情報数理科学 VII　確率的取り扱い

松島　慎

2020 年 5 月 27 日

1　確率に関する基本的な概念の復習

本講義では確率の厳密な定義は気にしないで確率変数 X が x という値をとる確率を $P(x)$ と書き、どの確率変数のことであるか明記しない。まずは取りうる値が離散である場合を考える。

- 取りうる値が離散である場合、P は x の確率質量関数であるという。
- 確率変数が複数ある場合、例えば $P(x,y)$ を x と y の同時確率という。
- ある確率分布 $P(x,y,z)$ を y について周辺化するとは $P(x,z) = \sum_{y\in\mathcal{Y}} P(x,y,z)$ を求めることである。ここで $\mathcal{Y}$ は y の取りうる値全体。
- x の周辺分布とは、x 以外の確率変数すべてを周辺化した確率分布 $P(x)$ のこと。
- $P(x|y)$ は y があたえられたもとでの x の条件付き分布といい、$P(x|y) = P(x,y)/P(y)$ と定義する。
- 与えられたデータを確率変数ととらえて、確率分布の族 $\mathcal{M} = \{P(\mathbf{x};\theta)|\theta\in\Theta\}$ を考えることが多い。この時 θ は $P(\mathbf{x}:\theta)$ のパラメータと呼ぶ。
- $P(x,y) = P(x)P(y)$ が成り立つとき x と y は独立であるという。

- $P(x,y|z) = P(x|z)P(y|z)$ が成り立つとき x と y は z が与えられたもとで条件付独立であるという。

確率変数がとりうる値が連続の場合も同様の言葉遣いをする。本講義では $P(x)$ は常に 0。

- x が集合 R に含まれる確率が $\int_{x\in R} p(x)dx$ と書けるとき、$p(x)$ は x の確率密度関数であるという。
- $p(x,y)$ を x と y の同時確率密度関数という（明白であるとき同時確率という）。
- 周辺分布は和 $\sum_{y\in\mathcal{Y}}$ を積分 $\int_{y\in\mathcal{Y}} dy$ に変える
- 条件付確率密度関数 $p(x|y) = p(x,y)/p(y)$ と定義する。

1.1　グラフィカルモデル

複雑な確率分布の独立性を表すための図式としてグラフィカルモデルを導入する。本講義は有向グラフィカルモデルのみを扱う。有向グラフィカルモデルはベイジアンネットワークとも呼ばれる。各確率変数をノードとして確率変数 $\{w,x,y,z,\ldots\} = \mathcal{A}$ の分解特性を有向グラフで表す方法である。

- $\mathrm{parent}(\bullet) = \left\{\circ|\ \circ \text{から} \bullet \text{へ枝がある}\right\}$ とし、

$$p(\mathcal{A}) = \prod_{\bullet\in\mathcal{A}} p(\bullet|\mathrm{parent}(\bullet))$$

- 例:

$$p(x,y,z) = p(x|\mathrm{parent}(x))p(y|\mathrm{parent}(y))p(z|\mathrm{parent}(z))$$

- 分解されたそれぞれの因子がどのような確率分布であるか等は別で指定する
- プレートは繰り返しを表す。
- 影付きはデータ、すなわち観測できる変数を表す。
- 点はパラメータ（確率変数でない変数）を表す。
- 無向グラフィカルモデル（この講義では扱わない）とは全く別物なので注意。

1.2　具体的な確率分布族

- ベルヌイ分布
  $x\in\{0,1\}, \theta = q$ とし
  $P(x;\theta) = \mathrm{Bernoulli}(x;q) = q^{(1-x)}(1-q)^x$
- ベルヌイ分布 (別の表現)
  $x\in\{0,1\}, \theta\in\mathbb{R}$ とし
  $P(x;\theta) = \mathrm{Bernoulli}(x;\theta) \propto \exp(\theta x) = \frac{\exp(\theta x)}{\exp(\theta)+\exp(0)} = (1+\exp((2x-1)\theta x))^{-1}$
- 符号付きベルヌイ分布
  $x\in\{-1,1\}, \theta\in\mathbb{R}$ とし
  $P(x;\theta) = \mathrm{SignedBernoulli}(x;\theta) \propto \exp(\theta x) = \frac{\exp(\theta x)}{\exp(\theta)+\exp(-\theta)} = (1+\exp(-2\theta x))^{-1}$
- 正規分布
  $x\in\mathbb{R}, \theta = (\mu,\sigma)$ とし
  $p(x;\theta) = \mathrm{Normal}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$

- ラプラス分布
  $x\in\mathbb{R}, \theta = (\mu,b)$ とし
  $p(x;\theta) = \mathrm{Laplace}(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$

2　扱ってきた問題の確率的導出

教師あり学習では $X\in\mathbb{R}^{n\times d}$ と $y\in\mathbb{R}^n$ が与えられていた。教師なし学習では $X\in\mathbb{R}^{n\times d}$ が与えられていた。

統計モデルとはデータの確率分布が特定の確率分布族に入っている確率分布で記述しようすること。教師なし学習の場合 $\mathcal{M} = \{P(\mathbf{x};\theta)|\theta\in\Theta\}$ を統計モデルと呼ぶ。特に生成モデルという。

教師あり学習の場合 $\mathcal{M} = \{P(\mathbf{x},y;\theta)|\theta\in\Theta\}$ を統計モデルと呼ぶ。特に生成モデルという。

$\mathcal{M} = \{P(y|\mathbf{x};\theta)|\theta \in \Theta\}$ を統計モデルと呼ぶ。特に識別モデルという。

データをうまく説明するパラメータを求める。

2.1 最尤推定とその例

2.1.1 最尤推定の定義

識別モデルの場合

$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmax}} \underbrace{p(y|X;\theta)}_{\text{尤度}} \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} \prod_i p(y_i|\mathbf{x}_i;\theta) \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_i -\log p(y_i|\mathbf{x}_i;\theta)\end{aligned}$$

生成モデルの場合、

$$\hat{\theta} = \operatorname{argmax} p(X, y;\theta),$$

または

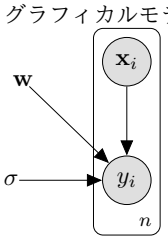
$$\hat{\theta} = \operatorname{argmax} p(X;\theta).$$

2.2 経験損失最小化の例

2.2.1 線形回帰の場合

線形回帰は識別モデルの最尤推定とみなせ、モデルのグラフィカルモデルは以下のように示すことができる。

- グラフィカルモデル



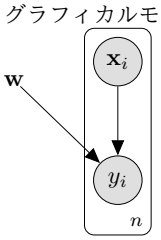
- 線形回帰 ($\theta = (\mathbf{w}, \sigma)$)

$$\begin{aligned}p(y_i|\mathbf{x}_i;\mathbf{w},\sigma) &= \text{Normal}(y_i|\mathbf{w}^\top\mathbf{x}_i,\sigma) \\ &= \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma}\left(\mathbf{w}^\top\mathbf{x}_i - y_i\right)^2\right) \\ \Rightarrow J(\mathbf{w}) &= -\log p(y|X;\mathbf{w}) \propto \sum_i \left(\mathbf{w}^\top\mathbf{x}_i - y_i\right)^2\end{aligned}$$

σ を固定し $\mathbf{w}$ を最尤推定または $(\sigma, \mathbf{w})$ の最尤推定は最小二乗法の解となる

2.2.2 2 値分類の場合

- グラフィカルモデル



- 二値分類 ($\theta = \mathbf{w}$)

$$\begin{aligned}P(y_i|\mathbf{x}_i;\mathbf{w}) &= \text{SignedBernoulli}(y_i;\mathbf{w}^\top\mathbf{x}_i) \propto \exp\left(y_i\mathbf{w}^\top\mathbf{x}_i\right) \\ \Leftrightarrow P(y_i|\mathbf{x}_i;\mathbf{w}) &= \frac{\exp\left(y_i\mathbf{w}^\top\mathbf{x}_i\right)}{\exp\left(-\mathbf{w}^\top\mathbf{x}_i\right) + \exp\left(\mathbf{w}^\top\mathbf{x}_i\right)} \\ &= \left(1 + \exp(-y_i\mathbf{w}^\top\mathbf{x}_i)\right)^{-1} \\ \Leftrightarrow J(\mathbf{w}) &= -\log P(y|X;\mathbf{w}) \\ &= \sum_i \log\left(1 + \exp(-y_i\mathbf{w}^\top\mathbf{x}_i)\right)\end{aligned}$$

$\mathbf{w}$ の最尤推定量はロジスティック回帰の解となる。

2.3 事後分布最大化（MAP 推定）とその例

2.3.1 事後分布最大化（MAP 推定）の定義

生成モデルの場合

生成モデルのパラメータ θ を確率変数と考えることで $p(X|\theta)$ を考えることができる。さらに $p(\theta)$ を設定すると、データとの同時分布などあらゆる分布を考えることができる。

- 事前分布：パラメータに関する事前知識

$$P(\theta)$$

- 事後分布：X が与えられた後でのパラメータに関する知識

$$P(\theta|X) = \frac{p(X,\theta)}{\int d\theta' p(X,\theta)} = \frac{\left(\prod_i P(\mathbf{x}_i|\theta)\right)p(\theta)}{\int d\theta' \left(\prod_i P(y_i|\mathbf{x}_i,\theta')\right)p(\theta')}$$

- MAP 推定（事後確率最大化）

$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmax}} P(\theta|X) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} \left(\prod_i P(\mathbf{x}|\theta)\right)p(\theta) \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}} -\sum_i \log P(\mathbf{x}|,\theta) + (-\log p(\theta))\end{aligned}$$

識別モデルの場合

識別モデルのパラメータ θ を確率変数と考えることで $p(y|X,\theta)$ を考えることができる。さらに $p(\theta)$ を設定して、X と θ は独立であると考える ($p(\theta) = p(\theta|X)$) データとの同時分布すると X が与えられたもとでのあらゆる分布（$p(y,\theta|X)$ や $p(\theta|X,y)$）を考えることができる。

- 事前分布

$$P(\theta)$$

- 事後分布

$$P(\theta|X,y) = \frac{p(y,\theta|X)}{\int d\theta' p(y,\theta|X)p(\theta)} = \frac{\left(\prod_i P(y_i|\mathbf{x}_i,\theta)\right)p(\theta)}{\int d\theta' \left(\prod_i P(y_i|\mathbf{x}_i,\theta')\right)p(\theta')}$$

- MAP 推定（事後確率最大化）

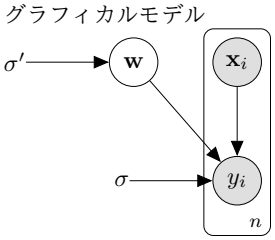
$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmax}} P(\theta|X,y) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} P(\theta,y|X) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} \left(\prod_i P(y_i|\mathbf{x}_i,\theta)\right)p(\theta) \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}} -\sum_i \log P(y_i|\mathbf{x}_i,\theta) + (-\log p(\theta))\end{aligned}$$

2.4 正則化付き経験損失最小化の例

2.4.1 線形回帰

- 線形回帰 $\theta = \mathbf{w}$

  グラフィカルモデル



各因子

$$
\begin{aligned}
p(\mathbf{w}) &= \mathrm{Normal}(\mathbf{w}|\mathbf{0}, \sigma' I) \\
&= \frac{1}{\sqrt{2\pi\sigma'}^d} \exp\left(-\frac{1}{2\sigma'} \|\mathbf{w}\|^2\right) \\
p(y_i|\mathbf{x}_i, \mathbf{w}) &= \mathrm{Normal}(y_i|\mathbf{w}^\top \mathbf{x}_i, \sigma) \\
&= \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma} \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2\right)
\end{aligned}
$$

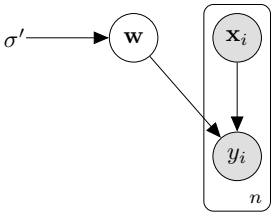
(σ, σ') を固定した $\mathbf{w}$ の MAP 推定量はリッジ回帰の解となる。この時事後分布は解析的に得られる。

$$
\begin{aligned}
&P(\mathbf{w}|X, y) \\
&= C \left(\prod_i \mathrm{Normal}(y_i|\mathbf{w}^\top \mathbf{x}_i, \sigma)\right) \mathrm{Normal}(\mathbf{w}|\mathbf{0}, \sigma' I) \\
&= C' \exp\left(-\frac{1}{2}\left(\sigma^{-1} \sum_i \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 + \sigma'^{-1} \|\mathbf{w}\|_2^2\right)\right) \\
&= C'' \exp\left(-\frac{1}{2} \mathbf{w}^\top \left(\sigma^{-1} X^\top X + \sigma'^{-1} I\right) \mathbf{w} - 2\sigma^{-1} X^\top y\right) \\
&= \mathrm{Normal}(\mathbf{w}; \sigma^{-1} \hat{\Sigma}^{-1} X^\top y, \hat{\Sigma})
\end{aligned}
$$

ここで $\hat{\Sigma} = (\sigma^{-1} X^\top X + \sigma'^{-1} I)^{-1}$ かつ $\mathrm{Normal}(\mathbf{w}; \mu, \Sigma) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\mu - x) \Sigma^{-1} (\mu - x)\right)$

2.4.2 二値分類

- グラフィカルモデル $\theta = \mathbf{w}$



- 各因子

$$
\begin{aligned}
p(\mathbf{w}) &= \mathrm{Normal}(\mathbf{w}|\mathbf{0}, \sigma' I; \sigma') \\
&= \frac{1}{\sqrt{2\pi\sigma'}^d} \exp\left(-\frac{1}{2\sigma'} \|\mathbf{w}\|^2\right) \\
P(y_i|\mathbf{x}_i, \mathbf{w}) &= \left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)\right)^{-1}
\end{aligned}
$$

- MAP 推定

$$
\begin{aligned}
p(\mathbf{w}|X, y) &= C P(y|X, \mathbf{w}) p(\mathbf{w}) \\
&= \prod_i 1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)^{-1} \frac{1}{\sqrt{2\pi\sigma'}^d} \exp\left(-\frac{1}{2\sigma'} \|\mathbf{w}\|^2\right) \\
\Rightarrow J(\mathbf{w}) &= -\log p(\mathbf{w}|X, y) \\
&= \sum_i \log\left(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)\right) + \frac{1}{2\sigma'} \|\mathbf{w}\|^2 + \text{const.}
\end{aligned}
$$

σ' を固定した $\mathbf{w}$ の MAP 推定量は L_2 ロジスティック回帰の解と対応する

2.5 潜在モデル（の最尤推定）とその例

2.5.1 潜在変数モデル

教師なし学習の多くは潜在変数モデルで記述される。

$$
p(\mathbf{x}_i; \theta) = \sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i; \theta) \text{ または } p(\mathbf{x}_i; \theta) = \int_{\mathbf{z}_i} d\mathbf{z}_i p(\mathbf{x}_i, \mathbf{z}_i; \theta)
$$

最尤推定は

$$
\begin{aligned}
\hat{\theta} &= \mathrm{argmax}\, p(X; \theta) = \mathrm{argmax} \prod_i p(\mathbf{x}_i; \theta) \\
&= \mathrm{argmin} \sum -\log\left(\sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i; \theta)\right)
\end{aligned}
$$

MAP 推定は

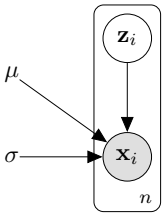
$$
\begin{aligned}
\hat{\theta} &= \mathrm{argmax}\, P(\theta|X) \\
&= \mathrm{argmax}\, P(\theta, X) \\
&= \mathrm{argmax}\left(\prod_i P(\mathbf{x}_i|\theta)\right) p(\theta) \\
&= \mathrm{argmin} \sum -\log\left(\sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i; \theta)\right) - \log p(\theta)
\end{aligned}
$$

となる。

2.5.2 クラスタリングの例

- GMM（簡単な例）$\theta = (\mu, \sigma)$

  グラフィカルモデル

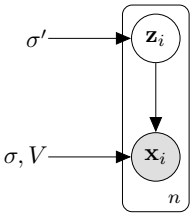


$$
\begin{aligned}
p(\mathbf{z}_i) &= K^{-1} \\
p(\mathbf{x}_i|\mathbf{z}_i; \theta) &= \prod_k \left(\mathrm{Normal}(\mathbf{x}_i|\mu_k, \sigma)\right)^{z_{ik}} \\
\Rightarrow p(\mathbf{x}_i; \theta) &= K^{-1} \sum_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma} \|\mu_k - \mathbf{x}_i\|^2\right) \\
\Rightarrow J(\theta) &= \sum_i \log\left(K^{-1} \sum_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma} \|\mu_i - \mathbf{x}_i\|^2\right)\right)
\end{aligned}
$$

σ を固定して μ の最尤推定量をもとめるのがソフト K-平均法の目的関数（$\sigma \to 0$ で K-平均法の目的関数に漸近）

2.5.3 確率的 PCA の例

- グラフィカルモデル



- 各因子

$$
\begin{aligned}
p(\mathbf{z}_i;\sigma') &= \mathrm{Normal}(\mathbf{z}_i|\mathbf{0},\sigma' I) \\
&= \frac{1}{\sqrt{2\pi\sigma'}^K}\exp\left(-\frac{1}{2\sigma}\left\|\mathbf{z}_i\right\|^2\right)
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{x}_i|\mathbf{z}_i;\sigma,V) &= \mathrm{Normal}(\mathbf{x}_i|V\mathbf{z}_i,\sigma I) \\
&= \frac{1}{\sqrt{2\pi\sigma}^d}\exp\left(-\frac{1}{2\sigma}\left\|\mathbf{x}_i - V\mathbf{z}_i\right\|^2\right) \\
\Rightarrow p(\mathbf{x}_i;\theta) &= \int d\mathbf{z}_i p(\mathbf{z}_i;\sigma')p(\mathbf{x}_i|\mathbf{z}_i;\sigma,V) \\
&= \int_{\mathbf{z}} d\mathbf{z}\mathrm{Normal}(\mathbf{x};V\mathbf{z},\sigma I)\mathrm{Normal}(\mathbf{z};0,\sigma' I) \\
&= \mathrm{Normal}(\mathbf{x}_i;\mathbf{0},\sigma'\sigma^{-1}V^\top V + I) \\
\Rightarrow J(V) &= \mathrm{tr}\left((\sigma'\sigma^{-1}V^\top V + I)^{-1}X^\top X\right)
\end{aligned}
$$

PCA と同様

$$
\begin{aligned}
&J(V) \\
&= -\sum_i \log p(\mathbf{x}_i;\theta) \\
&= \sum_i \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(|\sigma'\sigma^{-1}V^\top V + I|\right) + \mathbf{x}_i^\top(\sigma'\sigma^{-1}V^\top V + I)^{-1}\mathbf{x}_i \\
&= -\frac{n}{2}\log\left(|\sigma'\sigma^{-1}V^\top V + I|\right) + \mathrm{tr}\left((\sigma'\sigma^{-1}V^\top V + I)^{-1}XX^\top\right)
\end{aligned}
$$

正規分布の周辺化
$S = (\sigma^{-1}A^\top A + \sigma'^{-1}I)$, $T = (\sigma^{-1}AA^\top + \sigma'^{-1}I)$ とする

$$
\begin{aligned}
&\int_{\mathbf{z}} d\mathbf{z}\mathrm{Normal}(\mathbf{x}; A\mathbf{z}, \sigma I)\mathrm{Normal}(\mathbf{z}; 0, \sigma' I) \\
&\propto \int_{\mathbf{z}} d\mathbf{z} \exp\left(-\frac{1}{2}\sigma^{-1}\left\|A\mathbf{z}-\mathbf{x}\right\|^2\right)\exp\left(-\frac{1}{2}\sigma'^{-1}\left\|\mathbf{z}\right\|^2\right) \\
&\propto \int_{\mathbf{z}} d\mathbf{z} \exp\left(-\frac{1}{2}\left(\mathbf{z}^\top(\sigma^{-1}A^\top A + \sigma'^{-1}I)\mathbf{z} - 2(\sigma^{-1}A^\top\mathbf{x})^\top\mathbf{z} + \sigma^{-1}\mathbf{x}^\top\mathbf{x}\right)\right) \\
&\propto \int_{\mathbf{z}} d\mathbf{z} \exp\left(-\frac{1}{2}\left(\mathbf{z}^\top S\mathbf{z} - 2(\sigma^{-1}S^{-1}A^\top\mathbf{x})^\top S\mathbf{z} + \sigma^{-1}\mathbf{x}^\top\mathbf{x}\right)\right) \\
&\propto \int_{\mathbf{z}} d\mathbf{z} \exp\left(-\frac{1}{2}(\mathbf{z}-\sigma^{-1}S^{-1}A^\top\mathbf{x})^\top S(\mathbf{z}-\sigma^{-1}S^{-1}A^\top\mathbf{x})\right)\cdot\exp\left(-\frac{1}{2}\left(\sigma^{-1}\mathbf{x}^\top\mathbf{x} - (\sigma^{-1}A^\top\mathbf{x})^\top S^{-1}\sigma^{-1}A^\top\mathbf{x}\right)\right) \\
&= \int_{\mathbf{z}} d\mathbf{z} \exp\left(-\frac{1}{2}\mathbf{z}^\top S\mathbf{z}\right)\cdot\exp\left(-\frac{1}{2}\left(\sigma^{-1}\mathbf{x}^\top\mathbf{x} - \sigma(\sigma^{-1}A^\top\mathbf{x})^\top S^{-1}\sigma^{-1}A^\top\mathbf{x}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\mathbf{x}^\top\left(\sigma^{-1}I - \sigma^{-2}AS^{-1}A^\top\right)\mathbf{x}\right) \\
&= \exp\left(-\frac{1}{2}\mathbf{x}^\top\left(\sigma' T\right)^{-1}\mathbf{x}\right) \\
&\propto \mathrm{Normal}(\mathbf{x}; \mathbf{0}, \sigma' T)
\end{aligned}
$$

ここで $f(\mathbf{x}) \propto g(\mathbf{z})$ とは $\frac{f(\mathbf{x})}{\int_{\mathbf{x}} d\mathbf{x} f(\mathbf{x})} = \frac{g(\mathbf{x})}{\int_{\mathbf{x}} d\mathbf{z} g(\mathbf{x})} = p(\mathbf{x};\theta)$ を表すことに注意する。最後に現れる等式については A の特異値分解 $(U, V, \mathrm{diag}((s_i)_i))$ を用いて次式から分かる。$S = V\,\mathrm{diag}\left(\left(\sigma^{-1}s_i^2 + \sigma'^{-1}\right)_i\right)V^\top$, $T = U\,\mathrm{diag}\left(\left(\sigma^{-1}s_i^2 + \sigma'^{-1}\right)_i\right)U^\top$ なので、

$$
\begin{aligned}
&\sigma^{-1}I - \sigma^{-2}AS^{-1}A^\top \\
&= \sigma^{-1}I - \sigma U\,\mathrm{diag}((s_i)_i)V^\top V\,\mathrm{diag}\left(\left(\frac{1}{\sigma^{-1}s_i^2 + \sigma'^{-1}}\right)_i\right)V^\top V\,\mathrm{diag}((s_i)_i)U^\top \\
&= U\,\mathrm{diag}\left(\left(\sigma^{-1} - \frac{\sigma^{-2}s_i^2}{\sigma^{-1}s_i^2 + \sigma'^{-1}}\right)_i\right)U^\top \\
&= U\,\mathrm{diag}\left(\left(\frac{\sigma'^{-1}}{\sigma^{-1}s_i^2 + \sigma'^{-1}}\right)_i\right)U^\top \\
&= \sigma'^{-1}T^{-1}
\end{aligned}
$$

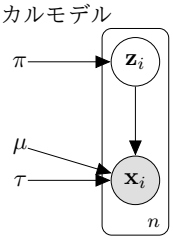
3 推定理論

3.1 混合ガウス分布モデル（GMM）

$$p(\mathbf{z}_i;\pi) = \mathrm{Catergorical}((z_{ik})_k|\pi) = \prod_k \pi_k^{z_{ik}}$$
$$p(\mathbf{x}_i|\mathbf{z}_i;\mu,\tau) = \prod_k \left(\mathrm{Normal}(\mathbf{x}_i|\mu_k,\tau_k^{-1}I)\right)^{z_{ik}}$$
$$= \prod_k \left(\sqrt{\frac{\tau_k}{2\pi}}^d \exp\left(-\frac{\tau_k}{2}\left\|\mu_k - \mathbf{x}_i\right\|^2\right)\right)^{z_{ik}}$$

$$\Rightarrow J(\theta) = -\log\left(p(\mathbf{X};\theta)\right)$$
$$-\sum_i \log\left(\sum_k \pi_k \sqrt{\frac{\tau_k}{2\pi}}^d \exp\left(-\frac{\tau_k}{2}\left\|\mu_k - \mathbf{x}_i\right\|^2\right)^{z_{ik}}\right)$$

グラフィカルモデル



3.2 KL ダイバージェンス

- 定義
  $\mathrm{KL}\left(q(x)\,\|\,p(x)\right) = \int_x dx q(x)\log\frac{q(x)}{p(x)}$
  離散の場合は適宜積分を和に変える
- 性質 $\mathrm{KL}\left(q(x)\,\|\,p(x)\right) \geq 0$
  等号は $p(x) = q(x)$ の時のみ成立
- 距離ではない（対称でない。三角不等式を満たさない。）が統計多様体に計量を与えられる

3.3 EM アルゴリズム再訪

次の目的関数を最小化する θ を求めることを考える。ただし、

$$J(\theta) = -\log\left(p(\mathbf{X};\theta)\right) = -\mathcal{L}(q,\theta) - \mathrm{KL}\left(q(\mathbf{Z})\,\|\,p(\mathbf{Z}|\mathbf{X};\theta)\right)$$

ここで、

$$\mathcal{L} = \sum q(\mathbf{Z})\log p(\mathbf{X},\mathbf{Z};\theta) - \sum q(\mathbf{Z})\log q(\mathbf{Z})$$

であり、目的関数は q によらないことに注意。ソフト K 平均法における γ は確率単体上にあったが、これはこの文脈における $q(\mathbf{Z})$ に相当する。EM アルゴリズムは完全データにおける最尤推定が可能な時適用できるアルゴリズムである。
EM アルゴリズム

1. θ を適当に定める。
2. E ステップ: $\mathcal{L}(q,\theta)$ を q について最小化

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\theta)$$

   目的関数の値は変わらない。実質的には $\gamma_i = q(\mathbf{z}_i)$ を計算する。
   $p(\mathbf{z}_i|\mathbf{x}_i,\theta) = \gamma^\top \mathbf{z}_i$ となるような γ を計算する。
   $q(\mathbf{Z})$ は事後確率分布における $\mathbf{Z}$ の期待値とみることもできる。
3. M ステップ: $\mathcal{L}(q,\theta)$ を θ について最小化

$$\begin{aligned}\theta &= \mathrm{argmin} - \mathcal{L}(q,\theta)\\ &= \mathrm{argmin} - \sum q(\mathbf{Z})\log p(\mathbf{X},\mathbf{Z};\theta)\\ &= \mathrm{argmin} - \sum_i\sum_k q(\mathbf{z}_i)\log p(\mathbf{x}_i,\mathbf{z}_i;\theta)\\ &= \mathrm{argmin} - \sum_i\sum_k \gamma_{ik}\log p(\mathbf{x}_i,\mathbf{z}_i;\theta)\end{aligned}$$

   （ほぼ）完全データにおける最尤推定
4. 値が変わらなくなるまで 2. と 3. を繰り返す。

3.3.1 GMM の例：E ステップ

$$\begin{aligned}q(\mathbf{Z}) &= p(\mathbf{Z}|\mathbf{X},\theta)\\ &= \prod_i p(\mathbf{z}_i|\mathbf{x}_i,\theta)\\ &= \prod_i \frac{\mathrm{Normal}(\mathbf{x}_i;\mu^\top\mathbf{z}_i,(\tau^\top\mathbf{z}_i)^{-1}I)\mathrm{Categorical}(\mathbf{z}_i;\pi)}{\sum_{z'}\mathrm{Normal}(\mathbf{x}_i;\mu^\top\mathbf{z}_i,\tau^\top\mathbf{z}_i)\mathrm{Categorical}(\mathbf{z}'_i;\pi)}\end{aligned}$$

$$\begin{aligned}\gamma_{ik} &= \mathrm{Normal}(\mathbf{x}_i;\mu_k,\tau_k^{-1})\mathrm{Categorical}(\mathbf{e}(k);\pi)\\ &= \frac{\sqrt{\frac{\tau_k}{2\pi}}^d \exp\left(-\frac{\tau_k}{2}\left\|\mu_k - \mathbf{x}_i\right\|^2\right)}{\sum_{k'}\sqrt{\frac{\tau_{k'}}{2\pi}}^d \exp\left(-\frac{\tau_{k'}}{2}\left\|\mu_{k'} - \mathbf{x}_i\right\|^2\right)}\end{aligned}$$

これをもとに $p(\mathbf{z}_i|\mathbf{x}_i,\theta) = \gamma^\top\mathbf{z}_i$ となるような γ を計算する。

3.3.2 GMM の例：M ステップ

$$\begin{aligned}&-\gamma_{ik}\log p(\mathbf{x}_i,\mathbf{z}_i;\theta)\\ &= \gamma^\top\mathbf{z}_i\left(-\log\pi^\top\mathbf{z}_i + \frac{d}{2}\left(\log 2\pi - \log\tau\right) + \frac{\tau_k}{2}\left\|\mu_k - \mathbf{x}_i\right\|^2\right)\\ &\Rightarrow \nabla_{\mu_k}\mathcal{L}(q,\theta) = \sum_i \gamma_{ik}\tau_k\left(\mu_k - \mathbf{x}_i\right)\\ &\nabla_{\tau_k}\mathcal{L}(q,\theta) = \sum_i \gamma_{ik}\left(\tau^{-1} + \frac{1}{2}\left\|\mu_k - \mathbf{x}_i\right\|_2^2\right)\\ &\nabla_{\pi_k}\mathcal{L}(q,\theta) = \sum_i \gamma_{ik}\end{aligned}$$

から

$$\begin{aligned}\mu_k &= \frac{\sum_i \gamma_{ik}\mathbf{x}_i}{\sum_i \gamma_{ik}}\\ \tau_k &= \frac{\sum_i \gamma_{ik}}{\sum_i \gamma_{ik}\frac{1}{2}\left\|\mu_k - \mathbf{x}_i\right\|^2}\\ \pi_k &= \frac{\sum_i \gamma_{ik}}{n}\end{aligned}$$

となる。π に確率単体制約があることに注意。

3.4 事後分布の近似推論

事後分布を求めることができると一般により複雑な予測が可能になる。

次の目的関数を最小化することを考える。

$$J(q_\theta, q_\mathbf{Z}) = \mathrm{KL}\left(q_\theta(\theta)q_\mathbf{Z}(\mathbf{Z})\,\|\,p(\mathbf{Z},\theta|\mathbf{X};\alpha)\right)$$

これは $\mathbf{X}$ の対数周辺尤度の下界。

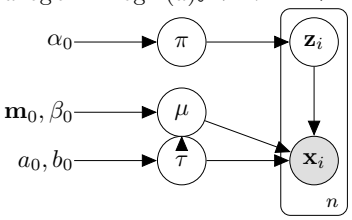
これにより、事後分布を次のように近似する。

$$p(\theta|\mathbf{X};\alpha) = \sum_{\mathbf{Z}} p(Z,\theta|\mathbf{X};\alpha) \approx q_\theta$$
$$p(\mathbf{Z}|\mathbf{X};\alpha) = \int_\theta d\theta p(\mathbf{Z},\theta|\mathbf{X};\alpha) \approx q_{\mathbf{Z}}$$

前節の GMM の場合、事前分布 $p(\theta;\zeta_0)$ を次のように置くとよい。$\zeta_0 = ((\alpha_{0k})_k, (a_{0k})_k, (b_{0k})_k, (\mathbf{m}_{0k})_k, (\beta_{0k})_k) \in \mathbb{R}^K \times \mathbb{R}^K \times \mathbb{R}^K \times \mathbb{R}^{K\times d} \times \mathbb{R}^K$ で、

$$\begin{aligned}
\log p(\theta;\zeta_0) &= p(\pi;\alpha_0)\prod_k \log p(\tau_k;a_{0k},b_{0k}) \log p(\mu_k|\tau_k) \\
\log p(\pi;\alpha_0) &= \log \mathrm{Dirichlet}(\pi;\alpha_0) \\
&= \sum_k (1-\alpha_{0k})\log(\pi_k) + \log Z(\alpha_0) \\
\log p(\tau_k;a_{0k},b_{0k}) &= \log \mathrm{Gamma}(\tau_k;a_{0k},b_{0k}) \\
&= (a_{0k}-1)\log\tau_k - b_{0k}\tau_k + \log Z(a_{0k},b_{0k}) \\
\log p(\mu_k|\tau_k;\mathbf{m}_k,\beta_k) &= \log \mathrm{Normal}(\mu_k|\mathbf{m}_{0k},\tau_k^{-1}{\beta_{0k}}^{-1}I) \\
&= \frac{d}{2}\log\frac{\beta_{0k}\tau_k}{2\pi} - \frac{\beta_{0k}\tau_k}{2}\left\|\mu_{0k}-\mathbf{m}_k\right\|^2
\end{aligned}$$

ここで $Z(\alpha) = \prod_k \Gamma(\alpha_k)/\Gamma(\sum_k \alpha_k)$ かつ $\log Z(\alpha,\beta) = a\log b - \log\Gamma(a)$。グラフィカルモデルを以下に示す。



VB-EM アルゴリズム

1. $q_\theta(\theta) = p(\theta;\zeta_0)$ とおく。
2. E ステップ: $J(q_{\mathbf{Z}}, q_\theta)$ を $q_{\mathbf{Z}}$ について最小化

$$\log q_{\mathbf{Z}}(\mathbf{Z}) \propto \int_\theta d\theta q_\theta(\theta)\log p(\mathbf{Z},\theta|\mathbf{X};\alpha)$$

目的関数の値は変わらない。実質的には $\gamma_i = q(\mathbf{z}_i)$ を計算する。
$\int_\theta \log p(\mathbf{z}_i|\mathbf{x}_i,\theta)q_\theta(\theta) \propto \gamma^\top \mathbf{z}_i$ となるような γ を計算する。
$q_{\mathbf{Z}}(\mathbf{Z})$ は $\mathbf{Z}$ の事後確率分布に基づく $\mathbf{Z}$ の期待値とみることもできる。

3. M ステップ: $J(q_{\mathbf{Z}}, q_\theta)$ を θ について最小化

$$\log q_\theta(\theta) \propto \sum_{\mathbf{Z}} \log p(\theta,\mathbf{Z}|\mathbf{X};\zeta_0) q_{\mathbf{Z}}(\mathbf{Z})$$

4. 値が変わらなくなるまで 2. と 3. を繰り返す。

$J(q_{\mathbf{Z}}, q_\theta)$ を $q_{\mathbf{Z}}$ について最小化

$$\begin{aligned}
&\mathrm{KL}\left(q_{\mathbf{Z}}q_\theta \,\|\, p(\mathbf{Z},\theta|X;\alpha)\right) \\
&= \int_\theta d\theta \sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z})q_\theta(\theta)\log\frac{p(\mathbf{Z},\theta|\mathbf{X},\alpha)}{q_\theta(\theta)q_{\mathbf{Z}}(\mathbf{Z})} \\
&= \int_\theta d\theta \sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z})q_\theta(\theta)\log p(\mathbf{Z},\theta|\mathbf{X},\alpha) \\
&- \sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z})\log q_{\mathbf{Z}}(\mathbf{Z}) - \int_\theta q_\theta(\theta)\log q_\theta(\theta) \\
&= \sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z})\log\exp\left(\int_\theta d\theta q_\theta(\theta)\log p(\mathbf{Z},\theta|\mathbf{X},\alpha)\right) \\
&- \sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z})\log q_{\mathbf{Z}}(\mathbf{Z}) - \int_\theta q_\theta(\theta)\log q_\theta(\theta) \\
&= \mathrm{KL}\left(q_{\mathbf{Z}}\middle\|\frac{\exp\left(\int_\theta d\theta q_\theta(\theta)\log p(\mathbf{Z},\theta|\mathbf{X},\alpha)\right)}{\sum_{\mathbf{Z}'}\exp\left(\int_\theta d\theta q_\theta(\theta')\log p(\mathbf{Z}',\theta|\mathbf{X},\alpha)\right)}\right) \\
&+ \sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z})\log\sum_{\mathbf{Z}'}\exp\left(\int_\theta d\theta q_\theta(\theta)\log p(\mathbf{Z}',\theta|\mathbf{X},\alpha)\right) \\
&- \int_\theta q_\theta(\theta)\log q_\theta(\theta)
\end{aligned}$$

第二項以降は $q_{\mathbf{Z}}$ に関して定数であるので、第一項の KL を最小化すればよい。M ステップも同様の構造を持つ。

3.4.1 GMM の例:E ステップの結果

$$q_\theta(\theta) = p(\theta;\zeta) = p(\theta;(\alpha,a,b,\mathbf{m},\beta))$$

である場合、 $q_{\mathbf{Z}}(\mathbf{Z}) = \prod_i q_{\mathbf{z}_i}(\mathbf{z}_i)$ で、

$$q_{\mathbf{z}_i}(\mathbf{z}_i) = \frac{\mathrm{Categorical}(\mathbf{z}_i|\tilde{\pi})\mathrm{Normal}\left(\mathbf{x}_i;\tilde{\mathbf{m}}^\top\mathbf{z}_i,(\tilde{\tau}^\top\mathbf{z}_i)^{-1}I\right)}{\sum_{\mathbf{z}_i'}\mathrm{Categorical}(\mathbf{z}_i'|\tilde{\pi})\mathrm{Normal}\left(\mathbf{x}_i;\tilde{\mathbf{m}}^\top\mathbf{z}_i',(\tilde{\tau}^\top\mathbf{z}_i')^{-1}I\right)}$$

となる。ここで、

$$\begin{aligned}
\tilde{\pi}_k &= \psi(\alpha_k) - \psi(\alpha^\top\mathbf{1}) + \frac{d}{2}\left(-\log a_k + \psi(a_k) - \beta_k^{-1}\right) \\
\tilde{\tau}_k &= \frac{a_k}{b_k} \\
\tilde{\mathbf{m}}_k &= \mathbf{m}_k
\end{aligned}$$

3.4.2 GMM の例:M ステップの結果

$$\begin{aligned}
q_{\mathbf{Z}}(\mathbf{Z}) &= \prod_i q_{\mathbf{z}_i}(\mathbf{z}_i) \\
&= \prod_i \gamma_i^\top \mathbf{z}_i
\end{aligned}$$

である場合、

$$q_\theta(\theta) = p(\theta;(\hat{\alpha},\hat{a},\hat{b},\hat{\mathbf{m}},\hat{\beta}))$$

となる。ここで、

$$
\begin{aligned}
\hat{\alpha}_k &= \alpha_{0k} + \sum_i \gamma_{ik} \\
\hat{\mathbf{m}}_k &= \frac{\beta_{0k}\mathbf{m}_{0k} + \sum \gamma_{ik}\mathbf{x}_i}{\beta_{0k} + \sum_i \gamma_{ik}} \\
\hat{a}_k &= a_{0k} + \sum_i \gamma_{ik} \\
\hat{b}_k &= b_{0k} + \frac{1}{2}\left(\beta_{0k}\left\|\mathbf{m}_{0k}\right\|^2 + \sum_i \gamma_{ik}\left\|\mathbf{x}_i\right\|^2 - \frac{\left\|\beta_{0k}\mathbf{m}_{0k} + \sum \gamma_{ik}\mathbf{x}_i\right\|^2}{\beta_{0k} + \sum_i \gamma_{ik}}\right) \\
&= b_{0k} + \frac{1}{2}\frac{\beta_k \sum_i \gamma_{ik}}{\beta_{0k} + \sum_i \gamma_{ik}}\left\|\beta_{0k}\mathbf{m}_{0k} - \frac{\sum_i \gamma_{ik}\mathbf{x}_i}{\sum_i \gamma_{ik}}\right\|^2 \\
&\quad + \sum_i \gamma_{ik}\left(\frac{\sum_i \gamma_{ik}\left\|\mathbf{x}_i\right\|^2}{\sum_i \gamma_{ik}} - \left\|\frac{\sum_i \gamma_{ik}\mathbf{x}_i}{\sum_i \gamma_{ik}}\right\|^2\right) \\
\hat{\beta}_k &= \beta_{0k} + \sum_i \gamma_{ik}
\end{aligned}
$$

3.5 E ステップの計算

$$
\begin{aligned}
&q_{\mathbf{Z}}(\mathbf{Z}) \\
&= \log p(\mathbf{Z}, \theta; \zeta_0) \\
&= \log p(\mathbf{X}, \mathbf{Z}, \theta; \zeta_0) - \log p(\mathbf{X}; \zeta_0) \\
&\propto \log p(\mathbf{X}, \mathbf{Z}, \theta; \zeta_0) \\
&= \log p(\mathbf{Z}, \pi; \zeta_0) + \log p(\mathbf{X}|\mathbf{Z}, \mu, \tau) + \log p(\mu, \tau; \zeta_0) + \log p(\pi; \zeta_0) \\
&\propto \log p(\mathbf{Z}|\pi) + \log p(\mathbf{X}|\mathbf{Z}, \mu, \tau)
\end{aligned}
$$

ここでは $f(\mathbf{Z}) \propto g(\mathbf{Z}) \Leftrightarrow q(\mathbf{Z}) = \frac{\exp(f(\mathbf{Z}))}{\sum_{\mathbf{Z}} \exp(f(\mathbf{Z}))} = \frac{\exp(g(\mathbf{Z}))}{\sum_{\mathbf{Z}} \exp(g(\mathbf{Z}))}$ である。したがって、

$$
\begin{aligned}
\log q_{\mathbf{Z}}(\mathbf{Z}) &= \int_\theta d\theta q_\theta(\theta) \log p(\mathbf{Z}, \theta|\mathbf{X}, \zeta_0) \\
&\propto \int_\theta d\theta q_\theta(\theta) \log p(\mathbf{Z}|\pi) + \int_\theta d\theta q_\theta(\theta) \log p(\mathbf{X}|\mathbf{Z}, \mu, \tau)
\end{aligned}
$$

第一項に関して、

$$
\begin{aligned}
&\int_\theta d\theta q_\theta(\theta) \log p(\mathbf{Z}|\pi) \\
&= \int_\pi d\pi \mathrm{Dirichlet}(\pi; \alpha)\left(\sum_i \log \mathrm{Categorical}(\mathbf{z}_i|\pi)\right) \\
&= \sum_i \sum_k \int_\pi d\pi z_{ik} \mathrm{Dirichlet}(\pi; \alpha') \log \pi_k \\
&= \sum_i \sum_k z_{ik}(\psi(\alpha'_k) - \psi({\alpha'}^\top \mathbf{1}))
\end{aligned}
$$

ここで $\int_\pi d\pi \mathrm{Dirichlet}(\pi; \alpha') \log \pi_k = \psi(\alpha'_k) - \psi\left({\alpha'}^\top \mathbf{1}\right)$ であることを用いた。第二項に関して

$$
\begin{aligned}
&\int_\theta d\theta q_\theta(\theta) \log p(\mathbf{X}|\mathbf{Z}, \mu, \tau) \\
&= \iint_{\tau,\mu} d\tau d\mu \prod_k \mathrm{Gamma}(\tau_k; a_k, b_k)\mathrm{Normal}(\mu_k|\mathbf{m}_k, (\beta_k\tau_k)^{-1}I) \\
&\qquad \cdot \sum_i \left(\log \mathrm{Normal}(\mathbf{x}_i; \mu_k, \tau_k^{-1}I)\right)^{z_{ik}} \\
&= \iint_{\tau,\mu} d\tau d\mu \prod_k \mathrm{Gamma}(\tau_k; a_k, b_k)\mathrm{Normal}(\mu_k|\mathbf{m}_k, (\beta_k\tau_k)^{-1}I) \\
&\qquad \cdot \sum_i \left(\frac{d}{2}\log\frac{\tau_k}{2\pi} - \frac{\tau_k}{2}\left\|\mathbf{x}_i - \mu_k\right\|^2\right)^{z_{ik}}
\end{aligned}
$$

正規分布に関する積分は

$$
\begin{aligned}
&\prod\left(\int_{\mu_k} d\mu_k \mathrm{Normal}(\mu_k|\mathbf{m}_k, (\beta_k\tau_k)^{-1}I)\left\|\mathbf{x}_i - \mu_k\right\|^2\right)^{z_{ik}} \\
&= \prod_k\left(\int_{\mu_k} d\mu_k \mathrm{Normal}(\mu_k|\mathbf{m}_k, (\beta_k\tau_k)^{-1}I)\left(\mathbf{x}_i^\top\mathbf{x}_i - 2\mathbf{x}_i^\top\mu_k + \mu_k^\top\mu_k\right)\right)^{z_{ik}} \\
&= \left(\left\|\mathbf{x}_i - \mathbf{m}_k\right\|^2 + d(\beta_k\tau_k)^{-1}\right)^{z_{ik}}
\end{aligned}
$$

である。よって第二項は

$$
\begin{aligned}
&\iint_{\mu,\tau} d\tau d\mu \prod_k \mathrm{Gamma}(\tau_k, a_k, b_k)\mathrm{Normal}(\mu_k|\mathbf{m}_k, (\beta_k\tau_k)^{-1}I) \\
&\qquad \cdot \left(\frac{d}{2}\log\frac{\tau_k}{2\pi} - \frac{\tau_k}{2}\left\|\mathbf{x}_i - \mu_k\right\|^2\right)^{z_{ik}} \\
&= \int_\tau d\tau \prod_k \mathrm{Gamma}(\tau_k; \bar{a}, \bar{b})\left(\frac{d}{2}\log\frac{\tau_k}{2\pi} - \frac{\tau_k\left\|\mathbf{x}_i - \mathbf{m}_k\right\|^2 + d\beta_k^{-1}}{2}\right)^{z_i k} \\
&= \sum_k z_{ik}\left(\frac{d}{2}\psi(a) + \frac{d}{2}\log\frac{1}{2\pi b_k} - \frac{1}{2}\left(\frac{a_k}{b_k}\left\|\mathbf{x}_i - \mathbf{m}_k\right\|^2 + d\beta_k^{-1}\right)\right)
\end{aligned}
$$

となる。なお、最後に

$$
\begin{aligned}
\int_{\tau_k} d\tau_k \mathrm{Gamma}(\tau_k; a, b)\log\tau_k &= \psi(a) - \log b \\
\int_{\tau_k} d\tau_k \mathrm{Gamma}(\tau_k; \alpha, \beta)\tau_k &= \frac{a}{b}
\end{aligned}
$$

であることを用いた。これらをまとめると、

$$
\begin{aligned}
&\log q_{\mathbf{Z}}(\mathbf{Z}) \\
&= \int_\theta d\theta q_\theta(\theta) \log p(\mathbf{Z}, \theta|\mathbf{X}; \alpha) \\
&\propto \int_\theta d\theta q_\theta(\theta) \log p(\mathbf{Z}|\pi) + \int_\theta d\theta q_\theta(\theta) \log p(\mathbf{X}|\mathbf{Z}, \mu, \tau) \\
&= \sum_i \sum_k z_{ik}\left(\psi(\alpha_k) - \psi(\alpha^\top\mathbf{1}) + \frac{d}{2}\left(-\log a_k + \psi(a_k) - \beta_k^{-1}\right)\right) \\
&+ \sum_i \sum_k z_{ik}\left(\frac{d}{2}\log\frac{a_k}{2\pi b_k} - \frac{a_k}{2b_k}\left\|\mathbf{x}_i - \bar{\mathbf{m}}_k\right\|^2\right)
\end{aligned}
$$

であるので

$$
\gamma_{ik} = \frac{\tilde{\pi}_k \mathrm{Normal}(\mathbf{x}; \bar{\mathbf{m}}_k, \frac{\bar{b}_k}{\bar{a}_k}I)}{\sum_{k'} \tilde{\pi}_{k'} \mathrm{Normal}(\mathbf{x}; \bar{\mathbf{m}}_{k'}, \frac{\bar{b}_{k'}}{\bar{a}_{k'}}I)}
$$

となる。ここで、

$$\bar{\pi}_k = \exp\left(\psi(\bar{\alpha'}_k) - \psi(\bar{\alpha'}^\top \mathbf{1}) + \frac{d}{2}\left(-\log a_k + \psi(a_k) - \beta_k^{-1}\right)\right)$$

である。

3.5.1 GMM の例:M ステップの計算

記号の簡潔さのため $\zeta_0 = \zeta = (\alpha, a, b, \mathbf{m}, \beta)$ とする。

$$\begin{aligned}
&\sum_{\mathbf{Z}} p(\theta, \mathbf{Z}|\mathbf{X}; \zeta) q_{\mathbf{Z}}(\mathbf{Z}) \\
&\propto \sum_{\mathbf{Z}} \log p(\theta, \mathbf{Z}, \mathbf{X}; \zeta) q_{\mathbf{Z}}(\mathbf{Z}) \\
&\propto \sum_{\mathbf{Z}} \log p(\mathbf{Z}, \pi; \alpha) + \log p(\mathbf{X}, \mu, \tau|\mathbf{Z}; a, b, \mathbf{m}, \beta)
\end{aligned}$$

よって $q_\theta(\theta) = q_\pi(\pi) q_{\mu,\tau}(\mu, \tau)$ となることがわかる。ここで、

$$\begin{aligned}
q_\pi(\pi) &= \frac{\exp\left(\sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z}) \log p(\mathbf{Z}, \pi; \alpha)\right)}{\int_\pi d\pi \exp\left(\sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z}) \log p(\mathbf{Z}, \pi; \alpha)\right)} \\
q_{\mu,\tau}(\mu, \tau) &= \frac{\exp\left(\sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z}) \log p(\mathbf{X}, \mu, \tau|\mathbf{Z}; a, b, \mathbf{m}, \beta)\right)}{\int_\pi d\pi \exp\left(\sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z}) \log p(\mathbf{X}, \mu, \tau|\mathbf{Z}; a, b, \mathbf{m}, \beta)\right)}
\end{aligned}$$

q_π について

$$\begin{aligned}
&\sum_{\mathbf{Z}} q_{\mathbf{Z}}(\mathbf{Z}) \log p(\mathbf{Z}, \pi; \alpha) \\
&= \left(\sum_i \sum_{\mathbf{z}_i} q_{\mathbf{z}_i}(\mathbf{z}_i) \log \mathrm{Categorical}(\mathbf{z}_i; \pi)\right) + \log \mathrm{Dirichlet}(\pi; \alpha) \\
&= \left(\sum_i \sum_k \gamma_{ik} \log \pi_k\right) - \sum_k (\alpha_k - 1) \log \pi_k - \log Z(\alpha) \\
&\propto \sum_k \left(\alpha_k - 1 + \sum_i \gamma_{ik}\right) \log \pi_k
\end{aligned}$$

ここでは $f(\pi) \propto g(\pi) \Leftrightarrow q_\pi = \frac{\exp(f(\pi))}{\int_\pi d\pi \exp(f(\pi))} = \frac{\exp(g(\pi))}{\int_\pi d\pi \exp(g(\pi))}$
よって

$$q_\pi = \mathrm{Dirichlet}(\pi; (\alpha_k + \sum \gamma_{ik})_k)$$

である。

$q_{\mu,\tau}$ について、

$$\begin{aligned}
&\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mu, \tau|\mathbf{Z}; \zeta) \\
&= \left(\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}, \mu, \tau)\right) + \log p(\tau, \mu; a, b, \mathbf{m}) \\
&= \left(\sum_i \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \log p(\mathbf{x}_i|\mathbf{z}_i, \mu, \tau))\right) + \log p(\tau, \mu; a, b, \mathbf{m})
\end{aligned}$$

さらに第一項は

$$\begin{aligned}
&\sum_i \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \log p(\mathbf{x}_i|\mathbf{z}_i, \mu, \tau) \\
&= \sum_i \sum_k \gamma_{ik} \left(\frac{d}{2} \log \frac{\tau_k}{2\pi} - \frac{\tau_k}{2} \|\mathbf{x}_i - \mu_k\|^2\right)
\end{aligned}$$

第二項は

$$\begin{aligned}
&\log p(\tau, \mu; \zeta) \\
&= \log p(\tau; a, b) + \log p(\mu|\tau; \mathbf{m}, \beta) \\
&= \sum_k \log \mathrm{Gamma}(\tau_k; a, b) + \log \mathrm{Normal}(\mu_k; \mathbf{m}_k, (\beta_k \tau_k)^{-1} I) \\
&= \sum_k (a-1) \log \tau_k - b\tau_k + \log Z(a, b) + \frac{d}{2} \log \frac{\beta_k \tau_k}{2\pi} - \frac{\beta_k \tau_k}{2} \|\mu_k - \mathbf{m}_k\|^2
\end{aligned}$$

と計算できる。μ_k についての項を整理すると、

$$\begin{aligned}
&-\frac{\beta_k \tau_k}{2} \|\mu_k - \mathbf{m}_k\|^2 - \sum_i \gamma_{ik} \frac{\tau_k}{2} \|\mathbf{x}_i - \mu_k\|^2 \\
&= -\frac{\tau_k}{2}\left(\left(\beta_k + \sum_i \gamma_{ik}\right) \|\mu_k\|^2 - 2\mu_k^\top \left(\beta_k \mathbf{m}_k + \sum_i \gamma_{ik} \mathbf{x}_i\right)\right) \\
&\quad - \frac{\tau_k}{2}\left(\beta_k \|\mathbf{m}_k\|^2 + \sum_i \gamma_{ik} \|\mathbf{x}_i\|^2\right) \\
&= -\frac{\left(\beta_k + \sum_i \gamma_{ik}\right) \tau_k}{2} \left\|\mu_k - \frac{\beta_k \mathbf{m}_k + \sum \gamma_{ik} \mathbf{x}_i}{\beta_k + \sum_i \gamma_{ik}}\right\|^2 \\
&\quad - \frac{\tau_k}{2}\left(\beta_k \|\mathbf{m}_k\|^2 + \sum_i \gamma_{ik} \|\mathbf{x}_i\|^2 - \frac{\left\|\beta_k \mathbf{m}_k + \sum \gamma_{ik} \mathbf{x}_i\right\|^2}{\beta_k + \sum_i \gamma_{ik}}\right)
\end{aligned}$$

となる。よって

$$\begin{aligned}
\hat{\beta}_k &= \beta_k + \sum_i \gamma_{ik} \\
\hat{\mathbf{m}}_k &= \frac{\beta_k \mathbf{m} + \sum \gamma_{ik} \mathbf{x}_i}{\beta_k + \sum_i \gamma_{ik}} \\
\hat{b}_k &= \frac{1}{2}\left(\beta_k \|\mathbf{m}\|_k^2 + \sum_i \gamma_{ik} \|\mathbf{x}_i\|^2 - \frac{\left\|\mathbf{m}_k + \sum \gamma_{ik} \mathbf{x}_i\right\|^2}{\beta_k + \sum_i \gamma_{ik}}\right) + b_k
\end{aligned}$$

と定義すると、$q_{\mu,\tau}(\mu, \tau)$ は以下のように整理できる。

$$\begin{aligned}
&\sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mu, \tau|\mathbf{Z}; \zeta) \\
&\propto \sum_k \left(\frac{d}{2} \log \frac{\hat{\tau}_k}{2\pi} - \frac{\hat{\tau}_k}{2} \|\mu_k - \hat{\mathbf{m}}_k\|\right) \\
&+ \left(a_k + \frac{d}{2} \sum_i \gamma_{ik} - 1\right) \log \tau_k - b_k \tau_k - (\hat{b}_k - b_k) \tau_k \\
&\propto \log \mathrm{Gamma}(\tau_k; \hat{a}_k, \hat{b}_k) + \log \mathrm{Normal}(\mu_k; \hat{\mathbf{m}}_k, (\hat{\beta}_k \tau_k)^{-1} I)
\end{aligned}$$

ここで $f(\mu, \tau) \propto g(\mu, \tau) \Leftrightarrow q_{\mu,\tau}(\mu, \tau) = \frac{\exp(f(\mu,\tau))}{\iint_{\mu,\tau} d\mu d\tau \exp(f(\mu,\tau))} = \frac{\exp(g(\mu,\tau))}{\iint_{\mu,\tau} d\mu d\tau \exp(g(\mu,\tau))}$。よって $q_{\mu,\tau}(\mu, \tau)$ は事前分布の超パラメータを別の値に設定した確率分布で表せる。