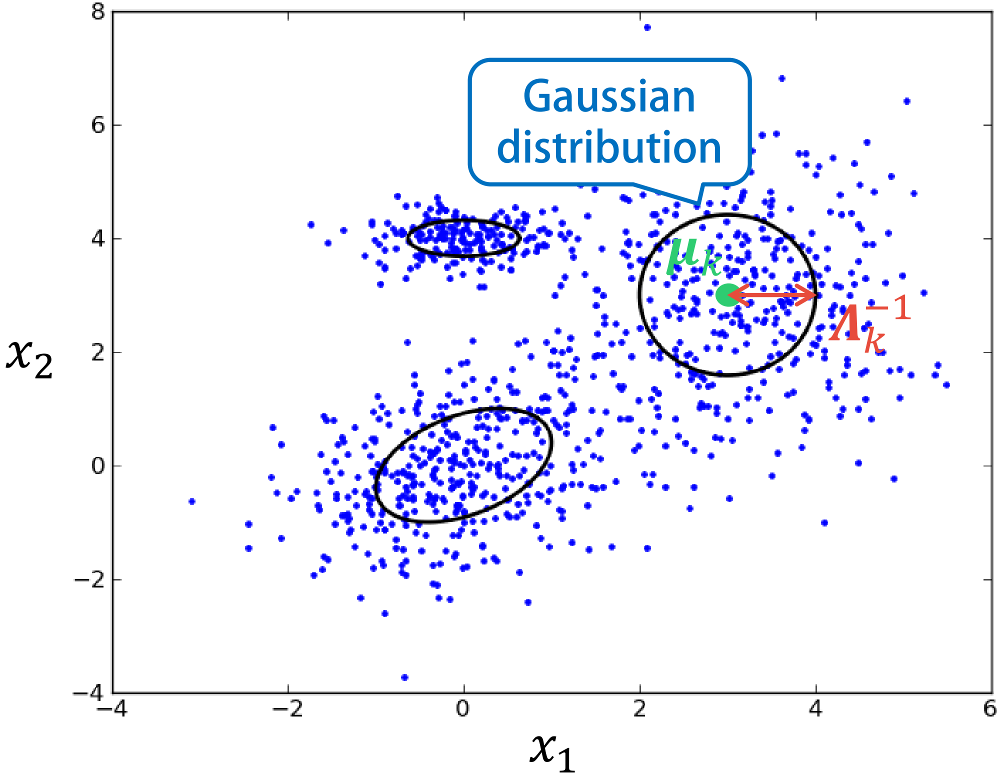
Learning Algorithms for Gaussian Mixture Models

Department of Intelligent Science and Technology
Graduate School of Informatics
Kyoto University
Kazuyoshi Yoshii
yoshii@kuis.kyoto-u.ac.jp

The Gaussian Mixture Model

- The GMM is used for representing how multi-dimensional vectors (e.g., feature vectors) are distributed stochastically



Probability distribution:

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

Parameters to be estimated:

Mixing ratios

$$\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$$

Mean vectors

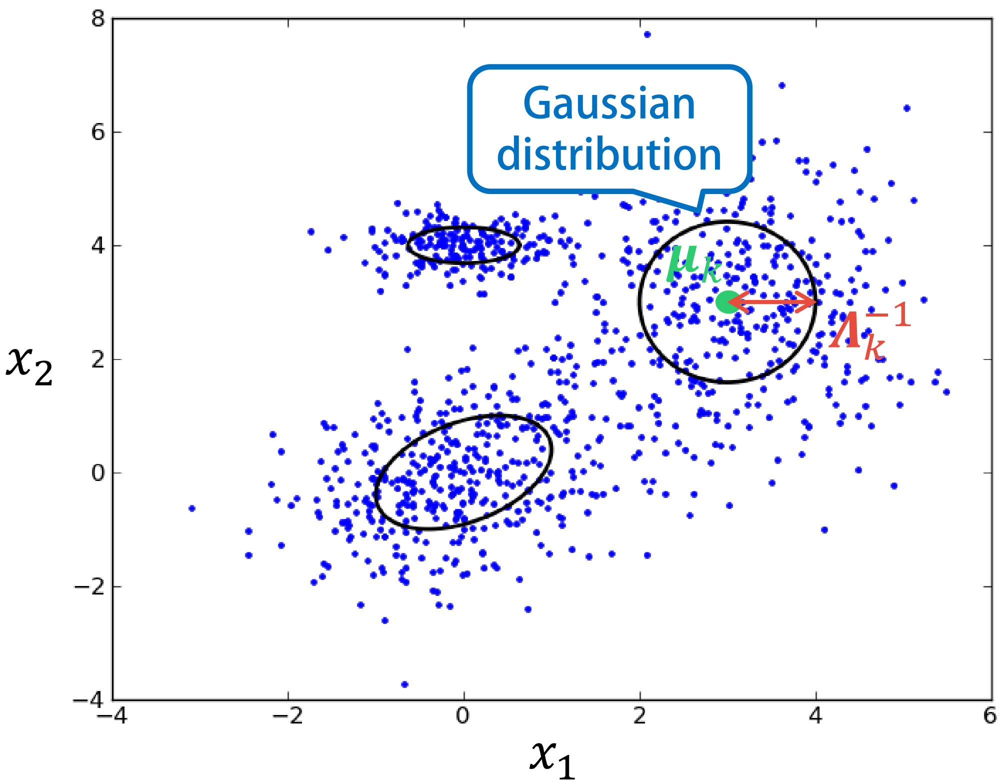
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K]$$

Precision matrices

$$\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_K]$$

Generative Story of GMM

- The GMM is a probabilistic model for clustering
  - Each vector (sample) exclusively belongs to one of K classes



Probability distribution:

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

Generative story:

Class indicator

Draw a latent variable

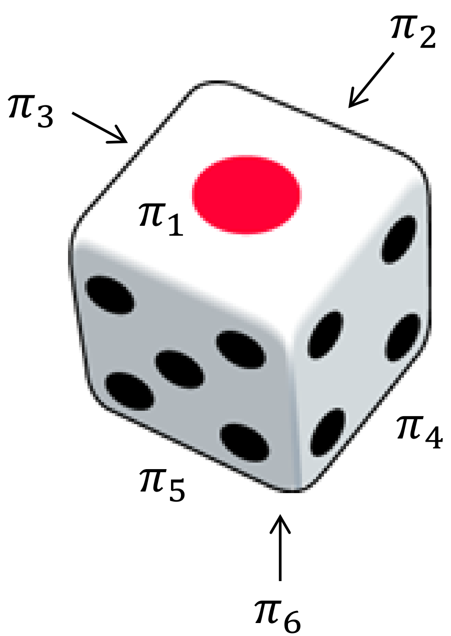
$$\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$$

Draw an observed variable

$$\boldsymbol{x}_n \sim \prod_{k=1}^{K} N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$$

Draw Latent Variables

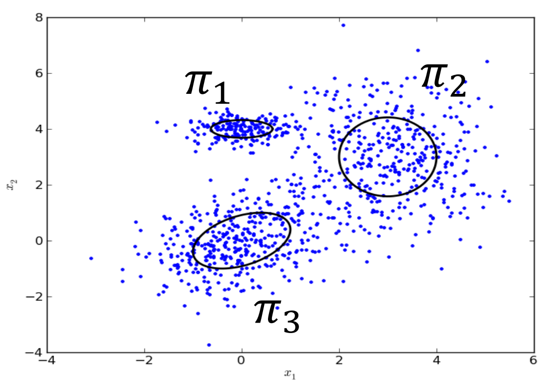
- Latent variables are categorical distributed
  - Draw each latent variable: $\boldsymbol{z}_n \sim \text{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$ $(\boldsymbol{\pi} = [\pi_1, \cdots, \pi_K])$
  - Use an one-of-K representation: $\boldsymbol{z}_n = [z_{n1}, z_{n2}, z_{n3}, \cdots, z_{nK}]$



Suppose we cast a K-sided die defined by $\boldsymbol{\pi}$

If we get "3" for the n^{th} trial, we say $\boldsymbol{z}_n = [0, 0, 1, 0, 0, 0]$

Only one of the elements takes the value of 1



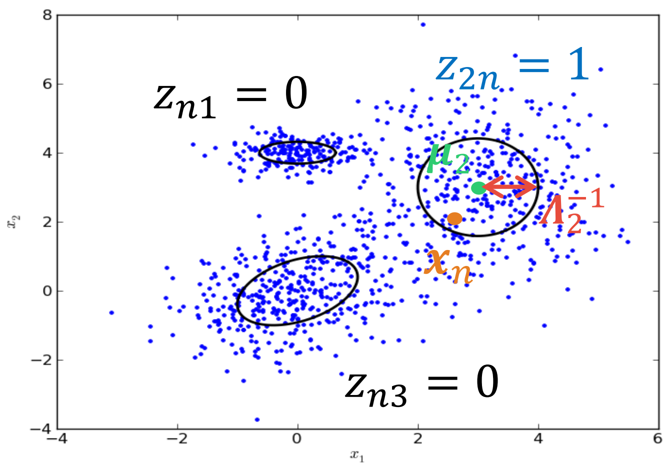
In the generative story of GMM, a class to which each sample belongs is stochastically determined by casting the die

Draw Observed Variables

- Observed variables are Gaussian distributed
  - Draw each observed variable: $\boldsymbol{x}_n \sim \prod_{k=1}^{K} N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$
  - Use the k^{th} Gaussian distribution when $z_{nk} = 1$

Expand the product:

$$\boldsymbol{x}_n \sim \prod_{k=1}^{3} N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}} = N(\boldsymbol{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1^{-1})^{z_{n1}} N(\boldsymbol{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Lambda}_2^{-1})^{z_{n2}} N(\boldsymbol{x}_n | \boldsymbol{\mu}_3, \boldsymbol{\Lambda}_3^{-1})^{z_{n3}}$$



Suppose $\boldsymbol{z}_n = [0, 1, 0]$

$$\boldsymbol{x}_n \sim N(\boldsymbol{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Lambda}_2^{-1})$$

The one-of-K representation can be used as a class indicator (selector)

This makes the derivation of learning algorithms easy (explained later)

Important Tips

- There are several kinds of K-dimensional values
  - Random variables
    - Mixing ratios: $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_k, \cdots, \pi_K]$
    - Latent variables: $\boldsymbol{z}_n = [z_{n1}, z_{n2}, \cdots, z_{nk}, \cdots, z_{nK}]$
  - Categorical probabilities
    - Posteriors: $\boldsymbol{\gamma}_n = [\gamma_{n1}, \gamma_{n2}, \cdots, \gamma_{nk}, \cdots, \gamma_{nK}]$

The values sum to unity

$$\sum_{k=1}^{K} \pi_k = 1 \qquad \sum_{k=1}^{K} z_{nk} = 1 \qquad \sum_{k=1}^{K} \gamma_{nk} = 1$$

$0 < \pi_k < 1$

Only one of the values is 1
The other values are 0

$0 < \gamma_{nk} < 1$

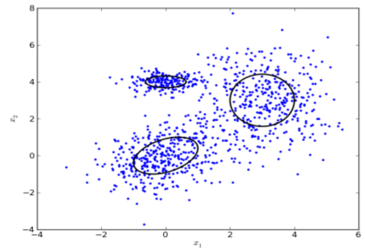
Probabilistic Modeling

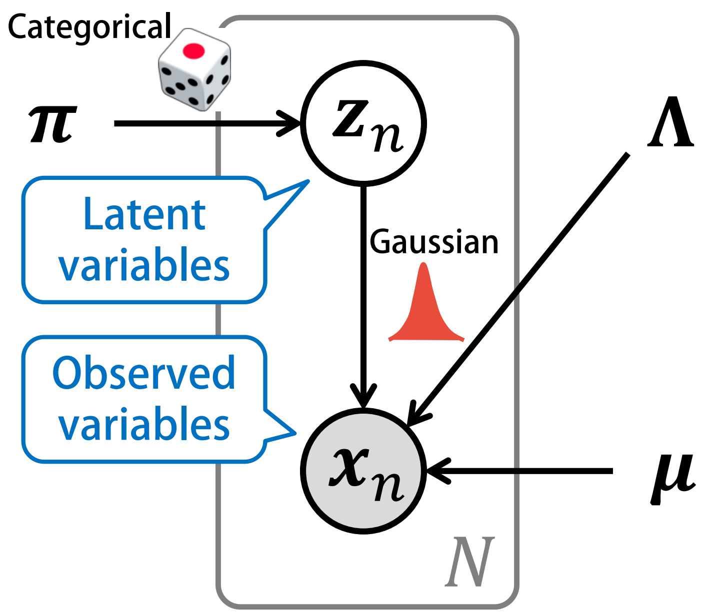
- Generative story of the GMM
  - Draw each latent variable: $\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$
  - Draw each observed variable: $\boldsymbol{x}_n \sim \prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$
- Two major approaches

	Maximum likelihood (ML) estimation	Bayesian estimation
Probabilistic model	$p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\mu},\boldsymbol{\Lambda}) = p(\boldsymbol{X}\|\boldsymbol{Z};\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z};\boldsymbol{\pi})$	$p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = p(\boldsymbol{X}\|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z},\boldsymbol{\pi})p(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$
Latent variables $\boldsymbol{Z}$	Posterior calculation $p(\boldsymbol{Z}\|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$	Posterior calculation $p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}\|\boldsymbol{X})$
Parameters $\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}$	Point estimation $\boldsymbol{\pi}^*,\boldsymbol{\mu}^*,\boldsymbol{\Lambda}^* = \mathrm{argmax}\, p(\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$	

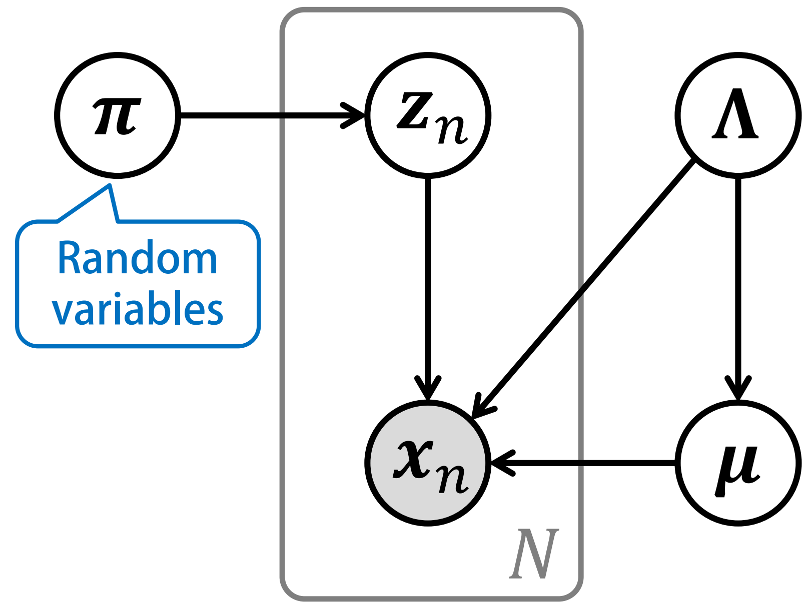
Graphical Representation

- Visualize dependency structures
  - Nodes: random variables (shaded = observable)
  - Edges: conditional dependencies





Likelihood model



Bayesian model

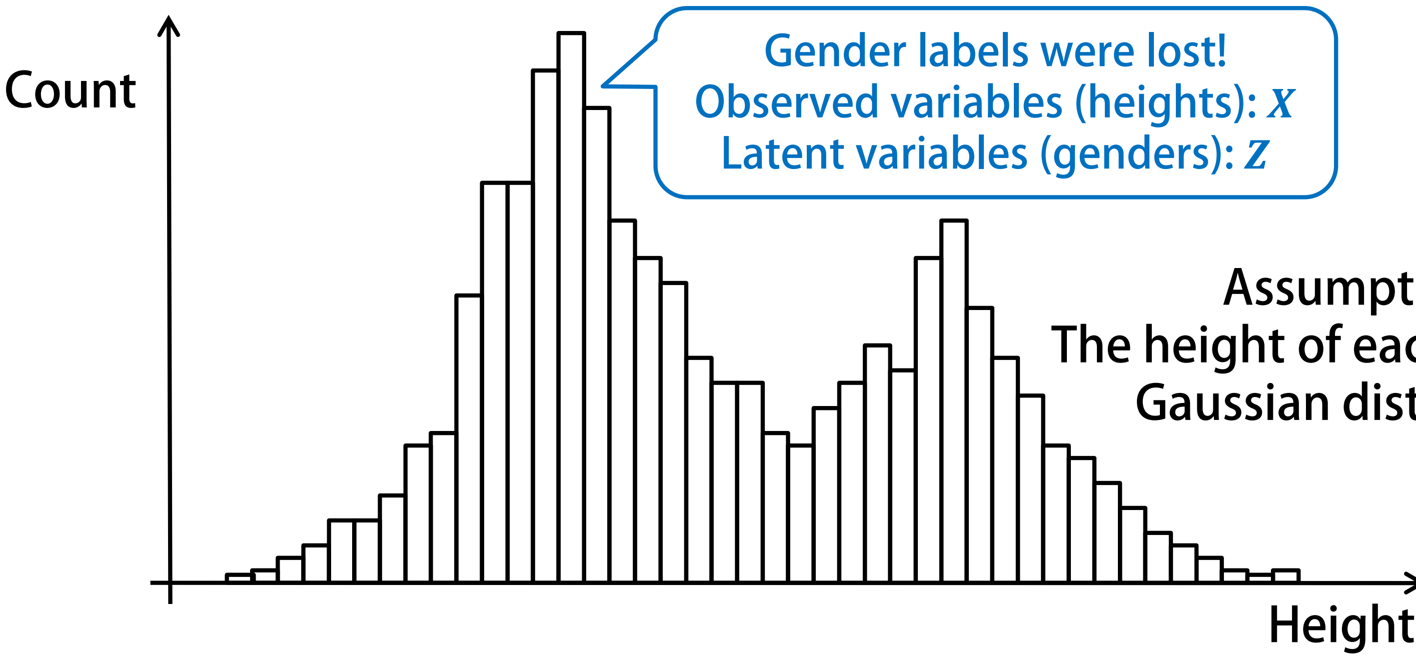
Maximum Likelihood Estimation of Finite Gaussian Mixture Models

Expectation-Maximization Algorithm

K-means Algorithm (Hard EM)

Unsupervised Learning for Unlabeled Data

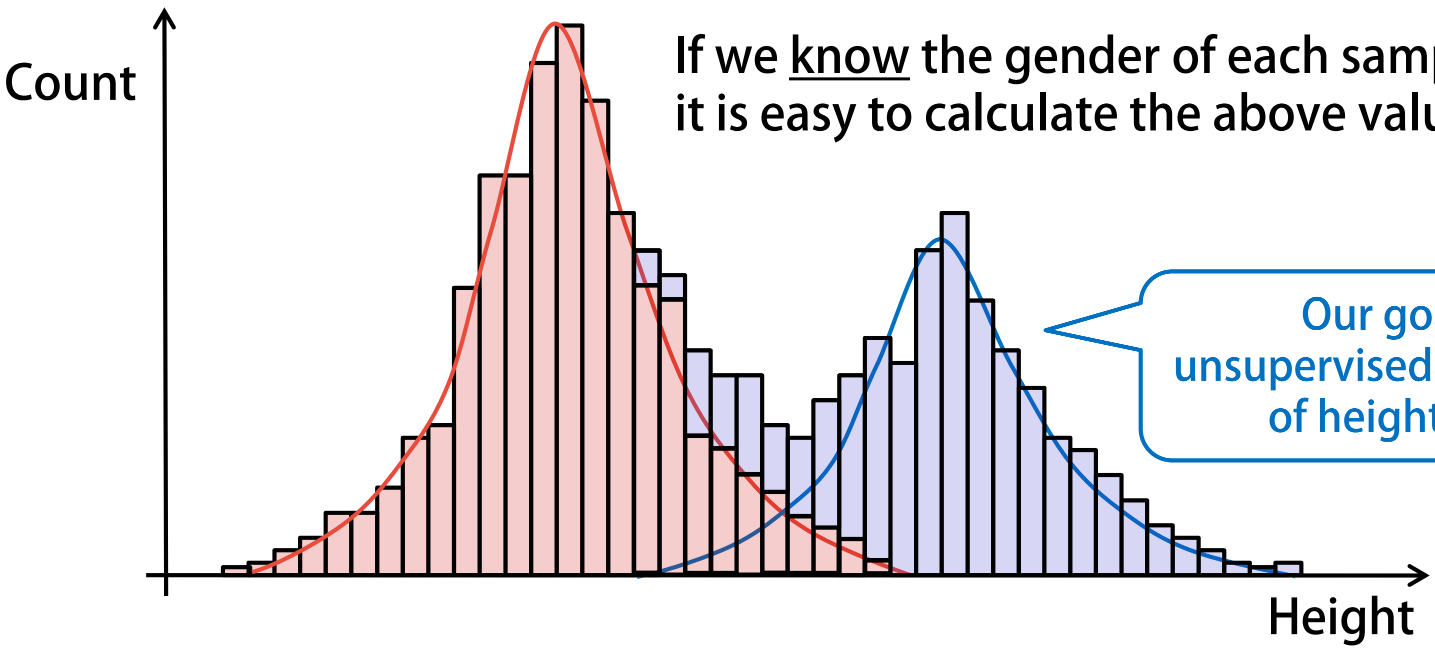
- Suppose we have unlabeled height data
  - We want to estimate
    - the averages μ and precisions Λ of the heights of male and female
    - the ratios π of male and female



Assumption:
The height of each gender is Gaussian distributed

Finite Gaussian Mixture Model

- Suppose we have unlabeled height data
  - We want to estimate
    - the averages μ and variances Λ of the heights of male and female
    - the ratios π of male and female



If we know the gender of each sample, it is easy to calculate the above values

Our goal is unsupervised clustering of height data

Probabilistic Modeling

- Generative story of the GMM
  - Draw each latent variable: $\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$
  - Draw each observed variable: $\boldsymbol{x}_n \sim \prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$
- Two major approaches

<table>
<tr><th></th><th>Maximum likelihood (ML) estimation</th><th>Bayesian estimation</th></tr>
<tr><td>Probabilistic model</td><td>$p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\mu},\boldsymbol{\Lambda})$
$= p(\boldsymbol{X}|\boldsymbol{Z};\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z};\boldsymbol{\pi})$</td><td>$p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})$
$= p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z},\boldsymbol{\pi})p(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$</td></tr>
<tr><td>Latent variables
$\boldsymbol{Z}$</td><td>Posterior calculation
$p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$</td><td rowspan="2">Posterior calculation
$p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X})$</td></tr>
<tr><td>Parameters
$\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}$</td><td>Point estimation
$\boldsymbol{\pi}^*,\boldsymbol{\mu}^*,\boldsymbol{\Lambda}^* = \mathrm{argmax}\, p(\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$</td></tr>
</table>

Parameter Estimation: Genders Known

- Estimate the ratios, averages, and variances

$$
\begin{array}{ll}
 & \quad\;\; k=1 \;\; k=2 \\
\boldsymbol{x}_1 = 180cm & \boldsymbol{z}_1 = [1, 0] \\
\boldsymbol{x}_2 = 170cm & \boldsymbol{z}_2 = [0, 1] \\
\boldsymbol{x}_3 = 166cm & \boldsymbol{z}_3 = [1, 0] \\
\boldsymbol{x}_4 = 175cm & \boldsymbol{z}_4 = [1, 0] \\
\boldsymbol{x}_5 = 160cm & \boldsymbol{z}_5 = [1, 0] \\
\boldsymbol{x}_6 = 155cm & \boldsymbol{z}_6 = [0, 1] \\
\boldsymbol{x}_7 = 165cm & \boldsymbol{z}_7 = [0, 1] \\
\boldsymbol{x}_8 = 162cm & \boldsymbol{z}_8 = [1, 0] \\
\boldsymbol{x}_9 = 150cm & \boldsymbol{z}_9 = [0, 1]
\end{array}
$$

Sufficient statistics for each class k (male or female)

$$S_k[1] = \sum_{n=1}^{N} z_{nk}$$

Count

$$S_k[\boldsymbol{x}] = \sum_{n=1}^{N} z_{nk}\, \boldsymbol{x}_n$$

Sum

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} z_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T$$

Ratio: $\pi_k = \dfrac{S_k[1]}{S_\cdot[1]}$ Average: $\boldsymbol{\mu}_k = \dfrac{S_k[\boldsymbol{x}]}{S_k[1]}$ Variance: $\boldsymbol{\Lambda}_k^{-1} = \dfrac{S_k[\boldsymbol{x}\boldsymbol{x}]}{S_k[1]} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T$

Parameter Estimation: Genders Unknown

- Use posteriors instead of latent variables

	$k=1 \quad k=2$	$k=1 \quad k=2$
$\boldsymbol{x}_1 = 180cm$	$\boldsymbol{z}_1 = [?,?]$	$p(\boldsymbol{z}_1 \mid \boldsymbol{X}) = [0.99, 0.01]$
$\boldsymbol{x}_2 = 170cm$	$\boldsymbol{z}_2 = [?,?]$	$p(\boldsymbol{z}_2 \mid \boldsymbol{X}) = [0.90, 0.10]$
$\boldsymbol{x}_3 = 166cm$	$\boldsymbol{z}_3 = [?,?]$	$p(\boldsymbol{z}_3 \mid \boldsymbol{X}) = [0.60, 0.40]$
$\boldsymbol{x}_4 = 175cm$	$\boldsymbol{z}_4 = [?,?]$	$p(\boldsymbol{z}_4 \mid \boldsymbol{X}) = [0.95, 0.05]$
$\boldsymbol{x}_5 = 160cm$	$\boldsymbol{z}_5 = [?,?]$	$p(\boldsymbol{z}_5 \mid \boldsymbol{X}) = [0.10, 0.90]$
$\boldsymbol{x}_6 = 155cm$	$\boldsymbol{z}_6 = [?,?]$	$p(\boldsymbol{z}_6 \mid \boldsymbol{X}) = [0.05, 0.95]$
$\boldsymbol{x}_7 = 165cm$	$\boldsymbol{z}_7 = [?,?]$	$p(\boldsymbol{z}_7 \mid \boldsymbol{X}) = [0.50, 0.50]$
$\boldsymbol{x}_8 = 162cm$	$\boldsymbol{z}_8 =$	$p(\boldsymbol{z}_8 \mid \boldsymbol{X}) = [0.30, 0.70]$
$\boldsymbol{x}_9 = 150cm$	$\boldsymbol{z}_9 =$	$p(\boldsymbol{z}_9 \mid \boldsymbol{X}) = [0.01, 0.99]$

Responsibility
$\gamma_n = [\gamma_{n1}, \gamma_{n2}]$

We cannot say $z_{nk} = 1$ for some k with absolute certainty

To deal with uncertainty, we estimate the posterior of $z_{nk} = 1$

Calculation of Sufficient Statistics

- Use posteriors instead of latent variables
  - Take into account the uncertainty of latent variables (genders)

Genders known

$$S_k[1] = \sum_{n=1}^{N} z_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} z_{nk}\,\boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} z_{nk}\,\boldsymbol{x}_n\boldsymbol{x}_n^T$$

Genders unknown

$$S_k[1] = \sum_{n=1}^{N} \gamma_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} \gamma_{nk}\,\boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} \gamma_{nk}\,\boldsymbol{x}_n\boldsymbol{x}_n^T$$

Ratio: $\pi_k^* = \frac{S_k[1]}{S_\cdot[1]}$ Average: $\boldsymbol{\mu}_k^* = \frac{S_k[\boldsymbol{x}]}{S_k[1]}$ Variance: ${\boldsymbol{\Lambda}_k^{-1}}^* = \frac{S_k[\boldsymbol{x}\boldsymbol{x}]}{S_k[1]} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$

How to estimate z or γ

K-means algorithm (hard EM)
(deterministic hard assignment)

EM algorithm
(deterministic soft assignment)

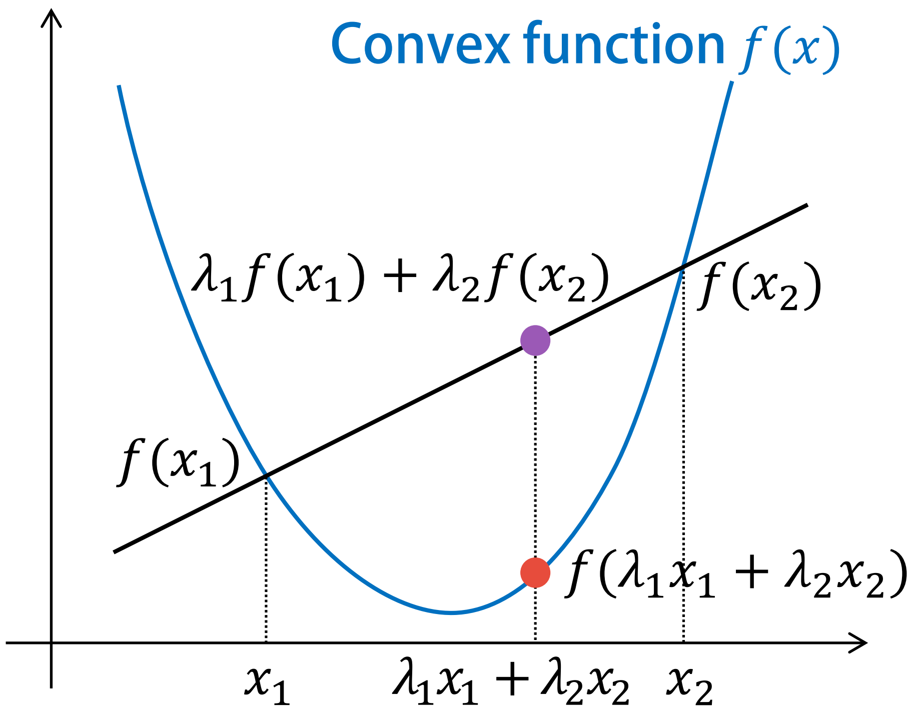
Probabilistic Modeling

- Generative story of the GMM
  - Draw each latent variable: $\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$
  - Draw each observed variable: $\boldsymbol{x}_n \sim \prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$
- Two major approaches

	Maximum likelihood (ML) estimation	Bayesian estimation
Probabilistic model	$p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ $= p(\boldsymbol{X}\|\boldsymbol{Z}; \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}; \boldsymbol{\pi})$	$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ $= p(\boldsymbol{X}\|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}, \boldsymbol{\pi})p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$
Latent variables $\boldsymbol{Z}$	Posterior calculation $p(\boldsymbol{Z}\|\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$	Posterior calculation $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\|\boldsymbol{X})$
Parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$	Point estimation $\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\Lambda}^* = \mathrm{argmax}\, p(\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$	

Jensen's Inequality

- A basic inequality for convex functions
  - Forms the basis of the EM and VB algorithms



$$f\left(\sum_{k=1}^{K} \lambda_k x_k\right) \leq \sum_{k=1}^{K} \lambda_k f(x_k)$$

for auxiliary variables $\boldsymbol{\lambda}$

such that $\sum_{k=1}^{K} \lambda_k = 1$

$$f\left(\int q(x)\, x\, dx\right) \leq \int q(x) f(x) dx$$

for auxiliary distribution $q(x)$

such that $\int q(x) dx = 1$

How to Use Jensen's Inequality

- Change the order of "sum" and "convex function"
  - Example: negative log of sum → sum of negative log

$$-\log\left(\sum_{k=1}^{K} x_k\right) = -\log\left(\sum_{k=1}^{K} \lambda_k \frac{x_k}{\lambda_k}\right) \le -\sum_{k=1}^{K} \lambda_k \log\left(\frac{x_k}{\lambda_k}\right) \stackrel{\text{def}}{=} U(\boldsymbol{\lambda})$$

Upper bound

When does the equality holds true (when is $U(\boldsymbol{x}, \boldsymbol{\lambda})$ minimized)?
→ Optimization problem with a constraint
→ Method of Lagrange multipliers

$$\sum_{k=1}^{K} \lambda_k = 1$$

$$F(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} U(\boldsymbol{\lambda}) + \omega\left(1 - \sum_{k=1}^{K} \lambda_k\right) \longrightarrow \frac{\partial F(\boldsymbol{\lambda})}{\partial \lambda_k} = -\log x_k + \log \lambda_k + 1 - \omega$$

Solving $\frac{\partial F(\boldsymbol{\lambda})}{\partial \lambda_k} = 0$, we get $\lambda_k = x_k e^{\omega - 1}$ $\longrightarrow e^{\omega-1} = \frac{1}{\sum_{k=1}^{K} x_k} \longrightarrow$

Equality condition

$$\boxed{\lambda_k = \frac{x_k}{\sum_{k=1}^{K} x_k}}$$

How to Use Jensen's Inequality

- Change the order of "sum" and "convex function"
  - Example: negative log of sum → sum of negative log

$$-\log \int p(x,z)dz = -\log \int q(z)\frac{p(x,z)}{q(z)}dz \le -\int q(z)\log\frac{p(x,z)}{q(z)} \stackrel{\text{def}}{=} U(q(x))$$

Upper bound

When does the equality holds true (when is $U(q(x))$ minimized)?

→ Optimization problem with a constraint

→ Method of Lagrange multipliers

$$\sum_{k=1}^{K} q(x) = 1$$

$$F(q(x)) \stackrel{\text{def}}{=} U(q(x)) + \omega\left(1 - \int q(x)dx\right)$$ → Minimize as in the previous slide

Equality condition

$$q(z) = \frac{p(x,z)}{\int p(x,z)dz} = \frac{p(x,z)}{p(x)} = p(z|x)$$

The Expectation-Maximization Algorithm

- A deterministic algorithm for ML estimation
  - Suppose a probabilistic model $p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{Z}; \boldsymbol{\theta})p(\boldsymbol{Z}; \boldsymbol{\theta})$
    - $\boldsymbol{X}$: observed variables $\boldsymbol{Z}$: latent variables $\boldsymbol{\theta}$: parameters
  - We aim to get ML estimates $\boldsymbol{\theta}^* = \operatorname{argmax} p(\boldsymbol{X}; \boldsymbol{\theta})$ Intractable!

$$\log p(\boldsymbol{X}; \boldsymbol{\theta}) = \log \int p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) dZ$$

$$= \log \int q(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z}$$

Introduce an arbitrary distribution $q(\boldsymbol{Z})$ called a variational distribution

$$\geq \int q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z}$$

Jensen's inequality

The equality holds true when $q^*(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\theta})$

E-step

$$= \int q(\boldsymbol{Z}) \log p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) dZ - \int q(\boldsymbol{Z}) \log q(\boldsymbol{Z})$$

M-step $\longrightarrow$ Maximize lower bound with respect to $\boldsymbol{\theta}$

Hard EM: $q^*(\boldsymbol{Z}) = \delta_{\boldsymbol{Z}^*}(\boldsymbol{Z})$

$\boldsymbol{Z}^* = \operatorname{argmax} p(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\theta})$

E Step for GMM

- Iterate E-step and M-step alternately
  - E-step: Calculate a posterior distribution over latent variables $\boldsymbol{Z}$

$$\boldsymbol{x}_1 = 180cm \quad \boldsymbol{z}_1 = [?\,,?\,] \qquad \boldsymbol{\gamma}_1 = p(\boldsymbol{z}_1|\boldsymbol{X}) = [0.99, 0.01]$$
$$\boldsymbol{x}_2 = 170cm \quad \boldsymbol{z}_2 = [?\,,?\,] \qquad \boldsymbol{\gamma}_2 = p(\boldsymbol{z}_2|\boldsymbol{X}) = [0.90, 0.10]$$
$$\boldsymbol{x}_3 = 166cm \quad \boldsymbol{z}_3 = [?\,,?\,] \qquad \boldsymbol{\gamma}_3 = p(\boldsymbol{z}_3|\boldsymbol{X}) = [0.60, 0.40]$$

$$q^*(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \gamma_{nk}^{z_{nk}}$$

Responsibility

$$q^*(z_{nk}=1) = p(z_{nk}=1|\boldsymbol{x}_n;\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$$
$$= \frac{p(\boldsymbol{x}_n, z_{nk}=1;\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}{\sum_{k'=1}^{K} p(\boldsymbol{x}_n, z_{nk'}=1;\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}$$
$$= \frac{\pi_k N\left(\boldsymbol{x}_n\middle|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)}{\sum_{k'=1}^{K} \pi_{k'} N\left(\boldsymbol{x}_n\middle|\boldsymbol{\mu}_{k'}, \boldsymbol{\Lambda}_{k'}^{-1}\right)} = \gamma_{nk}$$

How well the sample $\boldsymbol{x}_n$ is explained by each cluster

||

How likely the sample $\boldsymbol{x}_n$ was to be generated from each cluster

M Step for GMM

- Iterate E-step and M-step alternately
  - M-step: Update parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$
    - Calculate sufficient statistics

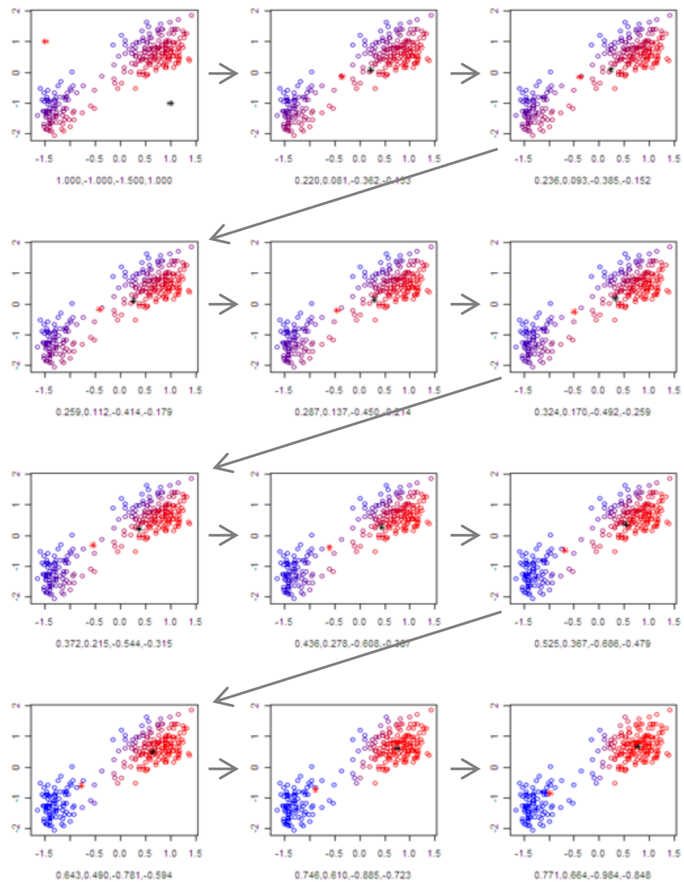
$$S_k[1] = \sum_{n=1}^{N} \gamma_{nk} \qquad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T$$

    - Estimate parameters

Ratio: $\pi_k^* = \dfrac{S_k[1]}{S_.[1]}$ Mean: $\boldsymbol{\mu}_k^* = \dfrac{S_k[\boldsymbol{x}]}{S_k[1]}$

Variance: $\boldsymbol{\Lambda}_k^{-1^*} = \dfrac{S_k[\boldsymbol{xx}]}{S_k[1]} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T$

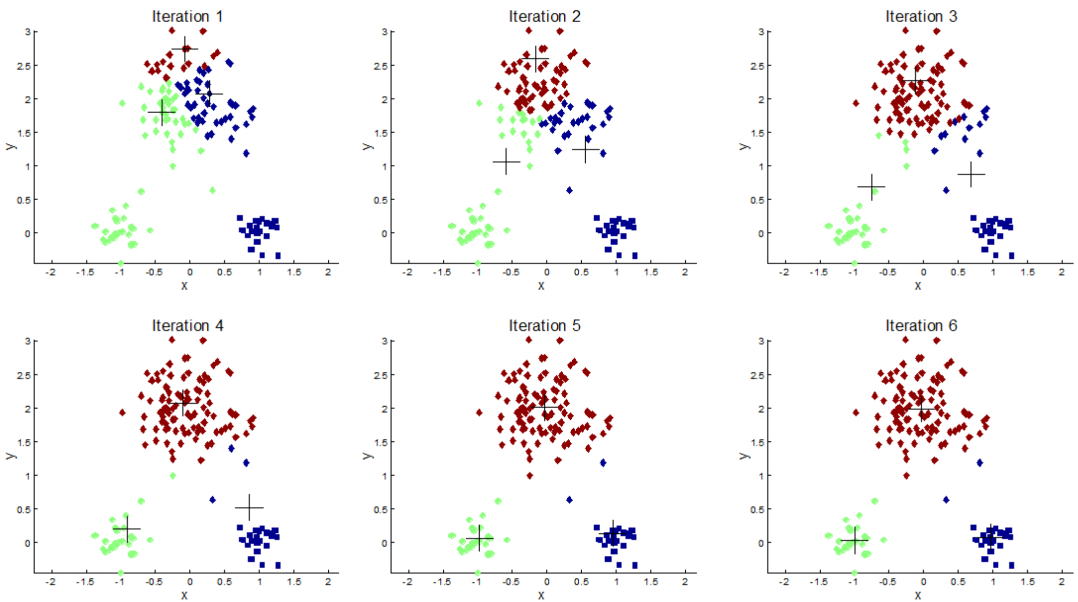


Hard EM Algorithm for GMM

- Iterate M-step and M-step alternately
  - M-step: Update latent variables $\boldsymbol{Z}$

    Hard assignment

    - $\boldsymbol{Z}^* = \operatorname{argmax} p(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \operatorname{argmax} p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$
  - M-step: Update parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$
    - $\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\Lambda}^* = \operatorname{argmax} p(\boldsymbol{X}|\boldsymbol{Z}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \operatorname{argmax} p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$



If the all $\boldsymbol{\Lambda}_k$'s are same, the hard EM for GMM reduces to the *k*-means algorithm

Hard EM vs. EM

- A key difference lies in how to deal with uncertainty

	K-means algorithm	EM algorithm
Latent variables $\boldsymbol{Z}$	Optimizing	Marginalizing out
Parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$	Optimizing	Optimizing

$$S_k[1] = \sum_{n=1}^{N} z_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} z_{nk}\,\boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} z_{nk}\,\boldsymbol{x}_n\boldsymbol{x}_n^T$$

$$S_k[1] = \sum_{n=1}^{N} \gamma_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} \gamma_{nk}\,\boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} \gamma_{nk}\,\boldsymbol{x}_n\boldsymbol{x}_n^T$$

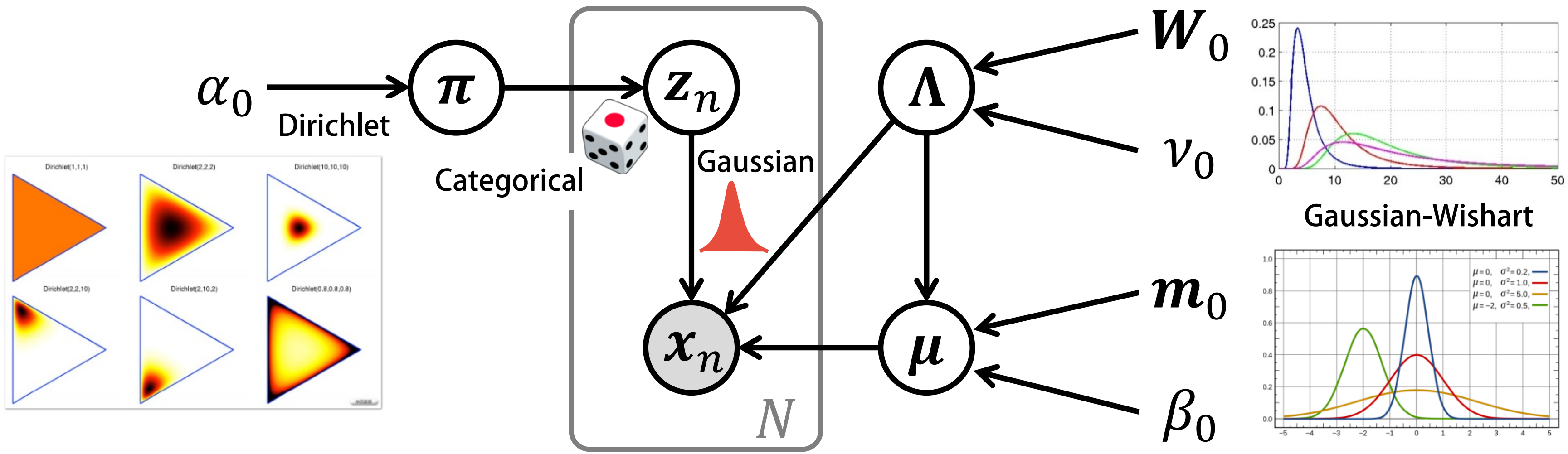
Ratio: $\pi_k^* = \frac{S_k[1]}{S_.[1]}$ Mean: $\boldsymbol{\mu}_k^* = \frac{S_k[\boldsymbol{x}]}{S_k[1]}$ Variance: ${\boldsymbol{\Lambda}_k^{-1}}^* = \frac{S_k[\boldsymbol{x}\boldsymbol{x}]}{S_k[1]} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$

Bayesian Estimation of Finite Gaussian Mixture Models

(Collapsed) Gibbs Sampling
(Collapsed) Variational Bayes

Bayesian Approach

- Regard parameters as random variables
  - Introduce prior distributions on parameters
    - The Dirichlet distribution
      - A conjugate prior on categorical distributions
    - The Gaussian-Wishart distribution
      - A conjugate prior on Gaussian distributions



Bayesian Approach

- Regard parameters as random variables
  - Introduce prior distributions on parameters
  - Calculate posterior distributions on random variables

Maximum likelihood estimation

Latent variables: $p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$

Ratio: $\pi_k^* = \frac{S_k[1]}{S_\cdot[1]}$ Mean: $\boldsymbol{\mu}_k^* = \frac{S_k[\boldsymbol{x}]}{S_k[1]}$ Variance: ${\boldsymbol{\Lambda}_k^{-1}}^* = \frac{S_k[\boldsymbol{x}\boldsymbol{x}]}{S_k[1]} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$

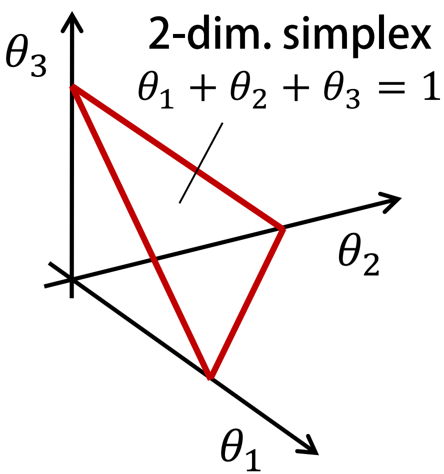
Bayesian estimation

$$\underbrace{p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})}_{\text{Likelihood}} \times \underbrace{p(\boldsymbol{\pi})p(\boldsymbol{\mu},\boldsymbol{\Lambda})}_{\text{Prior}} \longrightarrow \underbrace{p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X})}_{\text{Posterior}}$$

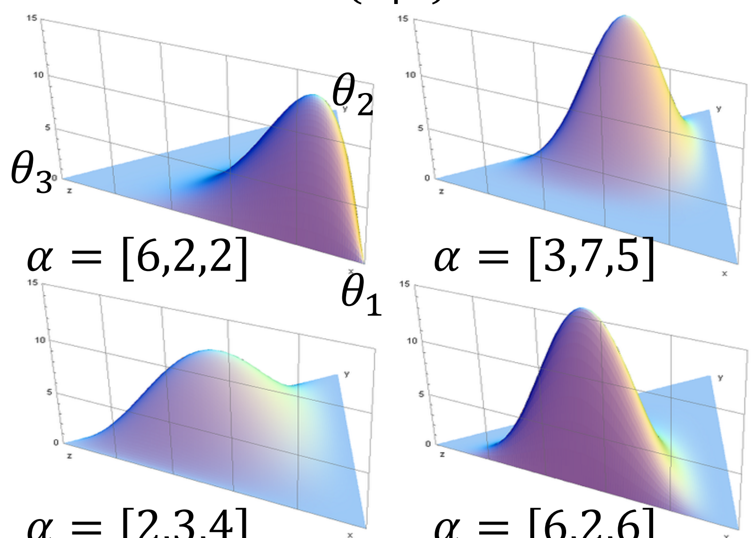
Bayes' theorem: $p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu},\boldsymbol{\Lambda})}{p(\boldsymbol{X})}$

Conjugate Priors

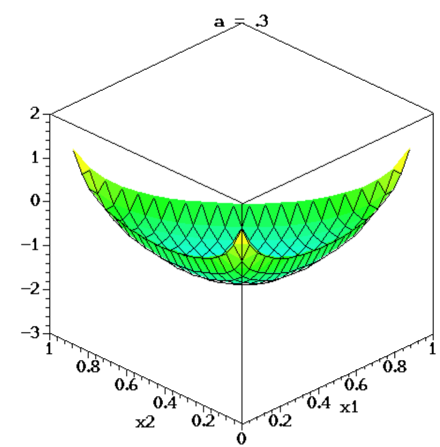
- Widely used for mathematical convenience
  - The posterior $p(\boldsymbol{\theta}|\boldsymbol{X})$ takes the same form of the prior $p(\boldsymbol{\theta})$ for a particular type of the likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$
    - $p(\boldsymbol{\pi}), p(\boldsymbol{\pi}|\boldsymbol{Z})$: Dirichlet $\qquad p(\boldsymbol{Z}|\boldsymbol{\pi})$: Categorical
    - $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}), p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z})$: Gaussian-Wishart $\qquad p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$: Gaussian



$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$



Changing α from 0 to 2



Probabilistic Modeling

- Generative story of the GMM
  - Draw each latent variable: $\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$
  - Draw each observed variable: $\boldsymbol{x}_n \sim \prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$
- Two major approaches

	Maximum likelihood (ML) estimation	Bayesian estimation
Probabilistic model	$p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ $= p(\boldsymbol{X}\|\boldsymbol{Z}; \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}; \boldsymbol{\pi})$	$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ $= p(\boldsymbol{X}\|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}, \boldsymbol{\pi})p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$
Latent variables $\boldsymbol{Z}$	Posterior calculation $p(\boldsymbol{Z}\|\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$	Posterior calculation $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\|\boldsymbol{X})$
Parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$	Point estimation $\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\Lambda}^* = \mathrm{argmax}\, p(\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$	

Posterior Calculation: Genders Known

- Estimate the ratios, averages, and variances

$k=1 \quad k=2$

$\boldsymbol{x}_1 = 180cm$	$\boldsymbol{z}_1 = [1, 0]$
$\boldsymbol{x}_2 = 170cm$	$\boldsymbol{z}_2 = [0, 1]$
$\boldsymbol{x}_3 = 166cm$	$\boldsymbol{z}_3 = [1, 0]$
$\boldsymbol{x}_4 = 175cm$	$\boldsymbol{z}_4 = [1, 0]$
$\boldsymbol{x}_5 = 160cm$	$\boldsymbol{z}_5 = [1, 0]$
$\boldsymbol{x}_6 = 155cm$	$\boldsymbol{z}_6 = [0, 1]$
$\boldsymbol{x}_7 = 165cm$	$\boldsymbol{z}_7 = [0, 1]$
$\boldsymbol{x}_8 = 162cm$	$\boldsymbol{z}_8 = [1, 0]$
$\boldsymbol{x}_9 = 150cm$	$\boldsymbol{z}_9 = [0, 1]$

Sufficient statistics for each cluster k (male or female)

$$S_k[1] = \sum_{n=1}^{N} z_{nk}$$

Count

$$S_k[\boldsymbol{x}] = \sum_{n=1}^{N} z_{nk}\, \boldsymbol{x}_n$$

Sum

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} z_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T$$

How to calculate the posterior distribution $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{X}, \boldsymbol{Z})$?

Bayesian Estimation for Categorical Distribution

- Calculate a posterior distribution on parameters $\boldsymbol{\pi}$
  - The generative story
    - Prior: $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}_0)$
    - Likelihood: $\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$

Posterior count

Prior count

Actual count

$$\alpha_k = \alpha_{0k} + S_k[1]$$

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_{0k}\right)}{\prod_{k=1}^{K} \Gamma(\alpha_{0k})} \prod_{k=1}^{K} \pi_k^{\alpha_{0k}-1}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

Bayes' theorem:

$$p(\boldsymbol{\pi}|\boldsymbol{Z}) = \frac{p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{p(\boldsymbol{Z})} \propto p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})$$

$$p(\boldsymbol{\pi}|\boldsymbol{Z}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \propto \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_{0k}\right)}{\prod_{k=1}^{K} \Gamma(\alpha_{0k})} \prod_{k=1}^{K} \pi_k^{\alpha_{0k}+S_k[1]-1}$$

We do not need to directly calculate the normalizing factor

Bayesian Estimation for Gaussian Distribution

- Calculate a posterior distribution on parameters $\boldsymbol{\mu}, \boldsymbol{\Lambda}$
  - The generative story
    - Prior: $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \sim N(\boldsymbol{\mu}_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0)$
    - Likelihood: $\boldsymbol{x}_n \sim \prod_{k=1}^{K} N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0)$$

$$p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

Bayes' theorem

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{X}, \boldsymbol{Z}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k | \boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k | \mathbf{W}_k, \nu_k)$$

Posterior count

Prior count

Actual count

$$\beta_k = \beta_0 + S_k[1]$$

$$\boldsymbol{m}_k = \frac{\beta_0 m_k + S_k[\boldsymbol{x}]}{\beta_0 + S_k[1]}$$

$$\nu_k = \nu_0 + S_k[1]$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + S_k[\boldsymbol{x}\boldsymbol{x}^T] - \beta_k \boldsymbol{m}_k \boldsymbol{m}_k^T$$

Posterior Estimation: Genders Unknown

- Use posteriors instead of latent variables

	$k=1\ \ k=2$	$k=1\quad k=2$
$\boldsymbol{x}_1 = 180cm$	$\boldsymbol{z}_1 = [?,?]$	$p(\boldsymbol{z}_1 \mid \boldsymbol{X}) = [0.99, 0.01]$
$\boldsymbol{x}_2 = 170cm$	$\boldsymbol{z}_2 = [?,?]$	$p(\boldsymbol{z}_2 \mid \boldsymbol{X}) = [0.90, 0.10]$
$\boldsymbol{x}_3 = 166cm$	$\boldsymbol{z}_3 = [?,?]$	$p(\boldsymbol{z}_3 \mid \boldsymbol{X}) = [0.60, 0.40]$
$\boldsymbol{x}_4 = 175cm$	$\boldsymbol{z}_4 = [?,?]$	$p(\boldsymbol{z}_4 \mid \boldsymbol{X}) = [0.95, 0.05]$
$\boldsymbol{x}_5 = 160cm$	$\boldsymbol{z}_5 = [?,?]$	$p(\boldsymbol{z}_5 \mid \boldsymbol{X}) = [0.10, 0.90]$
$\boldsymbol{x}_6 = 155cm$	$\boldsymbol{z}_6 = [?,?]$	$p(\boldsymbol{z}_6 \mid \boldsymbol{X}) = [0.05, 0.95]$
$\boldsymbol{x}_7 = 165cm$	$\boldsymbol{z}_7 = [?,?]$	$p(\boldsymbol{z}_7 \mid \boldsymbol{X}) = [0.50, 0.50]$
$\boldsymbol{x}_8 = 162cm$	$\boldsymbol{z}_8 = [?,?]$	$p(\boldsymbol{z}_8 \mid \boldsymbol{X}) = [0.30, 0.70]$
$\boldsymbol{x}_9 = 150cm$	$\boldsymbol{z}_9 =$	$p(\boldsymbol{z}_9 \mid \boldsymbol{X}) = [0.01, 0.99]$

Responsibility
$\gamma_n = [\gamma_{n1}, \gamma_{n2}]$

We cannot say $z_{nk} = 1$ for some k with absolute certainty

To deal with uncertainty, we estimate the posterior of $z_{nk} = 1$

Calculation of Sufficient Statistics

- Use posteriors instead of latent variables
  - Take into account the uncertainty of latent variables (genders)

Hard assignment

$$S_k[1] = \sum_{n=1}^{N} z_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} z_{nk}\, \boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} z_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T$$

Soft assignment

$$S_k[1] = \sum_{n=1}^{N} \gamma_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T$$

How to estimate z or γ

Gibbs sampling
(stochastic algorithm)

Variational Bayes
(deterministic algorithm)

Probabilistic Modeling

- Generative story of the GMM
  - Draw each latent variable: $\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi})$
  - Draw each observed variable: $\boldsymbol{x}_n \sim \prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$
- Two major approaches

	Maximum likelihood (ML) estimation	Bayesian estimation
Probabilistic model	$p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ $= p(\boldsymbol{X}\|\boldsymbol{Z}; \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}; \boldsymbol{\pi})$	$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ $= p(\boldsymbol{X}\|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}, \boldsymbol{\pi})p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$
Latent variables $\boldsymbol{Z}$	Posterior calculation $p(\boldsymbol{Z}\|\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$	Posterior calculation $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\|\boldsymbol{X})$
Parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$	Point estimation $\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\Lambda}^* = \mathrm{argmax}\, p(\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$	

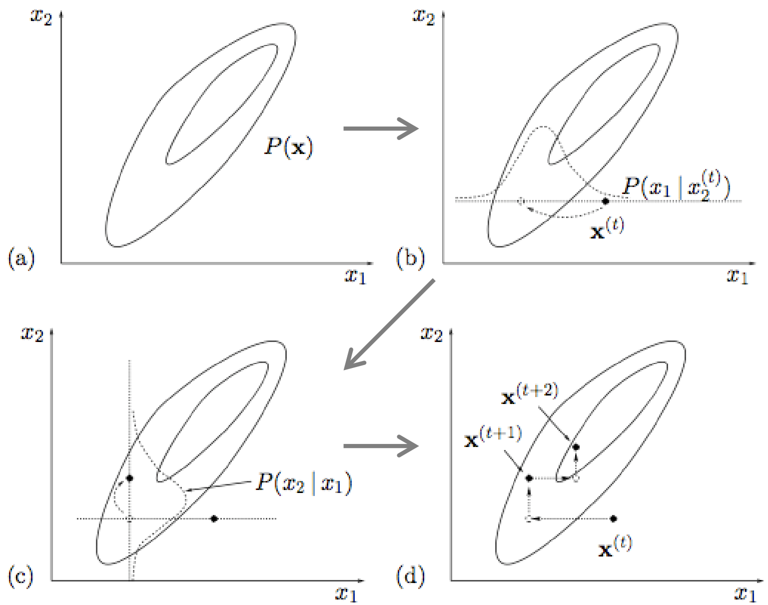
Gibbs Sampling vs. Variational Bayes

- Choose an appropriate approach according to situations
  - Each approach has pros and cons
  - In general, Gibbs sampling is easy to implement

	Gibbs sampling	Variational Bayes
Convergence to true posterior	Yes	No
Judgment of convergence	Difficult	Easy
Convergence speed	Slow	Fast
Quality of estimation results	High	Moderate

The Gibbs Sampling

- A popular variant of Markov chain Monte Carlo (MCMC)
  - Generate random samples from a probability distribution $p(\boldsymbol{X}) = \frac{f(\boldsymbol{X})}{Z}$ even if the normalizing factor Z is intractable
  - The acceptance ratio is 100%

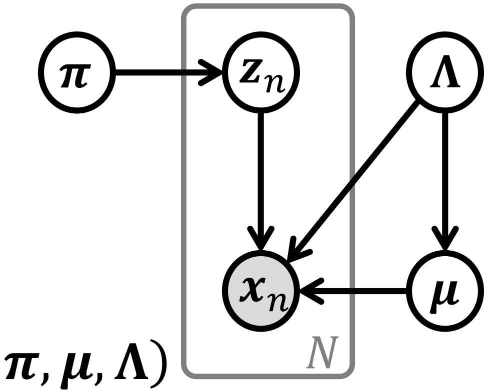


Objective: Generate independent samples from a probability distribution $p(\boldsymbol{X})$

1. Divide $\boldsymbol{X}$ into several groups $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_M$
2. for $t = 1{:}T$
   for $m = 1{:}M$
   Sample $\boldsymbol{X}_m^{(t+1)}$
   $\sim p\left(\boldsymbol{X}_m^{(t+1)} \middle| \boldsymbol{X}_1^{(t+1)}, \cdots, \boldsymbol{X}_{m-1}^{(t+1)}, \boldsymbol{X}_{m+1}^{(t)}, \cdots, \boldsymbol{X}_M^{(t)}\right)$
3. Pick up $\boldsymbol{X}^{(t)}$ with a certain interval

This sampling needs to be done easily

Gibbs Sampling for GMM



- Generate samples from $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$
  - Divide $\{\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ into $\{\boldsymbol{z}_1\}, \{\boldsymbol{z}_2\}, \cdots, \{\boldsymbol{z}_N\}, \{\boldsymbol{\pi}\}, \{\boldsymbol{\mu}, \boldsymbol{\Lambda}\}$
  - Iterate until convergence
    - for $n = 1{:}\,N$
      - Sample $\boldsymbol{z}_n \sim p(\boldsymbol{z}_n|\boldsymbol{X}, \boldsymbol{Z}_{-n}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$
    - Sample $\boldsymbol{\pi} \sim p(\boldsymbol{\pi}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\pi}|\boldsymbol{Z})$
    - Sample $\boldsymbol{\mu}, \boldsymbol{\Lambda} \sim p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}) = p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z})$

$$p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)}{\sum_{k'=1}^{K} \pi_{k'} N(\boldsymbol{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Lambda}_{k'})}$$

EM algorithm: soft assignment

Gibbs sampling: hard assignment

$$p(\boldsymbol{\pi}|\boldsymbol{Z}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

See "Posterior Calculation: Genders Known"

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)$$

The Variational Bayes

- A Bayesian extension of the EM algorithm
  - We aim to approximate a true posterior $p(\boldsymbol{Z}|\boldsymbol{X}) = p(\boldsymbol{Z}|\boldsymbol{X})/p(\boldsymbol{X})$ as a factorizable distribution $q(\boldsymbol{Z}) = \prod_{m=1}^{M} q(\boldsymbol{Z}_m)$

Intractable!

$$\log p(\boldsymbol{X}) = \log \int p(\boldsymbol{X},\boldsymbol{Z})d\boldsymbol{Z} = \log \int q(\boldsymbol{Z}) \frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \geq \int q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z}$$

Jensen's inequality

$$= \int \left(\prod_{m=1}^{M} q(\boldsymbol{Z}_m) \right) \left(\log p(\boldsymbol{X},\boldsymbol{Z}) - \sum_{m=1}^{M} \log q(\boldsymbol{Z}_m) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M$$

$$= \sum_{m=1}^{M} \left(\int q(\boldsymbol{Z}_m) \left(\int q(\boldsymbol{Z}_{-m}) \log p(\boldsymbol{X},\boldsymbol{Z}) d\boldsymbol{Z}_{-m} \right) d\boldsymbol{Z}_m - \int q(\boldsymbol{Z}_m) \log q(\boldsymbol{Z}_m) d\boldsymbol{Z}_m \right)$$

The lower bound is maximized when $\log q^*(\boldsymbol{Z}_m) = \langle \log p(\boldsymbol{X},\boldsymbol{Z}) \rangle_{q(\boldsymbol{Z}_{-m})} + \text{const.}$

The equality does <u>NOT</u> hold true!

VB-E step

VB-M step

Accuracy of Lower Bounding

- VB just approximates a true posterior $p(\boldsymbol{Z}|\boldsymbol{X})$
  - The accuracy depends on how to factorize a variational posterior $q(\boldsymbol{Z})$

$$\log p(\boldsymbol{X}) = \int q(\boldsymbol{Z}) \log p(\boldsymbol{X}) dZ = \int q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z}) p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z}) p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z}$$

$$= \int q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} + \int q(\boldsymbol{Z}) \log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z}$$

$$= \mathrm{LowerBound}(q) + \mathrm{KL}(q||p)$$

Maximize = Minimize

Kullback-Leibler (KL) divergence
between
a variational posterior $q(\boldsymbol{Z})$
and a true posterior $p(\boldsymbol{Z}|\boldsymbol{X})$

The KD divergence is 0 when $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ (intractable!)

If $q(\boldsymbol{Z})$ is assumed to be factorized, the KD divergence cannot be 0!

VB for GMM

- Approximate a true posterior $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{X})$
  - Assume a variational distribution $q(\boldsymbol{Z})q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \approx p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{X})$
  - Iteratively update (optimize) each factor
    - VB-E step

$$
\begin{aligned}
\log q^*(\boldsymbol{Z}) &= \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \mathrm{const.} \\
&= \langle \log p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z} | \boldsymbol{\pi}) \rangle_{q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \mathrm{const.}
\end{aligned}
$$

    - VB-M step

$$
\begin{aligned}
\log q^*(\boldsymbol{\pi}) &= \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \mathrm{const.} \\
&= \langle \log p(\boldsymbol{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}) \rangle_{q(\boldsymbol{Z})} + \mathrm{const.}
\end{aligned}
$$

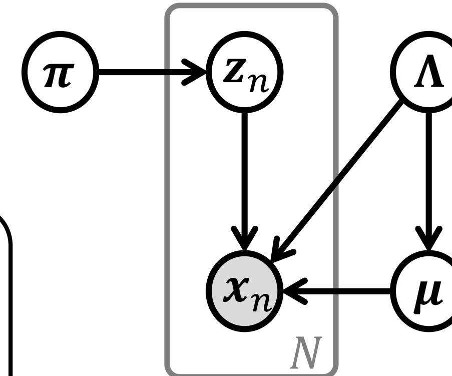
$$
\begin{aligned}
\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{\pi})} + \mathrm{const.} \\
&= \langle \log p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z})} + \mathrm{const.}
\end{aligned}
$$

Tractable posteriors: Use responsibilities instead of latent variables

Formulation of GMM

- Formulate a full joint distribution

$$p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu},\boldsymbol{\Lambda})$$



$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}$$

Likelihood functions

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_{0k}\right)}{\prod_{k=1}^{K}\Gamma(\alpha_{0k})}\prod_{k=1}^{K}\pi_k^{\alpha_{0k}-1}$$

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\Lambda}_k|\mathbf{W}_0,\nu_0)$$

Prior distributions

VB-E Step for GMM

- Invoke the updating formula of VB
  - Take the expectation of the full joint probability distribution under "factorized" variational posteriors over other variables
  - Focus on only terms including $\boldsymbol{Z}$
    (other terms can be absorbed into the normalization factor)

$$\begin{aligned}\log q^*(\boldsymbol{Z}) &= \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.} \\ &= \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.} \\ &= \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{\pi}) \rangle_{q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.}\end{aligned}$$

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \text{Categorical}(\boldsymbol{z}_n | \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

VB-E Step for GMM

- Proceed the calculation according the updating rule

$$\langle \log p(\boldsymbol{Z}|\boldsymbol{\pi}) \rangle_{q(\boldsymbol{\pi})} = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \langle \log \pi_k \rangle_{q(\boldsymbol{\pi})}$$

$$\langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})} = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\langle \log N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) \right\rangle_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}$$

$$\begin{aligned} \log q^*(\boldsymbol{Z}) &= \langle \log p(\boldsymbol{Z}|\boldsymbol{\pi}) \rangle_{q(\boldsymbol{\pi})} + \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})} + \text{const.} \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\langle \log \pi_k \rangle_{q(\boldsymbol{\pi})} + \left\langle \log N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) \right\rangle_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)} \right) + \text{const.} \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \rho_{nk} + \text{const.} \end{aligned}$$

VB-E Step for GMM

- Calculate the variational posterior over latent variables $\boldsymbol{Z}$
  - The normalization factor is automatically determined

$$\log q^*(\boldsymbol{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \log \rho_{nk} + \text{const.}$$

The distribution should be appropriately normalized

$$\gamma_{nk} = \frac{\rho_{nk}}{\sum_{k'=1}^{K} \rho_{nk'}}$$

$$\log q^*(\boldsymbol{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \log \gamma_{nk}$$

Latent variables are categorical distributed!

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \gamma_{nk}^{z_{nk}} = \prod_{n=1}^{N} \text{Categorical}\left(\boldsymbol{z}_n \middle| \boldsymbol{\gamma}_n\right)$$

VB-M Step for GMM

- Invoke the updating formula of VB
  - Take the expectation of the full joint probability distribution under "factorized" variational posteriors over other variables
  - Focus on only terms including $\boldsymbol{Z}$ (other terms can be absorbed into the normalization factor)

$$
\begin{aligned}
\log q^*(\boldsymbol{\pi}) &= \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.} \\
&= \langle \log p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) \rangle_{q(\boldsymbol{Z})} + \text{const.} \\
&= \log p(\boldsymbol{\pi}) + \langle \log p(\boldsymbol{Z}|\boldsymbol{\pi}) \rangle_{q(\boldsymbol{Z})} + \text{const.}
\end{aligned}
$$

$$
\begin{aligned}
\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{\pi})} + \text{const.} \\
&= \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z})} + \text{const.} \\
&= \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z})} + \text{const.}
\end{aligned}
$$

Bayesian estimation in simple conjugate models!
(Use responsibilities $q(\boldsymbol{Z})$ instead of latent variables $\boldsymbol{Z}$)

VB-M Step for GMM

- Calculate the variational posteriors over parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$
  - The posteriors take the same forms of the priors

$$S_k[1] = \sum_{n=1}^{N} \gamma_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n \quad S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T$$

Sufficient statistics

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0)$$

$$\downarrow$$

$$q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

$$\alpha_k = \alpha_{0k} + S_k[1]$$

Posterior count

Prior count

Actual count

$$\beta_k = \beta_0 + S_k[1]$$

$$\boldsymbol{m}_k = \frac{\beta_0 m_k + S_k[\boldsymbol{x}]}{\beta_0 + S_k[1]}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0)$$

$$\downarrow$$

$$q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)$$

$$\nu_k = \nu_0 + S_k[1]$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + S_k[\boldsymbol{x}\boldsymbol{x}^T] - \beta_k \boldsymbol{m}_k \boldsymbol{m}_k^T$$

EM vs. VB

- Both methods have similar updating formulas
  - EM: Using the values of parameters

$$\log p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(\log \pi_k + \log N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)\right) + \text{const.}$$

  - VB: Using the geometric means of parameters

$$\log q^*(\boldsymbol{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(\langle\log \pi_k\rangle_{q(\boldsymbol{\pi})} + \left\langle\log N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)\right\rangle_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}\right) + \text{const.}$$

$$\left\langle\log N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)\right\rangle_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)} = -\frac{D}{2}\log(2\pi) + \frac{1}{2}\langle\log\boldsymbol{\Lambda}_k\rangle - \frac{1}{2}\left(\frac{D}{\beta_k^{-1}} + \nu_k(\boldsymbol{x}_m - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\boldsymbol{x}_m - \boldsymbol{m}_k)\right)$$

$$\langle\log|\boldsymbol{\Lambda}_k|\rangle_{q(\boldsymbol{\pi})} = \sum_{d=1}^{D}\psi\left(\frac{c_k + 1 - d}{2}\right) + D\log 2 + \log|\boldsymbol{W}_k|$$

Log Function vs. Digamma Function

- The digamma function results in sparsifying effect
  - Example: Dirichlet distribution

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

- Mean

$$\mathrm{E}[\pi_k] = \frac{\alpha_k}{\sum_{k'=1}^{K} \alpha_{k'}}$$

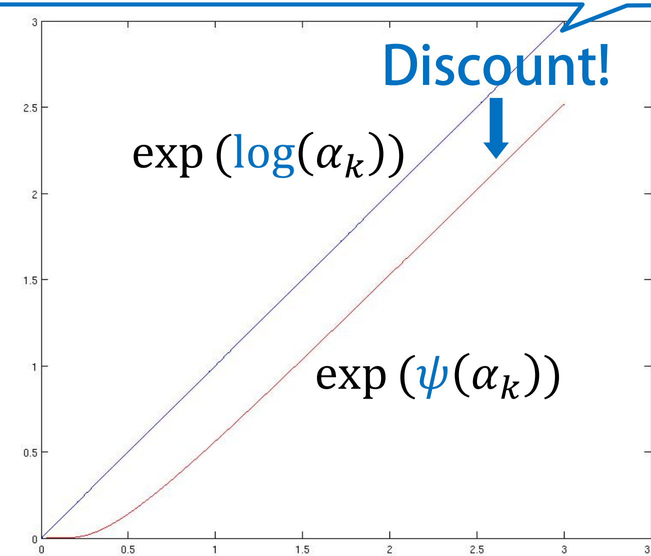
$$= \exp\left(\log(\alpha_k) - \log\left(\sum_{k'=1}^{K} \alpha_{k'}\right)\right)$$

- Geometric mean

$$\mathrm{G}[\pi_k] = \exp(\mathrm{E}[\log \pi_k])$$

$$= \exp\left(\psi(\alpha_k) - \psi\left(\sum_{k'=1}^{K} \alpha_{k'}\right)\right)$$

Small components tend to be degenerated in Bayesian mixture modeling



$\exp(\psi(0.1)) = 0.00003$

$\exp(\psi(0.5)) = 0.140$

$\exp(\psi(0.9)) = 0.470$

$\exp(\psi(1)) = 0.561$

$\exp(\psi(10)) = 9.504$

$\exp(\psi(100)) = 99.5004$

$\exp(\psi(1000)) = 999.500$

GS vs. EM vs. VB

- Both methods are based on similar updating formulas
  - GS: Stochastic hard assignment

$$p(z_{nk}=1|\boldsymbol{x}_n,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=\frac{\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}{\sum_{k'=1}^{K}\pi_{k'}N(\boldsymbol{x}_n|\boldsymbol{\mu}_{k'},\boldsymbol{\Lambda}_{k'})}$$

  - EM: Deterministic soft assignment

$$q^*(z_{nk}=1|\boldsymbol{x}_n,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=\frac{\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}{\sum_{k'=1}^{K}\pi_{k'}N(\boldsymbol{x}_n|\boldsymbol{\mu}_{k'},\boldsymbol{\Lambda}_{k'})}$$

  - VB: Deterministic soft assignment

$$q^*(z_{nk}=1)=\frac{\mathrm{G}[\pi_k]\mathrm{G}[N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)]}{\sum_{k'=1}^{K}\mathrm{G}[\pi_{k'}]\mathrm{G}[N(\boldsymbol{x}_n|\boldsymbol{\mu}_{k'},\boldsymbol{\Lambda}_{k'})]}$$

$$S_k[1]=\sum_{n=1}^{N}z_{nk}$$

$$S_k[\boldsymbol{x}]=\sum_{n=1}^{N}z_{nk}\,\boldsymbol{x}_n$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T]=\sum_{n=1}^{N}z_{nk}\,\boldsymbol{x}_n\boldsymbol{x}_n^T$$

Replace z_{nk} with γ_{nk}

"Collapsed" Algorithms

- Reduce the number of variables for fast/better estimation
  - The parameters can be marginalized out due to conjugacy

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \Rightarrow p(\boldsymbol{X}|\boldsymbol{Z}) = p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z})$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}$$

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0)$$

Conjugacy holds true (Dirichlet-Categorical)

Marginalization over $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ is analytically tractable!

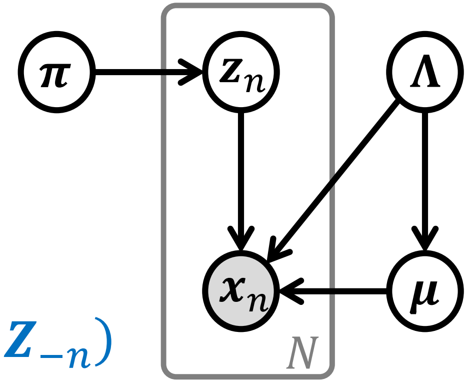
$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0)$$

Conjugacy holds true (Gaussian-Wishart-Gaussian)

Collapsed Gibbs Sampling for GMM

- Generate samples from $p(\boldsymbol{Z}|\boldsymbol{X})$
  - Divide $\{\boldsymbol{Z}\}$ into $\{\boldsymbol{z}_1\}, \{\boldsymbol{z}_2\}, \cdots, \{\boldsymbol{z}_N\}$
  - for $t = 1{:}\,T$
    - for $n = 1{:}\,N$
      - Sample $\boldsymbol{z}_n \sim p(\boldsymbol{z}_n|\boldsymbol{X}, \boldsymbol{Z}_{-n}) = p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})$



$$p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) \propto p(z_{nk} = 1, \boldsymbol{x}_n|\boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})$$

$$= p(z_{nk} = 1|\boldsymbol{Z}_{-n})p(\boldsymbol{x}_n|z_{nk} = 1, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})$$

$$= \int p(z_{nk} = 1|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{Z}_{-n})d\boldsymbol{\pi} \int p(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})\, d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

$$= \frac{\alpha_k^{(-n)}}{\sum_{k'=1}^{K} \alpha_{k'}^{(-n)}} \mathrm{St}\left(\boldsymbol{x}_n \middle| \boldsymbol{m}_k^{(-n)}, \boldsymbol{L}_k^{(-n)}, \nu_k^{(-n)} + 1 - D\right)$$

Product of two predictive distributions

Marginalizing Parameters Out

- Calculate predictive distributions
  - Marginalize likelihood functions under posteriors

$$\int \underbrace{p(z_{nk}=1|\boldsymbol{\pi})}_{\text{Likelihood}} \underbrace{p(\boldsymbol{\pi}|\boldsymbol{Z}_{-n})}_{\text{Posterior}} d\boldsymbol{\pi} = \int \pi_k \mathrm{Dir}\left(\boldsymbol{\pi}_k \middle| \boldsymbol{\alpha}^{(-n)}\right) d\boldsymbol{\pi} = \frac{\alpha_k^{(-n)}}{\sum_{k'=1}^{K} \alpha_{k'}^{(-n)}}$$

$$\int \underbrace{p(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)}_{\text{Likelihood}} \underbrace{p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})}_{\text{Posterior}} \, d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

$$= \int N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) N\left(\boldsymbol{\mu}_k \middle| \boldsymbol{m}_k^{(-n)}, \left(\beta_k^{(-n)} \boldsymbol{\Lambda}_k\right)^{-1}\right) W\left(\boldsymbol{\Lambda}_k \middle| \boldsymbol{W}_k^{(-n)}, \nu_k^{(-n)}\right) d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

$$= \mathrm{St}\left(\boldsymbol{x}_n | \boldsymbol{m}_k^{(-n)}, \boldsymbol{L}_k^{(-n)}, \nu_k^{(-n)} + 1 - D\right)$$

$$\boldsymbol{L}_k^{(-n)} = \frac{\nu_k^{(-n)} + 1 - D}{1 + \beta_k^{(-n)}} \boldsymbol{W}_k^{(-n)}$$

$$S_k[1] = \sum_{n' \neq N} z_{n'k} \quad S_k[\boldsymbol{x}] = \sum_{n' \neq n} z_{n'k} \boldsymbol{x}_{n'}$$

$$S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n' \neq n} z_{n'k} \boldsymbol{x}_{n'} \boldsymbol{x}_{n'}^T$$

Collapsed VB for GMM

- Approximate a posterior $p(\boldsymbol{Z}|\boldsymbol{X})$
  - Assume a variational distribution $\prod_{n=1}^{N} q(\boldsymbol{z}_n) \approx p(\boldsymbol{Z}|\boldsymbol{X})$
  - Iteratively update (optimize) each factor
    - CVB-E step: Invoke the updating formula of VB

$$\log q^*(\boldsymbol{z}_n) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}) \rangle_{q(\boldsymbol{Z}_{-n})} + \text{const.}$$

$$= \langle \log p(\boldsymbol{z}_n | \boldsymbol{X}, \boldsymbol{Z}_{-n}) p(\boldsymbol{X}|\boldsymbol{Z}_{-n}) p(\boldsymbol{Z}_{-n}) \rangle_{q(\boldsymbol{Z}_{-n})} + \text{const.}$$

$$= \langle \log p(\boldsymbol{z}_n | \boldsymbol{X}, \boldsymbol{Z}_{-n}) \rangle_{q(\boldsymbol{Z}_{-n})} + \text{const.}$$

$$p(z_{nk} = 1 | \boldsymbol{X}, \boldsymbol{Z}_{-n}) \propto p(z_{nk} = 1, \boldsymbol{x}_n | \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})$$

$$= \frac{\alpha_k^{(-n)}}{\sum_{k'=1}^{K} \alpha_{k'}^{(-n)}} \mathrm{St}\left(\boldsymbol{x}_n | \boldsymbol{m}_k^{(-n)}, \boldsymbol{L}_k^{(-n)}, \nu_k^{(-n)} + 1 - D \right)$$

Same as collapsed Gibbs sampling

CVB-E Step for GMM

- Calculate the variational posterior over latent variables $\boldsymbol{Z}$
  - The normalization factor is automatically determined

$$\log q^*(z_{nk} = 1) = \langle \log p(\boldsymbol{z}_n | \boldsymbol{X}, \boldsymbol{Z}_{-n}) \rangle_{q(\boldsymbol{Z}_{-n})} + \text{const.}$$

$$= \left\langle \log \frac{\alpha_k^{(-n)}}{\sum_{k'=1}^{K} \alpha_{k'}^{(-n)}} + \log \mathrm{St}\left(\boldsymbol{x}_n | \boldsymbol{m}_k^{(-n)}, \boldsymbol{L}_k^{(-n)}, \nu_k^{(-n)} + 1 - D\right) \right\rangle + \text{const.}$$

$$\approx \log \left\langle \alpha_k^{(-n)} \right\rangle - \log \sum_{k'=1}^{K} \left\langle \alpha_{k'}^{(-n)} \right\rangle$$

0-th order approximation (CVB0)

$$\mathrm{E}[\log x] \approx \log \mathrm{E}[x]$$

$$+ \log St\left(\boldsymbol{x}_n \middle| \left\langle \boldsymbol{m}_k^{(-n)} \right\rangle, \left\langle \boldsymbol{L}_k^{(-n)} \right\rangle, \left\langle \nu_k^{(-n)} \right\rangle + 1 - D\right) + \text{const.}$$

$$S_k[1] = \sum_{n' \neq N} \gamma_{n'k} \quad S_k[\boldsymbol{x}] = \sum_{n' \neq n} \gamma_{n'k} \boldsymbol{x}_{n'} \quad S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n' \neq n} \gamma_{n'k} \boldsymbol{x}_{n'} \boldsymbol{x}_{n'}^T$$

CGS vs. CVB

- Both methods are based on similar updating formulas
  - CGS: Stochastic hard assignment

$$p(z_{nk}=1|\boldsymbol{x}_n,\boldsymbol{X}_{-n},\boldsymbol{Z}_{-n})=\frac{\alpha_k^{(-n)}}{\sum_{k'=1}^K\alpha_{k'}^{(-n)}}\mathrm{St}\left(\boldsymbol{x}_n\middle|\boldsymbol{m}_k^{(-n)},\boldsymbol{L}_k^{(-n)},\nu_k^{(-n)}+1-D\right)$$

$$S_k[1]=\sum_{n'\neq N}z_{n'k}\quad S_k[\boldsymbol{x}]=\sum_{n'\neq n}z_{n'k}\boldsymbol{x}_{n'}\quad S_k[\boldsymbol{x}\boldsymbol{x}^T]=\sum_{n'\neq n}z_{n'k}\boldsymbol{x}_{n'}\boldsymbol{x}_{n'}^T$$

  - CVB: Deterministic soft assignment

$$q(z_{nk}=1)=\frac{\left\langle\alpha_k^{(-n)}\right\rangle}{\sum_{k'=1}^K\left\langle\alpha_{k'}^{(-n)}\right\rangle}\mathrm{St}\left(\boldsymbol{x}_n\middle|\left\langle\boldsymbol{m}_k^{(-n)}\right\rangle,\left\langle\boldsymbol{L}_k^{(-n)}\right\rangle,\left\langle\nu_k^{(-n)}\right\rangle+1-D\right)$$

$$S_k[1]=\sum_{n'\neq N}\gamma_{n'k}\quad S_k[\boldsymbol{x}]=\sum_{n'\neq n}\gamma_{n'k}\boldsymbol{x}_{n'}\quad S_k[\boldsymbol{x}\boldsymbol{x}^T]=\sum_{n'\neq n}\gamma_{n'k}\boldsymbol{x}_{n'}\boldsymbol{x}_{n'}^T$$

Comparison

- All methods are based on similar updating formulas
  - GS: Stochastic hard assignment
    - $p(z_{nk} = 1 | \boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \pi_k N(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$
  - CGS: Stochastic hard assignment
    - $p(z_{nk} = 1 | \boldsymbol{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) = \frac{\alpha_k^{(-n)}}{\sum_{k'=1}^{K} \alpha_{k'}^{(-n)}} \mathrm{St}\left(\boldsymbol{x}_n \middle| \boldsymbol{m}_k^{(-n)}, \boldsymbol{L}_k^{(-n)}, \nu_k^{(-n)} + 1 - D\right)$
  - EM: Deterministic soft assignment
    - $q^*(z_{nk} = 1 | \boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \pi_k N(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$
  - VB: Deterministic soft assignment
    - $q^*(z_{nk} = 1) \propto \mathrm{G}[\pi_k] \mathrm{G}[N(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)]$
  - CVB: Deterministic soft assignment
    - $q(z_{nk} = 1) = \frac{\left\langle \alpha_k^{(-n)} \right\rangle}{\sum_{k'=1}^{K} \left\langle \alpha_{k'}^{(-n)} \right\rangle} \mathrm{St}\left(\boldsymbol{x}_n \middle| \left\langle \boldsymbol{m}_k^{(-n)} \right\rangle, \left\langle \boldsymbol{L}_k^{(-n)} \right\rangle, \left\langle \nu_k^{(-n)} \right\rangle + 1 - D\right)$

All formulas are like:
Mixing ratio
×
Component distribution

Learning Algorithms

- Learning algorithm can be categorized with respect to how to deal with uncertainty

		Latent variables z		
		Point estimates	Posteriors	Sampled values
Parameters π, μ, Λ	Point estimates	K-means (maximization-maximization)	EM (expectation-maximization)	
	Posteriors	Bayesian K-means (maximization-expectation)	VB (expectation-expectation)	
	Sampled values			Gibbs sampling (sampling-sampling)

Implementation Example in C++

- Implement basic functions for updating posteriors
  - Input: prior + statistics   Output: posterior

posterior.h

```
void update_dirichlet
(mcl::Dirichlet& dirichlet,
 mcl::Dirichlet& dirichlet0,
 const std::vector<double>& s);
```

```
void update_gaussian_wishart
(mcl::Gaussian& gaussian,
 mcl::Wishart& wishart,
 const mcl::Gaussian& gaussian0,
 const mcl::Wishart& wishart0,
 double s,
 const std::vector<double>& sx,
 const std::vector<double>& sxx);
```

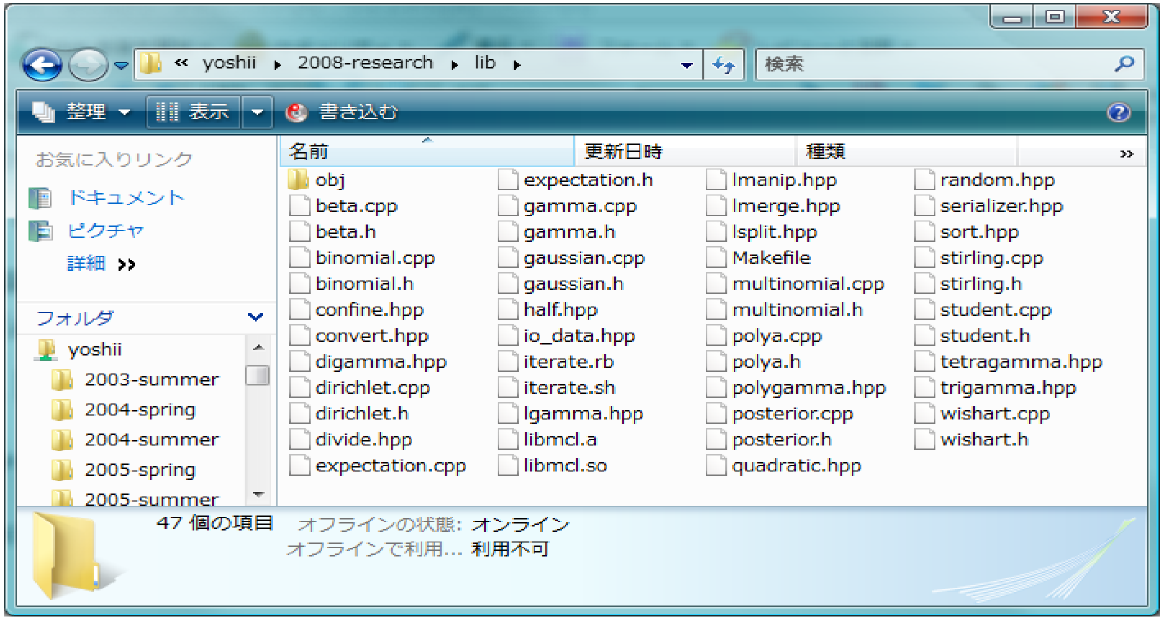
```
void update_student
(mcl::Student& student,
 const mcl::Gaussian& gaussian,
 const mcl::Wishart& wishart);
```

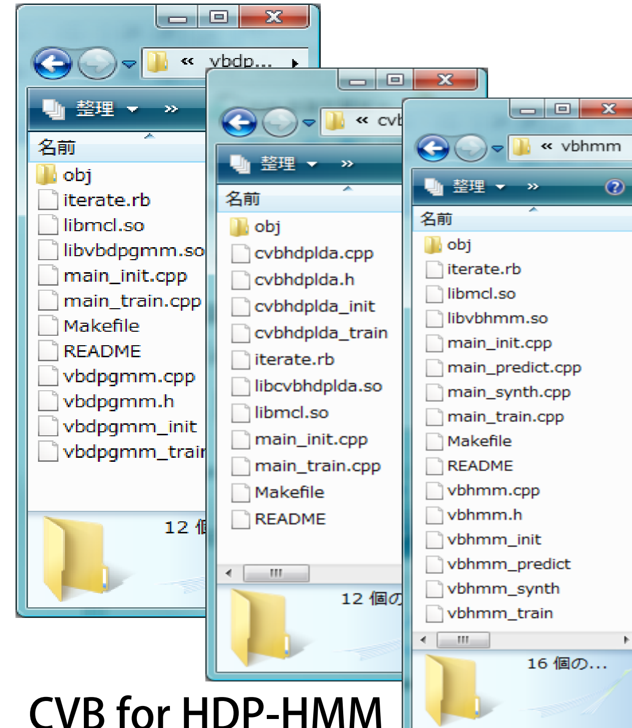
Predictive distribution
(used for collapsed inference)

Implementation Example in C++

- Combine appropriate functions for your model
  - Use conjugate priors as much as possible

VB for DP-GMM





Library

CVB for HDP-HMM

VB for HMM

MapReduce-type parallelization is easy

Implementation Example in C++

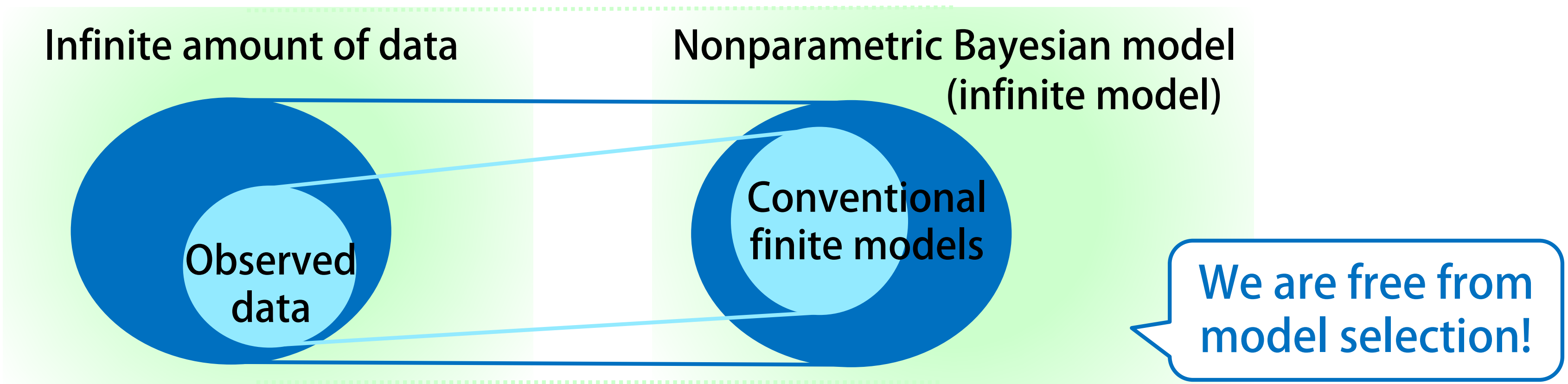
- Implement HTK-like commands
  - `vbgmm_init [model.xml] [K]`
    - Make an initial model with K components
  - `vbgmm_train [model.xml] [data.csv] ([#iterations])`
    - Update the model using the data
    - Overwrite the model file
- Parallelization based on boost::mpi
  - MapReducing EM algorithm for Master-Slave architecture
    - E-step: *Master* distributes the data to *Slaves*
      - Each *Slave* calculates the responsibilities for the given data
    - M-step: *Master* gathers the responsibilities from *Slaves*
      - *Master* updates the posteriors

Bayesian Estimation of Infinite Gaussian Mixture Models

Collapsed Gibbs Sampling
Variational Bayes

Nonparametric Bayesian Models

- Bayesian models with infinite complexity
  - "Nonparametric" means having an infinite number of parameters
  - Excellent generalization capability
    - If we have an infinite amount of data, all an infinite number of parameters are required
    - If we have a finite amount of data, only a finite subset is required



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

- An infinite-dimensional prior distribution
  - Capable of generating infinite-dimensional distributions

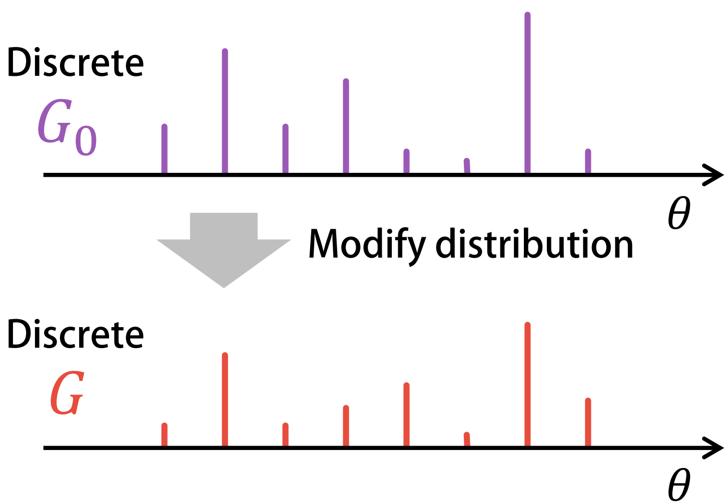
$$G \sim \mathrm{DP}(\alpha, G_0)$$

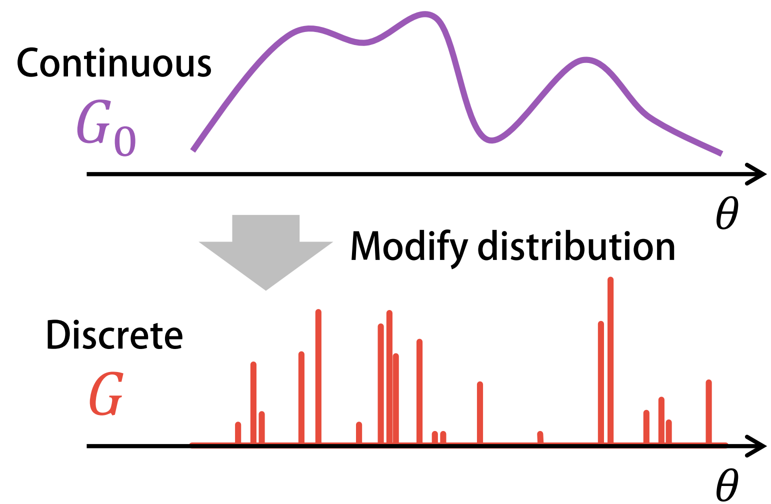
Concentration parameter

Base measure

The DP can be explicitly rewritten as

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \qquad \theta_k \sim G_0$$



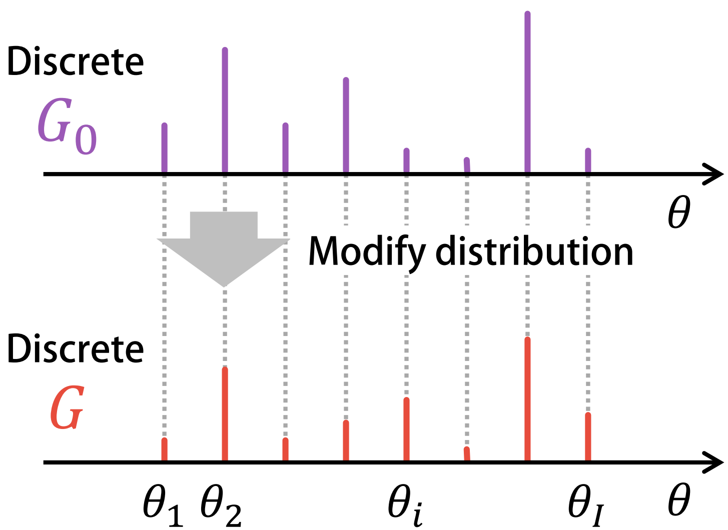


Discrete Base Measure G_0

- The DP always generates discrete distributions
  - The positions of "atoms" are shared with the discrete base measure

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \quad \theta_k \sim G_0$$

Each θ_k is one of $\{\theta_1, \theta_2, \cdots, \theta_I\}$



$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$
$$= \sum_{i=1}^{I} \left(\sum_{k:\theta_k=\theta_i} \pi_k \right) \delta_{\theta_i}(\theta)$$

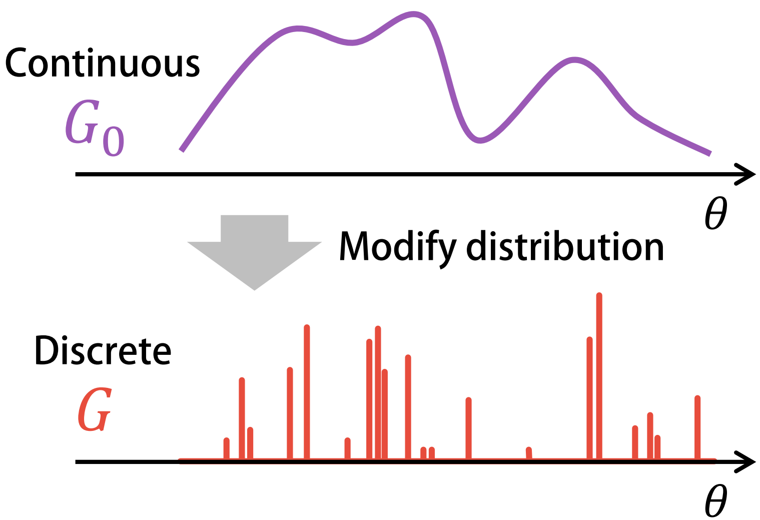
I-dimensional discrete distribution

Continuous Base Measure G_0

- The DP always generates discrete distributions
  - The number of "atoms" are countably infinite

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \quad \theta_k \sim G_0$$

θ_k's are almost surely disjoint



If we use a continuous prior distribution as a base measure G_0, we can generate an infinite-dim. discrete distribution!

If G_0 is a Gaussian-Wishart distribution (the probability space is over $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Lambda}\}$)

↓

G consists of infinitely many Gaussians $\{\theta_1, \cdots, \theta_\infty\}$ with weights $\{\pi_1, \cdots, \pi_\infty\}$

Stick Breaking Process

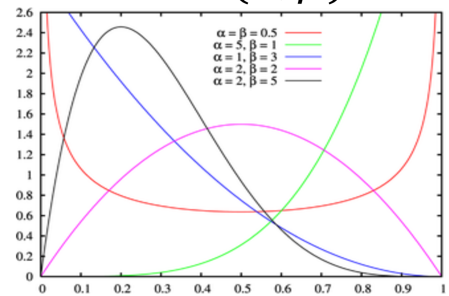
- Stochastically generate the weights $\{\pi_1, \cdots, \pi_\infty\}$
  - a.k.a. Griffiths-Engen-McCloskey distribution

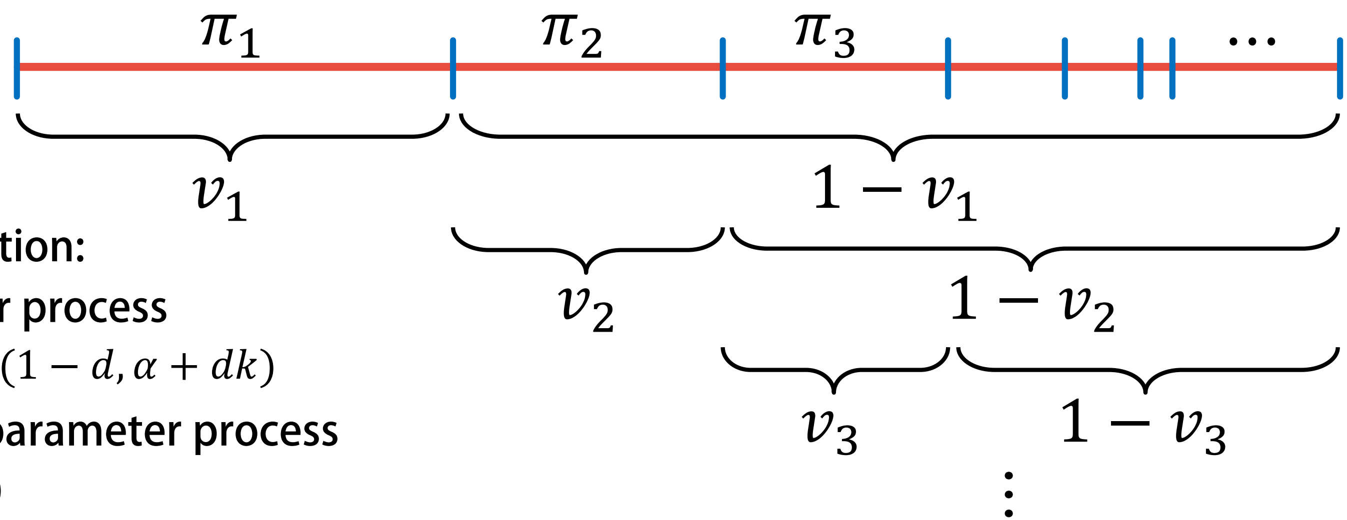
$$\boldsymbol{\pi} \sim \mathrm{SBP}(\alpha) \text{ or } \mathrm{GEM}(\alpha)$$

$$\downarrow$$

$$v_k \sim \mathrm{Beta}(1, \alpha) \quad \pi_k = v_k \prod_{k'=1}^{k-1} (1 - v_{k'})$$

Beta(α, β)





Generalization:

Pitman-Yor process

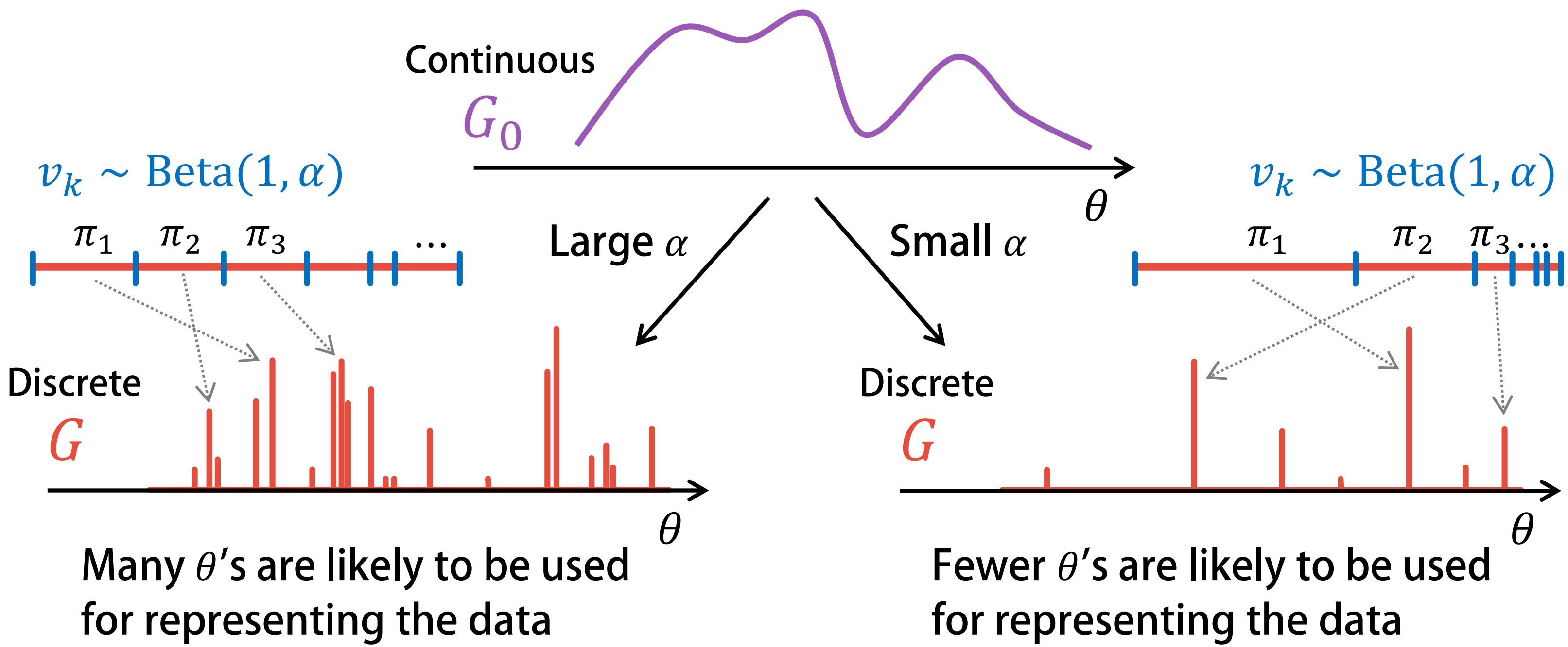
$v_k \sim \mathrm{Beta}(1 - d, \alpha + dk)$

Beta two-parameter process

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

Concentration Parameter α

- The concentration parameter controls the sparseness
  - The value of α is unknown → Introduce a hyper prior on α



Assume $\alpha \sim \mathrm{Gamma}(a, b)$ for taking into account uncertainty

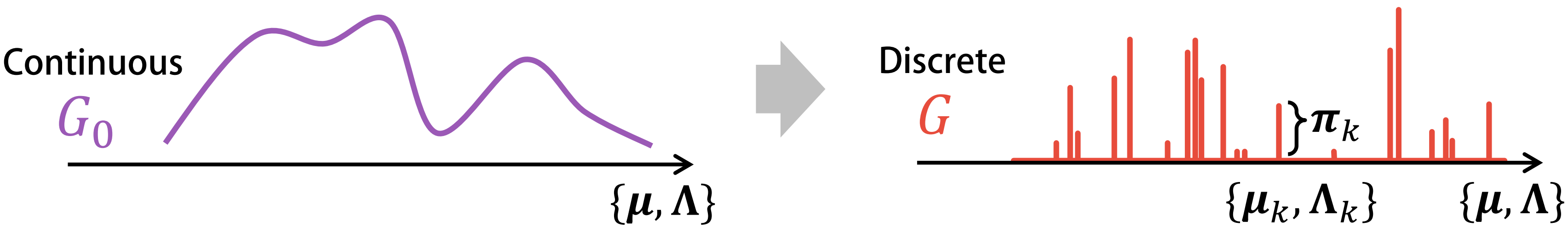
Generative Story of iGMM

- Generate infinitely many Gaussians using a DP

$$G(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$$

$\boldsymbol{\pi} \sim \mathrm{SBP}(\alpha)$ SBP prior

$\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \sim G_0(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ Gaussian-Wishart prior



- Generate samples independently

for $n = 1{:}N$

$\quad \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n \sim G(\boldsymbol{\mu}, \boldsymbol{\Lambda})$

$\quad \boldsymbol{x}_n \sim N(\boldsymbol{x}_n | \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n^{-1})$

end

Equivalent representation

$$\boldsymbol{z}_n \sim \mathrm{Categorical}(\boldsymbol{z}_n | \boldsymbol{\pi})$$

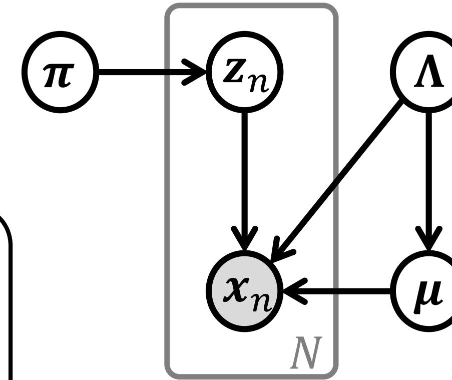
$$\boldsymbol{x}_n \sim \prod_{k=1}^{\infty} N(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)^{z_{nk}}$$

Infinite GMM!

Formulation of GMM

- Formulate a full joint distribution

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$$



$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \mathrm{Categorical}(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}$$

Likelihood functions

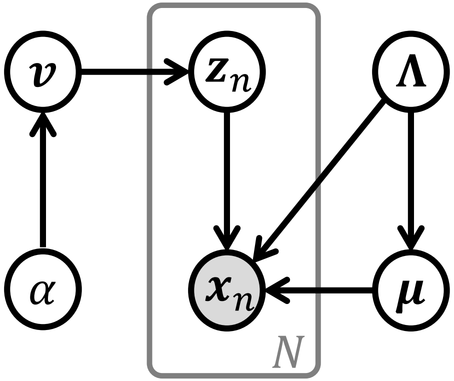
$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_{0k}\right)}{\prod_{k=1}^{K}\Gamma(\alpha_{0k})}\prod_{k=1}^{K}\pi_k^{\alpha_{0k}-1}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0)$$

Prior distributions

Formulation of iGMM

- Use a SBP prior instead of a Dirichlet prior



$$p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda},\alpha) = p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{v})p(\boldsymbol{v}|\alpha)p(\alpha)p(\boldsymbol{\mu},\boldsymbol{\Lambda})$$

Likelihood functions

$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{\infty} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

$$p(\boldsymbol{Z}|\boldsymbol{v}) = \prod_{n=1}^{N}\prod_{k=1}^{\infty}\left(\underbrace{v_k\prod_{k'=1}^{k-1}(1-v_{k'})}_{\pi_k}\right)^{z_{nk}}$$

Prior distributions

$$p(\boldsymbol{v}|\alpha) = \underbrace{\prod_{k=1}^{\infty}\mathrm{Beta}(v_k|1,\alpha)}_{\text{SBP prior}} \quad p(\alpha) = \mathrm{Gamma}(\alpha|a_0,b_0)$$

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{k=1}^{\infty} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\Lambda}_k|\mathbf{W}_0,\nu_0)$$

Hierarchical Conjugacy

- Beta-Bernoulli & Gamma-Exponential conjugacy
  - The VB is applicable for learning an iGMM

$$p(\boldsymbol{Z}|\boldsymbol{v}) = \prod_{n=1}^{N}\prod_{k=1}^{\infty}\left(v_k \prod_{k'=1}^{k-1}(1-v_{k'})\right)^{z_{nk}}$$

$$= \prod_{n=1}^{N}\prod_{k=1}^{\infty} v_k^{z_{nk}}(1-v_k)^{\sum_{k'=k+1}^{\infty} z_{nk'}}$$

The number of z_n's that pass k

$$= \prod_{k=1}^{\infty} v_k^{\sum_{n=1}^{N} z_{nk}}(1-v_k)^{\sum_{n=1}^{N}\sum_{k'>k} z_{nk'}}$$

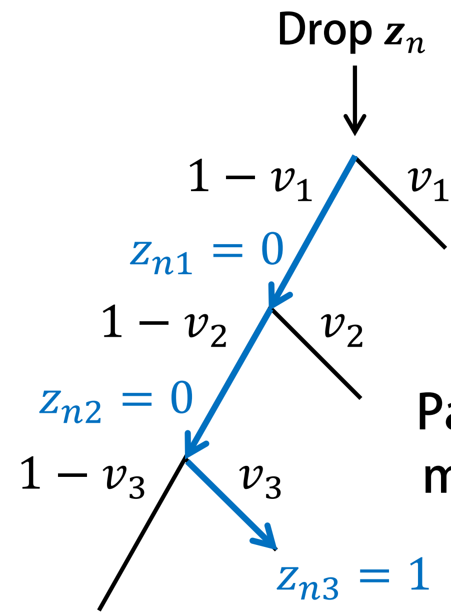
The number of z_n's that stop at k

Conjugate

$$p(\boldsymbol{v}|\alpha) = \prod_{k=1}^{\infty} \alpha v_k^{1-1}(1-v_k)^{\alpha-1}$$

Conjugate

$$p(\alpha) = \frac{b_0^{a_0}}{\Gamma(a_0)}\alpha^{a_0-1}e^{-b_0\alpha}$$



Pachinko machine

$p(z_{n3}=1|\boldsymbol{v}) = (1-v_1)(1-v_2)v_3$

VB for iGMM

- Approximate a posterior $p(\boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha | \boldsymbol{X})$
  - Use a variational distribution $q(\boldsymbol{Z})q(\boldsymbol{v})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\alpha) \approx p(\boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha | \boldsymbol{X})$
  - Iteratively update (optimize) each factor
    - VB-E step
      - $\log q^*(\boldsymbol{Z}) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha) \rangle_{q(\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha)} + \text{const.}$
        $= \langle \log p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z} | \boldsymbol{v}) \rangle_{q(\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.}$
    - VB-M step
      - $\log q^*(\boldsymbol{v}) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha)} + \text{const.}$
        $= \langle \log p(\boldsymbol{Z} | \boldsymbol{v}) p(\boldsymbol{v} | \alpha) \rangle_{q(\boldsymbol{Z}, \alpha)} + \text{const.}$
      - $\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{v}, \alpha)} + \text{const.}$
        $= \langle \log p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z})} + \text{const.}$
      - $\log q^*(\alpha) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.}$
        $= \langle \log p(\boldsymbol{v} | \alpha) p(\alpha) \rangle_{q(\boldsymbol{v})} + \text{const.}$

VB-E Step for iGMM

- Invoke the updating formula of VB
  - Take the expectation of the full joint probability distribution under variational posteriors over other variables
  - Focus on only terms including $\boldsymbol{Z}$
    (other terms can be absorbed into the normalization factor)

$$
\begin{aligned}
\log q^*(\boldsymbol{Z}) &= \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha) \rangle_{q(\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha)} + \text{const.} \\
&= \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{v}) p(\boldsymbol{v}|\boldsymbol{\alpha}) p(\alpha) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha)} + \text{const.} \\
&= \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{v}) \rangle_{q(\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.}
\end{aligned}
$$

$$
p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{\infty} N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}
$$

$$
p(\boldsymbol{Z}|\boldsymbol{v}) = \prod_{n=1}^{N} \prod_{k=1}^{\infty} \left(v_k \prod_{k'=1}^{k-1} (1 - v_{k'}) \right)^{z_{nk}}
$$

VB-E Step for iGMM

- Proceed the calculation according the updating rule

$$\langle \log p(\boldsymbol{Z}|\boldsymbol{v}) \rangle_{q(\boldsymbol{v})} = \sum_{n=1}^{N} \sum_{k=1}^{\infty} z_{nk} \left(\langle \log v_k \rangle_{q(v_k)} + \sum_{k'=1}^{k-1} \langle \log(1 - v_{k'}) \rangle_{q(v_{k'})} \right)$$

$$\langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{\mu}, \boldsymbol{\Lambda})} = \sum_{n=1}^{N} \sum_{k=1}^{\infty} z_{nk} \left\langle \log N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) \right\rangle_{q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)}$$

$$\log q^*(\boldsymbol{Z}) = \langle \log p(\boldsymbol{Z}|\boldsymbol{v}) \rangle_{q(\boldsymbol{v})} + \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.}$$

Infinite GMM

$$= \sum_{n=1}^{N} \sum_{k=1}^{\infty} z_{nk} \left(\boxed{\langle \log v_k \rangle_{q(v_k)} + \sum_{k'=1}^{k-1} \langle \log(1 - v_{k'}) \rangle_{q(v_{k'})}} + \left\langle \log N\left(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) \right\rangle_{q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)} \right) + \text{const.}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{\infty} z_{nk} \log \rho_{nk} + \text{const.}$$

Finite GMM

$$\langle \log \pi_k \rangle_{q(\boldsymbol{\pi})}$$

VB-E Step for iGMM

- Calculate the variational posterior over latent variables $\boldsymbol{Z}$
  - The normalization factor is automatically determined

$$\log q^*(\boldsymbol{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{\infty} z_{nk} \log \rho_{nk} + \text{const.}$$

$$\gamma_{nk} = \frac{\rho_{nk}}{\sum_{k'=1}^{K} \rho_{nk'}}$$

$$\log q^*(\boldsymbol{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{\infty} z_{nk} \log \gamma_{nk}$$

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{\infty} \gamma_{nk}^{z_{nk}} = \prod_{n=1}^{N} \text{Categorical}\left(\boldsymbol{z}_n | \boldsymbol{\gamma}_n\right)$$

Truncate the variational posterior at the level K ***i.e.***, $q(z_{nk>K}) = 0$
The larger K becomes, the more accurate the approximation is

Latent variables are categorical distributed!

VB-M Step for iGMM

- Invoke the updating formula of VB
  - Take the expectation of the full joint probability distribution under variational posteriors over other variables
  - Focus on only terms including $\boldsymbol{Z}$
    (other terms can be absorbed into the normalization factor)

$$\log q^*(\boldsymbol{v}) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha)} + \text{const.}$$
$$= \log p(\boldsymbol{v}|\boldsymbol{\alpha}) + \langle \log p(\boldsymbol{Z}|\boldsymbol{v}) \rangle_{q(\boldsymbol{Z})} + \text{const.}$$

$$\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{\pi}, \alpha)} + \text{const.}$$
$$= \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \langle \log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z})} + \text{const.}$$

Same as finite GMM

$$\log q^*(\alpha) = \langle \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \rangle_{q(\boldsymbol{Z}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} + \text{const.}$$
$$= \log p(\alpha) + \langle \log p(\boldsymbol{v}|\alpha) \rangle_{q(\boldsymbol{v})} + \text{const.}$$

Bayesian estimation in simple conjugate models!
(Use responsibilities $q(\boldsymbol{Z})$ instead of latent variables $\boldsymbol{Z}$)

VB-M Step for iGMM

- Calculate the variational posterior over parameters $\boldsymbol{v}$
  - The posteriors take the same forms of the priors

$$S_k[1] = \sum_{n=1}^{N} \gamma_{nk} \quad S_k[\boldsymbol{x}] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n \quad S_k[\boldsymbol{x}\boldsymbol{x}^T] = \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T$$

Sufficient statistics

$$\begin{cases} p(\boldsymbol{v}|\alpha) = \prod_{k=1}^{\infty} \mathrm{Beta}(v_k|1,\alpha) = \prod_{k=1}^{\infty} \alpha v_k^{1-1}(1-v_k)^{\alpha-1} \\ p(\boldsymbol{Z}|\boldsymbol{v}) = \prod_{k=1}^{\infty} v_k^{\sum_{n=1}^{N} z_{nk}} (1-v_k)^{\sum_{n=1}^{N}\sum_{k'=k+1}^{\infty} z_{nk'}} \\ p(\boldsymbol{v}|\boldsymbol{Z},\alpha) = \prod_{k=1}^{\infty} \mathrm{Beta}\left(v_k \middle| 1 + \sum_{n=1}^{N} z_{nk}, \alpha + \sum_{n=1}^{N}\sum_{k'=k+1}^{\infty} z_{nk'}\right) \end{cases}$$

Bayes' theorem

Replace z_{nk} with γ_{nk}

$q^*(\boldsymbol{v})$

VB-M Step for iGMM

- Calculate the variational posterior over parameter α
  - The posterior takes the same forms of the prior
  - Use that fact that if $x \sim \text{Beta}(1, \alpha)$, then $-\log(1-x) \sim \text{Exponential}(\alpha)$
  - $q^*(\alpha)$ is analytically tractable in case of iGMM

$$\begin{cases} p(\alpha) = \text{Gamma}(\alpha | a_0, b_0) = \dfrac{b_0^{a_0}}{\Gamma(a_0)} \alpha^{a_0 - 1} e^{-b_0 \alpha} \\ p(\boldsymbol{v}|\alpha) = \prod_{k=1}^{\infty} \text{Beta}(v_k | 1, \alpha) = \alpha^K \prod_{k=1}^{\infty} (1 - v_k)^{\alpha - 1} \\ p(\alpha|\boldsymbol{v}) = \text{Gamma}\left(\alpha \middle| a_0 + K, b_0 - \sum_{k=1}^{K} \log(1 - v_k)\right) \end{cases}$$

Bayes' theorem

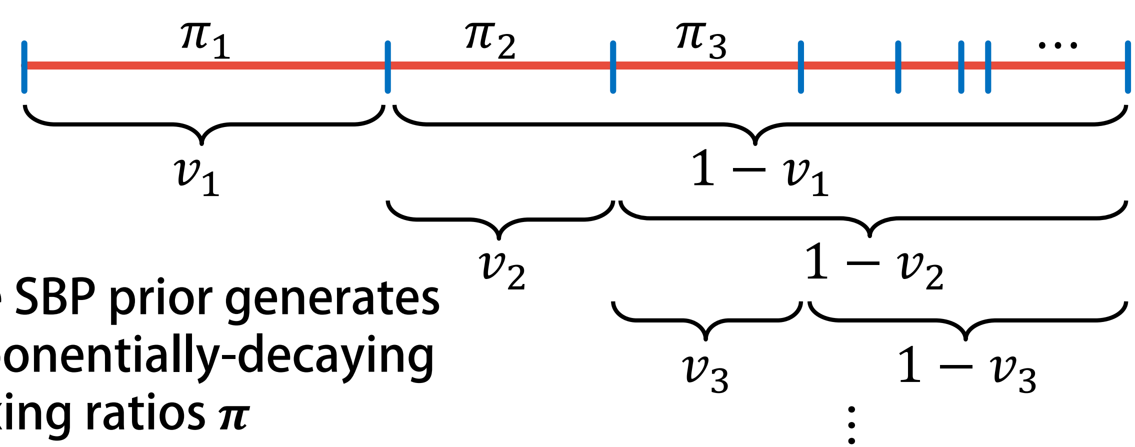
Replace $\log(1 - v_k)$ with $\langle \log(1 - v_k) \rangle_{q(v_k)}$

$\Rightarrow q^*(\alpha)$

Some Tricks

- Truncate the variational poster $q(\boldsymbol{Z})$
  - The infinite-dimensional true posterior $p(\boldsymbol{Z}|\boldsymbol{X})$ is NOT truncated!
  - $q(\boldsymbol{z}_n)$ is truncated at a sufficiently large level K ***i.e.***, $q(z_{nk>K}) = 0$
  - K corresponds to how accurately $q(\boldsymbol{Z})$ approximates $p(\boldsymbol{Z}|\boldsymbol{X})$
- Sort K clusters in descending order before VB-M step
  - Remove unnecessary cluster k with $S_k[1] \approx 0$

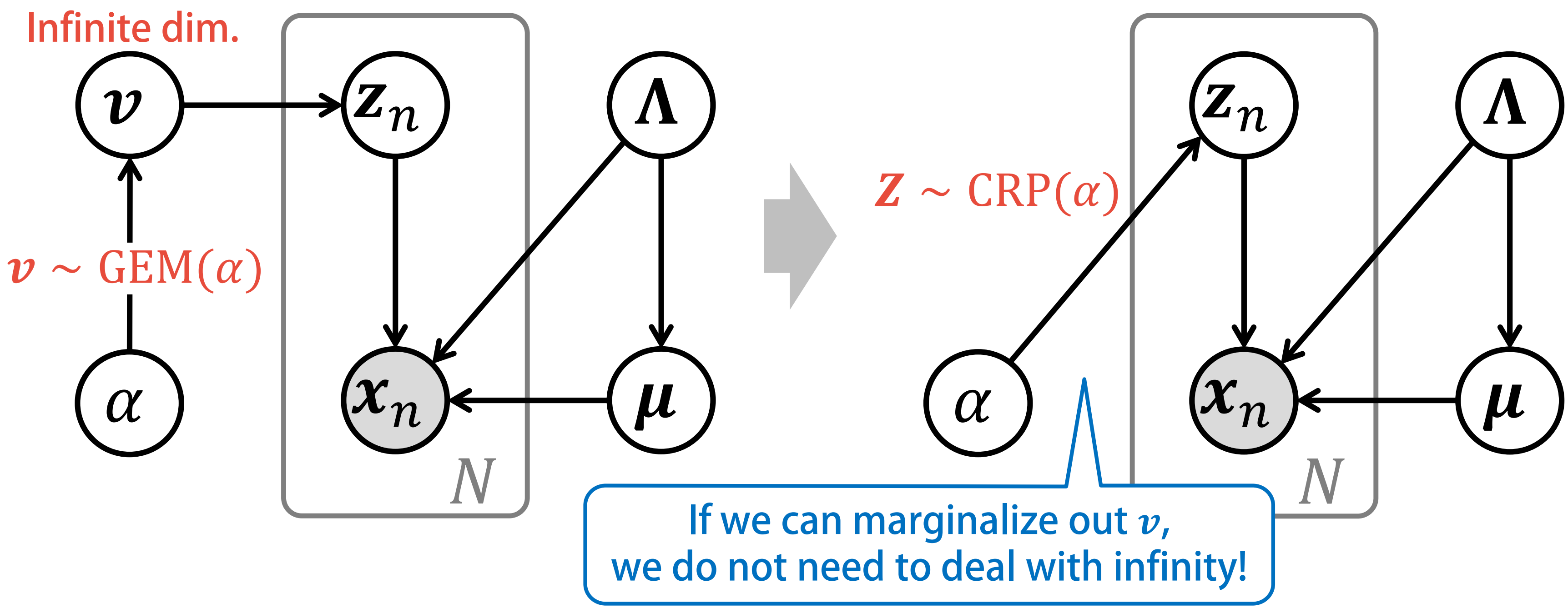
This is effective for:
1. accelerating the convergence
2. avoiding poor local maxima



The SBP prior generates exponentially-decaying mixing ratios $\boldsymbol{\pi}$

Limitation of VB

- Finite truncation at a certain level K is required for VB
  - A large amount of computational power is wasted
  - K should be sufficiently large even if only a few clusters are required for representing the data



Truncation-Free Approach

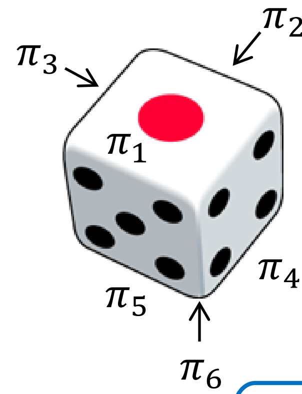
- Marginalize out infinite-dimensional parameters $\boldsymbol{\pi}$ or $\boldsymbol{v}$
  - Take the infinite limit of a Dirichlet-Categorical model

K-dimensional Dirichlet prior

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\pi}|\alpha\boldsymbol{\beta}_K) \quad \boldsymbol{\beta}_K = \underbrace{\left[\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right]}_{K}$$

Likelihood

$$\boldsymbol{z}_{1:N} \sim \mathrm{Categorical}(\boldsymbol{z}|\boldsymbol{\pi})$$



Infinite-sided die

Given $\boldsymbol{Z}_{-n}$ as observed data, $\boldsymbol{z}_n$ is predicted as:

$$p(z_{nk} = 1|\boldsymbol{Z}_{-n}) = \int \underbrace{p(z_{nk} = 1|\boldsymbol{\pi})}_{\text{Likelihood}} \underbrace{p(\boldsymbol{\pi}|\boldsymbol{Z}_{-n})}_{\text{Posterior}} d\boldsymbol{Z}_{-n}$$

$$= \int \pi_k \mathrm{Dir}\left(\boldsymbol{\pi}|\alpha\boldsymbol{\beta}_K + \textstyle\sum_{n' \neq n} \boldsymbol{z}_{n'}\right) d\boldsymbol{Z}_{-n} = \frac{\frac{\alpha}{K} + \overbrace{\sum_{n' \neq n} z_{n'k}}^{n_k^{(-n)}}}{\sum_{k'=1}^{K}\left(\frac{\alpha}{K} + \underbrace{\sum_{n' \neq n} z_{n'k'}}_{n_{k'}^{(-n)}}\right)} \xrightarrow{K \to \infty} \frac{n_k^{(-n)}}{(N-1)+\alpha}$$

The number of samples belonging to cluster k among $N-1$ samples

Truncation-Free Approach

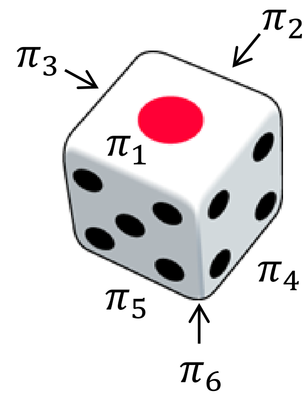
- Focus on the probability that a new cluster is selected
  - Accumulate the probabilities that existing clusters are selected

K-dimensional Dirichlet prior

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\pi}|\alpha\boldsymbol{\beta}_K) \qquad \boldsymbol{\beta}_K = \underbrace{\left[\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right]}_{K}$$

Likelihood

$$\boldsymbol{z}_{1:N} \sim \mathrm{Categorical}(\boldsymbol{z}|\boldsymbol{\pi})$$



Infinite-sided die

Given $\boldsymbol{Z}_{-n}$ consisting of K clusters, $\boldsymbol{z}_n$ is predicted as:

Sum: $\frac{N-1}{N-1+\alpha}$

$$p(z_{nk} = 1|\boldsymbol{Z}_{-n}) = \begin{cases} \dfrac{n_k^{(-n)}}{(N-1)+\alpha} & \text{Existing cluster } k \ (1 \le k \le K) \text{ is selected} \\ \dfrac{\alpha}{(N-1)+\alpha} & \text{New cluster } k \ (k > K) \text{ is created} \end{cases}$$

We index the new cluster as $K+1$

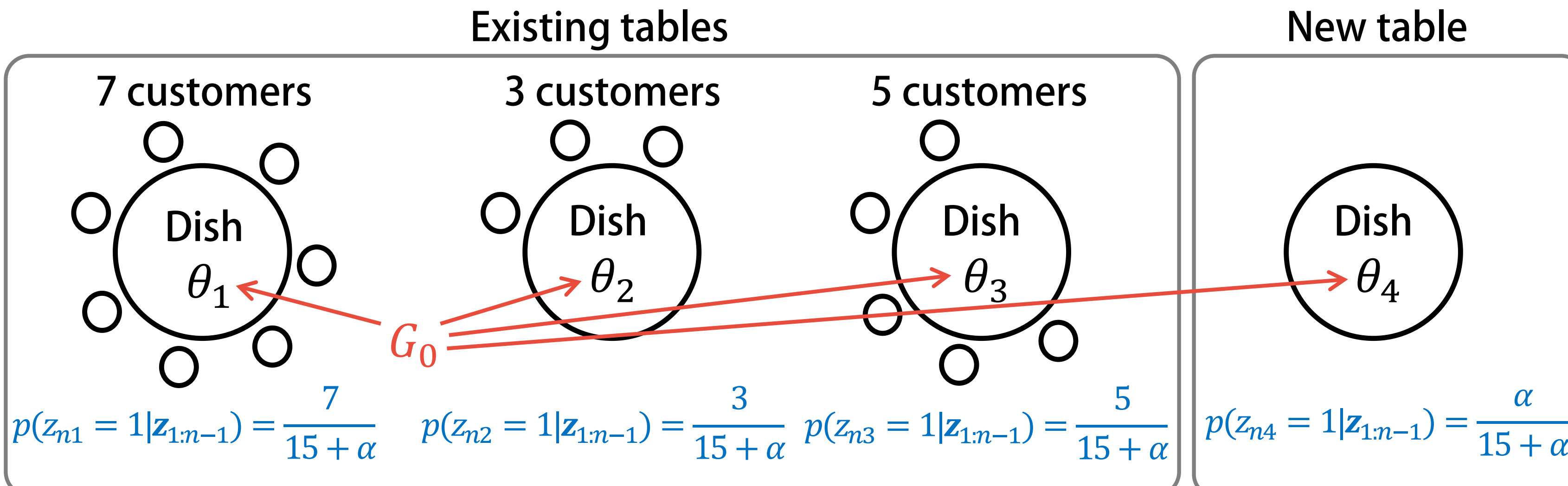
Chinese Restaurant Process

- Sequentially generate samples s.t. "the rich get richer"
  - Used as a prior on latent variables $\boldsymbol{Z}$ $(= \boldsymbol{z}_{1:N})$

$\boldsymbol{z}_{1:N} \sim \mathrm{CRP}(\alpha)$ $\theta_k \sim G_0(\theta)$ if a new cluster is created

Suppose $n-1$ customers $\boldsymbol{z}_{1:n-1}$ are already seated in restaurant G_0

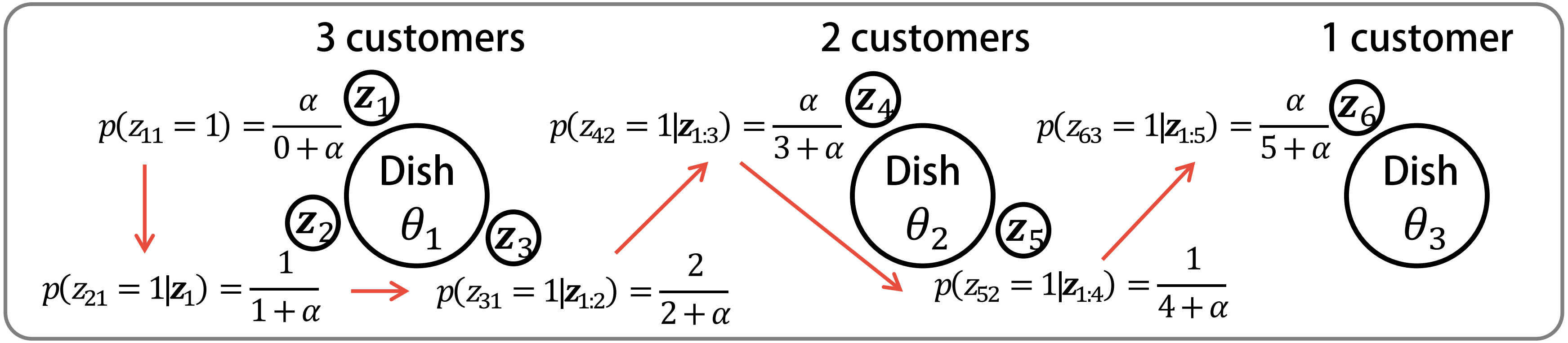
The next customer $\boldsymbol{z}_n$ stocastically selects a table as follows:



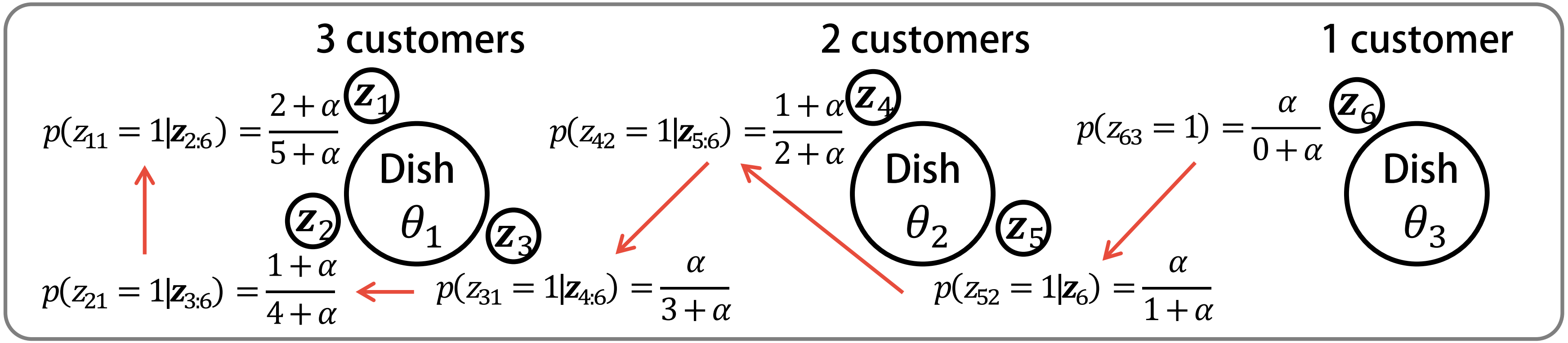
Exchangeability

- The customer order does not change the CRP probability

$$\mathrm{CRP}(\boldsymbol{Z}|\alpha) = p(\boldsymbol{z}_1)p(\boldsymbol{z}_2|\boldsymbol{z}_1)p(\boldsymbol{z}_3|\boldsymbol{z}_{1:2})p(\boldsymbol{z}_4|\boldsymbol{z}_{1:3})p(\boldsymbol{z}_5|\boldsymbol{z}_{1:4})p(\boldsymbol{z}_6|\boldsymbol{z}_{1:5})$$

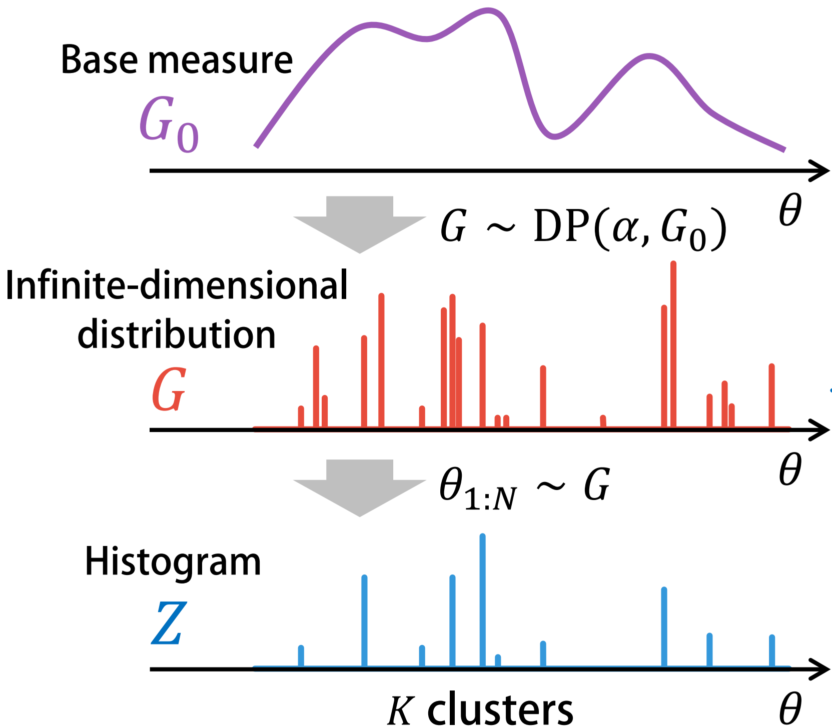


$$\mathrm{CRP}(\boldsymbol{Z}|\alpha) = p(\boldsymbol{z}_6)p(\boldsymbol{z}_5|\boldsymbol{z}_6)p(\boldsymbol{z}_4|\boldsymbol{z}_{5:6})p(\boldsymbol{z}_3|\boldsymbol{z}_{4:6})p(\boldsymbol{z}_2|\boldsymbol{z}_{3:6})p(\boldsymbol{z}_1|\boldsymbol{z}_{2:6})$$



Relationships between DP, SBP, and CRP

- Two major approaches to representing the DP
  - SBP: Represent how a distribution G is drawn from the DP
  - CRP: Represent how samples Z are drawn from the DP



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

$$\pi_{1:\infty} \sim \mathrm{SBP}(\alpha) \quad \theta_{1:\infty} \sim G_0(\theta)$$

$$\boldsymbol{z}_{1:N} \sim \mathrm{CRP}(\alpha) \quad \theta_{1:K} \sim G_0(\theta)$$

"Collapsed" Algorithms

- Reduce the number of variables for fast/better estimation
  - The parameters can be marginalized out because of conjugacy

$$p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda},\alpha)=p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z}|\alpha)p(\alpha)p(\boldsymbol{\mu},\boldsymbol{\Lambda}) \Rightarrow p(\boldsymbol{X}|\boldsymbol{Z})=p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z}|\alpha)p(\alpha)$$

$$p(\boldsymbol{Z}|\alpha)=\mathrm{CRP}(\boldsymbol{Z}|\alpha)$$

Marginal likelihood for $\boldsymbol{Z}$ (mixing ratios are marginalized out)

$$p(\boldsymbol{Z}|\alpha)\propto\lim_{k\to\infty}\int p(\boldsymbol{Z}|\boldsymbol{\pi})\mathrm{Dir}(\boldsymbol{\pi}|\alpha\boldsymbol{\beta}_K)d\boldsymbol{\pi}$$

$$p(\alpha)=\mathrm{Gamma}(\alpha|a_0,b_0)$$

Hyper prior on α

$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})=\prod_{n=1}^{N}\prod_{k=1}^{K}N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda})=\prod_{k=1}^{K}N(\boldsymbol{\mu}_k|\boldsymbol{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\Lambda}_k|\mathbf{W}_0,\nu_0)$$

Marginalization over $\boldsymbol{\mu},\boldsymbol{\Lambda}$ is analytically tractable!

Conjugacy holds true (Gaussian-Wishart-Gaussian)

Collapsed Gibbs Sampling for iGMM

- Generate samples from $p(\boldsymbol{Z}, \alpha | \boldsymbol{X})$
  - Divide $\{\boldsymbol{Z}, \alpha\}$ into $\{\boldsymbol{z}_1\}, \{\boldsymbol{z}_2\}, \cdots, \{\boldsymbol{z}_N\}, \{\alpha\}$
  - for $n = 1: N$
    - Sample $\boldsymbol{z}_n \sim p(\boldsymbol{z}_n | \boldsymbol{X}, \boldsymbol{Z}_{-n}, \alpha) = p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}, \alpha)$

$$p(z_{nk} = 1 | \boldsymbol{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}, \alpha) \propto p(z_{nk} = 1, \boldsymbol{x}_n | \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}, \alpha)$$

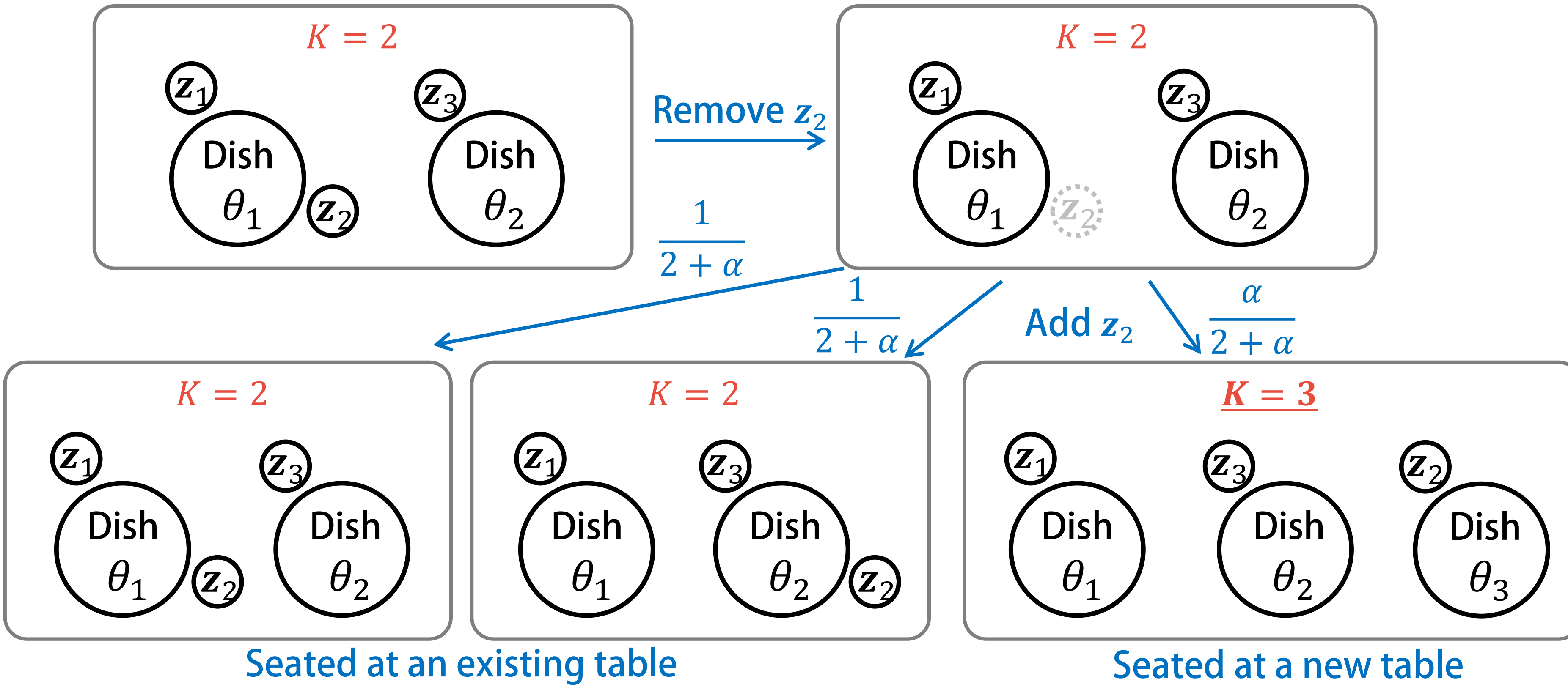
$$= p(z_{nk} = 1 | \boldsymbol{Z}_{-n}, \alpha) p(\boldsymbol{x}_n | z_{nk} = 1, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})$$

$$= \mathrm{CRP}(z_{nk} = 1 | \boldsymbol{Z}_{-n}, \alpha) \int p(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k | \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})\, d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

$$= \begin{cases} \dfrac{n_k^{(-n)}}{N - 1 + \alpha} \mathrm{St}\left(\boldsymbol{x}_n \middle| \boldsymbol{m}_k^{(-n)}, \boldsymbol{L}_k^{(-n)}, \nu_k^{(-n)} + 1 - D\right) & \text{for existing cluster } k \ (1 \le k \le K) \\ \dfrac{\alpha}{N - 1 + \alpha} \mathrm{St}(\boldsymbol{x}_n | \boldsymbol{m}_0, \boldsymbol{L}_0, \nu_0 + 1 - D) & \text{for new cluster } K + 1 \end{cases}$$

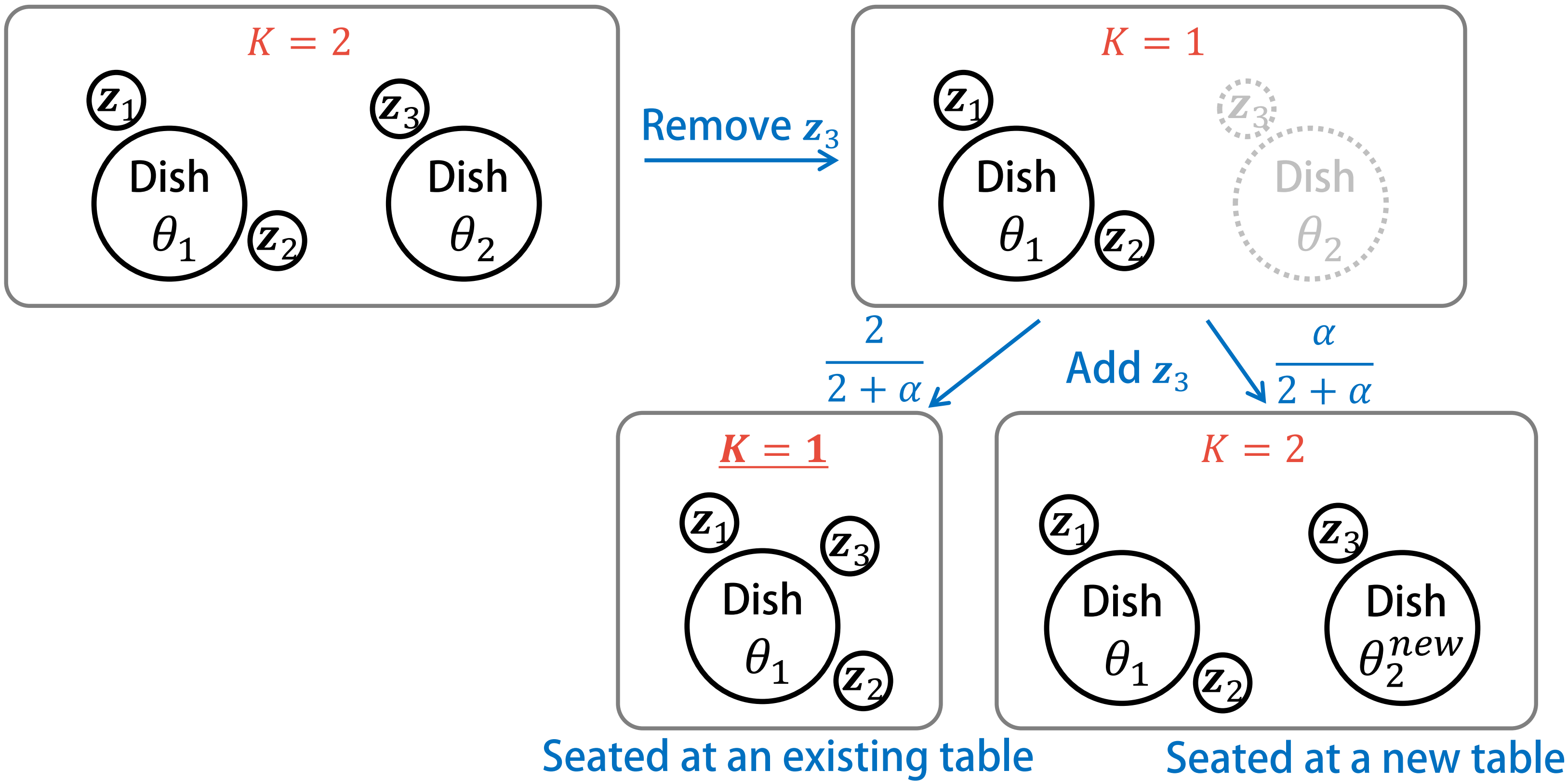
Remove-and-Add Scheme

- Update z_n using the remove-and-add scheme
  - The number of tables K can be increased



Remove-and-Add Scheme

- Update z_n using the remove-and-add scheme
  - The number of tables K can be decreased



CRP Probability

- Calculate the probability of seating arrangement

$$p(\boldsymbol{Z}|\alpha) = \frac{1}{\sum_{i=1}^{N}(i-1+\alpha)}\prod_{k=1}^{K}\alpha(n_k-1)! = \alpha^K\frac{\Gamma(\alpha)}{\Gamma(\alpha+N)}\prod_{k=1}^{K}(n_k-1)!$$

$$p(\boldsymbol{Z}|\alpha) = \int p(\boldsymbol{Z},\eta|\alpha)\,d\eta$$

Data augmentation

$$p(\boldsymbol{Z},\eta|\alpha) = \frac{\alpha^{K-1}(\alpha+n)}{\Gamma(N)}\eta^{\alpha}(1-\eta)^{N-1}\prod_{k=1}^{K}(n_k-1)!$$

$$1 = \int \mathrm{Beta}(\eta|\alpha+1,N)\,d\eta$$

$$= \frac{\Gamma(\alpha+N+1)}{\Gamma(\alpha+1)\Gamma(N)}\int \eta^{\alpha}(1-\eta)^{N-1}d\eta$$

$$\Gamma(x+1) = x\Gamma(x)$$

3 customers

2 customers

1 customer

$$p(z_{11}=1) = \frac{\alpha}{0+\alpha}$$

z_1, z_2, z_3 — Dish θ_1

$$p(z_{21}=1|\boldsymbol{z}_1) = \frac{1}{1+\alpha}$$

$$p(z_{31}=1|\boldsymbol{z}_{1:2}) = \frac{2}{2+\alpha}$$

$$p(z_{42}=1|\boldsymbol{z}_{1:3}) = \frac{\alpha}{3+\alpha}$$

z_4, z_5 — Dish θ_2

$$p(z_{52}=1|\boldsymbol{z}_{1:4}) = \frac{1}{4+\alpha}$$

$$p(z_{63}=1|\boldsymbol{z}_{1:5}) = \frac{\alpha}{5+\alpha}$$

z_6 — Dish θ_3

Collapsed Gibbs Sampling for iGMM

- Generate samples from $p(\boldsymbol{Z}, \alpha, \eta | \boldsymbol{X})$
  - Sample $\alpha \sim p(\alpha | \boldsymbol{X}, \boldsymbol{Z}, \eta) \propto p(\boldsymbol{Z}, \eta | \alpha) p(\alpha)$
  - Sample $\eta \sim p(\eta | \boldsymbol{X}, \boldsymbol{Z}, \alpha) \propto p(\boldsymbol{Z}, \eta | \alpha)$

$$p(\alpha) = \mathrm{Gamma}(\alpha | a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} \alpha^{a_0 - 1} e^{-b_0 \alpha}$$

$$p(\boldsymbol{Z}, \eta | \alpha) = \frac{\alpha^{K-1}(\alpha + n)}{\Gamma(N)} \eta^{\alpha} (1 - \eta)^{N-1} \prod_{k=1}^{K} (n_k - 1)!$$

$$\propto \alpha^{K} \eta^{\alpha} + n \alpha^{K-1} \eta^{\alpha}$$

Sampling from beta

$$p(\eta | \boldsymbol{Z}, \alpha) = \mathrm{Beta}(\alpha + 1, N)$$

Bayes' theorem

$$p(\alpha | \boldsymbol{Z}, \eta) \propto \alpha^{a_0 + K - 1} e^{-(b_0 - \log \eta)\alpha} + n \alpha^{a_0 + K - 2} e^{-(b_0 - \log \eta)\alpha}$$

$$\propto \omega \mathrm{Gamma}(a_0 + K, b_0 - \log \eta) + (1 - \omega) \mathrm{Gamma}(a_0 + K - 1, b_0 - \log \eta)$$

$$\frac{\omega}{1 - \omega} = \frac{a_0 + K - 1}{N(b_0 - \log \eta)}$$

Sampling from gamma mixture

Summary

- Maximum likelihood estimation for finite GMM
  - EM algorithm and hard EM (k-means)
- Bayesian estimation for finite GMM
  - (Collapsed) Gibbs sampling
  - (Collapsed) variational Bayes
- Bayesian estimation for infinite GMM
  - Collapsed Gibbs sampling with Chinese restaurant process
  - Variational Bayes with stick breaking process
- Other topics
  - Hierarchical Dirichlet process
    - HMM, PCFG (sequential data), LDA (grouped data)
  - Beta process, gamma process, Gaussian process
    - (Nonnegative) matrix factorization

GS is feasible with SBP

CVB is feasible with CRP

References

- Bayesian modeling
  - C. Bishop: Pattern Recognition and Machine Learning, Springer, 2010 (Ch. 9-11 & Appendix B).
  - 石井 健一郎, 上田 修功: 続・分かりやすいパターン認識 ～教師なし学習入門～, オーム社, 2014.
  - 中島 伸一: 変分ベイズ学習, 講談社, 2016.
- Nonparametric Bayesian modeling
  - 佐藤 一誠: ノンパラメトリックベイズ 点過程と統計的機械学習の数理, 講談社, 2016.
  - D. Blei, M. Jordan: Variational Inference for Dirichlet Process Mixtures, Bayesian Analysis, Vol. 1, No. 1, pp.121-144, 2006.
  - J. Sung, Z. Ghahramani: Latent-Space Variational Bayes, IEEE Trans. on PAMI, Vol. 30, No. 12, 2008.
- Concentration parameter modeling
  - M. Escobar, M. West: Bayesian Density Estimation and Inference Using Mixtures, Journal of the American Statistical Association, Vol. 90, No. 430, pp. 577-588, 1995.
  - T. Stepleton: Understanding the Antoniak equation, 2008. http://www.cs.cmu.edu/~tss/antoniak.pdf

Assignments

- ML estimation
  - Derive the update formulas of the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ (p. 22) by letting the partial derivative of the lower bound (p. 20) w.r.t. each parameter equal to zero.
  - Implement the EM algorithm by using your favorite language.
- Bayesian estimation
  - Derive the variational posteriors of the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ (p. 47) by using the formulas (p. 46)
  - Try one of the following at least:
    - Implement the VB algorithm
    - Implement the GS algorithm
  - Optional:
    - Implement the other algorithms for finite/infinite GMMs.

How to Submit

- Report submission
  - Deadline: 7/21 (Fri.)
  - "Assignments" → "Assignments 6/7 (Yoshii)"
  - Upload two files
    - PDF file: Report document
    - Zip file: Codes and instructions (README)
- Program specification
  - *your_program_or_script* x.csv z.csv params.dat
  - Show the value of the likelihood or lower bound at each iteration
  - Output z.csv and params.dat
    - z.csv: Posterior probabilities of z_n

```
0.2, 0.3, 0.5
0.5, 0.1, 0.4
0.1, 0.8, 0.1
  ・ ・ ・
```