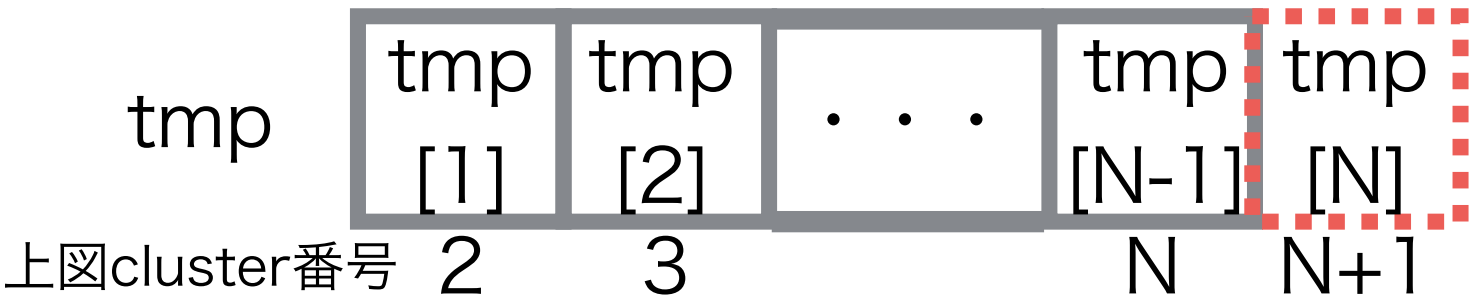
tmp	tmp [1]	tmp [2]	・・・	tmp [N-1]
上図cluster番号	2	3		N

Collapsed Gibbs Sampling for Dirichlet Process Gaussian Mixture Models

4) 実装のイメージ (概要)



新クラスターK+1について割当て確率 (ア) × (イ−2)を計算し、結果を tmp[K+1]に格納する

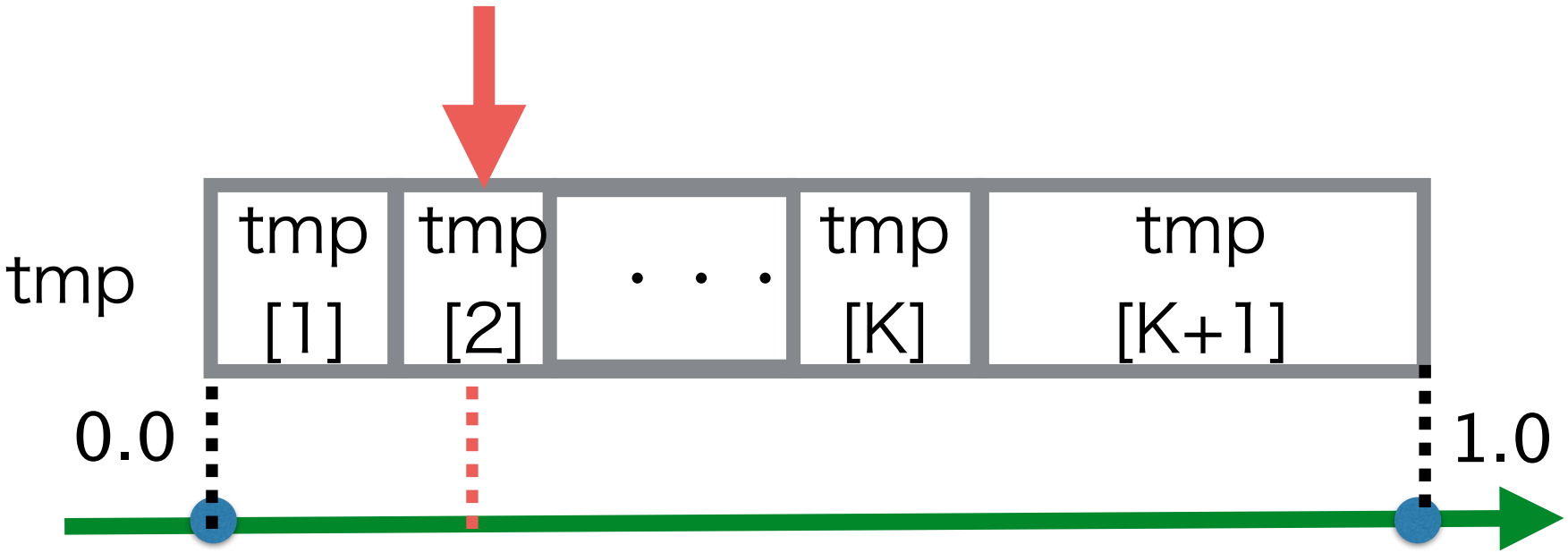


Collapsed Gibbs Sampling
for Dirichlet Process Gaussian Mixture Models

4) 実装のイメージ (概要)

tmpの要素を全て足すと１になるよう規格化

[0, 1]の一様乱数で得た値からランダムにクラスタを割り当てる



Reference

- ベイズ統計分析ハンドブック: 繁桝算男他, D.K.デイ, C.R.ラオ 2011 (p.292~300)
- ノンパラメトリックベイズモデル 上田修功,　山田武士 – 応用数理, 2007
- Machine Learning A Probabilistic Perspective Kelvin P. Murphy, 2012 (p.879~887)
- Bayesian Nonparametrics入門–点過程と統計的機械学習の接点– http://www.slideshare.net/issei_sato/ppml
- http://snippyhollow.github.io/blog/2013/03/10/collapsed-gibbs-sampling-for-dirichlet-process-gaussian-mixture-models/