2.3.5　多次元ガウス分布

分布関数

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{2.72}$$

($\boldsymbol{\Sigma}$ は正定値対称行列)

規格化定数の求め方

$$C = \int d\boldsymbol{x}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{A.5.1}$$

を考える。$\boldsymbol{\Sigma}$ が正定値であるため、$\boldsymbol{\Sigma}^{-1}$ も正定値だから、下三角行列 $\boldsymbol{L}$ を使ってコレスキー分解

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}\boldsymbol{L}^T \tag{A.5.2}$$

できる。積分変数の変数変換

$$\boldsymbol{y} = \boldsymbol{L}^T(\boldsymbol{x}-\boldsymbol{\mu}) \tag{A.5.3}$$

とすると、

$$d\boldsymbol{x} = \frac{1}{|\boldsymbol{L}^T|}d\boldsymbol{y} = \sqrt{|\boldsymbol{\Sigma}|}d\boldsymbol{y} \tag{A.5.4}$$

より

$$\begin{aligned} C &= \sqrt{|\boldsymbol{\Sigma}|}\int d\boldsymbol{y}\exp\left\{-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right\} \\ &= \sqrt{|\boldsymbol{\Sigma}|}\sqrt{(2\pi)^D} \\ &= \sqrt{(2\pi)^D|\boldsymbol{\Sigma}|} \end{aligned} \tag{A.5.5}$$

対数表示

$$\ln\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}\left\{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \ln|\boldsymbol{\Sigma}| + D\ln 2\pi\right\} \tag{2.73}$$

$\boldsymbol{\Sigma}$ が対角行列の場合

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_D^2 \end{pmatrix} \tag{2.74}$$

を代入すると

$$\begin{aligned} \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\sum_{d=1}^{D}\left\{\frac{(x_d-\mu_d)^2}{\sigma_d^2} + \ln\sigma_d^2 + \ln 2\pi\right\} \\ &= \ln\prod_{d=1}^{D}\mathcal{N}(x_d|\mu_d, \sigma_d^2) \end{aligned} \tag{2.75}$$

となる。ここで、右辺は1次元ガウス分布の式

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2.64}$$

で、計算には以下を使った。

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_D^2} \end{pmatrix}$$

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \sum_{d=1}^{D}\frac{(x_d-\mu_d)^2}{\sigma_d^2}$$

$$\ln|\boldsymbol{\Sigma}| = \sum_{d=1}^{D}\ln\sigma_d^2$$

基本的な期待値

$$\begin{aligned} \langle\boldsymbol{x}\rangle &= \int d\boldsymbol{x}\boldsymbol{x}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \int d\boldsymbol{x}\boldsymbol{x}\frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &= \int d\tilde{\boldsymbol{x}}(\tilde{\boldsymbol{x}}+\boldsymbol{\mu})\frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\tilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}\right\} \\ &= \boldsymbol{\mu}\int d\tilde{\boldsymbol{x}}\frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\tilde{\boldsymbol{x}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}\right\} \\ &= \boldsymbol{\mu} \end{aligned} \tag{2.76}$$

$$
\begin{aligned}
\langle \boldsymbol{x}\boldsymbol{x}^T \rangle &= \int d\boldsymbol{x}\, \boldsymbol{x}\boldsymbol{x}^T \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \int d\boldsymbol{x}\, \boldsymbol{x}\boldsymbol{x}^T \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
&= \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \int d\tilde{\boldsymbol{x}} (\tilde{\boldsymbol{x}}+\boldsymbol{\mu})(\tilde{\boldsymbol{x}}+\boldsymbol{\mu})^T \exp\left\{-\frac{1}{2}\tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}\right\} \\
&= \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \int d\tilde{\boldsymbol{x}} (\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T + \boldsymbol{\mu}\tilde{\boldsymbol{x}}^T + \tilde{\boldsymbol{x}}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T) \\
&\exp\left\{-\frac{1}{2}\tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}\right\} \\
&= \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \int d\tilde{\boldsymbol{x}}\, \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \exp\left\{-\frac{1}{2}\tilde{\boldsymbol{x}}^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}\right\} + \boldsymbol{\mu}\boldsymbol{\mu}^T
\end{aligned}
$$

ここで、右辺第一項の i, j 成分を I_{ij} として、改めて $\tilde{\boldsymbol{x}}$ を $\boldsymbol{x}$ と書くと

$$
I_{ij} = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \int d\boldsymbol{x}\, x_i x_j \exp\left\{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right\}
$$

ここで、$\boldsymbol{\Sigma}$ の逆行列を $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$ とすると、

$$
\begin{aligned}
I_{ij} &= \sqrt{\frac{|\boldsymbol{M}|}{(2\pi)^D}} \int d\boldsymbol{x}\, x_i x_j \exp\left\{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{M}\boldsymbol{x}\right\} \\
&= \sqrt{\frac{|\boldsymbol{M}|}{(2\pi)^D}} (-2) \frac{\partial}{\partial M_{ij}} \int d\boldsymbol{x} \exp\left\{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{M}\boldsymbol{x}\right\} \\
&= \sqrt{\frac{|\boldsymbol{M}|}{(2\pi)^D}} (-2) \frac{\partial}{\partial M_{ij}} \sqrt{\frac{(2\pi)^D}{|\boldsymbol{M}|}} \\
&= -2|\boldsymbol{M}|^{\frac{1}{2}} \frac{\partial}{\partial M_{ij}} |\boldsymbol{M}|^{-\frac{1}{2}} \\
&= \frac{1}{|\boldsymbol{M}|} \frac{\partial |\boldsymbol{M}|}{\partial M_{ij}} \\
&= \frac{\tilde{M}_{ij}^T}{|\boldsymbol{M}|} \\
&= \Sigma_{ij}
\end{aligned}
$$

ここで、$\tilde{M}$ は $\boldsymbol{M}$ の余因子行列。

以上により

$$
\langle \boldsymbol{x}\boldsymbol{x}^T \rangle = \tilde{\boldsymbol{\mu}}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \tag{2.77}
$$

エントロピー

$$
\begin{aligned}
\mathrm{H}[\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})] &= -\langle \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\rangle \\
&= \frac{1}{2}\left\{\langle(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\rangle + \ln|\boldsymbol{\Sigma}| + D\ln 2\pi\right\}
\end{aligned} \tag{2.78}
$$

右辺中括弧内第一項は計算できる。（テキストとは 2 行目が違う？）

$$
\begin{aligned}
&\langle(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\rangle \\
&= \langle \mathrm{Tr}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}]\rangle \\
&= \langle \mathrm{Tr}[(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{x}\boldsymbol{\mu}^T-\boldsymbol{\mu}\boldsymbol{x}^T+\boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}]\rangle \\
&= \mathrm{Tr}[(\langle\boldsymbol{x}\boldsymbol{x}^T\rangle-\langle\boldsymbol{x}\rangle\boldsymbol{\mu}^T-\boldsymbol{\mu}\langle\boldsymbol{x}^T\rangle+\boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}] \\
&= \mathrm{Tr}[(\boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Sigma}-\boldsymbol{\mu}\boldsymbol{\mu}^T-\boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}] \\
&= \mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}] \\
&= \mathrm{Tr}(I_D) \\
&= D
\end{aligned} \tag{2.79}
$$

(2.78) に (2.79) を代入して

$$
\mathrm{H}[\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})] = \frac{1}{2}\left\{\ln|\boldsymbol{\Sigma}| + D(\ln 2\pi + 1)\right\} \tag{2.80}
$$

KL ダイバージェンス

$$
p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{A.5.6}
$$
$$
q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) \tag{A.5.7}
$$

として、KL ダイバージェンスの定義式 (2.13) を使うと

$$
\begin{aligned}
&\mathrm{KL}[q(\boldsymbol{x})][p(\boldsymbol{x})] \\
&= -\mathrm{H}[\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})] - \langle \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\rangle_{q(\boldsymbol{x})} \\
&= -\frac{1}{2}\left\{\ln|\hat{\boldsymbol{\Sigma}}| + D(\ln 2\pi + 1)\right\} \\
&\quad + \frac{1}{2}\left\{\langle(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\rangle_{q(\boldsymbol{x})} + \ln|\boldsymbol{\Sigma}| + D\ln 2\pi\right\} \\
&= \frac{1}{2}\left\{\langle(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\rangle_{q(\boldsymbol{x})} + \ln\frac{|\boldsymbol{\Sigma}|}{|\hat{\boldsymbol{\Sigma}}|} - D\right\}
\end{aligned} \tag{A.5.8}
$$

ここで、中括弧内第一項を考える。

$$
\begin{aligned}
&\langle(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\rangle_{q(\boldsymbol{x})} \\
&= \langle \mathrm{Tr}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right]\rangle_{q(\boldsymbol{x})} \\
&= \langle \mathrm{Tr}\left[(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{x}\boldsymbol{\mu}^T-\boldsymbol{\mu}\boldsymbol{x}^T+\boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}\right]\rangle_{q(\boldsymbol{x})} \\
&= \mathrm{Tr}\left[\left(\langle\boldsymbol{x}\boldsymbol{x}^T\rangle_{q(\boldsymbol{x})}-\langle\boldsymbol{x}\rangle_{q(\boldsymbol{x})}\boldsymbol{\mu}^T-\boldsymbol{\mu}\langle\boldsymbol{x}^T\rangle_{q(\boldsymbol{x})}+\boldsymbol{\mu}\boldsymbol{\mu}^T\right)\boldsymbol{\Sigma}^{-1}\right] \\
&= \mathrm{Tr}\left[\left(\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T+\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^T-\boldsymbol{\mu}\hat{\boldsymbol{\mu}}^T+\boldsymbol{\mu}\boldsymbol{\mu}^T\right)\boldsymbol{\Sigma}^{-1}\right] \\
&= \mathrm{Tr}\left[\left(\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T-\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^T-\boldsymbol{\mu}\hat{\boldsymbol{\mu}}^T+\boldsymbol{\mu}\boldsymbol{\mu}^T+\hat{\boldsymbol{\Sigma}}\right)\boldsymbol{\Sigma}^{-1}\right] \\
&= \mathrm{Tr}\left[\left((\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T+\hat{\boldsymbol{\Sigma}}\right)\boldsymbol{\Sigma}^{-1}\right] \qquad \text{(A.5.9)}
\end{aligned}
$$

よって、KL ダイバージェンスは

$$
\begin{aligned}
&\mathrm{KL}[q(\boldsymbol{x})][p(\boldsymbol{x})] \\
&= \frac{1}{2}\left\{\mathrm{Tr}\left[\left((\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T+\hat{\boldsymbol{\Sigma}}\right)\boldsymbol{\Sigma}^{-1}\right]+\ln\frac{|\boldsymbol{\Sigma}|}{|\hat{\boldsymbol{\Sigma}}|}-D\right\} \qquad (2.84)
\end{aligned}
$$

となる。