パターン認識と機械学習 演習問題解答

合同会社レスタス

目次

1 序論

1.1 （基本）多項式回帰の最小二乗和誤差

(1.2) を w_i で偏微分すると

$$\begin{aligned}\frac{\partial E(\boldsymbol{w})}{\partial w_i} &= \frac{\partial}{2\partial w_i}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w})-t_n\}^2 \\ &= \sum_{n=1}^{N}(x_n)^i\{y(x_n,\boldsymbol{w})-t_n\} \\ &= \sum_{j=0}^{M}\sum_{n=1}^{N}(x_n)^{i+j}w_j - \sum_{n=1}^{N}(x_n)^i t_n\end{aligned}$$

$E(\boldsymbol{w})$ を最小化する $\boldsymbol{w}$ では偏微分係数は 0 なので

$$\sum_{j=0}^{M}\sum_{n=1}^{N}(x_n)^{i+j}w_j = \sum_{n=1}^{N}(x_n)^i t_n$$

この式に (1.123) を用いると (1.122) と等価であることが分かる。

1.2 （基本）正則化項付き多項式回帰の最小二乗和誤差

(1.4) を w_i で偏微分すると

$$\begin{aligned}\frac{\partial \widetilde{E}(\boldsymbol{w})}{\partial w_i} &= \frac{\partial}{2\partial w_i}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w})-t_n\}^2 + \frac{\lambda\partial}{2\partial w_i}\|\boldsymbol{w}\|^2 \\ &= \sum_{n=1}^{N}(x_n)^i\{y(x_n,\boldsymbol{w})-t_n\} + \lambda w_i \\ &= \sum_{j=0}^{M}\sum_{n=1}^{N}(x_n)^{i+j}w_j + \lambda w_i - \sum_{n=1}^{N}(x_n)^i t_n\end{aligned}$$

$\widetilde{E}(\boldsymbol{w})$ を最小化する $\boldsymbol{w}$ では偏微分係数は 0 なので

$$\sum_{j=0}^{M}\widetilde{A}_{ij}w_j = T_i$$

ただし、

$$\widetilde{A}_{ij} = \sum_{n=1}^{N}(x_n)^{i+j} + I_{ij}\lambda, \quad T_i = \sum_{n=1}^{N}(x_n)^i t_n$$

ここで、I_{ij} は単位行列の要素、すなわち $i=j$ のとき 1、$i\neq j$ のとき 0 である。

1.3 （標準）条件付き確率

りんご・オレンジ・ライムをそれぞれ a, o, l と表す。りんごを選ぶ確率は

$$\begin{aligned}p(a) &= p(r,a)+p(b,a)+p(g,a) \\ &= p(a\,|\,r)p(r)+p(a\,|\,b)p(b)+p(a\,|\,g)p(g)\end{aligned}$$

$$
\begin{aligned}
&= 3/10 \cdot 0.2 + 1/2 \cdot 0.2 + 3/10 \cdot 0.6 \\
&= 0.34
\end{aligned}
$$

選んだ果物がオレンジであったとき、それが緑の箱から取り出された確率は

$$
\begin{aligned}
p(g \,|\, o) &= \frac{p(g, o)}{p(o)} \\
&= \frac{p(o \,|\, g)p(g)}{p(o \,|\, r)p(r) + p(o \,|\, b)p(b) + p(o \,|\, g)p(g)} \\
&= \frac{3/10 \cdot 0.6}{4/10 \cdot 0.2 + 1/2 \cdot 0.2 + 3/10 \cdot 0.6} \\
&= 0.5
\end{aligned}
$$

1.4 （標準）確率密度関数の変数変換

(1.27) の両辺を y で微分すると

$$
p'_y(y) = sp'_x(g(y))\, g'(y)^2 + sp_y(g(y))g''(y)
$$

ただし、

$$
s = \begin{cases} 1 & (g'(y) \geqslant 0) \\ -1 & (g'(y) < 0) \end{cases}
$$

ここで $p_x(x)$ を最大化する $\widehat{x}$ については $p'_x(\widehat{x}) = 0$ であるので、$g(\breve{y}) = \widehat{x}$ なる $\breve{y}$ について

$$
p'_y(\breve{y}) = sp_y(\widehat{x})g''(\breve{y})
$$

であり、$s \neq 0$、$p_y(\widehat{x}) > 0$ であるから、これは $g''(\breve{y}) \neq 0$ であれば 0 にはならない。

$g(y)$ が線形変換であればその二階微分係数はあらゆる点で 0 であるから、$\widehat{x} = g(\widehat{y})$ なる $\widehat{y}$ で最大値をとる。

1.5 （基本）確率変数の分散

$\mathbb{E}[f(x)]$ は確率変数ではなく、定数であることに注意すると

$$
\begin{aligned}
\mathrm{var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \\
&= \mathbb{E}\left[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\right] \\
&= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[2f(x)\mathbb{E}[f(x)]\right] + \mathbb{E}\left[\mathbb{E}[f(x)]^2\right] \\
&= \mathbb{E}\left[f(x)^2\right] - 2\mathbb{E}[f(x)]\mathbb{E}\left[f(x)\right] + \mathbb{E}[f(x)]^2 \\
&= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2
\end{aligned}
$$

1.6 （基本）独立な確率変数の共分散

x, y が離散分布なら

$$
\begin{aligned}
\mathbb{E}_{x,y}[xy] &= \sum_{x,y} p(xy)xy \\
&= \sum_x \sum_y p(x)p(y)xy
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_x \left\{ p(x) x \sum_y p(y) y \right\} \\
&= \left\{ \sum_x p(x) x \right\} \left\{ \sum_y p(y) y \right\} \\
&= \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}
$$

x, y が連続分布なら

$$
\begin{aligned}
\mathbb{E}_{x,y}[xy] &= \iint p(xy) xy \, \mathrm{d}x \, \mathrm{d}y \\
&= \iint p(x) p(y) xy \, \mathrm{d}x \, \mathrm{d}y \\
&= \int p(x) x \left\{ \int p(y) \, \mathrm{d}y \right\} \mathrm{d}x \\
&= \int p(x) \mathrm{d}x \int p(y) \, \mathrm{d}y \\
&= \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}
$$

したがって、いずれの場合でも

$$
\begin{aligned}
\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[x, y] - \mathbb{E}[x]\mathbb{E}[y] \\
&= \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] \\
&= 0
\end{aligned}
$$

1.7 （標準）ガウス分布の正規化条件

以下の変数変換を考える。

$$
\begin{aligned}
x &= r \cos\theta \\
y &= r \sin\theta
\end{aligned}
$$

この変換のヤコビアンは

$$
\begin{aligned}
\left| \frac{\partial(x, y)}{\partial(r, \theta)} \right| &= \begin{vmatrix} \partial x/\partial r & \partial x/\partial \theta \\ \partial y/\partial r & \partial y/\partial \theta \end{vmatrix} \\
&= \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} \\
&= r
\end{aligned}
$$

したがって (1.125) は以下のようになる。

$$
\begin{aligned}
I^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} x^2 - \frac{1}{2\sigma^2} y^2 \right) \mathrm{d}x \, \mathrm{d}y \\
&= \int_{-\pi}^{\pi} \int_{0}^{\infty} \exp\left(-\frac{1}{2\sigma^2} r^2 \right) r \, \mathrm{d}r \, \mathrm{d}\theta \\
&= 2\pi \int_{0}^{\infty} \exp\left(-\frac{1}{2\sigma^2} r^2 \right) r \, \mathrm{d}r
\end{aligned}
$$

ここで $u = \exp(-r^2/2\sigma^2)$ とすると、$du = -r/\sigma^2 \cdot u \, \mathrm{d}r$ であるから

$$
I^2 = 2\pi \int_{1}^{0} (-\sigma^2) \, \mathrm{d}u
$$

$$= 2\pi\sigma^2$$

つまり

$$I = (2\pi\sigma^2)^{1/2}$$

が得られる。次にガウス分布 $\mathcal{N}(x\,|\,\mu,\sigma^2)$ の密度関数の積分を考える。$v = x - \mu$ とすると、

$$\begin{aligned}\int_{-\infty}^{\infty}\mathcal{N}(x\,|\,\mu,\sigma^2)\,\mathrm{d}x &= \int_{-\infty}^{\infty}\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x\\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x\\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}v^2\right)\mathrm{d}v\\ &= \frac{I}{(2\pi\sigma^2)^{1/2}}\\ &= 1\end{aligned}$$

1.8 （標準）ガウス分布の 1 次モーメントと 2 次モーメント

$y = x - \mu$ とおくと、

$$\begin{aligned}\int_{-\infty}^{\infty}\mathcal{N}(x\,|\,\mu,\sigma^2)x\,\mathrm{d}x &= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x\,\mathrm{d}x\\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}y^2\right)(y+\mu)\,\mathrm{d}y\end{aligned}$$

ここで、$\exp(-y^2/2\sigma^2)y$ は奇関数のため、その積分値が 0 になることに注意すると、

$$\begin{aligned}\int_{-\infty}^{\infty}\mathcal{N}(x\,|\,\mu,\sigma^2)x\,\mathrm{d}x &= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}y^2\right)\mu\,\mathrm{d}y\\ &= \mu\int_{-\infty}^{\infty}\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}y^2\right)\mathrm{d}y\\ &= \mu\end{aligned}$$

次に、(1.127) から、

$$\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x = (2\pi\sigma^2)^{1/2}$$

この両辺を σ^2 で微分すると、

$$\int_{-\infty}^{\infty}\frac{1}{2\sigma^4}(x-\mu)^2\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x = \left(\frac{\pi}{2\sigma^2}\right)^{1/2}$$

したがって

$$\frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}(x^2-2\mu x+\mu^2)\,\mathrm{d}x = \sigma^2$$

ここで、$\frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}x\,\mathrm{d}x = \mu$、$\frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\,\mathrm{d}x = 1$ に注意すると、

$$\frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x^2\,\mathrm{d}x - 2\mu^2 + \mu^2 = \sigma^2$$

したがって、

$$\frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x^2\,\mathrm{d}x = \mu^2 + \sigma^2$$

この式は (1.50) と等価である。

次に x の分散は

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = (\mu^2 + \sigma^2) - \mu^2 = \sigma^2$$

1.9 （基本）ガウス分布のモード

ガウス分布 $\mathcal{N}(x \,|\, \mu, \sigma^2)$ の密度関数を微分する。

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}(x \,|\, \mu, \sigma^2) &= \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= -\frac{1}{\sigma^2(2\pi\sigma^2)^{1/2}}(x-\mu)\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\end{aligned}$$

したがって、$x < \mu$ なら微分係数は正、$x > \mu$ なら微分係数は負であることから、$x = \mu$ でモードが与えられることが分かる。

次に多変量ガウス変数の同次密度関数を偏微分する。

$$\frac{\partial}{\partial \mathbf{x}}\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \mathbf{x}}\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1a}$$

$$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\frac{\partial}{\partial \mathbf{x}}\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1b}$$

$$= \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\frac{\partial}{\partial \mathbf{x}}\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1c}$$

$$= -\frac{1}{2}\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\left\{\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}\right\}(\mathbf{x}-\boldsymbol{\mu}) \tag{1d}$$

$$= -\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{1e}$$

したがって $\frac{\partial}{\partial \mathbf{x}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ の要素がすべて 0 となるのは $\mathbf{x} = \boldsymbol{\mu}$ のときで、このときにモードが与えられる。なお、(1c) から (1d) への計算では $\frac{\partial \mathbf{x}^{\mathrm{T}}A\mathbf{x}}{\partial \mathbf{x}} = (A + A^{\mathrm{T}})\mathbf{x}$ を利用し、(1d) から (1e) への計算では共分散行列は $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathrm{T}}$ が成り立つことを利用している。

1.10 （基本）独立な確率変数の和の平均・分散

x, z が連続分布なら

$$\begin{aligned}\mathbb{E}[x+z] &= \iint p(x)p(z)(x+z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \iint p(x)p(z)x\,\mathrm{d}x\,\mathrm{d}z + \iint p(x)p(z)z\,\mathrm{d}x\,\mathrm{d}z \\ &= \int p(z)\int p(x)x\,\mathrm{d}x\,\mathrm{d}z + \int p(x)\int p(z)z\,\mathrm{d}z\,\mathrm{d}x \\ &= \int p(z)\mathbb{E}[x]\,\mathrm{d}z + \int p(x)\mathbb{E}[z]\,\mathrm{d}x \\ &= \mathbb{E}[x]\int p(z)\,\mathrm{d}z + \mathbb{E}[z]\int p(x)\,\mathrm{d}x \\ &= \mathbb{E}[x] + \mathbb{E}[z]\end{aligned}$$

x, z の両方、もしくは、片方が離散分布であっても $\int(...)\mathrm{d}x$ を $\sum_x(...)$ に置き換えれば同様の変形が可能。

分散についてはこの和の平均の性質と

$$\mathrm{var}[x+z] = \mathbb{E}[(x+z)^2] - \mathbb{E}[x+z]^2$$

$$
\begin{aligned}
&= \mathbb{E}[x^2 + 2xz + z^2] - \mathbb{E}[x + z]^2 \\
&= \mathbb{E}[x^2] + 2\mathbb{E}[xz] + \mathbb{E}[z^2] - (\mathbb{E}[x] + \mathbb{E}[z])^2 \\
&= \mathbb{E}[x^2] - \mathbb{E}[x]^2 + \mathbb{E}[z^2] - \mathbb{E}[z]^2 + 2\mathbb{E}[xz] - 2\mathbb{E}[x]\mathbb{E}[z] \\
&= \mathrm{var}[x] + \mathrm{var}[z] + 2\,\mathrm{cov}[x, z]
\end{aligned}
$$

x, z が独立なとき $\mathrm{cov}[x, z] = 0$ であるから、

$$\mathrm{var}[x + z] = \mathrm{var}[x] + \mathrm{var}[z]$$

1.11 （基本）ガウス分布のパラメータの最尤推定

まず、(1.54) を μ で偏微分すると

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \ln p(\mathbf{x} \,|\, \mu, \sigma^2) &= \frac{\partial}{\partial \mu} \left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \right\} \\
&= \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu) \\
&= \frac{1}{\sigma^2} \left(\sum_{n=1}^{N} x_n - N\mu \right)
\end{aligned}
$$

つまり、μ が以下の値をとるときに偏微分係数が 0 となる。

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

(1.54) を σ^2 で偏微分すると

$$
\begin{aligned}
\frac{\partial}{\partial (\sigma^2)} \ln p(\mathbf{x} \,|\, \mu, \sigma^2) &= \frac{\partial}{\partial (\sigma^2)} \left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \right\} \\
&= \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2\sigma^2} \\
&= \frac{1}{2\sigma^4} \left\{ \sum_{n=1}^{N} (x_n - \mu)^2 - N\sigma^2 \right\}
\end{aligned}
$$

μ の最尤解 μ_{ML} は分かっているので、それを利用すると分散の最尤解は以下のようになる。

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2$$

1.12 （基本）最尤推定値の期待値

$n = m$ のときは $x_n x_m = x_n^2$ であるから (1.50) より

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n^2] = \mu^2 + \sigma^2$$

$n \neq m$ なら x_n と x_m は独立なので (1.49) より

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2$$

これらの式を 1 つにまとめると次式が得られる。

$$\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2$$

つぎに、

$$\begin{aligned}
\mathbb{E}[\mu_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] \\
&= \mu \\
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(x_n - \frac{1}{N}\sum_{m=1}^{N} x_m\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbb{E}\left[x_n^2\right] - \frac{2}{N}\sum_{m=1}^{N}\mathbb{E}\left[x_n x_m\right] + \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{m=1}^{N} x_m\right)^2\right]\right\} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\mu^2 + \sigma^2 - \frac{2}{N}\sum_{m=1}^{N}\mathbb{E}\left[x_n x_m\right] + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbb{E}\left[x_m x_l\right]\right\} \\
&= \mu^2 + \sigma^2 - \frac{2}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}\left[x_n x_m\right] + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbb{E}\left[x_m x_l\right] \\
&= \mu^2 + \sigma^2 - \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}\left[x_n x_m\right] \\
&= \mu^2 + \sigma^2 - \frac{1}{N^2}\left\{N(\mu^2 + \sigma^2) + N(N-1)\mu^2\right\} \\
&= \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned}$$

1.13 （基本）平均が既知のときの分散の最尤推定値の期待値

$\widehat{\sigma}^2 = \sum_n (x_n - \mu)^2/N$ とすると、

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2x_n\mu + \mu^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}\left[x_n^2\right] - 2\mu\mathbb{E}\left[x_n\right] + \mu^2\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{(\mu^2 + \sigma^2) - 2\mu^2 + \mu^2\right\} \\
&= \sigma^2
\end{aligned}$$

1.14 （標準）多項式の 2 次の項の独立パラメータの数

以下のように $w_{ij}^S = (w_{ij}+w_{ji})/2,\ w_{ij}^A = (w_{ij}-w_{ji})/2$ とすると、$w_{ij}^S = w_{ji}^S,\ w_{ij}^A = -w_{ji}^A,\ w_{ij} = w_{ij}^S + w_{ji}^A$ がすべて成り立つ。このとき

$$\begin{aligned}
\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^S + w_{ij}^A)x_ix_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_ix_j + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_ix_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_ix_j + \sum_{i<j} w_{ij}^A x_ix_j + \sum_{i>j} w_{ij}^A x_ix_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_ix_j + \sum_{i<j} w_{ij}^A x_ix_j - \sum_{i>j} w_{ji}^A x_ix_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_ix_j + \sum_{i<j} w_{ij}^A x_ix_j - \sum_{j>i} w_{ij}^A x_jx_i \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_ix_j
\end{aligned}$$

w_{ij} の要素のうち $i<j$ となる要素はその対角要素と同じ値であるから、w_{ij} の要素のうち独立な要素の数は $i \geqslant j$ となる $N(N-1)$ 要素である。

1.15 （難問）多項式の M 次の項の独立パラメータの数

多項式 (1.133) の D^M 個の係数のうち、$i_1 \geqslant i_2 \geqslant \cdots \geqslant i_M$ となる項はすべて独立であり、また、これを満たさない項は置換対称性によりこれを満たす項に従属であることから (1.133) からこの条件を満たすもののみを抜き出せば係数の冗長性を取り除くことができる。そのようにして書き直したものが (1.134) である。

(1.134) から、i 次元の $M-1$ 次の項は以下のように書ける。（i の添字を 2 から始めていることに注意）

$$\sum_{i_2=1}^{i}\sum_{i_3=1}^{i_2}\cdots\sum_{i_M=1}^{i_M-1}\widetilde{w}_{i_2i_3\cdots i_M}x_{i_2}x_{i_3}\cdots x_{i_M}$$

この式は (1.134) の 2 つ目の $\sum$ 以降と一致しているので、D 次元の M 次の独立なパラメータの数 $n(D,M)$ はこの項の数 $n(i,M-1)$ の和で表すことができる。すなわち、

$$n(D,M) = \sum_{i=1}^{D} n(i,M-1)$$

(1.136) の左辺において、$D=1$ とおくと、

$$\sum_{i=1}^{1}\frac{(i+M-2)!}{(i-1)!\,(M-1)!} = \frac{(M-1)!}{0!\,(M-1)!} = 1$$

また、(1.136) の左辺において、$D=1$ とおくと、

$$\frac{(1+M-1)!}{(1-1)!\,M!} = \frac{M!}{0!\,M!} = 1$$

であるから、(1.136) 式は $D=1$ のとき、任意の M に対して成り立つ。仮に任意の M に対して D 次元のときに (1.136) 式が成り立つと仮定すると、

$$\begin{aligned}\sum_{i=1}^{D+1}\frac{(i+M-1)!}{(i-1)!\,(M-1)!} &= \sum_{i=1}^{D}\frac{(i+M-1)!}{(i-1)!\,(M-1)!} + \frac{\{(D+1)+M-2\}!}{\{(D+1))-1\}!\,(M-1)!}\\ &= \frac{(D+M-1)!}{(D-1)!\,M!} + \frac{(D+M-1)!}{D!\,(M-1)!}\\ &= \frac{D\cdot(D+M-1)!+M\cdot(D+M-1)!}{D!\,M!}\\ &= \frac{(D+M)\cdot(D+M-1)!}{D!\,M!}\\ &= \frac{(D+M)!}{D!\,M!}\\ &= \frac{\{(D+1)+M-1\}!}{\{(D+1)-1\}!\,M!}\end{aligned}$$

となり、(1.136) 式は任意の M に対して $(D+1)$ 次元でも成り立つ。したがって数学的帰納法により、任意の M と $D\geqslant 1$ なる D について (1.136) は成り立つ。

演習問題 1.14 の結果から

$$n(D,2)=\frac{D(D+1)}{2}=\frac{(D+1)!}{(D-1)!\,2!}$$

であるから、(1.137) は $M=2$ と任意の $D\geqslant 1$ に対して成り立つ[*1]。ここで $M-1$ 次の項で (1.137) が成り立つ、すなわち、

$$n(D,M-1)=\frac{(D+M-2)!}{(D-1)!\,(M-1)!}$$

が成り立つと仮定すると、(1.135) と (1.136) を使って、

$$\begin{aligned}n(D,M) &= \sum_{i=1}^{D} n(i,M-1)\\ &= \sum_{i=1}^{D}\frac{(i+M-2)!}{(i-1)!\,(M-1)!}\\ &= \frac{(D+M-1)!}{(D-1)!\,M!}\end{aligned}$$

となり、M 次の項でも (1.137) が成り立つ。したがって数学的帰納法により、任意の $M\geqslant 2,\ D\geqslant 1$ について (1.136) は成り立つ。また、0 次の項は定数のみで構成されるので、$n(D,0)=1$、1 次の項は変数の数だけ独立したパラメータが存在することから $n(D,1)=D$ であるから $M\geqslant 0,\ D\geqslant 1$ について (1.137) が成り立つ。

1.16 （難問）M 次の多項式の独立パラメータの数

M_1 次の項のパラメータと M_2 次の項のパラメータは $M_i\neq M_2$ であればすべて独立なパラメータであるので、M 次の多項式の独立なパラメータの数は M 次以下の項のパラメータの数の総和である。すなわち、

$$N(D,M)=\sum_{m=0}^{M} n(D,m)$$

である。

[*1] 問題の指示に従い $M=2$ の場合を利用したが、後述する $M=0$ の場合を利用したほうがスマート。

次に、

$$N(D,0) = n(D,0) = 1 = \frac{(D+0)!}{D!\,0!}$$

と書けるから、(1.139) は $M = 0$ のとき、任意の $D \geqslant 1$ について成り立つ。M 次の多項式について (1.139) が成り立つと仮定すると、

$$\begin{aligned}
N(D,M+1) &= \sum_{m=1}^{M+1} n(D,m) \\
&= n(D,M+1) + \sum_{m=1}^{M} n(D,m) \\
&= \frac{(D+M)!}{(D-1)!\,(M+1)!} + \frac{(D+M)!}{D!\,M!} \\
&= \frac{D\cdot(D+M)! + (M+1)\cdot(D+M)!}{D!\,(M+1)!} \\
&= \frac{(D+M+1)\cdot(D+M)!}{D!\,(M+1)!} \\
&= \frac{(D+M+1)!}{D!\,(M+1)!}
\end{aligned}$$

であるから、$M+1$ 次の多項式についても (1.139) が成り立つことを利用している。したがって数学的帰納法により、任意の $M \geqslant 0,\ D \geqslant 1$ に対して (1.139) は成り立つ。

$D \gg M$ であるなら、

$$\begin{aligned}
N(D,M) &= \frac{(D+M)!}{D!\,M!} \\
&\simeq \frac{(D+M)^{D+M}e^{-(D+M)}}{D^D e^{-D} M!} \\
&= \frac{(D+M)^D (D+M)^M}{D^D e^M M!} \\
&= \frac{(1+\frac{M}{D})^D D^M (1+\frac{M}{D})^M}{e^M M!} \\
&\simeq \frac{(1+M)(1+\frac{M^2}{D})}{e^M M!} D^M \\
&\simeq \frac{1+M}{e^M M!} D^M
\end{aligned}$$

となるので、D^M のオーダーで大きくなる。また、(1.139) は M と D について対称であるから、$M \gg D$ であれば $N(D,M) \simeq \{(1+D)/(e^D D!)\} M^D$ であり、M^D のオーダーで大きくなる。

3 次の多項式について $D = 10$ のとき、

$$N(10,3) = \frac{13!}{10!\,3!} = \frac{13\times 12\times 11}{3\times 2\times 1} = 286$$

3 次の多項式について $D = 100$ のとき、

$$N(100,3) = \frac{103!}{100!\,3!} = \frac{103\times 102\times 101}{3\times 2\times 1} = 176851$$

1.17 （標準）ガンマ関数

ガンマ関数の定義から

$$\Gamma(x+1) = \int_0^\infty u^x e^{-u}\,\mathrm{d}u$$

$$
\begin{aligned}
&= \int_0^\infty u^x(-e^{-u})'\,\mathrm{d}u \\
&= \left[u^x(-e^{-u})\right]_0^\infty - \int_0^\infty xu^{x-1}(-e^{-u})\,\mathrm{d}u \\
&= 0 + x\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u = x\Gamma(x)
\end{aligned}
$$

これで、(1.141) が示せた。

次に、$\Gamma(1)$ をガンマ関数の定義にしたがって計算すると

$$\Gamma(1) = \int_0^\infty e^{-u}\,\mathrm{d}u = \left[-e^{-u}\right]_0^\infty = 1$$

つまり $\Gamma(1) = 0!$ である。ある非負の整数 x について $\Gamma(x+1) = x!$ であると仮定すると、

$$
\begin{aligned}
\Gamma(x+2) &= (x+1)\Gamma(x+1) \\
&= (x+1)\cdot x! \\
&= (x+1)!
\end{aligned}
$$

であるから、数学的帰納法により、任意の非負の整数 x について $\Gamma(x+1) = x!$ である。

1.18 （標準）D 次元単位球の表面積と体積

(1.126) で $\sigma = 1/\sqrt{2}$ とすると

$$\int_{-\infty}^\infty e^{-x^2}\,\mathrm{d}x = \pi^{1/2}$$

これを利用すると (1.142) の左辺は

$$\prod_{i=1}^D \int_{-\infty}^\infty e^{-x_i^2}\,\mathrm{d}x_i = \pi^{D/2}$$

一方 (1.142) の右辺は $s = r^2$ とおくと $\mathrm{d}s/\mathrm{d}r = 2r$ であるから

$$
\begin{aligned}
S_D\int_0^\infty e^{-r^2}r^{D-1}\,\mathrm{d}r &= S_D\int_0^\infty e^{-r^2}r^{2\left(\frac{D}{2}-1\right)}r\,\mathrm{d}r \\
&= S_D\int_0^\infty e^{-s}s^{\frac{D}{2}-1}\,\mathrm{d}s \\
&= S_D\Gamma(D/2)
\end{aligned}
$$

したがって、(1.142) は以下のように書ける。

$$\pi^{D/2} = S_D\Gamma(D/2)$$

つまり、

$$S_D = \frac{\pi^{D/2}}{\Gamma(D/2)}$$

D 次元の球の表面積は半径の $D-1$ 乗に比例することから、D 次元の単位球の体積は

$$V_D = \int_0^1 S_D r^{D-1}\,\mathrm{d}r = \left[\frac{S_D r^D}{D}\right]_{r=0}^{r=1} = \frac{S_D}{D}$$

$D = 2$（単位円）であれば、S_2（円周の長さ）と V_2（面積）は

$$S_2 = \frac{2\pi}{\Gamma(1)} = 2\pi$$

$$V_2 = \frac{S_2}{2} = \frac{2\pi}{2} = \pi$$

$D = 3$（単位球）であれば、S_3（表面積）と V_3（体積）は

$$S_3 = \frac{2\pi^{3/2}}{\Gamma(3/2)} = \frac{2\pi^{3/2}}{\sqrt{\pi}/2} = 4\pi^2$$
$$V_3 = \frac{S_3}{3} = \frac{4}{3}\pi^2$$

1.19 （標準）D 次元単位球と超立方体の体積比

演習問題 1.18 から D 次元の半径 a の球の体積は

$$V_D = \frac{S_D}{D}a^D = \frac{2\pi^{D/2}a^D}{D\,\Gamma(D/2)}$$

一方、超立方体の体積は $(2a)^D$ であるから、その比は

$$\frac{\text{球の体積}}{\text{立方体の体積}} = \frac{2\pi^{D/2}a^D}{D\,\Gamma(D/2)(2a)^D} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}$$

ここで (1.146) を使うと $D \gg 1$ のときに、

$$\begin{aligned}\frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} &\simeq \frac{\pi^{D/2}}{D2^{D-1}(2\pi)^{1/2}e^{-(D/2-1)}(D/2-1)^{D/2-1/2}} \\ &= \frac{\pi^{D/2}2e^{D/2}(D/2-1)^{1/2}}{D2^D(2\pi)^{1/2}e(D/2-1)^{D/2}} \\ &= \frac{(\pi e)^{D/2}(2D-4)^{1/2}}{(2\pi)^{1/2}eD(2D-4)^{D/2}} \\ &= \frac{1}{(2\pi)^{1/2}e}\left(\frac{\pi e}{2D-4}\right)^{D/2}\left(\frac{2-4/D}{D}\right)^{1/2}\end{aligned}$$

したがって、$D \to \infty$ のとき (1.145) は 0 に収束する。

超立方体の頂点の座標はすべての要素が $\pm a$ であることから、中心からの距離は $\sqrt{Da^2}$ である。中心から側面までの距離は a であることから、その比は $\sqrt{D}$ であり、$D \to \infty$ のとき、∞ に発散する。

1.20 （標準）D 次元ガウス分布の密度分布

半径 r の単位球の表面積は $S_D r^{D-1}$ であるから、半径 r にある暑さ ϵ の薄皮の体積は $S_D r^{D-1}\epsilon$ である。また、$p(\mathbf{x})$ はこの薄皮上で一様の確率密度 $p(\|\mathbf{x}\| = r)$ をもつことから、$p(\mathbf{x})$ をこの薄皮 shell 部分で積分すると

$$\begin{aligned}\int_{\text{shell}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} &\simeq p(\|\mathbf{x}\| = r)S_D r^{D-1}\epsilon \\ &= \frac{1}{(2\pi\sigma)^{D/2}}\exp\left(-\frac{r^2}{2\sigma^2}\right)S_D r^{D-1}\epsilon \\ &= \frac{S_D r^{D-1}}{(2\pi\sigma)^{D/2}}\exp\left(-\frac{r^2}{2\sigma^2}\right)\epsilon\end{aligned}$$

ここで、(1.148) を使うと $\int_{\text{shell}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \simeq p(r)\epsilon$ が示される。

次に $p(r)$ を r で微分すると

$$\frac{\mathrm{d}}{\mathrm{d}r}p(r) = \frac{S_D}{(2\pi\sigma^2)^{D/2}}\left\{(D-1)r^{D-2} - r^{D-1}\frac{r}{\sigma^2}\right\}\exp\left(-\frac{r^2}{2\sigma^2}\right)$$

$$= \frac{S_D}{(2\pi\sigma^2)^{D/2}} \frac{\sigma^2(D-1) - r^2}{\sigma^2} \; r^{D-2} \exp\left(-\frac{r^2}{2\sigma^2}\right)$$

この式を 0 にする $r > 0$ なる r は $\widehat{r} = \sqrt{D-1}\sigma$ でだけあり、$D \gg 1$ であるなら、$\widehat{r} \simeq \sqrt{D}\sigma$ である。

$\epsilon \ll \widehat{r}$ であり、D が十分に大きいとき

$$p(\widehat{r}+\epsilon) = \frac{S_D}{(2\pi\sigma^2)^{D/2}} (\widehat{r}+\epsilon)^{D-1} \exp\left\{-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2}\right\} \tag{2a}$$

$$= \frac{S_D}{(2\pi\sigma^2)^{D/2}} \; \widehat{r}^{D-1} \left(1+\frac{\epsilon}{\widehat{r}}\right)^{D-1} \exp\left(-\frac{\widehat{r}^2}{2\sigma^2}\right) \exp\left\{-\frac{2\widehat{r}\epsilon+\epsilon^2}{2\sigma^2}\right\} \tag{2b}$$

$$= \frac{S_D \widehat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\widehat{r}^2}{2\sigma^2}\right) \cdot \left(1+\frac{\epsilon}{\widehat{r}}\right)^{D-1} \exp\left\{-\frac{2\widehat{r}\epsilon+\epsilon^2}{2\sigma^2}\right\} \tag{2c}$$

$$= p(\widehat{r}) \exp\left\{(D-1)\ln\left(1+\frac{\epsilon}{\widehat{r}}\right) - \frac{2\widehat{r}\epsilon+\epsilon^2}{2\sigma^2}\right\} \tag{2d}$$

$$\simeq p(\widehat{r}) \exp\left\{D\ln\left(1+\frac{\epsilon}{\widehat{r}}\right) - \frac{2\widehat{r}\epsilon+\epsilon^2}{2\sigma^2}\right\} \tag{2e}$$

$$\simeq p(\widehat{r}) \exp\left\{D\left(\frac{\epsilon}{\widehat{r}} - \frac{\epsilon^2}{2\widehat{r}^2}\right) - \frac{2\widehat{r}\epsilon+\epsilon^2}{2\sigma^2}\right\} \tag{2f}$$

$$\simeq p(\widehat{r}) \exp\left\{D\left(\frac{\epsilon}{\sqrt{D}\sigma} - \frac{\epsilon^2}{2D\sigma^2}\right) - \frac{2\sqrt{D}\sigma\epsilon+\epsilon^2}{2\sigma^2}\right\} \tag{2g}$$

$$= p(\widehat{r}) \exp\left\{\frac{2\sqrt{D}\sigma\epsilon-\epsilon^2}{2\sigma^2} - \frac{2\sqrt{D}\sigma\epsilon+\epsilon^2}{2\sigma^2}\right\} \tag{2h}$$

$$= p(\widehat{r}) \exp\left(-\frac{\epsilon^2}{\sigma^2}\right) \tag{2i}$$

ここで、(2e) から (2f) への近似には $x \ll 1$ のとき $\ln(1+x) \simeq x - \frac{x^2}{2}$ であることを、(2f) から (2g) への近似には $\widehat{r} \simeq \sqrt{D}\sigma$ を使った。

最後に原点での確率密度は

$$p(\mathbf{x}=\mathbf{0}) = \frac{1}{(2\pi\sigma^2)^{D/2}}$$

半径 $\widehat{r}$ の球面上の確率密度は

$$\begin{aligned} p(\|\mathbf{x}\| = \widehat{r}) &= \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\widehat{r}^2}{2\sigma^2}\right) \\ &\simeq \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{(\sqrt{D}\sigma)^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{D}{2}\right) \end{aligned}$$

となり、$p(\|\mathbf{x}\| = \widehat{r})$ は $p(\mathbf{x}=\mathbf{0})$ の $\exp(-D/2)$ 倍大きい。

1.21 （標準）誤識別率を最小化する 2 クラス分類問題の解

$a > 0$ で $a \leqslant b$ であるのなら $a^2 \leqslant ab$。両辺の平方根を取ると $a \leqslant (ab)^{1/2}$ となる。

このことから以下の式が成り立つことがわかる。

$$\begin{cases} p(\mathbf{x},\mathcal{C}_1) \leqslant \{p(\mathbf{x},\mathcal{C}_1)p(\mathbf{x},\mathcal{C}_2)\}^{1/2} & (p(\mathbf{x},\mathcal{C}_1) \leqslant p(\mathbf{x},\mathcal{C}_2) \text{ のとき}) \\ p(\mathbf{x},\mathcal{C}_2) \leqslant \{p(\mathbf{x},\mathcal{C}_1)p(\mathbf{x},\mathcal{C}_2)\}^{1/2} & (p(\mathbf{x},\mathcal{C}_2) \leqslant p(\mathbf{x},\mathcal{C}_1) \text{ のとき}) \end{cases}$$

ここで誤識別率を最小にする決定領域 $\mathcal{C}_1, \mathcal{C}_2$ では真の分類 $\mathcal{R}_1, \mathcal{R}_2$ について、$\mathbf{x} \in \mathcal{R}_1$ のとき $p(\mathbf{x}, \mathcal{C}_2) \leqslant p(\mathbf{x}, \mathcal{C}_1)$、$\mathbf{x} \in \mathcal{R}_2$ のとき $p(\mathbf{x}, \mathcal{C}_1) \leqslant p(\mathbf{x}, \mathcal{C}_2)$ であることに注意すると、

$$\begin{aligned} p(\text{誤り}) &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x} \\ &\leqslant \int_{\mathcal{R}_1} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}\,\mathrm{d}\mathbf{x} \\ &= \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}\,\mathrm{d}\mathbf{x} \end{aligned}$$

1.22 （基本）損失行列 $1 - I_{kj}$ のクラス分類問題

式 (1.81) に $L_{kj} = 1 - I_{kj}$ を適用すると、

$$\begin{aligned} \sum_k L_{kj} p(\mathcal{C}_k \,|\, \mathbf{x}) &= \sum_k (1 - I_{kj}) p(\mathcal{C}_k \,|\, \mathbf{x}) \\ &= \sum_{k \neq j} p(\mathcal{C}_k \,|\, \mathbf{x}) \\ &= 1 - p(\mathcal{C}_j \,|\, \mathbf{x}) \end{aligned}$$

となり、単に事後確率を最大化するクラスを求める問題に帰着することがわかる。この損失行列は正しい分類にはペナルティを与えず、誤りに対して等しいペナルティを与えるものであり、誤りの種類を問わずに誤識別率を最小化する損失関数と解釈できる。

1.23 （標準）損失行列とクラスに対する事前確率が与えられたときの期待損失

期待損失を最小にするには、各 $\mathbf{x}$ ごとに $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$ を最小化にする j を選べばよいが、これは

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) = \sum_k L_{kj} p(\mathbf{x} \,|\, \mathcal{C}_k) p(\mathcal{C}_k)$$

と書けるから、損失行列と事前確率から求められる $L^*_{kj} = L_{kj} p(\mathcal{C}_k)$ を新たな損失行列と考え、$L^*_{kj} p(\mathbf{x} \,|\, \mathcal{C}_k)$ が最小になるクラス j に割り当てればよい。

1.24 （標準）棄却オプションがあるときの決定基準

入力ベクトル $\mathbf{x}$ が与えられたとき、クラス j に割り当てた際の損失は $\sum_k L_{kj} p(\mathcal{C}_k \,|\, \mathbf{x})$ であるから、これが λ を下回ればクラス j に割り当てたほうが棄却しないより損失は小さくなり、逆にすべての j についてこれが λ を上回れば棄却したほうが損失は小さくなる。したがって損失を最小化する決定基準は以下の通りである。

$$\begin{cases} \sum_k L_{kj} p(\mathcal{C}_k \,|\, \mathbf{x}) \text{ が最小になるクラス } j \text{ を採用する} & \left(\min_l \sum_k L_{kl} p(\mathcal{C}_k \,|\, \mathbf{x}) < \lambda \text{ のとき}\right) \\ \text{棄却する} & \left(\min_l \sum_k L_{kl} p(\mathcal{C}_k \,|\, \mathbf{x}) \geqslant \lambda \text{ のとき}\right) \end{cases}$$

損失行列が $1 - I_{kj}$ であれば、演習問題 1.22 からこれは次のように書ける。

$$\begin{cases} 1 - p(\mathcal{C}_j \,|\, \mathbf{x}) \text{ が最小になるクラス } j \text{ を採用する} & \left(\min_l \{1 - p(\mathcal{C}_l \,|\, \mathbf{x})\} < \lambda \text{ のとき}\right) \\ \text{棄却する} & \left(\max_l \{1 - p(\mathcal{C}_l \,|\, \mathbf{x})\} \geqslant \lambda \text{ のとき}\right) \end{cases}$$

この決定基準は以下の決定基準と等価である。

$$\begin{cases} p(\mathcal{C}_j \,|\, \mathbf{x}) \text{ が最大になるクラス } j \text{ を採用する} & (p(\mathcal{C}_l \,|\, \mathbf{x}) > 1 - \lambda \text{ となる } l \text{ があるとき}) \\ \text{棄却する} & (\text{すべての } l \text{ について } p(\mathcal{C}_l \,|\, \mathbf{x}) \leqslant 1 - \lambda \text{ のとき}) \end{cases}$$

これは棄却しきい値を用いた表現そのものであり、棄却しきい値 θ は棄却時の損失 λ を用いて $\theta = 1 - \lambda$ と書ける。

1.25 （基本）目標変数がベクトルのときの期待損失

(1.151) の変分を計算すると

$$\frac{\delta \mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))]}{\delta \mathbf{y}(\mathbf{x})} = 2 \int \{\mathbf{y}(\mathbf{x}) - \mathbf{t}\} \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t}$$

期待損失を最小化する $\mathbf{y}(\mathbf{x})$ ではこの変分が $\mathbf{x}$ によらず零ベクトルになるので、

$$\int \{\mathbf{y}(\mathbf{x}) - \mathbf{t}\} \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} = \mathbf{0}$$

つまり

$$\int \mathbf{y}(\mathbf{x}) \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} = \int \mathbf{t} \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t}$$

ここで、$\int p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} = p(\mathbf{x})$ であることと、$p(\mathbf{x}, \mathbf{t}) = p(\mathbf{t} \,|\, \mathbf{x}) \, p(\mathbf{x})$ であることに注意すると、この式は

$$\mathbf{y}(\mathbf{x}) \, p(\mathbf{x}) = \int \mathbf{t} \, p(\mathbf{t} \,|\, \mathbf{x}) \, p(\mathbf{x}) \, \mathrm{d}\mathbf{t}$$

と書ける。したがって、

$$\mathbf{y}(\mathbf{x}) = \int \mathbf{t} \, p(\mathbf{t} \,|\, \mathbf{x}) \, \mathrm{d}\mathbf{t} = \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]$$

単一の目標変数のときは $\mathbf{t}$, $\mathbf{y}(\mathbf{x})$ はそれぞれ 1 次元のベクトル、すなわちスカラー t, $y(\mathbf{x})$ と考えれば良いので、この式は (1.89) に帰着されることがわかる。

1.26 （基本）目標変数がベクトルのときの期待損失 2

損失関数は次のように書くことができる。

$$\begin{aligned} \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] + \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}\|^2 && \text{(3a)} \\ &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\|^2 + 2\{\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\}^{\mathrm{T}} \{\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}\} + \|\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}\|^2 && \text{(3b)} \end{aligned}$$

この式の中間の項にでてくる $\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}$ に $\mathbf{x}$, $\mathbf{t}$ の同時確率をかけて $\mathbf{t}$ で積分した値が

$$\begin{aligned} \int (\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}) \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} &= \int \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} - \int \mathbf{t} \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} \\ &= \int \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} - \int \mathbf{t} \, p(\mathbf{t} \,|\, \mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{t} \\ &= \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] \int p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} - p(\mathbf{x}) \int \mathbf{t} \, p(\mathbf{t} \,|\, \mathbf{x}) \, \mathrm{d}\mathbf{t} \quad (\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] \text{ は } \mathbf{t} \text{ の関数ではない}) \\ &= \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] \, p(\mathbf{x}) - p(\mathbf{x}) \, \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] = \mathbf{0} \end{aligned}$$

となることに注意すると、(3b) の中間の項に $\mathbf{x}$, $\mathbf{t}$ の同時確率をかけて $\mathbf{x}$, $\mathbf{t}$ で積分した値は

$$\iint 2\{\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\}^{\mathrm{T}} \{\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}\} \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} \, \mathrm{d}\mathbf{x}$$

$$= 2 \int \{\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\}^{\mathrm{T}} \left\{ \int (\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} \right\} \mathrm{d}\mathbf{x} = 0$$

(3b) の最初の項と最後の項も同様に $\mathbf{x}$, $\mathbf{t}$ の同時確率をかけて $\mathbf{x}$, $\mathbf{t}$ で積分すると

$$\begin{aligned} \iint \|\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\|^2 p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} \, \mathrm{d}\mathbf{x} &= \int \|\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\|^2 \left\{ \int p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} \right\} \mathrm{d}\mathbf{x} \\ &= \int \|\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\|^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \end{aligned}$$

$$\begin{aligned} \iint \|\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} \, \mathrm{d}\mathbf{x} &= \iint \|\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}\|^2 p(\mathbf{t} \,|\, \mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{t} \, \mathrm{d}\mathbf{x} \\ &= \int \left\{ \int \|\mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}] - \mathbf{t}\|^2 p(\mathbf{t} \,|\, \mathbf{x}) \, \mathrm{d}\mathbf{t} \right\} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \int \mathrm{var}[\mathbf{t} \,|\, \mathbf{x}] \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \end{aligned}$$

したがって損失関数の期待値は以下のようになる。

$$\begin{aligned} \mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))] &= \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 \, p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{t} \, \mathrm{d}\mathbf{x} \\ &= \int \|\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t} \,|\, \mathbf{x}]\|^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \mathrm{var}[\mathbf{t} \,|\, \mathbf{x}] \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \end{aligned}$$

この式で $\mathbf{y}(\mathbf{x})$ を含むのは最初項だけで、損失関数の期待値を最小化する $\mathbf{y}(\mathbf{x})$ は $\mathbf{t}$ の条件付き期待値であることがわかる。

1.27 （標準）ミンコフスキー損失

$\mathbf{x}$ の値ごとに $\mathbb{E}[L_q]$ を最小にする $y(\mathbf{x})$ を探す問題なので、期待損失 (1.91) の最小化は条件付き確率を考えた以下の式の最小化と等価である。

$$\int |y(\mathbf{x}) - t|^q p(t \,|\, \mathbf{x}) \, \mathrm{d}t \tag{4}$$

(4) の変分は $y(\mathbf{x}) - t \geqslant 0$ となる t の範囲では $\int q \, |y(\mathbf{x}) - t|^{q-1} p(t \,|\, \mathbf{x}) \, \mathrm{d}t$、$y(\mathbf{x}) - t < 0$ となる t の範囲では $-\int q \, |y(\mathbf{x}) - t|^{q-1} p(t \,|\, \mathbf{x}) \, \mathrm{d}t$ であるから、(4) の変分は次のように書ける。

$$q \int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1} p(t \,|\, \mathbf{x}) \, \mathrm{d}t - q \int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1} p(t \,|\, \mathbf{x}) \, \mathrm{d}t$$

期待損失を最小化する $y(\mathbf{x})$ ではこの変分が $\mathbf{x}$ によらず 0 になるので、

$$\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1} p(t \,|\, \mathbf{x}) \, \mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1} p(t \,|\, \mathbf{x}) \, \mathrm{d}t$$

これが、期待損失を最小化する $y(\mathbf{x})$ の満たす条件である。

$q = 1$ であれば、

$$\int_{-\infty}^{y(\mathbf{x})} p(t \,|\, \mathbf{x}) \, \mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} p(t \,|\, \mathbf{x}) \, \mathrm{d}t$$

つまり、$t < y(\mathbf{x})$ となる条件付き確率と $t > y(\mathbf{x})$ となる条件付き確率が等しい、すなわち、期待損失を最小化する $y(\mathbf{x})$ が条件付きメディアンになることを示している。

次に、$q \to 0$ の場合を考える。(4) の $|y(\mathbf{x}) - t|^q$ は $y(\mathbf{x}) = t$ となる場所の近傍以外では至るところでほぼ 1 であり、$y(\mathbf{x}) = t$ の近傍でのみ 0 に近い値をとる。$p(t|\mathbf{x})$ は正規化されているため、(4) の値はほぼ 1 になるが、この $y(\mathbf{x}) = t$ の近傍での $|y(\mathbf{x}) - t|^q$ の落ち込みの分だけ、この値は 1 よりわずかに小さくなる。この落ち込み幅を多くするには $p(t \,|\, \mathbf{x})$ が大きい場所で $y(\mathbf{x}) = t$ となるような $y(\mathbf{x})$ を選べばよい。言い換えると $y(\mathbf{x})$ を条件付きモードにすることにより、期待損失を最小化できる。

1.28 （基本）情報量と確率の関係

$h(x)$ は確率 $p(x)$ の単調な情報量を表す関数であるので、ここでは

$$h(x) = f(p(x)) \tag{5}$$

と書くことにする。独立な変数 x, y について、$h(x, y)$ を (5) の表記を使うと以下のように表せる。

$$h(x, y) = f(p(x, y)) = f(p(x)p(y)) \tag{6}$$

一方、$h(x) + h(y)$ は (5) の表記を使うと

$$h(x) + h(y) = f(p(x)) + f(p(y)) \tag{7}$$

情報量が加法的であるなら (6) と (7) は等しいので、

$$f(p(x)p(y)) = f(p(x)) + f(p(y))$$

これは、独立事象の確率の積の情報量は、各事象の確率の情報量の和になるという性質を情報量 $h(\cdot)$ が持つことを示している。したがって、

$$h(p^2) = h(p) + h(p) = 2h(p)$$

が成り立つ。

次に $n = 1$ のとき $h(p^n) = nh(p)$ は明らかに成り立つ。ある正の整数 k に対して $h(p^k) = kh(p)$ が成り立つと仮定すると、

$$h(p^{k+1}) = h(p^k\, p) = h(p^k) + h(p) = kh(p) + h(p) = (k + 1)h(p) = (n/m)mh(p^{1/m})$$

であるから、任意の正の整数 n について $h(p^n) = nh(p)$ が成り立つ。

また、正の整数 m, n に対して、

$$h(p^{n/m}) = nh(p^{1/m}) = (n/m)mh(p^{1/m}) = (n/m)h(p^{m/m}) = (n/m)h(p)$$

である。

最後に正の実数 x に対して $h(p^x) = xh(p)$ が成り立つならば、確率 $0 < p < 1$ において $-\ln p$ は正の実数なので、

$$h(p) = h(e^{(-1)\cdot(-\ln p)}) = -h(e^{-1}) \ln p$$

つまり、$h(p)$ は $\ln p$ に定数 $-h(e^{-1})$ をかけた値に等しい。すなわち $h(p) \propto \ln p$ である。

1.29 （基本）離散分布のエントロピー

関数 $-\ln x$ は凸関数なので、イェンセンの不等式 (1.115) から

$$-\ln\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leqslant -\sum_{i=1}^{M} \lambda_i \ln x_i \tag{8}$$

ここで λ_i は $\lambda_i \geqslant 0, \sum_{i=1}^{M} \lambda_i = 1$ であるから、離散確率変数 x の各事象の発生確率 p_i とみなすことができる。また、$x_i = 1/p_i$ とすると (8) の左辺は

$$-\ln\left(\sum_{i=1}^{M} p_i \frac{1}{p_i}\right) = -\ln\left(\sum_{i=1}^{M} 1\right) = -\ln M$$

一方右辺は

$$-\sum_{i=1}^{M} p_i \ln\left(\frac{1}{p_i}\right) = -\left(-\ln \sum_{i=1}^{M} p_i \ln p_i\right) = -H[x]$$

したがって、(8) から $-\ln M \leqslant -H[x]$ 、すなわち $H[x] \leqslant \ln M$ が成り立つことがわかる。

1.30 （標準）KL ダイバージェンス

まず、$q(x)$ と $p(x)$ の比の対数を計算する。

$$\begin{aligned}
\ln\left\{\frac{q(x)}{p(x)}\right\} &= \ln\left[\frac{\frac{1}{(2\pi s^2)^{1/2}}\exp\left\{-\frac{1}{2s^2}(x-m)^2\right\}}{\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}\right] \\
&= \ln\left[\frac{\sigma}{s}\exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2s^2}(x-m)^2\right\}\right] \\
&= \ln\left(\frac{\sigma}{s}\right) + \frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2s^2}(x-m)^2 \\
&= \left(\frac{1}{2\sigma^2} - \frac{1}{2s^2}\right)x^2 - \left(\frac{\mu}{\sigma^2} - \frac{m}{s^2}\right)x + \left(\frac{\mu^2}{2\sigma^2} - \frac{m^2}{2s^2}\right) + \ln\left(\frac{\sigma}{s}\right)
\end{aligned}$$

カルバック-ライブラーダイバージェンスを (1.113) の定義に従って計算すると

$$\mathrm{KL}(p\|q) = -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}\mathrm{d}x \tag{9a}$$

$$= -\int \mathcal{N}(x|\mu,\sigma^2)\left\{\left(\frac{1}{2\sigma^2} - \frac{1}{2s^2}\right)x^2 - \left(\frac{\mu}{\sigma^2} - \frac{m}{s^2}\right)x + \left(\frac{\mu^2}{2\sigma^2} - \frac{m^2}{2s^2}\right) + \ln\left(\frac{\sigma}{s}\right)\right\}\mathrm{d}x \tag{9b}$$

$$= -\left\{\left(\frac{1}{2\sigma^2} - \frac{1}{2s^2}\right)(\mu^2+\sigma^2) - \left(\frac{\mu}{\sigma^2} - \frac{m}{s^2}\right)\mu + \left(\frac{\mu^2}{2\sigma^2} - \frac{m^2}{2s^2}\right) + \ln\left(\frac{\sigma}{s}\right)\right\} \tag{9c}$$

$$= \frac{\sigma^2 - s^2 + (m-\mu)^2}{2s^2} - \ln\left(\frac{\sigma}{s}\right) \tag{9d}$$

なお、(9c) に至る計算においてはガウス関数のモーメントの性質 (1.48)、(1.49)、(1.50) を利用した。

1.31 （標準）同時分布の微分エントロピー

(1.152) の右辺と左辺の差を計算すると

$$\begin{aligned}
\mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{x},\mathbf{y}] &= \mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}\,|\,\mathbf{x}] - \mathrm{H}[\mathbf{x}] \\
&= \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}\,|\,\mathbf{x}] \\
&= \mathrm{I}[\mathbf{x},\mathbf{y}]
\end{aligned}$$

ここで、相互情報量 $\mathrm{I}[\mathbf{x},\mathbf{y}]$ は非負であるので $\mathrm{H}[\mathbf{x},\mathbf{y}] \leqslant \mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}]$ が成り立つ。相互情報量が 0 となるのは $\mathbf{x}$ と $\mathbf{y}$ が独立なときであり、また独立なときに限って 0 となることから、(1.152) の等号が成り立つのは $\mathbf{x}$ と $\mathbf{y}$ が統計的に独立なとき、またそのときに限る。

1.32 （基本）線形変換された確率変数のエントロピー

連続変数のエントロピーの定義 (1.104) から

$$\begin{aligned}
\mathrm{H}[\mathbf{y}] &= -\int p(\mathbf{y})\ln p(\mathbf{y})\,\mathrm{d}\mathbf{y} \\
&= -\int \frac{p(\mathbf{x})}{|\det(\mathbf{A})|}\ln\left\{\frac{p(\mathbf{x})}{|\det(\mathbf{A})|}\right\}|\det(\mathbf{A})|\,\mathrm{d}\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
&= -\int p(\mathbf{x}) \left\{ \ln p(\mathbf{x}) - \ln |\det(\mathbf{A})| \right\} \mathrm{d}\mathbf{x} \\
&= -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \ln |\det(\mathbf{A})| \int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&= \mathrm{H}[\mathbf{x}] + \ln |\det(\mathbf{A})|
\end{aligned}
$$

1.33 （標準）離散確率変数の条件付きエントロピー

離散確率変数の条件付きエントロピーは

$$
\begin{aligned}
\mathrm{H}[y \,|\, x] &= -\sum_x \sum_y p(x, y) \ln p(y \,|\, x) \\
&= -\sum_x \sum_y p(x) p(y \,|\, x) \ln p(y \,|\, x) \\
&= -\sum_x p(x) \left\{ \sum_y p(y \,|\, x) \ln p(y \,|\, x) \right\} \qquad (10)
\end{aligned}
$$

ここで、$p(y \,|\, x)$ は 0 以上、1 以下であることから、任意の x, y について $p(y \,|\, x) \ln p(y \,|\, x) \leqslant 0$ である。しかし、$p(y \,|\, x) \ln p(y \,|\, x) < 0$ なる x, y があったとすると (10) は正になることから、$\mathrm{H}[y \,|\, x] = 0$ であるなら、任意の x, y について $p(y \,|\, x) \ln p(y \,|\, x) = 0$ となるはずである。ここで、$p(y_j \,|\, x_i) \neq 0$ となる x_i, y_j の組み合わせが存在するならそのときは $\ln p(y_j \,|\, x_i) = 0$ すなわち $p(y_j \,|\, x_i) = 1$ である。しかし、任意の x_i に対し $\sum_j p(y_j \,|\, x_i) = 1$ であるから、各 x_i に対し $\ln p(y_j \,|\, x_i) \neq 0$ となる y_j は一つだけである。

1.34 （標準）エントロピーを最大化する連続確率分布

(1.108) の上にある式は

$$
\int_{-\infty}^{\infty} \left\{ -p(x) \ln p(x) + \lambda_1 p(x) + \lambda_2 x \, p(x) + \lambda_3 (x - \mu)^2 p(x) \right\} \mathrm{d}x - \lambda_1 - \lambda_2 \mu - \lambda_3 \sigma^2
$$

と書けるので、オイラー-ラグランジュ方程式 (D.8) は

$$
-\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 = 0
$$

これを $p(x)$ について解くと以下の (1.108) 式が得られる。

$$
p(x) = \exp \left\{ -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 \right\}
$$

仮に $x \to \pm\infty$ のとき $p(x)$ が発散、もしくは 0 以外の値に収束したとすると、任意の x に対して $p(x) \geqslant 0$ であるから $\int_{-\infty}^{\infty} p(x) \, \mathrm{d}x$ は発散し、(1.105) と矛盾する。したがって、$x \to \pm\infty$ のとき $p(x) \to 0$ であり、$\lambda_3 < 0$ であることがわかる。ここで (1.106) の左辺は

$$
\begin{aligned}
\int_{-\infty}^{\infty} x \, p(x) \, \mathrm{d}x &= \frac{1}{2\lambda_3} \int_{-\infty}^{\infty} \{\lambda_2 + 2\lambda_3 (x - \mu)\} p(x) \, \mathrm{d}x + \left(\mu - \frac{\lambda_2}{2\lambda_3} \right) \int_{-\infty}^{\infty} p(x) \, \mathrm{d}x \qquad (11a) \\
&= \frac{1}{2\lambda_3} \left[p(x) \right]_{-\infty}^{\infty} + \left(\mu - \frac{\lambda_2}{2\lambda_3} \right) \qquad (11b) \\
&= \mu - \frac{\lambda_2}{2\lambda_3} \qquad (11c)
\end{aligned}
$$

これが μ に等しくなるため、$\lambda_2 = 0$ である。なお、(11b) に至る計算では (1.105) を利用した。次に (1.107) の左辺に代入すると

$$
\int_{-\infty}^{\infty} (x - \mu)^2 \, p(x) \, \mathrm{d}x = \frac{1}{2\lambda_3} \int_{-\infty}^{\infty} (x - \mu) \{2\lambda_3 (x - \mu)\} p(x) \, \mathrm{d}x \qquad (12a)
$$

$$= \frac{1}{2\lambda_3}\int_{-\infty}^{\infty}(x-\mu)p'(x)\,\mathrm{d}x \tag{12b}$$

$$= \frac{1}{2\lambda_3}\left[(x-\mu)p(x)\right]_{-\infty}^{\infty} - \frac{1}{2\lambda_3}\int_{-\infty}^{\infty}p(x)\,\mathrm{d}x \tag{12c}$$

$$= \frac{1}{2\lambda_3}\left[(x-\mu)\exp\left\{-1+\lambda_1+\lambda_3(x-\mu)^2\right\}\right]_{-\infty}^{\infty} - \frac{1}{2\lambda_3} \tag{12d}$$

$$= -\frac{1}{2\lambda_3} \tag{12e}$$

これが σ^2 に等しくなるため、$\lambda_3 = -1/(2\sigma^2)$ である。ここでも、(12d) に至る計算では (1.105) を利用した。これらの結果を (1.105) の左辺に代入すると

$$\int_{-\infty}^{\infty}p(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\exp\left\{-1+\lambda_1-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x \tag{13a}$$

$$= \exp(-1+\lambda_1)\int_{-\infty}^{\infty}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x \tag{13b}$$

$$= \exp(-1+\lambda_1)\cdot(2\pi\sigma^2)^{1/2} \tag{13c}$$

これが 1 に等しくなるため、$\lambda_1 = 1 + \frac{1}{2}\ln(2\pi\sigma^2)$ である。ここで (13c) に至る計算では (1.48) を利用した。したがって、$p(x)$ は以下のように書ける。

$$p(x) = \exp\left\{\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

つまり、$p(x)$ はガウス分布である。

1.35 （基本）1 変数ガウス分布のエントロピー

微分エントロピーの定義 (1.104) から

$$\begin{aligned}
\mathrm{H}[x] &= -\int_{-\infty}^{\infty}p(x)\ln p(x)\,\mathrm{d}x \\
&= -\int_{-\infty}^{\infty}p(x)\ln\left[\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\right]\mathrm{d}x \\
&= -\int_{-\infty}^{\infty}p(x)\left\{\ln\frac{1}{(2\pi\sigma^2)^{1/2}} - \frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x \\
&= -\ln\frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}p(x)\,\mathrm{d}x + \frac{1}{2\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2p(x)\,\mathrm{d}x \\
&= -\ln\frac{1}{(2\pi\sigma^2)^{1/2}} + \frac{\sigma^2}{2\sigma^2} \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}\left\{1+\ln(2\pi\sigma^2)\right\}
\end{aligned}$$

1.36 （基本）凸関数と 2 階微分係数

関数 $f(x)$ が真に凸であるとき、$a < x_\lambda < b$ なる関数の定義域上の任意の a, b, x_λ について (1.114) から

$$f(x_\lambda) < \lambda f(a) + (1-\lambda)f(b) \tag{14}$$

が成り立つ。ただし、λ は $\lambda a + (1-\lambda)b = x_\lambda$ となる $0 < \lambda < 1$ の範囲の値である。この式は以下の式と等価である。

$$f(x_\lambda) - f(a) < (1-\lambda)\{f(b) - f(a)\}$$

この式の両辺を $x_\lambda - a$ すなわち $(1-\lambda)(b-a)$ で割ると

$$\frac{f(x_\lambda) - f(a)}{x_\lambda - a} < \frac{f(b) - f(a)}{b - a}$$

ここで $x_\lambda \to a$ とするとこの式の左辺は $x = a$ における微分係数と等しいので

$$f'(a) < \frac{f(b) - f(a)}{b - a} \tag{15}$$

また、(14) は以下の式とも等価である。

$$\lambda\{f(b) - f(a)\} < f(b) - f(x_\lambda)$$

この式の両辺を $b - x_\lambda$ すなわち $\lambda(b-a)$ で割ると

$$\frac{f(b) - f(a)}{b - a} < \frac{f(b) - f(x_\lambda)}{b - x_\lambda}$$

ここで $x_\lambda \to b$ とするとこの式の左辺は $x = b$ における微分係数と等しいので

$$\frac{f(b) - f(a)}{b - a} < f'(b) \tag{16}$$

(15) と (16) から任意の $a < b$ なる a, b に対して

$$f'(a) < f'(b)$$

これは導関数 $f'(x)$ が単調増加であることを表している。すなわち $f(x)$ が真に凸であれば任意の x について $f''(x) > 0$ である。

次に、任意の x について $f''(x) > 0$ のとき、導関数は単調増加なので、任意の $a < x_\lambda < b$ なる a, b, x_λ に対して $f'(a) < f'(x_\lambda) < f'(b)$ である。したがって、平均値の定理から $\frac{f(x_\lambda)-f(a)}{x_\lambda - a} = f'(c)$ となる c と $\frac{f(b)-f(x_\lambda)}{b-x_\lambda} = f'(d)$ となる d が、それぞれ $a < c < x_\lambda, x_\lambda < d < b$ の範囲に存在する。ここで、$f'(x)$ は単調増加で、$c < d$ であるから $f'(c) < f'(d)$ 、すなわち

$$\frac{f(x_\lambda) - f(a)}{x_\lambda - a} < \frac{f(b) - f(x_\lambda)}{b - x_\lambda}$$

が成り立つ。$x_\lambda = \lambda a + (1-\lambda)b$ とおいてこの式を整理すると、(14) が得られる。したがって、任意の x について $f''(x) > 0$ のとき、任意の $a < x_\lambda < b$ なる a, b, x_λ について、(14) が成り立ち、これは $f(x)$ が真に凸であることを示している。

1.37 （基本）同時分布のエントロピー

エントロピーの定義 (1.104) から

$$\begin{aligned}
\mathrm{H}[\mathbf{x}, \mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x} \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \ln\{p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x})\} \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x} \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \{\ln p(\mathbf{y} \mid \mathbf{x}) + \ln p(\mathbf{x})\} \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
&= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}\,|\,\mathbf{x})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x} - \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x} \\
&= \mathrm{H}[\mathbf{y}\,|\,\mathbf{x}] - \int \left\{ \int p(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} \right\} \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \mathrm{H}[\mathbf{y}\,|\,\mathbf{x}] - \int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \mathrm{H}[\mathbf{y}\,|\,\mathbf{x}] + \mathrm{H}[\mathbf{x}]
\end{aligned}
$$

1.38 （標準）イェンセンの不等式

(1.115) は $M = 1$ のとき $\lambda_1 = 1$ であることに注意すると明らかに成り立つことがわかる。凸関数 $f(x)$ に対して $\lambda_i \geqslant 0$ かつ $\sum_{i=1}^{m+1} \lambda_i = 1$ のとき、

$$
\begin{aligned}
f\left(\sum_{i=1}^{m+1} \lambda_i x_i \right) &= f\left(\sum_{i=1}^{m} \lambda_i x_i + \lambda_{m+1} x_{m+1} \right) \\
&= f(s_m \overline{x_m} + \lambda_{m+1} x_{m+1}) \\
&\leqslant s_m f(\overline{x_m}) + \lambda_{m+1} f(x_{m+1})
\end{aligned}
$$

ただし、$s_m = \sum_{i=1}^{m} \lambda_i$, $\overline{x_m} = \sum_{i=1}^{m} \lambda_i x_i / s_m$ であり、このとき $s_m + \lambda_{m+1} = 1$ であることを使っている。$\lambda_i / s_m \geqslant 0$, $\sum_{i=1}^{m} \lambda_i / s_m = 1$ であることに注意すると (1.115) が $M = m$ のとき成り立つのであれば

$$
\begin{aligned}
f(\overline{x_m}) &= f\left(\sum_{i=1}^{m} \frac{\lambda_i}{s_m} x_i \right) \\
&\leqslant \sum_{i=1}^{m} \frac{\lambda_i}{s_m} f(x_i)
\end{aligned}
$$

であるから、

$$
\begin{aligned}
f\left(\sum_{i=1}^{m+1} \lambda_i x_i \right) &\leqslant s_m f(\overline{x_m}) + \lambda_{m+1} f(x_{m+1}) \\
&\leqslant s_m \sum_{i=1}^{m} \frac{\lambda_i}{s_m} f(x_i) + \lambda_{m+1} f(x_{m+1}) \\
&= \sum_{i=1}^{m} \lambda_i f(x_i) + \lambda_{m+1} f(x_{m+1}) \\
&= \sum_{i=1}^{m+1} \lambda_i f(x_i)
\end{aligned}
$$

となり、(1.115) が $M = m + 1$ のときも成り立つことが示せる。したがって、数学的帰納法により任意の自然数 M について (1.115) が成り立つ。

1.39 （難問）エントロピーの計算

以下、$p(x = i)$ を $p(x_i)$ と書く。$p(y_j)$, $p(x_i, y_j)$, $p(x_i \,|\, y_j)$ 等も同様。

$$
\mathrm{H}[x] = -\sum_i p(x_i) \ln p(x_i) = -2/3 \cdot \ln(2/3) - 1/3 \cdot \ln(1/3) = 0.6365
$$

$$
\mathrm{H}[y] = -\sum_j p(y_j) \ln p(y_j) = -1/3 \cdot \ln(1/3) - 2/3 \cdot \ln(2/3) = 0.6365
$$

$$\mathrm{H}[y\,|\,x] = -\sum_{i,j} p(y_j, x_i) \ln p(y_j\,|\,x_i) = -1/3 \cdot \ln(1/2) - 1/3 \cdot \ln(1/2) - 1/3 \cdot \ln 1 = 0.4621$$
$$\mathrm{H}[x\,|\,y] = -\sum_{i,j} p(x_i, y_j) \ln p(x_i\,|\,y_j) = -1/3 \cdot \ln 1 - 1/3 \cdot \ln(1/2) - 1/3 \cdot \ln(1/2) = 0.4621$$
$$\mathrm{H}[x, y] = -\sum_{i,j} p(x_i, y_j) \ln p(x_i, y_j) = -1/3 \cdot \ln(1/3) - 1/3 \cdot \ln(1/3) - 1/3 \cdot \ln(1/3) = 1.0986$$
$$\mathrm{I}[x, y] = \mathrm{H}[x] - \mathrm{H}[x\,|\,y] = 0.6365 - 0.4621 = 0.1744$$

1.40 （基本）イェンセンの不等式の応用

$\lambda_i = 1/M$ および $f(x) = \ln x$ を (1.115) の左辺に適用すると、

$$\ln\left(\sum_{i=1}^{M} \frac{1}{M} x_i\right) = \ln\left(\frac{1}{M} \sum_{i=1}^{M} x_i\right)$$

$\lambda_i = 1/M$ および $f(x) = \ln x$ を (1.115) の右辺に適用すると、

$$\sum_{i=1}^{M} \frac{1}{M} \ln x_i = \ln\left(\prod_{i=1}^{M} x_i\right)^{1/M}$$

$\ln x$ は凹関数であるので、(1.115) の不等号の向きが逆になることに注意すると、

$$\ln\left(\frac{1}{M} \sum_{i=1}^{M} x_i\right) \geqslant \ln\left(\prod_{i=1}^{M} x_i\right)^{1/M}$$

と書ける。ここで、$\ln x$ は単調増加なので

$$\frac{1}{M} \sum_{i=1}^{M} x_i \geqslant \left(\prod_{i=1}^{M} x_i\right)^{1/M}$$

この式の左辺は算術平均、右辺は幾何平均であるから、算術平均は幾何平均よりも小さくならないことが示せた。

1.41 （基本）相互情報量

相互情報量 $\mathrm{I}(\mathbf{x}, \mathbf{y})$ の定義 (1.120) から

$$\begin{aligned}
\mathrm{I}[\mathbf{x}, \mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln\left\{\frac{p(\mathbf{x})\,p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})}\right\} \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \ln\left\{\frac{p(\mathbf{x})\,p(\mathbf{y})}{p(\mathbf{y})\,p(\mathbf{x}\,|\,\mathbf{y})}\right\} \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}\,|\,\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= -\int \left\{\int p(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y}\right\} \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} - \mathrm{H}[\mathbf{x}\,|\,\mathbf{y}] \\
&= -\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} - \mathrm{H}[\mathbf{x}\,|\,\mathbf{y}] \\
&= \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}\,|\,\mathbf{y}]
\end{aligned}$$

$\mathrm{I}(\mathbf{x}, \mathbf{y})$ の定義 (1.120) は $\mathbf{x}, \mathbf{y}$ について対称なので同様に $\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}\,|\,\mathbf{x}]$ も示せる。

2 確率分布

2.1 （基本）ベルヌーイ分布の性質

$p(x=0|\mu)=1-\mu\,,\ p(x=1|\mu)=\mu$ であるから

$$\sum_{x=0}^{1} p(x|\mu) = (1-\mu)+\mu = 1$$

$$\mathbb{E}[x] = \sum_{x=0}^{1} x\,p(x|\mu) = 0\cdot(1-\mu)+1\cdot\mu = \mu$$

$$\mathrm{var}[x] = \sum_{x=0}^{1}(x-\mathbb{E}[x])^2 p(x|\mu) = (0-\mu)^2(1-\mu)+(1-\mu)^2\mu = \mu(1-\mu)$$

$$\mathrm{H}[x] = -\sum_{x=0}^{1} p(x|\mu)\ln p(x|\mu) = -(1-\mu)\ln(1-\mu)-\mu\ln\mu$$

2.2 （標準）ベルヌーイ分布の対称な表現方法

(2.261) の表現を使うと $p(x=-1|\mu)=(1-\mu)/2\,,\ p(x=1|\mu)=(1+\mu)/2$ である。したがって

$$p(x=-1|\mu)+p(x=1|\mu) = \frac{1-\mu}{2}+\frac{1+\mu}{2} = 1$$

であり、 (2.261) は正規化されている。

次に平均、分散、エントロピーはそれぞれ

$$\mathbb{E}[x] = (-1)\cdot\frac{1-\mu}{2}+1\cdot\frac{1+\mu}{2} = \mu$$

$$\mathrm{var}[x] = (-1-\mu)^2\frac{1-\mu}{2}+(1-\mu)^2\frac{1+\mu}{2} = 1-\mu^2$$

$$\mathrm{H}[x] = -\frac{1-\mu}{2}\ln\left(\frac{1-\mu}{2}\right)-\frac{1+\mu}{2}\ln\left(\frac{1+\mu}{2}\right)$$

2.3 （標準）二項定理

(2.10) の定義を用いると (2.262) の左辺は

$$\begin{aligned}\binom{N}{m}+\binom{N}{m-1} &= \frac{N!}{(N-m)!m!}+\frac{N!}{(N-m+1)!(m-1)!}\\ &= \frac{(N-m+1)\cdot N!+m\cdot N!}{(N-m+1)!m!}\\ &= \frac{(N+1)\cdot N!}{(N-m+1)!m!}\\ &= \frac{(N+1)!}{(N-m+1)!m!}\\ &= \binom{N+1}{m}\end{aligned}$$

となり、(2.262) が示せる。

次に (2.263) を証明する。(2.263) は $N = 0$ のときに明らかに成り立つ。$N = n$ のときに (2.263) が成り立つとすると

$$\begin{aligned}
(1+x)^{n+1} &= (1+x)^n(1+x) \\
&= \sum_{m=0}^{n} \binom{n}{m} x^m (1+x) \\
&= \sum_{m=0}^{n} \binom{n}{m} x^m + \sum_{m=0}^{n} \binom{n}{m} x^{m+1} \\
&= \binom{n}{0} x^0 + \sum_{m=1}^{n} \binom{n}{m} x^m + \sum_{m=0}^{n} \binom{n}{m} x^{m+1} \\
&= \binom{n}{0} x^0 + \sum_{m=1}^{n} \left\{ \binom{n+1}{m} - \binom{n}{m-1} \right\} x^m + \sum_{m=0}^{n} \binom{n}{m} x^{m+1} \\
&= \binom{n+1}{0} x^0 + \sum_{m=1}^{n} \binom{n+1}{m} x^m - \sum_{m=1}^{n} \binom{n}{m-1} x^m + \sum_{m=0}^{n} \binom{n}{m} x^{m+1} \\
&= \sum_{m=0}^{n} \binom{n+1}{m} x^m - \sum_{m=0}^{n-1} \binom{n}{m} x^{m+1} + \sum_{m=0}^{n} \binom{n}{m} x^{m+1} \\
&= \sum_{m=0}^{n} \binom{n+1}{m} x^m + \binom{n}{n} x^{n+1} \\
&= \sum_{m=0}^{n} \binom{n+1}{m} x^m + \binom{n+1}{n+1} x^{n+1} \\
&= \sum_{m=0}^{n+1} \binom{n+1}{m} x^m
\end{aligned}$$

となり、(2.263) は $N = n + 1$ のときも成り立つ。したがって、数学的帰納法により $N \geqslant 0$ の整数 N に対して (2.263) は成り立つ。

最後に二項分布が正規化されていることを示す。二項分布の各事象の確率の和は

$$\begin{aligned}
\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} &= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{-m} \\
&= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \left(\frac{\mu}{1-\mu} \right)^m \\
&= (1-\mu)^N \left(1 + \frac{\mu}{1-\mu} \right)^N \\
&= (1-\mu)^N \left(\frac{1}{1-\mu} \right)^N \\
&= 1
\end{aligned}$$

となり、二項分布は正規化されていることが示された。

2.4 （標準）二項分布の平均と分散

(2.264) の左辺を μ で微分すると

$$\frac{\mathrm{d}}{\mathrm{d}\mu} \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = \sum_{m=0}^{N} \binom{N}{m} \left\{ m\mu^{m-1}(1-\mu)^{N-m} - (N-m)\mu^m (1-\mu)^{N-m-1} \right\}$$

$$= \sum_{m=0}^{N} \binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1} \{m(1-\mu) - (N-m)\mu\}$$
$$= \sum_{m=0}^{N} m\binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1} - N\mu \sum_{m=0}^{N} \binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1}$$

(2.264) の右辺を μ で微分すると 0 なので

$$\sum_{m=0}^{N} m\binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1} = N\mu \sum_{m=0}^{N} \binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1}$$

両辺に $\mu(1-\mu)$ をかけると

$$\sum_{m=0}^{N} m\binom{N}{m} \mu^{m}(1-\mu)^{N-m} = N\mu \sum_{m=0}^{N} \binom{N}{m} \mu^{m}(1-\mu)^{N-m}$$

左辺の和の部分は (2.264) から 1 であるので、

$$\sum_{m=0}^{N} m\binom{N}{m} \mu^{m}(1-\mu)^{N-m} = N\mu \tag{17}$$

この式の左辺は二項分布の平均 $\mathbb{E}[m]$ であるので、二項分布の平均 (2.11) が示された。

(17) の両辺を再度 μ で微分すると

$$\sum_{m=0}^{N} m^2\binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1} - N\mu \sum_{m=0}^{N} m\binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1} = N$$

両辺に $\mu(1-\mu)$ をかけると

$$\sum_{m=0}^{N} m^2\binom{N}{m} \mu^{m}(1-\mu)^{N-m} - N\mu \sum_{m=0}^{N} m\binom{N}{m} \mu^{m}(1-\mu)^{N-m} = N\mu(1-\mu)$$

この式の左辺の第 2 項に (17) を使って整理すると

$$\sum_{m=0}^{N} m^2\binom{N}{m} \mu^{m}(1-\mu)^{N-m} = N\mu(1-\mu) + N^2\mu^2$$

この式の左辺は二項分布の二次モーメント $\mathbb{E}[m^2]$ である。したがって、二項分布の分散 $\mathrm{var}[m]$ は

$$\mathrm{var}[m] = \mathbb{E}[m^2] - \mathbb{E}[m]^2 = N\mu(1-\mu) + N^2\mu^2 - (N\mu)^2 = N\mu(1-\mu)$$

2.5 （標準）ベータ分布の正規化の確認

問題の指示のとおり、(2.266) を計算すると

$$\Gamma(a)\Gamma(b) = \int_0^\infty \exp(-x)x^{a-1}\,\mathrm{d}x \int_0^\infty \exp(-y)y^{b-1}\,\mathrm{d}y \tag{18a}$$
$$= \int_0^\infty \int_0^\infty \exp(-x)x^{a-1}\exp(-y)y^{b-1}\,\mathrm{d}y\,\mathrm{d}x \tag{18b}$$
$$= \int_0^\infty \int_x^\infty \exp(-x)x^{a-1}\exp(x-t)(t-x)^{b-1}\,\mathrm{d}t\,\mathrm{d}x \tag{18c}$$
$$= \int_0^\infty \int_x^\infty \exp(-t)x^{a-1}(t-x)^{b-1}\,\mathrm{d}t\,\mathrm{d}x \tag{18d}$$

$$= \int_0^\infty \int_0^t \exp(-t) x^{a-1} (t-x)^{b-1} \,\mathrm{d}x \,\mathrm{d}t \tag{18e}$$

$$= \int_0^\infty \int_0^1 \exp(-t) (t\mu)^{a-1} (t-t\mu)^{b-1} t \,\mathrm{d}\mu \,\mathrm{d}t \tag{18f}$$

$$= \int_0^\infty \exp(-t) t^{a+b-1} \int_0^1 \mu^{a-1} (1-\mu)^{b-1} \,\mathrm{d}\mu \,\mathrm{d}t \tag{18g}$$

$$= \int_0^\infty \exp(-t) t^{a+b-1} \,\mathrm{d}t \int_0^1 \mu^{a-1} (1-\mu)^{b-1} \,\mathrm{d}\mu \tag{18h}$$

$$= \Gamma(a+b) \int_0^1 \mu^{a-1} (1-\mu)^{b-1} \,\mathrm{d}\mu \tag{18i}$$

ここで、(18e) において、積分の順序を t, x の順番から x, t の順番へと入れ替えているが、積分区間は図 1 の水色の領域で、積分順序を入れ替えることにより、積分区間が図 2 のように変わることに注意。したがって、

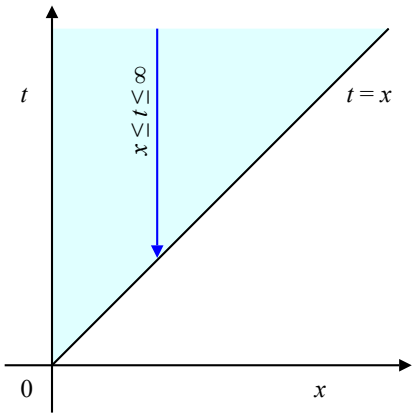

図 1　t, x の順に積分

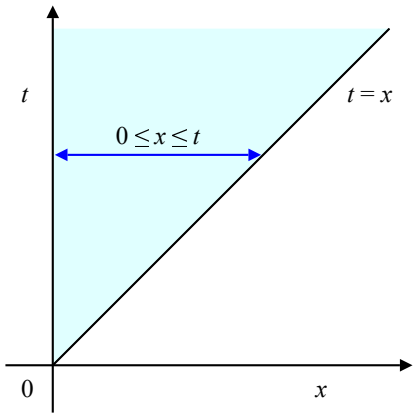

図 2　x, t の順に積分

$$\int_0^1 \mu^{a-1} (1-\mu)^{b-1} \,\mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

2.6　（基本）ベータ分布の性質

定義にしたがって、平均・分散を計算する。

$$\begin{aligned}
\mathbb{E}[\mu] &= \int_0^1 \mu \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \,\mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{(a+1)-1} (1-\mu)^{b-1} \,\mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \\
&= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \int_0^1 \mu^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \,\mathrm{d}\mu - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{(a+2)-1} (1-\mu)^{b-1} \,\mathrm{d}\mu - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} - \left(\frac{a}{a+b} \right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{(a+1)a\Gamma(a)\Gamma(b)}{(a+b+1)(a+b)\Gamma(a+b)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{(a+1)a}{(a+b+1)(a+b)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{ab}{(a+b+1)(a+b)^2}
\end{aligned}
$$

モードを求めるため、密度関数を μ で微分すると

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu}\mathrm{Beta}(\mu|\mathrm{a},\mathrm{b}) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\{(a-1)\mu^{a-2}(1-\mu)^{b-1} - (b-1)\mu^{a-1}(1-\mu)^{b-2}\} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-2}(1-\mu)^{b-2}\{(a-1)(1-\mu) - (b-1)\mu\} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-2}(1-\mu)^{b-2}\{(a-1) - (a+b-2)\mu\}
\end{aligned}
$$

モードでは微分係数が 0 になるので

$$
\mathrm{mode}[\mu] = \frac{a-1}{a+b-2}
$$

なお、$\mu = 0, \mu = 1$ でも微分係数は 0 になるが、ここではベータ分布の確率密度が 0 になるので、除外している。（細かいことを言うと、ここで求めた値がモードにならないケースもある。テキスト図 2.2 参照）

2.7 （標準）ベータ分布を事前分布とした二項分布の推定

事前平均は $\frac{a}{a+b}$、事後平均は $\frac{m+a}{m+a+l+b}$、最尤推定量は $\frac{m}{m+l}$ である。ここで、（問題文で指示された方法と異なるが）「事前平均と事後平均の差」と「事後平均と最尤推定量の差」の積を計算すると

$$
\left(\frac{a}{a+b} - \frac{m+a}{m+a+l+b}\right)\left(\frac{m+a}{m+a+l+b} - \frac{m}{m+l}\right) = \frac{(la-mb)^2}{(a+b)(m+a+l+b)^2(m+l)} \geqslant 0
$$

したがって、「事前平均と事後平均の差」と「事後平均と最尤推定量の差」は同符号である。つまり、「事前平均 $\leqslant$ 事後平均 $\leqslant$ 最尤推定量」もしくは 「事前平均 $\geqslant$ 事後平均 $\geqslant$ 最尤推定量」であり、事後平均は事前平均と最尤推定量の間にある。

2.8 （基本）条件付き期待値と条件付き分散

まず、x, y が連続変数の場合を考える。(2.270) の右辺を定義にしたがって計算すると

$$
\begin{aligned}
\mathbb{E}_y[\mathbb{E}_x[x|y]] &= \int \left\{ \int x\, p(x|y)\, \mathrm{d}x \right\} p(y)\, \mathrm{d}y \\
&= \iint x\, p(x|y)\, p(y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \iint x\, p(x,y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int x \left\{ \int p(x,y)\, \mathrm{d}y \right\} \mathrm{d}x \\
&= \int x\, p(x)\, \mathrm{d}x \\
&= \mathbb{E}[x]
\end{aligned}
$$

x, y の両方、もしくは一方が離散変数の場合は $\int(\cdots)\mathrm{d}x, \int(\cdots)\mathrm{d}y$ を $\sum_x(\cdots), \sum_y(\cdots)$ に書き換えれば同様に証明できる。

(2.271) については $\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ を利用して左辺を計算すると

$$\begin{aligned}
\mathbb{E}_y[\mathrm{var}_x[x|y]] + \mathrm{var}_y[\mathbb{E}_x[x|y]] &= (\mathbb{E}_y[\mathbb{E}_x[x^2|y] - \mathbb{E}_x[x|y]^2]) + (\mathbb{E}_y[\mathbb{E}_x[x|y]^2] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2) && \text{(19a)}\\
&= \mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[\mathbb{E}_x[x|y]^2] + \mathbb{E}_y[\mathbb{E}_x[x|y]^2] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2 && \text{(19b)}\\
&= \mathbb{E}[x^2] - \mathbb{E}[x]^2 && \text{(19c)}\\
&= \mathrm{var}[x] && \text{(19d)}
\end{aligned}$$

ここで、(19c) に至る計算では、先に証明した (2.270) を利用した。

2.9 （難問）ディリクレ分布の正規化

問題文にある通り、$M = 2$ のディリクレ分布はベータ分布と一致するため、正規化されていることが確認できている。次に、$M - 1$ 変数の場合に正規化されている、すなわち以下の式が成り立つと仮定する。

$$\int \cdots \int_{\mathcal{D}_{M-1}} \frac{\Gamma(\sum_{k=1}^{M-1} \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{M-1})} \prod_{k=1}^{M-1} \mu_k^{\alpha_k - 1} \, \mathrm{d}\mu_{M-2} \cdots \mathrm{d}\mu_1 = 1$$

ただし、領域 $\mathcal{D}_{M-1}$ は $\sum_{k=1}^{M-1} \mu_k = 1$, $\mu_k \geqslant 0$ である。ここで領域 $\mathcal{D}_{M-1}$ の制約を使って、この式から μ_{M-1} を消去すると、

$$\int \cdots \int_{\mathcal{D}_{M-1}} \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \cdot (1 - s)^{\alpha_{M-1} - 1} \, \mathrm{d}\mu_{M-2} \cdots \mathrm{d}\mu_1 = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{M-1})}{\Gamma(\sum_{k=1}^{M-1} \alpha_k)} \tag{20}$$

となることがわかる。ただし、$s = 1 - \sum_{k=1}^{M-2} \mu_k$ である。ここで、ディリクレ分布を (2.272) と書き、その積分を取ると

$$\begin{aligned}
&\int \cdots \int_{\mathcal{D}_M} C_M \prod_{k=1}^{M-1} \mu_k^{\alpha_k - 1} \left(1 - \sum_{j=1}^{M-1} \mu_j\right)^{\alpha_M - 1} \mathrm{d}\mu_{M-1} \cdots \mathrm{d}\mu_1 && \text{(21a)}\\
&= C_M \int \cdots \int_{\mathcal{D}_M} \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \cdot \mu_{M-1}^{\alpha_{M-1} - 1} (1 - s - \mu_{M-1})^{\alpha_M - 1} \, \mathrm{d}\mu_{M-1} \cdots \mathrm{d}\mu_1 && \text{(21b)}\\
&= C_M \int \cdots \int_{\mathcal{D}_{M-1}} \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \left\{ \int_0^{1-s} \mu_{M-1}^{\alpha_{M-1} - 1} (1 - s - \mu_{M-1})^{\alpha_M - 1} \mathrm{d}\mu_{M-1} \right\} \mathrm{d}\mu_{M-2} \cdots \mathrm{d}\mu_1 && \text{(21c)}\\
&= C_M \int \cdots \int_{\mathcal{D}_{M-1}} \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \left\{ (1 - s)^{\alpha_{M-1} + \alpha_M - 1} \int_0^1 t^{\alpha_{M-1} - 1} (1 - t)^{\alpha_M - 1} \mathrm{d}t \right\} \mathrm{d}\mu_{M-2} \cdots \mathrm{d}\mu_1 && \text{(21d)}\\
&= C_M \frac{\Gamma(\alpha_{M-1}) \Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)} \int \cdots \int_{\mathcal{D}_{M-1}} \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} (1 - s)^{\alpha_{M-1} + \alpha_M - 1} \, \mathrm{d}\mu_{M-2} \cdots \mathrm{d}\mu_1 && \text{(21e)}\\
&= C_M \frac{\Gamma(\alpha_{M-1}) \Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)} \cdot \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{M-2}) \cdot \Gamma(\alpha_{M-1} + \alpha_M)}{\Gamma(\sum_{k=1}^{M} \alpha_k)} && \text{(21f)}\\
&= C_M \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_M)}{\Gamma(\sum_{k=1}^{M} \alpha_k)} && \text{(21g)}
\end{aligned}$$

となる。ここで、(21d) に至る変形では $\mu_{M-1} = (1 - s)t$ とおいて変数変換を行い、(21e) に至る計算では (2.265) を、(21f) では (20) を使用した。(20) の使用については α_{M-1} の代わりに $\alpha_{M-1} + \alpha_M$ を使っていることに注意。

したがって $C_M = \Gamma(\sum_{k=1}^{M} \alpha_k) / \{\Gamma(\alpha_1) \cdots \Gamma(\alpha_M)\}$ であれば (2.272) 式は任意の $M \geqslant 2$ の整数の場合に正規化されていることが示せた。

2.10 （標準）ディリクレ分布の性質

$k \neq j$ なる k に対して $m_k = \mu_k/(1-\mu_j)$ とおくと m_k $(k \neq j)$ の和は 1 であり、その分布は明らかに $\prod_{k\neq j} \mu_k^{\alpha_k - 1}$ に比例するので、ディリクレ分布 $\mathrm{Dir}(\boldsymbol{m}|\boldsymbol{\alpha})$ である。そのことに注意すると、

$$\mathbb{E}[\mu_j] = \int_{\mathcal{D}} \mu_j \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}\,\mathrm{d}\boldsymbol{\mu} \tag{22a}$$

$$= \int_{\mathcal{D}} \mu_j^{\alpha_j} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \frac{\Gamma(\alpha_0-\alpha_j)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{j-1})\Gamma(\alpha_{j+1})\cdots\Gamma(\alpha_K)} \prod_{k\neq j} \mu_k^{\alpha_k - 1}\,\mathrm{d}\boldsymbol{\mu} \tag{22b}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \int_0^1 \mu_j^{\alpha_j}(1-\mu_j)^{\alpha_0-\alpha_j-1} \left\{ \int_{\mathcal{D}-\mu_j} \mathrm{Dir}(\boldsymbol{m}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{m} \right\} \mathrm{d}\mu_j \tag{22c}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \int_0^1 \mu_j^{\alpha_j}(1-\mu_j)^{\alpha_0-\alpha_j-1}\,\mathrm{d}\mu_j \tag{22d}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \frac{\Gamma(\alpha_j+1)\Gamma(\alpha_0-\alpha_j)}{\Gamma(\alpha_0+1)} \tag{22e}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \frac{\alpha_j\Gamma(\alpha_j)\Gamma(\alpha_0-\alpha_j)}{\alpha_0\Gamma(\alpha_0)} \tag{22f}$$

$$= \frac{\alpha_j}{\alpha_0} \tag{22g}$$

だたし、領域 $\mathcal{D}$ は $\sum_k \mu_k = 1,\ \mu_k \geqslant 0$ を、領域 $\mathcal{D}-\mu_j$ は $\sum_{k\neq j} \mu_k = 1-\mu_j,\ \mu_k \geqslant 0$ を表す。また、(22c) に至る計算においては $m_k = \mu_k/(1-\mu_j)$ の変数変換をしているが、$\boldsymbol{\mu}$ の K 個の要素のうち一つは従属変数であるので、$\boldsymbol{m}$ に含まれる独立変数の数は $K-2$ 個であり、$\mathrm{d}\boldsymbol{\mu} = (1-\mu_j)^{K-2}\mathrm{d}\boldsymbol{m}\,\mathrm{d}\mu_j$ であることに注意。また、(22e) に至る計算においては (2.265) を使用した。

分散を求める前に二次モーメント $\mathbb{E}[\mu_j^2]$ を先の計算に倣って計算すると、

$$\begin{aligned}
\mathbb{E}[\mu_j^2] &= \int_{\mathcal{D}} \mu_j^2 \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}\,\mathrm{d}\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \int_0^1 \mu_j^{\alpha_j+1}(1-\mu_j)^{\alpha_0-\alpha_j-1}\,\mathrm{d}\mu_j \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \frac{\Gamma(\alpha_j+2)\Gamma(\alpha_0-\alpha_j)}{\Gamma(\alpha_0+2)} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_j)} \frac{(\alpha_j+1)\alpha_j\Gamma(\alpha_j)\Gamma(\alpha_0-\alpha_j)}{(\alpha_0+1)\alpha_0\Gamma(\alpha_0)} \\
&= \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)}
\end{aligned}$$

したがって分散は

$$\begin{aligned}
\mathrm{var}[\mu_j] &= \mathbb{E}[\mu_j^2] - \mathbb{E}[\mu_j]^2 \\
&= \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} - \left(\frac{\alpha_j}{\alpha_0}\right)^2 \\
&= \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}
\end{aligned}$$

共分散を求める前に $\mathbb{E}[\mu_j\mu_l]$ を平均に倣って計算すると、

$$\mathbb{E}[\mu_j\mu_l] = \int_{\mathcal{D}} \mu_j\mu_l \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}\,\mathrm{d}\boldsymbol{\mu} \tag{23a}$$

$$= \int_{\mathcal{D}} \mu_j^{\alpha_j} \mu_l^{\alpha_l} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j-\alpha_l)\Gamma(\alpha_j)\Gamma(\alpha_l)} \frac{\Gamma(\alpha_0-\alpha_j-\alpha_l)}{\prod_{k\neq j,l}\Gamma(\alpha_k)} \prod_{k\neq j,l} \mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu} \tag{23b}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j-\alpha_l)\Gamma(\alpha_j)\Gamma(\alpha_l)} \int_0^1 \int_0^{1-\mu_j} \mu_j^{\alpha_j} \mu_l^{\alpha_l} (1-\mu_j-\mu_l)^{\alpha_0-\alpha_j-\alpha_l-1}\,\mathrm{d}\mu_l\,\mathrm{d}\mu_j \tag{23c}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j-\alpha_l)\Gamma(\alpha_j)\Gamma(\alpha_l)} \int_0^1 \mu_j^{\alpha_j} (1-\mu_j)^{\alpha_0-\alpha_j} \left\{ \int_0^1 t^{\alpha_l} (1-t)^{\alpha_0-\alpha_j-\alpha_l-1}\,\mathrm{d}t \right\} \mathrm{d}\mu_j \tag{23d}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j-\alpha_l)\Gamma(\alpha_j)\Gamma(\alpha_l)} \frac{\Gamma(\alpha_l+1)\Gamma(\alpha_0-\alpha_j-\alpha_l)}{\Gamma(\alpha_0-\alpha_j+1)} \int_0^1 \mu_j^{\alpha_j} (1-\mu_j)^{\alpha_0-\alpha_j}\,\mathrm{d}\mu_j \tag{23e}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j-\alpha_l)\Gamma(\alpha_j)\Gamma(\alpha_l)} \frac{\Gamma(\alpha_l+1)\Gamma(\alpha_0-\alpha_j-\alpha_l)}{\Gamma(\alpha_0-\alpha_j+1)} \frac{\Gamma(\alpha_j+1)\Gamma(\alpha_0-\alpha_j+1)}{\Gamma(\alpha_0+2)} \tag{23f}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0-\alpha_j-\alpha_l)\Gamma(\alpha_j)\Gamma(\alpha_l)} \frac{\alpha_l\Gamma(\alpha_l)\Gamma(\alpha_0-\alpha_j-\alpha_l)}{\Gamma(\alpha_0-\alpha_j+1)} \frac{\alpha_j\Gamma(\alpha_j)\Gamma(\alpha_0-\alpha_j+1)}{(\alpha_0+1)\alpha_0\Gamma(\alpha_0)} \tag{23g}$$

$$= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} \tag{23h}$$

なお、(23d) に至る計算では $t=\mu_l/(1-\mu_j)$ の変数変換をしている。したがって共分散は

$$\begin{aligned}
\mathrm{cov}[\mu_j,\mu_l] &= \mathbb{E}[\mu_j\mu_l] - \mathbb{E}[\mu_j]\mathbb{E}[\mu_j] \\
&= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j}{\alpha_0}\frac{\alpha_l}{\alpha_0} \\
&= -\frac{\alpha_j\alpha_l}{\alpha_0^2(\alpha_0+1)}
\end{aligned}$$

2.11 （基本）ディリクレ分布の対数の平均

まず、$\prod_{k=1}^M \mu_k^{\alpha_k-1}$ を α_j で偏微分する。

$$\begin{aligned}
\frac{\partial}{\partial\alpha_j}\prod_{k=1}^M \mu_k^{\alpha_k-1} &= \frac{\partial}{\partial\alpha_j}\left(\prod_{k\neq j}\mu_k^{\alpha_k-1}\cdot\mu_j^{\alpha_j-1}\right) \\
&= \frac{\partial}{\partial\alpha_j}\left(\prod_{k\neq j}\mu_k^{\alpha_k-1}\cdot e^{(\alpha_j-1)\ln\mu_j}\right) \\
&= \ln\mu_j \prod_{k\neq j}\mu_k^{\alpha_k-1}\cdot e^{(\alpha_j-1)\ln\mu_j} \\
&= \ln\mu_j \prod_{k\neq j}\mu_k^{\alpha_k-1}\cdot\mu_j^{\alpha_j-1} \\
&= \ln\mu_j \prod_{k=1}^M \mu_k^{\alpha_k-1}
\end{aligned} \tag{24}$$

また、ディリクレ分布を正規化するための係数 $\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}$ の逆数を α_j で偏微分すると

$$\begin{aligned}
\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)} &= \prod_{k\neq j}\Gamma(\alpha_k)\cdot\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_1+\cdots+\alpha_K)} \\
&= \prod_{k\neq j}\Gamma(\alpha_k)\frac{\Gamma'(\alpha_j)\Gamma(\alpha_1+\cdots+\alpha_K)-\Gamma(\alpha_j)\Gamma'(\alpha_1+\cdots+\alpha_K)}{\{\Gamma(\alpha_1+\cdots+\alpha_K)\}^2} \\
&= \prod_{k\neq j}\Gamma(\alpha_k)\frac{\Gamma'(\alpha_j)\Gamma(\alpha_0)-\Gamma(\alpha_j)\Gamma'(\alpha_0)}{\{\Gamma(\alpha_0)\}^2} \\
&= \prod_{k=1}^M\Gamma(\alpha_k)\frac{\Gamma'(\alpha_j)\Gamma(\alpha_0)-\Gamma(\alpha_j)\Gamma'(\alpha_0)}{\Gamma(\alpha_j)\{\Gamma(\alpha_0)\}^2}
\end{aligned}$$

$$= \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)}\left(\frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} - \frac{\Gamma'(\alpha_0)}{\Gamma(\alpha_0)}\right) \tag{25}$$

ここでディガンマ関数の定義から $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$ であるから、(25) は

$$\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)} = \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)}\{\psi(\alpha_j) - \psi(\alpha_0)\} \tag{26}$$

となる。これを利用すると、

$$\mathbb{E}[\ln\mu_j] = \int_{\mathcal{D}} \ln\mu_j \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu} \tag{27a}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\int_{\mathcal{D}}\frac{\partial}{\partial\alpha_j}\prod_{k=1}^{M}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu} \tag{27b}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\frac{\partial}{\partial\alpha_j}\int_{\mathcal{D}}\prod_{k=1}^{M}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu} \tag{27c}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)} \tag{27d}$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)}\{\psi(\alpha_j) - \psi(\alpha_0)\} \tag{27e}$$

$$= \psi(\alpha_j) - \psi(\alpha_0) \tag{27f}$$

だたし、領域 $\mathcal{D}$ は $\sum_k \mu_k = 1,\ \mu_k \geqslant 0$ である。また、(27b) に至る計算では (24) を、(27d) に至る計算ではディリクレ分布の正規化条件を、(27e) に至る計算では (26) をそれぞれ利用した。

2.12 （基本）連続変数の一様分布

まず、(2.278) が正規化されていることを確かめるために (2.278) を積分する。

$$\begin{aligned}\int_a^b \mathrm{U}(x|a,b)\,\mathrm{d}x &= \int_a^b \frac{\mathrm{d}x}{b-a}\\ &= \left[\frac{x}{b-a}\right]_{x=a}^{x=b}\\ &= \frac{b}{b-a} - \frac{a}{b-a}\\ &= 1\end{aligned}$$

したがって、$\mathrm{U}(x|a,b)$ は正規化されていることが確認できた。

次に平均を定義に従って計算する。

$$\begin{aligned}\mathbb{E}[x] &= \int_a^b \frac{x}{b-a}\,\mathrm{d}x\\ &= \left[\frac{x^2}{2(b-a)}\right]_{x=a}^{x=b}\\ &= \frac{b^2}{2(b-a)} - \frac{a^2}{2(b-a)}\\ &= \frac{a+b}{2}\end{aligned}$$

最後に分散は

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

$$
\begin{aligned}
&= \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x - \left(\frac{a+b}{2}\right)^2 \\
&= \left[\frac{x^3}{3(b-a)}\right]_{x=a}^{x=b} - \frac{(a+b)^2}{4} \\
&= \frac{b^3}{3(b-a)} - \frac{a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\
&= \frac{(b-a)^2}{12}
\end{aligned}
$$

2.13 （標準）2 つの多変量ガウス分布間の KL ダイバージェンス

まず、$q(\mathbf{x})$ と $p(\mathbf{x})$ の比の対数を計算する。

$$
\begin{aligned}
\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} &= \ln\left[\frac{\dfrac{1}{(2\pi)^{D/2}}\dfrac{1}{|\mathbf{L}|^{1/2}}\exp\left\{-\dfrac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right\}}{\dfrac{1}{(2\pi)^{D/2}}\dfrac{1}{|\mathbf{\Sigma}|^{1/2}}\exp\left\{-\dfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}}\right] \\
&= \ln\left[\left(\frac{|\mathbf{\Sigma}|}{|\mathbf{L}|}\right)^{1/2}\exp\left\{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right\}\right] \\
&= \frac{1}{2}\ln\frac{|\mathbf{\Sigma}|}{|\mathbf{L}|} + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) \\
&= \frac{1}{2}\left\{\mathbf{x}^{\mathrm{T}}(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{x} - 2(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}-\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1})\mathbf{x} + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} + \ln\frac{|\mathbf{\Sigma}|}{|\mathbf{L}|}\right\} \qquad (28)
\end{aligned}
$$

ここで、(28) に至る計算ではベクトル $\mathbf{x}, \mathbf{y}$ と行列 $\mathbf{A}$ について $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{x}$ であることと精度行列 $\mathbf{\Sigma}^{-1}, \mathbf{L}^{-1}$ が対称行列であることを使っている。(28) の最初の項である二次形式 $\mathbf{x}^{\mathrm{T}}(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{x}$ の期待値を計算すると

$$
\begin{aligned}
\int \mathbf{x}^{\mathrm{T}}(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{x}\,p(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \int \mathrm{Tr}\{\mathbf{x}^{\mathrm{T}}(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{x}\}p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \int \mathrm{Tr}\{(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}\,p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \mathrm{Tr}\left\{\int(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{x}\mathbf{x}^{\mathrm{T}}p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right\} \\
&= \mathrm{Tr}\left\{(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\int\mathbf{x}\mathbf{x}^{\mathrm{T}}p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right\} \\
&= \mathrm{Tr}\{(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}+\mathbf{\Sigma})\} \\
&= \mathrm{Tr}\{(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + (\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{\Sigma}\} \\
&= \mathrm{Tr}\{\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\boldsymbol{\mu} + \mathbf{I} - \mathbf{L}^{-1}\mathbf{\Sigma}\} \\
&= \boldsymbol{\mu}^{\mathrm{T}}(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\boldsymbol{\mu} + D - \mathrm{Tr}(\mathbf{L}^{-1}\mathbf{\Sigma}) \qquad (29)
\end{aligned}
$$

なおこの計算では二次形式 $\mathbf{x}^{\mathrm{T}}(\mathbf{\Sigma}^{-1}-\mathbf{L}^{-1})\mathbf{x}$ はスカラーなので、トレースをとっても値が変わらないこと、トレース作用素の循環性 (C.9)、および、ガウス分布の二次モーメント (2.62) を使っている。次に (28) の二番目の項の期待値を計算すると

$$
\begin{aligned}
\int(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}-\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1})\mathbf{x}\,p(\mathbf{x})\,\mathrm{d}\mathbf{x} &= (\boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}-\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1})\int\mathbf{x}\,p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= (\boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}-\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1})\boldsymbol{\mu} \\
&= \boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} \qquad (30)
\end{aligned}
$$

カルバック-ライブラーダイバージェンス (1.113) は (28) の $p(\mathbf{x})$ おける期待値の符号を変えたものであるので、(29) と (30) を利用すると

$$
\begin{aligned}
&\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}x \\
&= -\frac{1}{2} \int \left\{ \mathbf{x}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \mathbf{L}^{-1})\mathbf{x} - 2(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1})\mathbf{x} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} + \ln \frac{|\boldsymbol{\Sigma}|}{|\mathbf{L}|} \right\} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&= -\frac{1}{2} \left\{ \boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \mathbf{L}^{-1})\boldsymbol{\mu} + D - \mathrm{Tr}(\mathbf{L}^{-1}\boldsymbol{\Sigma}) - 2(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} + \ln \frac{|\boldsymbol{\Sigma}|}{|\mathbf{L}|} \right\} \\
&= \frac{1}{2} \left\{ \boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} - 2\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} + \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} - D + \mathrm{Tr}(\mathbf{L}^{-1}\boldsymbol{\Sigma}) - \ln \frac{|\boldsymbol{\Sigma}|}{|\mathbf{L}|} \right\} \\
&= \frac{1}{2} \left\{ \boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} + \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} - D + \mathrm{Tr}(\mathbf{L}^{-1}\boldsymbol{\Sigma}) - \ln \frac{|\boldsymbol{\Sigma}|}{|\mathbf{L}|} \right\} \\
&= \frac{1}{2} \left\{ (\boldsymbol{m} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}^{-1}(\boldsymbol{m} - \boldsymbol{\mu}) - D + \mathrm{Tr}(\mathbf{L}^{-1}\boldsymbol{\Sigma}) - \ln \frac{|\boldsymbol{\Sigma}|}{|\mathbf{L}|} \right\}
\end{aligned}
$$

2.14 （標準）エントロピーを最大化する多変量連続確率分布

ラグランジュ関数 (E.4) は以下のようになる。

$$
\begin{aligned}
L(\mathbf{x}, \lambda, \mathbf{m}, \mathbf{L}) = &- \int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \lambda \left\{ \int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} - 1 \right\} \\
&+ \mathbf{m}^{\mathrm{T}} \left\{ \int p(\mathbf{x}) \mathbf{x} \, \mathrm{d}\mathbf{x} - \boldsymbol{\mu} \right\} + \mathrm{Tr} \left[\mathbf{L} \left\{ \int p(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \, \mathrm{d}\mathbf{x} - \boldsymbol{\Sigma} \right\} \right] \\
= &\int \left\{ -p(\mathbf{x}) \ln p(\mathbf{x}) + \lambda \, p(\mathbf{x}) + \mathbf{m}^{\mathrm{T}}\mathbf{x} \, p(\mathbf{x}) + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x} - \boldsymbol{\mu}) p(\mathbf{x}) \right\} \mathrm{d}\mathbf{x} \\
&- \lambda - \mathbf{m}^{\mathrm{T}}\boldsymbol{\mu} - \mathrm{Tr}(\mathbf{L}\boldsymbol{\Sigma})
\end{aligned}
$$

ここで $\lambda, \mathbf{m}, \mathbf{L}$ はラグランジュ乗数で、それぞれ、スカラー・D 次元ベクトル・$D \times D$ 行列である。したがってオイラー-ラグランジュ方程式 (D.8) は

$$
-\ln p(\mathbf{x}) - 1 + \lambda + \mathbf{m}^{\mathrm{T}}\mathbf{x} + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x} - \boldsymbol{\mu}) = 0
$$

これを $p(\mathbf{x})$ について解くと

$$
p(\mathbf{x}) = \exp \left\{ \lambda - 1 + \mathbf{m}^{\mathrm{T}}\mathbf{x} + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x} - \boldsymbol{\mu}) \right\} \tag{31}
$$

この式を平方完成すると、

$$
p(\mathbf{x}) = \exp \left\{ \left(\mathbf{x} - \boldsymbol{\mu} + \frac{1}{2}\mathbf{L}^{-1}\mathbf{m} \right)^{\mathrm{T}} \mathbf{L} \left(\mathbf{x} - \boldsymbol{\mu} + \frac{1}{2}\mathbf{L}^{-1}\mathbf{m} \right) + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \frac{1}{4}\mathbf{m}^{\mathrm{T}}\mathbf{L}\mathbf{m} + \lambda - 1 \right\}
$$

ここで以下の関数を導入する。

$$
q(\mathbf{y}) = \exp \left(\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \frac{1}{4}\mathbf{m}^{\mathrm{T}}\mathbf{L}\mathbf{m} + \lambda - 1 \right)
$$

この関数は $p(\mathbf{x})$ に対して $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} + \frac{1}{2}\mathbf{L}^{-1}\mathbf{m}$ の変数変換を行った関数である。$q(\mathbf{y})$ は $p(\mathbf{x})$ を平行移動させただけの関数なので、これも正規化されており、その積分値は 1 である。また、$q(\mathbf{y})\mathbf{y}$ は奇関数であるのでその積分値は 0 である。これらのことに注意すると、(2.281) の左辺は

$$
\int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int q(\mathbf{y}) \left(\mathbf{y} + \boldsymbol{\mu} - \frac{1}{2}\mathbf{L}^{-1}\mathbf{m} \right) \mathrm{d}\mathbf{y}
$$

$$
\begin{aligned}
&= \int q(\mathbf{y})\mathbf{y}\,\mathrm{d}\mathbf{y} + \left(\boldsymbol{\mu} - \frac{1}{2}\mathbf{L}^{-1}\mathbf{m}\right)\int q(\mathbf{y})\,\mathrm{d}\mathbf{y} \\
&= \boldsymbol{\mu} - \frac{1}{2}\mathbf{L}^{-1}\mathbf{m}
\end{aligned}
$$

これが $\boldsymbol{\mu}$ に等しいので $-\frac{1}{2}\mathbf{L}^{-1}\mathbf{m} = \mathbf{0}$ 、すなわち $\mathbf{m} = \mathbf{0}$ である。この結果を (31) に代入すると (2.280) の左辺は

$$
\begin{aligned}
\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \int \exp\left\{\lambda - 1 + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathrm{d}\mathbf{x} \\
&= \ln(\lambda - 1)\int \exp\left\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathrm{d}\mathbf{x}
\end{aligned}
$$

ここで、$\mathbf{L} = -\frac{1}{2}\mathbf{M}^{-1}$ となる行列 $\mathbf{M}$ を導入すると、指数部分がガウス分布の指数部分に一致することを使い

$$
\begin{aligned}
\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \ln(\lambda - 1)\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{M}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathrm{d}\mathbf{x} \\
&= (2\pi)^{D/2}|\mathbf{M}|^{1/2}\ln(\lambda - 1)\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{M})\,\mathrm{d}\mathbf{x} \\
&= (2\pi)^{D/2}|\mathbf{M}|^{1/2}\ln(\lambda - 1)
\end{aligned}
$$

となる。これが 1 に等しくなるため、$\ln(\lambda - 1) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{M}|^{1/2}}$ である。これらの結果を (31) に代入すると

$$
p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{M}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{M}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}
$$

これは多変量ガウス分布 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{M})$ である。したがって、平均、共分散の制約のもとでエントロピーを最大化する多変量分布はガウス分布である。なお、(2.282) を用いると $\mathbf{M} = \boldsymbol{\Sigma}$ つまり、$\mathbf{L} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}$ であることがわかる。

2.15 （標準）多変量ガウス分布のエントロピー

エントロピーの定義 (1.104) から

$$
\begin{aligned}
\mathrm{H}[\mathbf{x}] &= -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\ln\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathrm{d}\mathbf{x} \\
&= -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right]\mathrm{d}\mathbf{x} \\
&= -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left\{-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathrm{d}\mathbf{x} \\
&= \left(\frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}|\right)\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathrm{d}\mathbf{x} + \frac{1}{2}\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{Tr}\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}\,\mathrm{d}\mathbf{x} \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}[\mathrm{Tr}\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}] \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}[\mathrm{Tr}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}} - 2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})] \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - 2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{x}] + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{Tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\Sigma}) - 2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\} \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \mathbf{I} - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{Tr}(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \mathbf{I} - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{Tr}(\mathbf{I}) \\
&= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{D}{2} \\
&= \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{D}{2}(1 + \ln(2\pi))
\end{aligned}
$$

だたし、$\mathbf{I}$ は $D \times D$ の単位行列である。

2.16 （難問）ガウス分布の和の分布のエントロピー

x の分布は次のようになる。

$$
\begin{aligned}
p(x) &= \int_{-\infty}^{\infty} p(x|x_2)p(x_2)\,\mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} p(x_1 = x - x_2)p(x_2)\,\mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} \frac{\tau_1^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2}\tau_1(x - x_2 - \mu_1)^2\right\} \frac{\tau_2^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2}\tau_2(x_2 - \mu_2)^2\right\} \mathrm{d}x_2 \\
&= \frac{(\tau_1\tau_2)^{1/2}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\tau_1(x - x_2 - \mu_1)^2 - \frac{1}{2}\tau_2(x_2 - \mu_2)^2\right\} \mathrm{d}x_2 \\
&= \frac{(\tau_1\tau_2)^{1/2}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1 x - \tau_1\mu_1 + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 - \frac{\tau_1\tau_2(x - \mu_1 - \mu_2)^2}{2(\tau_1 + \tau_2)}\right\} \mathrm{d}x_2 \\
&= \frac{(\tau_1\tau_2)^{1/2}}{2\pi} \exp\left\{-\frac{\tau_1\tau_2(x - \mu_1 - \mu_2)^2}{2(\tau_1 + \tau_2)}\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1 x - \tau_1\mu_1 + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2\right\} \mathrm{d}x_2 \\
&= \frac{(\tau_1\tau_2)^{1/2}}{2\pi} \left(\frac{2\pi}{\tau_1 + \tau_2}\right)^{1/2} \exp\left\{-\frac{\tau_1\tau_2(x - \mu_1 - \mu_2)^2}{2(\tau_1 + \tau_2)}\right\} \\
&= \left\{\frac{\tau_1\tau_2}{2\pi(\tau_1 + \tau_2)}\right\}^{1/2} \exp\left\{-\frac{\tau_1\tau_2(x - \mu_1 - \mu_2)^2}{2(\tau_1 + \tau_2)}\right\}
\end{aligned}
$$

したがって、$p(x)$ は分散 $\frac{\tau_1+\tau_2}{\tau_1\tau_2}$ のガウス分布であり、その微分エントロピーは (1.110) より

$$
\mathrm{H}[x] = \frac{1}{2}\left[1 + \ln\left\{\frac{2\pi(\tau_1 + \tau_2)}{\tau_1\tau_2}\right\}\right]
$$

2.17 （基本）ガウス分布の精度行列の対称性

$\mathbf{B}$ を反対称行列、$\mathbf{x}, \boldsymbol{\mu}$ をそれぞれ任意のベクトルとすると $(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{B}(\mathbf{x} - \boldsymbol{\mu})$ はスカラーなので

$$
\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{B}(\mathbf{x} - \boldsymbol{\mu}) &= \{(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{B}(\mathbf{x} - \boldsymbol{\mu})\}^{\mathrm{T}} \\
&= (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) \\
&= -(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{B}(\mathbf{x} - \boldsymbol{\mu})
\end{aligned}
$$

となるため、$(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{B}(\mathbf{x} - \boldsymbol{\mu})$ は 0 である。ここで、精度行列 $\boldsymbol{\Sigma}^{-1}$ が対称行列 $\mathbf{A}$ と反対称行列 $\mathbf{B}$ を使って $\boldsymbol{\Sigma}^{-1} = \mathbf{A} + \mathbf{B}$ と書けたとすると、ガウス分布の指数部分は次のようになる。

$$
\begin{aligned}
\exp\left\{(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} &= \exp\left\{(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{A} + \mathbf{B})(\mathbf{x} - \boldsymbol{\mu})\right\} \\
&= \exp\left\{(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{B}(\mathbf{x} - \boldsymbol{\mu})\right\}
\end{aligned}
$$

$$= \exp\left\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

となり、精度行列は対称行列であるとしても一般性は失われない。なお、任意の正方行列が対称行列と反対称行列の和で表せることは演習問題 1.14 ですでに示している。

2.18 （難問）実対称行列の固有値と固有ベクトル

複素数 λ_i と複素ベクトル $\mathbf{u}_i$ の共役複素数・共役複素ベクトルをそれぞれ $\overline{\lambda}_i, \overline{\mathbf{u}}_i$ と書くとする。$\boldsymbol{\Sigma}$ を実対称行列とし、(2.45) の複素共役をとると

$$\boldsymbol{\Sigma}\overline{\mathbf{u}}_i = \overline{\lambda}_i\overline{\mathbf{u}}_i$$

この式の両辺のベクトルと $\mathbf{u}_i$ との内積を計算すると

$$(\boldsymbol{\Sigma}\overline{\mathbf{u}}_i)^{\mathrm{T}}\mathbf{u}_i = (\overline{\lambda}_i\overline{\mathbf{u}}_i)^{\mathrm{T}}\mathbf{u}_i$$

$\boldsymbol{\Sigma}$ は対称行列であることと、λ_i はスカラーであることに注意すると

$$\overline{\mathbf{u}}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_i = \overline{\lambda}_i\overline{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_i$$

ここで、この式の左辺に (2.45) を代入すると

$$\lambda_i\overline{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_i = \overline{\lambda}_i\overline{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_i$$

$\mathbf{u}_i \neq \mathbf{0}$ であれば $\overline{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_i$ は非負の実数なので $\lambda_i = \overline{\lambda}_i$ 、すなわち λ_i は実数である。

次に 2 つの固有ベクトル $\mathbf{u}_i, \mathbf{u}_j$ の内積を考える。$\mathbf{u}_i, \mathbf{u}_j$ の固有値がともに 0 でないなら

$$\begin{aligned}
\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j &= \frac{1}{\lambda_i}(\lambda_i\mathbf{u}_i^{\mathrm{T}})\mathbf{u}_j \\
&= \frac{1}{\lambda_i}(\boldsymbol{\Sigma}\mathbf{u}_i)^{\mathrm{T}}\mathbf{u}_j \\
&= \frac{1}{\lambda_i}\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_j \\
&= \frac{1}{\lambda_i}\mathbf{u}_i^{\mathrm{T}}(\boldsymbol{\Sigma}\mathbf{u}_j) \\
&= \frac{\lambda_j}{\lambda_i}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j
\end{aligned} \tag{32}$$

したがって $\frac{\lambda_j}{\lambda_i} \neq 1$ すなわち $\lambda_i \neq \lambda_j$ であるなら、$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 0$ であり、$\mathbf{u}_i$ と $\mathbf{u}_j$ は直交する。

仮に、実対称行列 $\boldsymbol{\Sigma}$ のすべての固有値が 0 でないときは求めた固有ベクトル $\mathbf{u}_i$ を $|\mathbf{u}_i|$ 倍することにより、(2.46) が成り立つように $\mathbf{u}_i$ を決めることができる。実対称行列 $\boldsymbol{\Sigma}$ の固有値に一つだけ 0 が含まれているときは、固有値 0 に対する固有ベクトルを $\mathbf{u}_j$ とすると、$\lambda_j = 0$ の場合にも (32) は成り立つので、その場合は明らかに内積が 0 であり、$\mathbf{u}_j$ はその他すべての固有値と直交することがわかる。したがって、ここでも固有ベクトル $\mathbf{u}_i$ を $|\mathbf{u}_i|$ 倍することにより、(2.46) が成り立つように $\mathbf{u}_i$ を決めることができる。

最後に固有値のうち d 個が 0 であるなら、各固有値 0 に対する固有ベクトル $\mathbf{v}_1, \mathbf{v}_2 \cdots, \mathbf{v}_d$ は線形独立であので、これらの固有ベクトルの張る空間上の任意のベクトル $\mathbf{a}$ は a_i を実数として $\mathbf{a} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_d\mathbf{v}_d$ と書けるから、

$$\begin{aligned}
\boldsymbol{\Sigma}\mathbf{a} &= \boldsymbol{\Sigma}(a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_d\mathbf{v}_d) \\
&= a_1\boldsymbol{\Sigma}\mathbf{v}_1 + a_2\boldsymbol{\Sigma}\mathbf{v}_2 + \cdots + a_d\boldsymbol{\Sigma}\mathbf{v}_d \\
&= \mathbf{0}
\end{aligned}$$

となり、$\mathbf{v}_1, \mathbf{v}_2 \cdots, \mathbf{v}_d$ によって張られる空間上の零ベクトルでない任意のベクトルも固有値 0 に対する固有ベクトルであることがわかる。したがって、固有値 0 に対する固有ベクトルとして $\mathbf{v}_1, \mathbf{v}_2 \cdots, \mathbf{v}_d$ によって張

られる空間上の直交する d 個の単位ベクトルを選ぶことにより、(2.46) が成り立つように $\mathbf{u}_i$ を決めることができる。（これらのベクトルが固有値が 0 以外の固有ベクトルと直交することも先の議論から明らかである。）

2.19 （標準）共分散行列の固有ベクトルによる表現

$\mathbf{u}_i$ は互いに直交する単位ベクトルであることから、D 次の任意の実ベクトル $\mathbf{a}$ は実数 $\hat{a}_i$ を用いて $\mathbf{a} = \hat{a}_1\mathbf{u}_1 + \hat{a}_2\mathbf{u}_2 + \cdots + \hat{a}_D\mathbf{u}_D$ とかける。このとき、

$$\left(\sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right)\mathbf{a} = \left(\sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right)(\hat{a}_1\mathbf{u}_1 + \hat{a}_2\mathbf{u}_2 + \cdots + \hat{a}_D\mathbf{u}_D) \tag{33a}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D}\hat{a}_j\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j \tag{33b}$$

$$= \sum_{i\neq j}\hat{a}_j\mathbf{u}_i(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j) + \sum_{i=1}^{D}\hat{a}_i\mathbf{u}_i(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i) \tag{33c}$$

$$= \sum_{i=1}^{D}\hat{a}_i\mathbf{u}_i \tag{33d}$$

$$= \mathbf{a} \tag{33e}$$

ここで、(33d) に至る計算では (2.46) から $i \neq j$ のとき $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 0$ であることと、$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i = 1$ であることを利用した。この結果は任意のベクトル $\mathbf{a}$ を $\sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$ で線形変換すると元のベクトルになることを示しており、すなわち、$\sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$ が単位行列であることを示している。ここで、(2.45) の両辺に右から $\mathbf{u}_i^{\mathrm{T}}$ をかけると

$$\mathbf{\Sigma}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} = \lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

この両辺のすべての i での総和を取ると

$$\mathbf{\Sigma}\sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} = \sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

ここで、$\sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$ は単位行列であるから、

$$\mathbf{\Sigma} = \sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

次に、(2.49) 式の右辺に右から $\mathbf{\Sigma}$ をかけると

$$\begin{aligned}\left(\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right)\mathbf{\Sigma} &= \left(\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right)\left(\sum_{j=1}^{D}\lambda_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right)\\ &= \sum_{i=1}^{D}\sum_{j=1}^{D}\frac{\lambda_j}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\\ &= \sum_{i\neq j}\frac{\lambda_j}{\lambda_i}\mathbf{u}_i(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j)\mathbf{u}_j^{\mathrm{T}} + \sum_{i=1}^{D}\frac{\lambda_i}{\lambda_i}\mathbf{u}_i(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i^{\mathrm{T}}\\ &= \sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\\ &= \mathbf{I}\end{aligned}$$

したがって

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

2.20 （標準）正定値行列

任意の実ベクトル $\mathbf{a}$ は実数 $\hat{a}_i$ を用いて $\mathbf{a} = \hat{a}_1\mathbf{u}_1 + \hat{a}_2\mathbf{u}_2 + \cdots + \hat{a}_D\mathbf{u}_D$ とかけるので

$$\begin{aligned}
\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a} &= \left(\sum_{i=1}^{D} \hat{a}_i \mathbf{u}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma} \left(\sum_{j=1}^{D} \hat{a}_j \mathbf{u}_j\right) \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} \hat{a}_i \hat{a}_j \mathbf{u}_i^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{u}_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} \hat{a}_i \hat{a}_j \lambda_j \mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j \\
&= \sum_{i=1}^{D} \hat{a}_i^2 \lambda_i
\end{aligned}$$

であるから、(2.45) で定義される $\boldsymbol{\Sigma}$ のすべての固有値 λ_i が正であれば、任意の実ベクトル $\mathbf{a}$ について $\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}$ は正になる。また、ある固有値 λ_i が 0 または負であるとき、 $\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}$ が 0 または負になるような実ベクトル $\mathbf{a}$ が存在する（例えば $\mathbf{a} = \mathbf{u}_i$、すなわち $\hat{a}_i = 1, \hat{a}_j = 0\ (j \neq i)$）ことも明らかなので、任意の実ベクトル $\mathbf{a}$ について $\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}$ が正であれば、(2.45) で定義される $\boldsymbol{\Sigma}$ のすべての固有値 λ_i が正である。

2.21 （基本）実対称行列の独立なパラメータ数

$D \times D$ の実対称行列の (i, j) 成分を A_{ij} と書くことにする。この行列の D^2 個のパラメータのうち、$i = j$ となるパラメータは D 個、$i \neq j$ となるパラメータは $(D^2 - D)$ 個ある。この $i \neq j$ となるパラメータのうち、A_{ij} と A_{ji} は同じ値であるので、$i \neq j$ となるパラメータの中で独立なパラメータ数は $(D^2 - D)/2$ である。したがって、$D \times D$ の実対称行列の独立なパラメータ数は

$$D + (D^2 - D)/2 = D(D + 1)/2$$

2.22 （基本）対称行列の逆行列

対称行列 $\mathbf{M}$ の逆行列を考える。逆行列の定義から

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$$

この式の左辺の転置行列を考えると

$$(\mathbf{M}\mathbf{M}^{-1})^{\mathrm{T}} = (\mathbf{M}^{-1})^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} = (\mathbf{M}^{-1})^{\mathrm{T}}\mathbf{M}$$

単位行列の転置行列は単位行列なので $(\mathbf{M}^{-1})^{\mathrm{T}}$ もまた $\mathbf{M}$ の逆行列である。すなわち $(\mathbf{M}^{-1})^{\mathrm{T}} = \mathbf{M}^{-1}$ であり、これは $\mathbf{M}^{-1}$ が対称行列であることを意味している。

2.23 （標準）マハラノビス距離が一定以下の領域の体積

(2.52) を用いて変数変換を行うと、マハラノビス距離が一定のとなる点の集合はテキスト図 2.7 のとおり、原点を中心として座標軸 y_i の方向（$\mathbf{u}_i$ 方向）に軸を持つ超楕円面となる。（この変換はヤコビアンが 1 なの

で、体積の計算には影響がないことに注意。）そして、その超楕円面の各軸の半径は $\lambda_i^{1/2}\Delta$ である。ここで、$z_i = y_i/(\lambda_i^{1/2}\Delta)$ とおいて変数変換するとこの楕円面は超単位球面となる。この変換に関するヤコビアンは

$$\left|\frac{\partial \mathbf{y}}{\partial \mathbf{z}}\right| = \prod_{i=1}^{D}(\lambda_i^{1/2}\Delta) = \left(\prod_{i=1}^{D}\lambda_i^{1/2}\right)\Delta^D = |\mathbf{\Sigma}|^{1/2}\Delta^D$$

である。なお、最後に式変形では (2.55) を利用した。求める超楕円体内部の体積は、超単位球の体積にヤコビアンをかけることによって求められるので、その体積は $V_D|\mathbf{\Sigma}|^{1/2}\Delta^D$ である。

2.24 （標準）分割された行列の逆行列

(2.76) の右辺に (2.287) を左からかけると

$$\begin{aligned}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{M} & -\mathbf{MBD}^{-1}\\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{CMBD}^{-1}\end{pmatrix} &= \begin{pmatrix}(\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C})\mathbf{M} & (-\mathbf{AM}+\mathbf{I}+\mathbf{BD}^{-1}\mathbf{CM})\mathbf{BD}^{-1}\\ \mathbf{CM}-\mathbf{CM} & -\mathbf{CMBD}^{-1}+\mathbf{I}+\mathbf{CMBD}^{-1}\end{pmatrix}\\ &= \begin{pmatrix}(\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C})\mathbf{M} & \{\mathbf{I}-(\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C})\mathbf{M}\}\mathbf{BD}^{-1}\\ \mathbf{O} & \mathbf{I}\end{pmatrix}\end{aligned}$$

(2.77) から $(\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C})\mathbf{M} = \mathbf{I}$ であるので

$$\begin{aligned}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{M} & -\mathbf{MBD}^{-1}\\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{CMBD}^{-1}\end{pmatrix} &= \begin{pmatrix}\mathbf{I} & (\mathbf{I}-\mathbf{I})\mathbf{BD}^{-1}\\ \mathbf{O} & \mathbf{I}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{I} & \mathbf{O}\\ \mathbf{O} & \mathbf{I}\end{pmatrix}\end{aligned}$$

となり、計算結果が単位行列であることが確認できる。これにより、(2.76) が示せた。

2.25 （標準）多変量ガウス分布の周辺分布と条件付き分布

$\mathbf{x}_c$ を周辺消去したガウス関数の周辺分布の平均と共分散行列は (2.98) から以下のようになる。

$$\text{平均}: \begin{pmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{pmatrix}, \quad \text{共分散行列}: \begin{pmatrix}\mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab}\\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb}\end{pmatrix}$$

したがって条件付き分布 $p(\mathbf{x}_a|\mathbf{x}_b)$ は (2.96) そのものであり、

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a \,|\, \boldsymbol{\mu}_{a|b}, \mathbf{\Lambda}_{aa}^{-1})$$

ただし、$\mathbf{\Lambda}_{aa}, \boldsymbol{\mu}_{a|b}$ は以下のように定義する。

$$\begin{aligned}\begin{pmatrix}\mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab}\\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb}\end{pmatrix} &= \begin{pmatrix}\mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab}\\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb}\end{pmatrix}^{-1}\\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_a)\end{aligned}$$

2.26 （標準）Woodbury 行列反転公式

(2.289) の左辺に $(\mathbf{A}+\mathbf{BCD})$ を左からかけると

$$\begin{aligned}&(\mathbf{A}+\mathbf{BCD})\{\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}\}\\ &= \mathbf{I}-\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}+\mathbf{BCDA}^{-1}-\mathbf{BCDA}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}\\ &= \mathbf{I}-\mathbf{B}\{(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}-\mathbf{C}+\mathbf{CDA}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\}\mathbf{DA}^{-1}\\ &= \mathbf{I}-\mathbf{B}\{(\mathbf{I}+\mathbf{CDA}^{-1}\mathbf{B})(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}-\mathbf{C}\}\mathbf{DA}^{-1}\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{I} - \mathbf{B}\{\mathbf{C}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1} - \mathbf{C}\}\mathbf{D}\mathbf{A}^{-1} \\
&= \mathbf{I} - \mathbf{B}(\mathbf{C} - \mathbf{C})\mathbf{D}\mathbf{A}^{-1} \\
&= \mathbf{I}
\end{aligned}
$$

となるので、$(\mathbf{A} + \mathbf{BCD})$ の逆行列が (2.289) で表されることが示された。

2.27 （基本）独立な確率変数ベクトルの平均と共分散行列

確率変数ベクトルの各要素について考えると、$y_i = x_i + z_i$ の平均は x_i, z_i それぞれの平均の和であるので、$\mathbf{y} = \mathbf{x} + \mathbf{z}$ の平均は $\mathbf{x}, \mathbf{z}$ それぞれの平均の和になる。

$y_i = x_i + z_i$ の分散は x_i, z_i が独立なら、x_i, z_i それぞれの平均の和となる。また、$i \neq j$ のとき $y_i = x_i + z_i$ と $y_j = x_j + z_j$ の共分散は $\mathbf{x}, \mathbf{z}$ が独立であれば $\mathrm{cov}[x_i, z_j] = \mathrm{cov}[z_i, x_j] = 0$ なので

$$
\begin{aligned}
\mathrm{cov}[x_i + z_i, x_j + z_j] &= \mathbb{E}[(x_i + z_i)(x_j + z_j)] - \mathbb{E}[x_i + z_i]\mathbb{E}[x_j + z_j] \\
&= \mathbb{E}[x_i x_j + x_i z_j + z_i x_j + + z_i z_j] - (\mathbb{E}[x_i] + \mathbb{E}[z_i])(\mathbb{E}[x_j] + \mathbb{E}[z_j]) \\
&= (\mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j]) + (\mathbb{E}[x_i z_j] - \mathbb{E}[x_i]\mathbb{E}[z_j]) \\
&\quad + (\mathbb{E}[z_i x_j] - \mathbb{E}[z_i]\mathbb{E}[x_j]) + (\mathbb{E}[z_i z_j] - \mathbb{E}[z_i]\mathbb{E}[z_j]) \\
&= \mathrm{cov}[x_i, x_j] + \mathrm{cov}[x_i, z_j] + \mathrm{cov}[z_i, x_j] + \mathrm{cov}[z_i, z_j] \\
&= \mathrm{cov}[x_i, x_j] + \mathrm{cov}[z_i, z_j]
\end{aligned}
$$

したがって、分散・共分散ともに $\mathbf{x}, \mathbf{z}$ のそれぞれの分散・共分散の和になる。つまり、共分散行列はそれぞれの共分散行列の和になる。

2.28 （難問）結合されたガウス分布の周辺分布と条件付き分布

周辺分布 $p(\mathbf{x})$ について、(2.92) から平均が $\boldsymbol{\mu}$ であり、(2.93) から共分散行列が $\boldsymbol{\Lambda}^{-1}$ であることがわかるので、(2.99) は明らか。

条件付き分布 $p(\mathbf{y}|\mathbf{x})$ について、(2.81) から平均は

$$
\begin{aligned}
\mathbb{E}[\mathbf{y}|\mathbf{x}] &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\Lambda}^{-1})^{-1}(\mathbf{x} - \boldsymbol{\mu}) \\
&= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} + \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) \\
&= \mathbf{A}\mathbf{x} + \mathbf{b}
\end{aligned}
$$

(8.82) から共分散行列は

$$
\begin{aligned}
\mathrm{cov}[\mathbf{y}|\mathbf{x}] &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\
&= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\
&= \mathbf{L}^{-1}
\end{aligned}
$$

となり、(2.99) が示された。

2.29 （標準）ガウス分布とガウス分布の条件付き分布との同時分布の共分散行列

(2.77) に (2.104) を適用すると

$$
\mathbf{M} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A})^{-1} = \boldsymbol{\Lambda}^{-1}
$$

したがって (2.76) を用いて (2.104) の逆行列を求めると

$$
\begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} \\ \mathbf{L}^{-1}\mathbf{L}\mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{L}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} \end{pmatrix}
$$

$$= \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix}$$

2.30 （基本）ガウス分布とガウス分布の条件付き分布との同時分布の平均

問題の指示どおり (2.107) に (2.105) を適用すると

$$\begin{aligned}
\mathbb{E}[\mathbf{z}] &= \mathbf{R}^{-1} \begin{pmatrix} \mathbf{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{\Lambda}^{-1}(\mathbf{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}) + \mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b} \\ \mathbf{A}\mathbf{\Lambda}^{-1}(\mathbf{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}) + (\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})\mathbf{L}\mathbf{b} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}
\end{aligned}$$

2.31 （標準）線形ガウス分布を用いた和の周辺分布の算出

$\mathbf{x}$ を固定したとき、$\mathbf{y}$ の平均は $\mathbf{x} + \boldsymbol{\mu}_{\mathbf{z}}$、共分散行列は $\mathbf{\Sigma}_{\mathbf{z}}$ となる。すなわち $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{x} + \boldsymbol{\mu}_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}})$ であり、これは平均が $\mathbf{x}$ の線形関数で、共分散は $\mathbf{x}$ に独立なガウス分布である。したがって $p(\mathbf{y})$ の平均は (2.109) から

$$\mathbb{E}[\mathbf{y}] = \mathbf{I}\boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{z}} = \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{z}}$$

$p(\mathbf{y})$ の共分散行列は (2.110) から

$$\mathrm{cov}[\mathbf{y}] = \mathbf{\Sigma}_{\mathbf{z}} + \mathbf{I}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{I}^{\mathrm{T}} = \mathbf{\Sigma}_{\mathbf{x}} + \mathbf{\Sigma}_{\mathbf{z}}$$

2.32 （難問）線形ガウス分布の平方完成 1

$p(\mathbf{x}, \mathbf{y})$ は、$p(\mathbf{x})$ と $p(\mathbf{y}|\mathbf{x})$ の積で表されるので、周辺分布 $p(\mathbf{y})$ は $\int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x}$ で得られる。そこでまず、$p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$ の指数部分のみを考える。この指数部分は次のように表される。

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})$$

これを $\mathbf{x}$ で積分するために $\mathbf{x}$ について平方完成する。

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{\Lambda}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{x} \\
&\quad -\frac{1}{2}(\mathbf{A}\mathbf{x})^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}(\mathbf{A}\mathbf{x})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \frac{1}{2}(\mathbf{y} - \mathbf{b})^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{x} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b} + \frac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathrm{const} \\
&= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{\Lambda}\boldsymbol{\mu} + \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b} + \mathrm{const} \\
&= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\mathbf{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x} + \mathbf{x}^{\mathrm{T}}\{\mathbf{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})\} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b} + \mathrm{const} \\
&= -\frac{1}{2}(\mathbf{x} - \mathbf{m})^{\mathrm{T}}\mathbf{M}(\mathbf{x} - \mathbf{m}) + \frac{1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{M}\mathbf{m} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b} + \mathrm{const}
\end{aligned}$$

ただし、const は $\mathbf{x}, \mathbf{y}$ に無関係な定数、$\mathbf{M}, \mathbf{m}$ は以下の式で定義される値である。

$$\mathbf{M} = \mathbf{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A}$$

$$\mathbf{m} = \mathbf{M}^{-1}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})\}$$

ここで $p(\mathbf{y})$ を求めるため、この式の指数をとって $\mathbf{x}$ で積分をすると $\exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{m})\}$ は確率変数 $\mathbf{x}$ のガウス分布であるので

$$\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= C_1 \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{m}) + \frac{1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{M}\mathbf{m} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b}\right\}\mathrm{d}\mathbf{x} \\
&= C_1 \exp\left\{\frac{1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{M}\mathbf{m} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b}\right\} \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{m})\right\}\mathrm{d}\mathbf{x} \\
&= C_2 \exp\left\{\frac{1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{M}\mathbf{m} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b}\right\} \qquad (34)
\end{aligned}$$

ただし、C_i は分布を正規化するための定数である。つまり、この (34) の指数部分を平方完成すれば周辺分布 $p(\mathbf{y})$ の平均と分散が求められる。ここで、$\boldsymbol{\Lambda}, \mathbf{L}$ は対称行列なので

$$\mathbf{M}^{\mathrm{T}} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{\mathrm{T}} = \boldsymbol{\Lambda}^{\mathrm{T}} + \mathbf{A}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{A} = \boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} = \mathbf{M}$$

であるので、(34) の指数部分の最初の項は

$$\begin{aligned}
\frac{1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{M}\mathbf{m} &= \frac{1}{2}\left[\mathbf{M}^{-1}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})\}\right]^{\mathrm{T}}\mathbf{M}\mathbf{M}^{-1}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})\} \\
&= \frac{1}{2}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})\}^{\mathrm{T}}\mathbf{M}^{-1}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})\} \\
&= \frac{1}{2}\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{A} - \mathbf{b}^{\mathrm{T}}\mathbf{L}\mathbf{A}\}\mathbf{M}^{-1}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{y} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}\} \\
&= \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{M}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}) + \mathrm{const}
\end{aligned}$$

したがって (34) の指数部分を平方完成すると

$$\begin{aligned}
&\frac{1}{2}\mathbf{m}^{\mathrm{T}}\mathbf{M}\mathbf{m} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b} \\
&\quad= \frac{1}{2}\mathbf{y}^{\mathrm{T}}(\mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L} - \mathbf{L})\mathbf{y} + \mathbf{y}^{\mathrm{T}}\{\mathbf{L}\mathbf{A}\mathbf{M}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}) + \mathbf{L}\mathbf{b}\} + \mathrm{const} \\
&\quad= -\frac{1}{2}\mathbf{y}^{\mathrm{T}}(\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})\mathbf{y} + \mathbf{y}^{\mathrm{T}}\{\mathbf{L}\mathbf{A}\mathbf{M}^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} + (\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})\mathbf{b}\} + \mathrm{const} \\
&\quad= -\frac{1}{2}(\mathbf{y}-\mathbf{n})^{\mathrm{T}}(\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})(\mathbf{y}-\mathbf{n}) + \mathrm{const}
\end{aligned}$$

ただし、$\mathbf{n}$ は以下の式で定義される値である。

$$\mathbf{n} = (\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})^{-1}\{\mathbf{L}\mathbf{A}\mathbf{M}^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} + (\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})\mathbf{b}\} \qquad (35)$$

つまり、$p(\mathbf{y})$ の平均は $\mathbf{n}$、共分散は $(\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})^{-1}$ である。まずは共分散から計算する。Woodbury 行列反転公式 (2.289) を使うと

$$\begin{aligned}
(\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})^{-1} &= \{\mathbf{L} + (\mathbf{L}\mathbf{A})(-\mathbf{M}^{-1})(\mathbf{A}^{\mathrm{T}}\mathbf{L})\}^{-1} \\
&= \mathbf{L}^{-1} - \mathbf{L}^{-1}\mathbf{L}\mathbf{A}(-\mathbf{M} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} \\
&= \mathbf{L}^{-1} - \mathbf{A}(-\mathbf{M} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}} \\
&= \mathbf{L}^{-1} - \mathbf{A}(-(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A}) + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}} \\
&= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \qquad (36)
\end{aligned}$$

となり、(2.110) と一致していることが確認できる。次に平均を計算する。まず、(36) を使うと、

$$\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L} = (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1} \qquad (37)$$

$$\mathbf{LAM}^{-1} = \{\mathbf{L} - (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}\}\mathbf{L}^{-1}(\mathbf{A}^{\mathrm{T}})^{-1} \tag{38}$$

(36)、(37)、(38) を使うと (35) は

$$\begin{aligned}
&(\mathbf{L} - \mathbf{LAM}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})^{-1}[\mathbf{LAM}^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} + (\mathbf{L} - \mathbf{LAM}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L})\mathbf{b}] \\
&\quad = (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})[\{\mathbf{L} - (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}\}\mathbf{L}^{-1}(\mathbf{A}^{\mathrm{T}})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} + (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b}] \\
&\quad = (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})\{\mathbf{L} - (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}\}\mathbf{L}^{-1}(\mathbf{A}^{\mathrm{T}})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{b} \\
&\quad = (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})(\mathbf{A}^{\mathrm{T}})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{L}^{-1}(\mathbf{A}^{\mathrm{T}})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{b} \\
&\quad = \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{b} \\
&\quad = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}
\end{aligned}$$

となり、(2.109) と一致していることが確認できる。

2.33 （難問）線形ガウス分布の平方完成 2

$p(\mathbf{x}|\mathbf{y})$ は、$p(\mathbf{x})p(\mathbf{y}|\mathbf{x})/p(\mathbf{y})$ で得られる。そこでまず、$p(\mathbf{x})p(\mathbf{y}|\mathbf{x})/p(\mathbf{y})$ の指数部分のみを考える。この指数部分は前問の結果も用いると次のように表される。

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})-\frac{1}{2}(\mathbf{y}-\mathbf{Ax}-\mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{Ax}-\mathbf{b})+\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\mathrm{T}}(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})$$

これを $\mathbf{x}$ について平方完成するのだが、この式の 3 項目は $\mathbf{x}$ には無関係なので、最初の 2 つの項のみを考えればそれでよい。この平方完成は前問で計算しており、結果は次のようになる。

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})-\frac{1}{2}(\mathbf{y}-\mathbf{Ax}-\mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{Ax}-\mathbf{b}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{m}) + \mathrm{const}$$

ただし、const は $\mathbf{x}$ に無関係な定数、$\mathbf{M}, \mathbf{m}$ は以下の式で定義される値である。

$$\begin{aligned}
\mathbf{M} &= \boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA} \\
\mathbf{m} &= (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})\}
\end{aligned}$$

ここで、この分布は平均 $\mathbf{m}$、共分散 $\mathbf{M}^{-1}$ のガウス分布であり、(2.111) および (2.112) と一致することが確認できる。

2.34 （標準）ガウス分布の共分散の最尤推定

(2.118) を $\boldsymbol{\Sigma}$ で微分する。

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{\partial}{\partial\boldsymbol{\Sigma}}\left\{-\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right\} \\
&= -\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\sum_{n=1}^{N}\mathrm{Tr}\{(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\} \\
&= -\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\sum_{n=1}^{N}\mathrm{Tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\} \\
&= -\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathrm{Tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\}
\end{aligned} \tag{39}$$

ただし、$\mathbf{S}$ は以下の式で与えられる行列である。

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}$$

(39) の最初の項は (C.28) を使えば微分できる。2 番目の項については行列の各成分を書き下してみる。$\boldsymbol{\Sigma}$ の (i,j) 成分を Σ_{ij} と書くと、2 番目の項の微分部分の (i,j) 成分は

$$\frac{\partial}{\partial \Sigma_{ij}} \mathrm{Tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\} = \mathrm{Tr}\left\{\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \Sigma_{ij}}\mathbf{S}\right\} \tag{40a}$$

$$= \mathrm{Tr}\left\{-\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\} \tag{40b}$$

$$= -\mathrm{Tr}\left\{\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right\} \tag{40c}$$

なお、(40b) に至る計算では (C.21) を使っている。ここで、$\partial\boldsymbol{\Sigma}/\partial\Sigma_{ij}$ は (i,j) 成分のみが 1 で、その他の成分が 0 の行列であるから $\mathrm{Tr}(\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}\mathbf{A}) = A_{ji}$ となるので、

$$\frac{\partial}{\partial \Sigma_{ij}} \mathrm{Tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\} = -(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})_{ji}$$

これを行列にまとめると次のようになる。

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \mathrm{Tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\} = -(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}$$

この結果と (C.28) を使うと (39) は

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}(\boldsymbol{\Sigma}^{-1})^{\mathrm{T}} + \frac{N}{2}(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}$$

となる。この値を 0 とおくと

$$(\boldsymbol{\Sigma}^{-1})^{\mathrm{T}} = (\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}$$

両辺の転置行列をとり、左右からそれぞれ $\boldsymbol{\Sigma}$ をかけると

$$\boldsymbol{\Sigma} = \mathbf{S}$$

が得られる。なお、この最尤解の導出において $\boldsymbol{\Sigma}$ が対称行列であることを仮定していないことに注意。$\boldsymbol{\Sigma}$ が対称行列であることを仮定するなら、上記の偏微分における独立パラメータ数は $D \times D$ ではなく、$D(D-1)/2$ として考える必要があり、(C.28) は成り立たない。

2.35 （標準）ガウス分布の分散の最尤推定値の平均

(2.62) の証明についてはテキストに証明があるので割愛。

(2.291) について証明する。$n = m$ のときは (2.62) を使うと

$$\mathbb{E}[\mathbf{x}_n\mathbf{x}_m^{\mathrm{T}}] = \mathbb{E}[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} \tag{41}$$

$n \neq m$ のときは $\mathbf{x}_n, \mathbf{x}_m$ は独立なので (2.59) を使うと

$$\mathbb{E}[\mathbf{x}_n\mathbf{x}_m^{\mathrm{T}}] = \mathbb{E}[\mathbf{x}_n]\mathbb{E}[\mathbf{x}_m] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \tag{42}$$

(41) と (42) を一つにまとめると (2.291) が得られる。

次に (2.122) の平均を計算すると

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\{\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}_n\boldsymbol{\mu}_{\mathrm{ML}}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}\mathbf{x}_n^{\mathrm{T}}] + \mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}\boldsymbol{\mu}_{\mathrm{ML}}^{\mathrm{T}}]\}$$

ここで、和の内部の各項を (2.291) を使って計算すると、

$$
\begin{aligned}
\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}] &= \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} \\
\mathbb{E}[\mathbf{x}_n\boldsymbol{\mu}_{\mathrm{ML}}^{\mathrm{T}}] &= \mathbb{E}\left[\mathbf{x}_n\left(\frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N}\sum_{m=1}^{N}\mathbb{E}[\mathbf{x}_n\mathbf{x}_m^{\mathrm{T}}] \\
&= \frac{1}{N}(N\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}) \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma} \\
\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}\mathbf{x}_n^{\mathrm{T}}] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m\right)\mathbf{x}_n^{\mathrm{T}}\right] \\
&= \frac{1}{N}\sum_{m=1}^{N}\mathbb{E}[\mathbf{x}_m\mathbf{x}_n^{\mathrm{T}}] \\
&= \frac{1}{N}(N\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}) \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma} \\
\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}\boldsymbol{\mu}_{\mathrm{ML}}^{\mathrm{T}}] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m\right)\left(\frac{1}{N}\sum_{l=1}^{N}\mathbf{x}_l^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbb{E}[\mathbf{x}_m\mathbf{x}_l^{\mathrm{T}}] \\
&= \frac{1}{N^2}(N^2\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + N\boldsymbol{\Sigma}) \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}
\end{aligned}
$$

したがって

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{1}{N}\sum_{n=1}^{N}\left\{(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}) - \left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right) - \left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right) + \left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right)\right\} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{N-1}{N}\boldsymbol{\Sigma}\right\} \\
&= \frac{N-1}{N}\boldsymbol{\Sigma}
\end{aligned}
$$

となり、(2.124) が示された。

2.36 （標準）1 変数ガウス分布の分散の逐次推定

(2.292) を逐次推定の式に書き改めると

$$
\begin{aligned}
\sigma_{(N)}^2 &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2 \\
&= \frac{1}{N}(x_N - \mu)^2 + \frac{1}{N}\sum_{n=1}^{N-1}(x_n - \mu)^2 \\
&= \frac{1}{N}(x_N - \mu)^2 + \frac{N-1}{N}\sigma_{(N-1)}^2
\end{aligned}
$$

$$= \sigma^2_{(N-1)} + \frac{1}{N}\{(x_N - \mu)^2 - \sigma^2_{(N-1)}\} \tag{43}$$

(2.135) にガウス分布を代入し分散の逐次推定式をつくると

$$\begin{aligned}
\sigma^2_{(N)} &= \sigma^2_{(N-1)} - a_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}\left[-\ln\mathcal{N}(x_N|\mu,\sigma^2)\right] \\
&= \sigma^2_{(N-1)} - a_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}\left[\frac{1}{2}\ln(2\pi\sigma^2_{(N-1)}) + \frac{1}{2\sigma^2_{(N-1)}}(x_N-\mu)^2\right] \\
&= \sigma^2_{(N-1)} - a_{N-1}\left\{\frac{1}{2\sigma^2_{(N-1)}} - \frac{1}{2\sigma^4_{(N-1)}}(x_N-\mu)^2\right\} \\
&= \sigma^2_{(N-1)} + \frac{a_{N-1}}{2\sigma^4_{(N-1)}}\{(x_N-\mu)^2 - \sigma^2_{(N-1)}\}
\end{aligned} \tag{44}$$

となり、(43) と同じ形式になることが確認できる。(43) と (44) を比較すると次式が得られる。

$$a_{N-1} = \frac{2\sigma^4_{(N-1)}}{N}$$

2.37 （標準）多変量ガウス分布の分散の逐次推定

(2.122) を逐次推定の式に書き改めると

$$\begin{aligned}
\boldsymbol{\Sigma}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \frac{1}{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} + \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \frac{1}{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} + \frac{N-1}{N}\boldsymbol{\Sigma}^{(N-1)} \\
&= \boldsymbol{\Sigma}^{(N-1)} + \frac{1}{N}\{(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} - \boldsymbol{\Sigma}^{(N-1)}\}
\end{aligned} \tag{45}$$

ここで、$\boldsymbol{\Sigma}^{(N)}$ の (i,j) 成分を $\Sigma^{(N)}_{ij}$ と表し、(2.135) に多変量ガウス分布を代入し分散の逐次推定式をつくると

$$\begin{aligned}
\Sigma^{(N)}_{ij} &= \Sigma^{(N-1)}_{ij} - a_{N-1}\frac{\partial}{\partial\Sigma^{(N-1)}_{ij}}\left[-\ln\mathcal{N}(\mathbf{x}_N|\boldsymbol{\mu},\boldsymbol{\Sigma}^{(N-1)})\right] \\
&= \Sigma^{(N-1)}_{ij} - a_{N-1}\frac{\partial}{\partial\Sigma^{(N-1)}_{ij}}\left[\frac{1}{2}\ln|\boldsymbol{\Sigma}^{(N-1)}| + \frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right] \\
&= \Sigma^{(N-1)}_{ij} - \frac{1}{2}a_{N-1}\left[\frac{\partial}{\partial\Sigma^{(N-1)}_{ij}}\ln|\boldsymbol{\Sigma}^{(N-1)}| + \frac{\partial}{\partial\Sigma^{(N-1)}_{ij}}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right] \\
&= \Sigma^{(N-1)}_{ij} - \frac{1}{2}a_{N-1}\left[\mathrm{Tr}\left\{(\boldsymbol{\Sigma}^{(N-1)})^{-1}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma^{(N-1)}_{ij}}\right\} + (\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\frac{\partial(\boldsymbol{\Sigma}^{(N-1)})^{-1}}{\partial\Sigma^{(N-1)}_{ij}}(\mathbf{x}_n-\boldsymbol{\mu})\right] \\
&= \Sigma^{(N-1)}_{ij} - \frac{1}{2}a_{N-1}\left[\mathrm{Tr}\left\{(\boldsymbol{\Sigma}^{(N-1)})^{-1}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma^{(N-1)}_{ij}}\right\} + \mathrm{Tr}\left\{(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\frac{\partial(\boldsymbol{\Sigma}^{(N-1)})^{-1}}{\partial\Sigma^{(N-1)}_{ij}}(\mathbf{x}_n-\boldsymbol{\mu})\right\}\right] \\
&= \Sigma^{(N-1)}_{ij} - \frac{1}{2}a_{N-1}\,\mathrm{Tr}\left[(\boldsymbol{\Sigma}^{(N-1)})^{-1}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma^{(N-1)}_{ij}}\right. \\
&\qquad\left. -(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma^{(N-1)}_{ij}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right]
\end{aligned}$$

$$
\begin{aligned}
&= \Sigma_{ij}^{(N-1)} - \frac{1}{2}a_{N-1}\,\mathrm{Tr}\left[(\boldsymbol{\Sigma}^{(N-1)})^{-1}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma_{ij}^{(N-1)}}\right.\\
&\qquad\left.-(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma_{ij}^{(N-1)}}\right]\\
&= \Sigma_{ij}^{(N-1)} + \frac{1}{2}a_{N-1}\,\mathrm{Tr}\left[\left\{(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}-(\boldsymbol{\Sigma}^{(N-1)})^{-1}\right\}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma_{ij}^{(N-1)}}\right]
\end{aligned}
$$

ここで、$\boldsymbol{\Sigma}^{(N-1)}$ は対称行列であるので、$\partial\boldsymbol{\Sigma}^{(N-1)}/\partial\Sigma_{ij}^{(N-1)}$ は (i,j) 成分と (j,i) 成分だけが 1 でそれ以外の成分が 0 の行列であることに注意すると（演習問題 2.34 とは違う！）

$$
\mathrm{Tr}\left[\mathbf{A}\frac{\partial\boldsymbol{\Sigma}^{(N-1)}}{\partial\Sigma_{ij}^{(N-1)}}\right] = \begin{cases} A_{ij}+A_{ji} & (i\neq j\ \text{のとき})\\ A_{ij} & (i=j\ \text{のとき})\end{cases}
$$

であるから、

$$
\Sigma_{ij}^{(N)} = \begin{cases} \Sigma_{ij}^{(N-1)} + a_{N-1}[(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}-(\boldsymbol{\Sigma}^{(N-1)})^{-1}]_{ij} & (i\neq j\ \text{のとき})\\ \Sigma_{ij}^{(N-1)} + \frac{1}{2}a_{N-1}[(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}^{(N-1)})^{-1}-(\boldsymbol{\Sigma}^{(N-1)})^{-1}]_{ij} & (i=j\ \text{のとき})\end{cases}
$$

この式と (45) を比較すると a_{N-1} は以下の演算子を表していると言える。

$$
\begin{cases} \dfrac{1}{N}(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\cdot)(\boldsymbol{\Sigma}^{(N-1)})^{-1} & (i\neq j, \text{すなわち異なる変数間の共分散を算出するとき})\\ \dfrac{2}{N}(\boldsymbol{\Sigma}^{(N-1)})^{-1}(\cdot)(\boldsymbol{\Sigma}^{(N-1)})^{-1} & (i=j, \text{すなわち特定の変数の分散を算出するとき})\end{cases}
$$

2.38 （基本）ガウス分布の平均に対するベイズ推論

事後分布を求めるために $p(\mathbf{x}|\mu)p(\mu)$ を計算する。

$$
\begin{aligned}
p(\mathbf{x}|\mu)p(\mu) &= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}\frac{1}{(2\pi\sigma_0^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}\\
&= \frac{1}{(2\pi\sigma^2)^{N/2}(2\pi\sigma_0^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}
\end{aligned}
$$

ここで重要になるのは指数の内部だけなので、そこだけに注目すると、

$$
\begin{aligned}
-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 &= -\left(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)\mu^2+\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n+\frac{\mu_0}{\sigma_0^2}\right)\mu+\mathrm{const}\\
&= -\left(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)\left\{\mu^2-\frac{2\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(\frac{N\mu_{\mathrm{ML}}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\mu\right\}+\mathrm{const}\\
&= -\left(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)\left\{\mu^2-\frac{2(\sigma^2\mu_0+N\sigma_0^2\mu_{\mathrm{ML}})}{N\sigma_0^2+\sigma^2}\mu\right\}+\mathrm{const}\\
&= -\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\left\{\mu-\frac{\sigma^2\mu_0+N\sigma_0^2\mu_{\mathrm{ML}}}{N\sigma_0^2+\sigma^2}\right\}^2+\mathrm{const}
\end{aligned}
$$

ただし、μ_{ML} は以下の式で表される μ の最尤推定解、すなわちサンプル平均である。

$$
\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n
$$

したがって、この事後分布の平均と分散の逆数は以下のようになる。

$$\mu_N = \frac{\sigma^2\mu_0 + N\sigma_0^2\mu_{\mathrm{ML}}}{N\sigma_0^2 + \sigma^2} = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}$$
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

2.39 （標準）ガウス分布の平均に対するベイズ推論の逐次更新

まず、(2.142) を逐次更新の式に改める。

$$\begin{aligned}\frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \\ &= \frac{1}{\sigma_0^2} + \frac{N-1}{\sigma^2} + \frac{1}{\sigma^2} \\ &= \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}\end{aligned} \tag{46}$$

次に (2.141) を逐次更新の式に改めるのだが、(2.142) から $N\sigma_0^2 + \sigma^2 = \sigma^2\sigma_0^2/\sigma_N^2$ であることに注意すると

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}} = \frac{\sigma_N^2}{\sigma_0^2}\mu_0 + \frac{N\sigma_N^2}{\sigma^2}\mu_{\mathrm{ML}}$$

この式の μ_{ML} を展開し逐次更新の式に改めると

$$\begin{aligned}\mu_N &= \frac{\sigma_N^2}{\sigma_0^2}\mu_0 + \frac{\sigma_N^2}{\sigma^2}\left(\sum_{n=1}^{N-1} x_n + x_N\right) \\ &= \frac{\sigma_N^2}{\sigma_0^2}\mu_0 + \frac{(N-1)\sigma_N^2}{\sigma^2}\frac{1}{N-1}\sum_{n=1}^{N-1} x_n + \frac{\sigma_N^2}{\sigma^2}x_N \\ &= \frac{\sigma_N^2}{\sigma_{N-1}^2}\left\{\frac{\sigma_{N-1}^2}{\sigma_0^2}\mu_0 + \frac{(N-1)\sigma_{N-1}^2}{\sigma^2}\frac{1}{N-1}\sum_{n=1}^{N-1} x_n\right\} + \frac{\sigma_N^2}{\sigma^2}x_N \\ &= \frac{\sigma_N^2}{\sigma_{N-1}^2}\mu_{N-1} + \frac{\sigma_N^2}{\sigma^2}x_N\end{aligned} \tag{47}$$

したがって、$N-1$ 個のデータを観測後の事後分布 $p(\mu|x_1, x_2, \ldots, x_{N-1})$ に尤度関数 $p(x_N|\mu)$ を掛けると

$$\begin{aligned}&p(\mu|x_1,x_2,\ldots,x_{N-1})p(x_N|\mu) \\ &\quad= \frac{1}{(2\pi\sigma_{N-1}^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma_{N-1}^2}(\mu-\mu_{N-1})^2\right\}\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x_N-\mu)^2\right\} \\ &\quad= \frac{1}{2\pi\sigma_{N-1}\sigma}\exp\left\{-\frac{1}{2\sigma_{N-1}^2}(\mu-\mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N-\mu)^2\right\}\end{aligned}$$

ここで重要になるのは指数の内部だけなので、そこに注目すると

$$\begin{aligned}&-\frac{1}{2\sigma_{N-1}^2}(\mu-\mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N-\mu)^2 \\ &\quad= -\frac{1}{2}\left(\frac{1}{\sigma_{N-1}^2}+\frac{1}{\sigma^2}\right)\mu^2 + \left(\frac{\mu_{N-1}}{\sigma_{N-1}^2}+\frac{x_N}{\sigma^2}\right)\mu + \mathrm{const} \\ &\quad= -\frac{1}{2}\left(\frac{1}{\sigma_{N-1}^2}+\frac{1}{\sigma^2}\right)\left\{\mu^2 - 2\left(\frac{1}{\sigma_{N-1}^2}+\frac{1}{\sigma^2}\right)^{-1}\left(\frac{\mu_{N-1}}{\sigma_{N-1}^2}+\frac{x_N}{\sigma^2}\right)\mu\right\} + \mathrm{const} \\ &\quad= -\frac{1}{2}\left(\frac{1}{\sigma_{N-1}^2}+\frac{1}{\sigma^2}\right)\left\{\mu - \left(\frac{1}{\sigma_{N-1}^2}+\frac{1}{\sigma^2}\right)^{-1}\left(\frac{1}{\sigma_{N-1}^2}\mu_{N-1}+\frac{1}{\sigma^2}x_N\right)\right\}^2 + \mathrm{const}\end{aligned}$$

となり、N 個のデータを観測後の事後分布はガウス分布であり、その分散は (46) と一致することが確認できる。平均については $\left(\frac{1}{\sigma^2_{N-1}}+\frac{1}{\sigma^2}\right)^{-1}$ が分散、すなわち σ^2_N であることに注意すると (47) に一致していることが判る。

2.40 （標準）多次元ガウス分布の平均に対するベイズ推論

事後分布 $p(\boldsymbol{\mu}|\mathbf{X})$ は事前分布 $p(\boldsymbol{\mu})$ と $p(\mathbf{X})$ の積に比例する。そこで、まず $p(\boldsymbol{\mu})$ と $p(\mathbf{X})$ を考えると

$$
\begin{aligned}
p(\boldsymbol{\mu}) &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}_0|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\}\\
p(\mathbf{X}) &= \prod_{n=1}^{N}\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right\}\right]\\
&= \frac{1}{(2\pi)^{DN/2}}\frac{1}{|\boldsymbol{\Sigma}|^{N/2}}\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right\}
\end{aligned}
$$

これらの確率の積は指数部にのみ $\boldsymbol{\mu}$ の二次形式を含むので、その積はガウス分布になる。そこで、これらの確率の積の指数の内部に注目すると

$$
\begin{aligned}
&-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\\
&\quad= -\frac{1}{2}\left\{\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}-2\boldsymbol{\mu}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\mathbf{x}_n\right)+\mathrm{const}\right\}\\
&\quad= -\frac{1}{2}\{\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}-2\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+N\boldsymbol{\Sigma}^{-1}\mathbf{x}_{\mathrm{ML}})+\mathrm{const}\}
\end{aligned}
$$

ただし $\mathbf{x}_{\mathrm{ML}}$ は $\boldsymbol{\mu}$ の最尤推定量、すなわち $\mathbf{x}_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ である。したがって、事後分布 $p(\boldsymbol{\mu}|\mathbf{X})$ は以下の平均・精度を持つ D 次元ガウス分布である。

$$
\begin{aligned}
\boldsymbol{\mu}_N &= (\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+N\boldsymbol{\Sigma}^{-1}\mathbf{x}_{\mathrm{ML}})\\
\boldsymbol{\Sigma}_N^{-1} &= \boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1}
\end{aligned}
$$

2.41 （基本）ガンマ分布の正規化

ガンマ分布 (2.146) をガンマ分布の定義域で積分する。

$$
\begin{aligned}
\int_0^\infty \mathrm{Gam}(\lambda|a,b)\,\mathrm{d}\lambda &= \int_0^\infty \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)\,\mathrm{d}\lambda\\
&= \frac{1}{\Gamma(a)}b^a\int_0^\infty \lambda^{a-1}\exp(-b\lambda)\,\mathrm{d}\lambda
\end{aligned}
$$

ここで、$\lambda=u/b$ の変数変換を行うと、

$$
\begin{aligned}
\frac{1}{\Gamma(a)}b^a\int_0^\infty \lambda^{a-1}\exp(-b\lambda)\,\mathrm{d}\lambda &= \frac{1}{\Gamma(a)}b^a\int_0^\infty \left(\frac{u}{b}\right)^{a-1}\exp(-u)\frac{1}{b}\,\mathrm{d}u\\
&= \frac{1}{\Gamma(a)}\int_0^\infty u^{a-1}\exp(-u)\,\mathrm{d}u
\end{aligned}
$$

この式に現れる積分は (1.141) から ガンマ関数 $\Gamma(a)$ そのものであるので、この式は 1 に等しい。つまり、ガンマ分布の定義 (2.146) は正規化されている。

2.42 （標準）ガンマ分布の平均・分散・モード

前問と同様に平均と分散は $\lambda = u/b$ の変数変換とガンマ関数の定義 (1.141) から以下のように計算できる。

$$\begin{aligned}
\mathbb{E}[\lambda] &= \int_0^\infty \frac{1}{\Gamma(a)} b^a \lambda^a \exp(-b\lambda)\,\mathrm{d}\lambda \\
&= \frac{1}{\Gamma(a)} b^a \int_0^\infty \left(\frac{u}{b}\right)^a \exp(-u)\frac{1}{b}\,\mathrm{d}u \\
&= \frac{1}{b\Gamma(a)} \int_0^\infty u^a \exp(-u)\,\mathrm{d}u \\
&= \frac{1}{b\Gamma(a)} \Gamma(a+1) \\
&= \frac{a}{b} \\
\mathrm{var}[\lambda] &= \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda]^2 \\
&= \int_0^\infty \frac{1}{\Gamma(a)} b^a \lambda^{a+1} \exp(-b\lambda)\,\mathrm{d}\lambda - \mathbb{E}[\lambda]^2 \\
&= \frac{1}{\Gamma(a)} b^a \int_0^\infty \left(\frac{u}{b}\right)^{a+1} \exp(-u)\frac{1}{b}\,\mathrm{d}u - \mathbb{E}[\lambda]^2 \\
&= \frac{1}{b^2\Gamma(a)} \int_0^\infty u^{a+1} \exp(-u)\,\mathrm{d}u - \mathbb{E}[\lambda]^2 \\
&= \frac{1}{b^2\Gamma(a)} \Gamma(a+2) - \mathbb{E}[\lambda]^2 \\
&= \frac{a(a+1)}{b^2} - \left(\frac{a}{b}\right)^2 \\
&= \frac{a}{b^2}
\end{aligned}$$

次にモードを求めるため、ガンマ分布を微分する。

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\left\{\frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)\right\} &= \frac{1}{\Gamma(a)} b^a \{(a-1)\lambda^{a-2}\exp(-b\lambda) - b\lambda^{a-1}\exp(-b\lambda)\} \\
&= \frac{1}{\Gamma(a)} b^a \lambda^{a-2}(a-1-b\lambda)\exp(-b\lambda)
\end{aligned}$$

これを 0 とおくことによりモードは以下のように求められる。

$$\mathrm{mode}[\lambda] = \frac{a-1}{b}$$

2.43 （基本）1 変数ガウス分布の一般化

(2.293) は偶関数なのでその積分は以下のように計算できる。

$$\begin{aligned}
\int_{-\infty}^\infty \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left(-\frac{|x|^q}{2\sigma^2}\right)\mathrm{d}x &= \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \int_0^\infty \exp\left(-\frac{x^q}{2\sigma^2}\right)\mathrm{d}x \\
&= \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \int_0^\infty \exp(-u)\frac{(2\sigma)^{1/q}u^{1/q-1}}{q}\,\mathrm{d}u \\
&= \frac{1}{\Gamma(1/q)} \int_0^\infty u^{1/q-1}\exp(-u)\,\mathrm{d}u \\
&= \frac{1}{\Gamma(1/q)}\Gamma(1/q) \\
&= 1
\end{aligned}$$

したがって、(2.293) で表される分布は正規化されている。$q = 2$ のとき、(2.293) は次のようになる。

$$\begin{aligned} p(x|\sigma^2, 2) &= \frac{2}{2(2\sigma^2)^{1/2}\Gamma(1/2)} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \end{aligned}$$

これは平均 0、分散 σ^2 のガウス分布である。

次に問題で考える対数尤度関数は

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) &= \sum_{n=1}^{N} \ln\left\{\frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left(-\frac{|y(\mathbf{x}_n, \mathbf{w}) - t_n|^q}{2\sigma^2}\right)\right\} \\ &= \sum_{n=1}^{N}\left(-\frac{|y(\mathbf{x}_n, \mathbf{w}) - t_n|^q}{2\sigma^2}\right) + \sum_{n=1}^{N} \ln \frac{1}{(2\sigma^2)^{1/q}} + \sum_{n=1}^{N} \ln \frac{q}{2\Gamma(1/q)} \\ &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q}\ln(2\sigma^2) + N \ln \frac{q}{2\Gamma(1/q)} \end{aligned}$$

となり、(2.295) が示せた。

2.44 （標準）ガウス分布の共役事前分布と事後分布

問題文には「1 変数ガウス分布 $\mathcal{N}(x_n|\mu, \tau^{-1})$」とあるが、テキスト 2.3.6 節にしたがって精度を λ とし、「$\mathcal{N}(x_n|\mu, \lambda^{-1})$」について考えることとする。

事後分布は $p(\mathbf{x}|\mu, \lambda)p(\mu, \lambda)$ に比例するので、これを計算する。ただし、μ, λ に無関係な定数の係数部分は無視して計算する。

$$\begin{aligned} & p(\mathbf{x}|\mu, \lambda)p(\mu, \lambda) \\ &= \left\{\prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1})\right\} \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1}) \mathrm{Gam}(\lambda|a, b) \\ &\propto \left[\lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}\right] \left[\lambda^{1/2} \exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\}\right] \left[\lambda^{a-1}\exp(-b\lambda)\right] \\ &= \lambda^{(N+1)/2+a-1} \exp\left[-\frac{\lambda}{2}\left\{(N+\beta)\mu^2 - 2\left(\sum_{n=1}^{N} x_n + \beta\mu_0\right)\mu + \sum_{n=1}^{N} x_n^2 + \beta\mu_0^2 + 2b\right\}\right] \\ &= \lambda^{(N+1)/2+a-1} \exp\left[-\frac{\lambda}{2}\{(N+\beta)\mu^2 - 2(N\mu_{\mathrm{ML}} + \beta\mu_0)\mu + N(\lambda_{\mathrm{ML}}^{-1} + \mu_{\mathrm{ML}}^2) + \beta\mu_0^2 + 2b\}\right] \end{aligned} \tag{48}$$

ただし、$\mu_{\mathrm{ML}}, \lambda_{\mathrm{ML}}$ はそれぞれ μ, λ の最尤推定解、すなわち、サンプル平均とサンプル精度である。

$$\begin{aligned} \mu_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\ \lambda_{\mathrm{ML}}^{-1} &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \end{aligned}$$

ここで (48) の指数の内部を平方完成すると次のようになる。

$$-\frac{(N+\beta)\lambda}{2}\left(\mu - \frac{N\mu_{\mathrm{ML}} + \beta\mu_0}{N+\beta}\right)^2 - \frac{\lambda}{2}\left\{N(\lambda_{\mathrm{ML}}^{-1} + \mu_{\mathrm{ML}}^2) + \beta\mu_0^2 + 2b - (N+\beta)\left(\frac{N\mu_{\mathrm{ML}} + \beta\mu_0}{N+\beta}\right)^2\right\}$$

したがって $\beta_N = N + \beta,\ \mu_N = (N\mu_{\mathrm{ML}} + \beta\mu_0)/\beta_N$ とおくと、(48) は

$$\begin{aligned}
&\lambda^{(N+1)/2+a-1} \exp\left[-\frac{\lambda}{2}\{(N+\beta)\mu^2 - 2(N\mu_{\mathrm{ML}} + \beta\mu_0)\mu + N(\lambda_{\mathrm{ML}}^{-1} + \mu_{\mathrm{ML}}^2) + \beta\mu_0^2 + 2b\}\right] \\
&= \lambda^{(N+1)/2+a-1} \exp\left\{-\frac{\beta_N\lambda}{2}(\mu - \mu_N)^2\right\} \exp\left\{-\frac{\lambda}{2}(N(\lambda_{\mathrm{ML}}^{-1} + \mu_{\mathrm{ML}}^2) + \beta\mu_0^2 + 2b - \beta_N\mu_N^2)\right\} \\
&= \left[\lambda^{1/2} \exp\left\{-\frac{\beta_N\lambda}{2}(\mu - \mu_N)^2\right\}\right] \left[\lambda^{N/2+a-1} \exp\left\{-\frac{1}{2}(N(\lambda_{\mathrm{ML}}^{-1} + \mu_{\mathrm{ML}}^2) + \beta\mu_0^2 + 2b - \beta_N\mu_N^2)\lambda\right\}\right] \\
&\propto \mathcal{N}(\mu|\mu_N, (\beta_N\lambda)^{-1})\mathrm{Gam}(\lambda|a + N/2, b + \{N(\lambda_{\mathrm{ML}}^{-1} + \mu_{\mathrm{ML}}^2) + \beta\mu_0^2 - \beta_N\mu_N^2\}/2)
\end{aligned}$$

となり、この分布がガウス-ガンマ分布であることが確認できる。各パラメータを比較すると、事後分布の各パラメータは以下のようになっている。

$$\begin{aligned}
\beta_N &= N + \beta \\
\mu_N &= \frac{N\mu_{\mathrm{ML}} + \beta\mu_0}{\beta_N} \\
a_N &= a + \frac{N}{2} \\
b_N &= b + \frac{N(\lambda_{\mathrm{ML}}^{-1} + \mu_{\mathrm{ML}}^2) + \beta\mu_0^2 - \beta_N\mu_N^2}{2}
\end{aligned}$$

2.45 （基本）平均が既知で精度が未知の多変量ガウス分布の共役事前分布

平均 $\boldsymbol{\mu}$ が既知で精度行列 $\boldsymbol{\Lambda}$ が未知の多変量ガウス分布から N 個の観測値集合 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ が得られる確率は次のようになる。

$$\begin{aligned}
p(\mathbf{X}|\boldsymbol{\Lambda}) &= \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\Lambda}) \\
&= \frac{1}{(2\pi)^{DN/2}} |\boldsymbol{\Lambda}|^{N/2} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu})\right\}
\end{aligned}$$

精度行列 $\boldsymbol{\Lambda}$ の事前分布が (2.155) で与えらるとすると、事後分布は事前分布と $p(\mathbf{X}|\boldsymbol{\Lambda})$ の積に比例するのでこの積を計算すると

$$\begin{aligned}
&p(\mathbf{X}|\boldsymbol{\Lambda})\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) \\
&\quad\propto |\boldsymbol{\Lambda}|^{N/2} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu})\right\} |\boldsymbol{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right) \\
&\quad= |\boldsymbol{\Lambda}|^{(\nu+N-D-1)/2} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right\} \\
&\quad= |\boldsymbol{\Lambda}|^{(\nu+N-D-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}\right) - \frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right\} \\
&\quad= |\boldsymbol{\Lambda}|^{(\nu+N-D-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda} + \mathbf{W}^{-1}\boldsymbol{\Lambda}\right)\right\} \\
&\quad= |\boldsymbol{\Lambda}|^{(\nu+N-D-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\left\{\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} + \mathbf{W}^{-1}\right\}\boldsymbol{\Lambda}\right)\right\}
\end{aligned}$$

となり、以下の $\mathbf{W}_N, \nu_N$ をパラメータとするウィシャート分布であることがわかる。

$$\mathbf{W}_N = \left\{\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} + \mathbf{W}^{-1}\right\}^{-1}$$

$$\nu_N = \nu + N$$

したがって、平均が既知で精度行列が未知の多変量ガウス分布の共役事前分布はウィシャート分布である。

2.46 （基本）スチューデントの t 分布

(2.158) 式を変形する。

$$\begin{aligned}
\int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|\mathrm{a},\mathrm{b})\,\mathrm{d}\tau &= \int_0^\infty \frac{b^a\tau^{a-1}}{\Gamma(a)}\left(\frac{\tau}{2\pi}\right)^{1/2}\exp\left[-\left\{b+\frac{(x-\mu)^2}{2}\right\}\tau\right]\mathrm{d}\tau \\
&= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\int_0^\infty \tau^{a-1/2}\exp\left[-\left\{b+\frac{(x-\mu)^2}{2}\right\}\tau\right]\mathrm{d}\tau \\
&= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left\{b+\frac{(x-\mu)^2}{2}\right\}^{-a-1/2}\int_0^\infty z^{a-1/2}\exp(-z)\,\mathrm{d}z \\
&= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left\{b+\frac{(x-\mu)^2}{2}\right\}^{-a-1/2}\Gamma(a+1/2) \\
&= \frac{\Gamma(a+1/2)}{\Gamma(a)}\left(\frac{1}{2\pi b}\right)^{1/2}\left\{1+\frac{(x-\mu)^2}{2b}\right\}^{-a-1/2}
\end{aligned}$$

ただし、途中 $z = \left\{b+(x-\mu)^2/2\right\}\tau$ の変数変換を行っている。ここで、$\nu = 2a,\ \lambda = a/b$ 、すなわち $a = \nu/2,\ b = \nu/2\lambda$ とおくと

$$\begin{aligned}
\int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|\mathrm{a},\mathrm{b})\,\mathrm{d}\tau &= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{1}{2\pi\nu/2\lambda}\right)^{1/2}\left\{1+\frac{(x-\mu)^2}{2\nu/2\lambda}\right\}^{-\nu/2-1/2} \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left\{1+\frac{\lambda(x-\mu)^2}{\nu}\right\}^{-\nu/2-1/2}
\end{aligned}$$

となり、(2.159) が導出された。

2.47 （基本）ガウス分布とスチューデントの t 分布の関係

(2.159) のうち、x に無関係な係数部分を無視して極限を考える。$z = \nu/\lambda(x-\mu)^2$ とおくと

$$\begin{aligned}
\left\{1+\frac{\lambda(x-\mu)^2}{\nu}\right\}^{-\nu/2-1/2} &= \left\{1+\frac{1}{z}\right\}^{-z\lambda(x-\mu)^2/2-1/2} \\
&= \left\{\left(1+\frac{1}{z}\right)^z\right\}^{-\lambda(x-\mu)^2/2}\left(1+\frac{1}{z}\right)^{-1/2}
\end{aligned}$$

ここで、$\nu\to\infty$ のとき $z\to\infty$ であり、$\lim_{z\to\infty}(1+1/z)^z = e$ であるから、$\nu\to\infty$ のとき

$$\left\{\left(1+\frac{1}{z}\right)^z\right\}^{-\lambda(x-\mu)^2/2}\left(1+\frac{1}{z}\right)^{-1/2} \to \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\}$$

となり、これは平均 μ 、分散 λ^{-1} のガウス分布の密度関数に比例していることが判る。すなわちスチューデントの t 分布の $\nu\to\infty$ における極限は平均 μ 、分散 λ^{-1} のガウス分布である。

2.48 （基本）多変量へのスチューデントの t 分布の拡張

(2.161) を計算すると

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta$$

$$
\begin{aligned}
&= \int_0^\infty \frac{(\nu/2)^{\nu/2}\eta^{\nu/2-1}}{\Gamma(\nu/2)} \frac{|\eta\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{\nu}{2}\eta - \frac{\eta}{2}\Delta^2\right) \mathrm{d}\eta \\
&= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)(2\pi)^{D/2}} |\boldsymbol{\Lambda}|^{1/2} \int_0^\infty \eta^{\nu/2+D/2-1} \exp\left\{-\left(\frac{\nu+\Delta^2}{2}\right)\eta\right\} \mathrm{d}\eta \\
&= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)(2\pi)^{D/2}} |\boldsymbol{\Lambda}|^{1/2} \left(\frac{\nu+\Delta^2}{2}\right)^{-\nu/2-D/2} \int_0^\infty z^{\nu/2+D/2-1} \exp\{-z\} \mathrm{d}z \\
&= \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)} \frac{(\nu/2)^{\nu/2}}{(2\pi)^{D/2}} |\boldsymbol{\Lambda}|^{1/2} (\nu/2)^{-\nu/2-D/2} \left(1+\frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2} \\
&= \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}} \left(1+\frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2}
\end{aligned}
$$

となり、(2.162) が示せた。ただし、上式において $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$ であり、途中 $z = (\nu+\Delta^2)\eta/2$ の変数変換を行っている。

次に、この式が正規化されていることを確認する。領域 $\mathcal{D}$ を $\mathbf{x}$ の定義域全体とすると $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1})$, $\mathrm{Gam}(\eta|\nu/2, \nu/2)$ はともに正規化されているので、

$$
\begin{aligned}
\int_{\mathcal{D}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \mathrm{d}\mathbf{x} &= 1 \\
\int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2) \mathrm{d}\eta &= 1
\end{aligned}
$$

であるから

$$
\begin{aligned}
\int_{\mathcal{D}} \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) \mathrm{d}\mathbf{x} &= \int_{\mathcal{D}} \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \mathrm{Gam}(\eta|\nu/2, \nu/2) \mathrm{d}\eta \mathrm{d}\mathbf{x} \\
&= \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2) \int_{\mathcal{D}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \mathrm{d}\mathbf{x} \mathrm{d}\eta \\
&= \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2) \mathrm{d}\eta \\
&= 1
\end{aligned}
$$

となり、(2.161) で表される分布も正規化されていることが示せた。

2.49 （標準）多変量スチューデントの t 分布の性質

まずは平均を定義にしたがって求める。

$$
\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \int_{\mathcal{D}} \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) \mathbf{x} \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{D}} \left\{ \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \mathrm{Gam}(\eta|\nu/2, \nu/2) \mathrm{d}\eta \right\} \mathbf{x} \mathrm{d}\mathbf{x} \\
&= \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2) \int_{\mathcal{D}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \mathbf{x} \mathrm{d}\mathbf{x} \mathrm{d}\eta \\
&= \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2) \boldsymbol{\mu} \mathrm{d}\eta \\
&= \boldsymbol{\mu} \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2) \mathrm{d}\eta \\
&= \boldsymbol{\mu}
\end{aligned}
$$

分散も定義から計算する。

$$
\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}]
$$

$$
\begin{aligned}
&= \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}] \\
&= \int_{\mathcal{D}} \left\{ \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \mathrm{Gam}(\eta|\nu/2, \nu/2)\, \mathrm{d}\eta \right\} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\, \mathrm{d}\mathbf{x} \\
&= \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2) \int_{\mathcal{D}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\eta \\
&= \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2)(\eta\boldsymbol{\Lambda})^{-1}\, \mathrm{d}\eta \\
&= \boldsymbol{\Lambda}^{-1} \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2)\eta^{-1}\, \mathrm{d}\eta \\
&= \boldsymbol{\Lambda}^{-1} \int_0^\infty \frac{(\nu/2)^{\nu/2}\eta^{\nu/2-2}}{\Gamma(\nu/2)} \exp(-\nu\eta/2)\, \mathrm{d}\eta \\
&= \boldsymbol{\Lambda}^{-1} \int_0^\infty \frac{(\nu/2)^{\nu/2}(2z/\nu)^{\nu/2-2}}{\Gamma(\nu/2)} \exp(-z)\, \mathrm{d}z \\
&= \frac{\nu/2}{\Gamma(\nu/2)} \boldsymbol{\Lambda}^{-1} \int_0^\infty z^{\nu/2-2} \exp(-z)\, \mathrm{d}z \\
&= \frac{\nu/2}{\Gamma(\nu/2)} \boldsymbol{\Lambda}^{-1} \Gamma(\nu/2-1) \\
&= \frac{\nu/2}{\Gamma(\nu/2)} \boldsymbol{\Lambda}^{-1} \frac{\Gamma(\nu/2)}{\nu/2-1} \\
&= \frac{\nu}{\nu-2} \boldsymbol{\Lambda}^{-1}
\end{aligned}
$$

なお、ここでも途中で $z = \nu\eta/2$ の変数変換を行っている。

最後にモードを求めるために (2.162) を $\mathbf{x}$ で微分するのだが、微分係数が $\mathbf{0}$ となる $\mathbf{x}$ を求めるのであるから、(2.162) のうち $\mathbf{x}$ に無関係な係数を無視すると

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) &\propto \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} \left(1 + \frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2} \\
&= -\frac{\nu+D}{2} \left(1 + \frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2-1} \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} \left(1 + \frac{\Delta^2}{\nu}\right) \\
&\propto \left(1 + \frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2-1} \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} \Delta^2
\end{aligned}
$$

ここで、$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$ の変数変換を行うと、$\Delta^2 = \mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{y}$ であるから

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) &\propto \left(1 + \frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2-1} \frac{\mathrm{d}}{\mathrm{d}\mathbf{y}} \mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{y} \\
&= \left(1 + \frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2-1} \boldsymbol{\Lambda}\mathbf{y}
\end{aligned}
$$

これを $\mathbf{0}$ にするような $\mathbf{y}$ は $\mathbf{y} = \mathbf{0}$、すなわち、$\mathbf{x} = \boldsymbol{\mu}$ のときである。したがって、

$$
\mathrm{mode}[\mathbf{x}] = \boldsymbol{\mu}
$$

2.50 （基本）多変量スチューデントの t 分布と多変量ガウス分布の関係

(2.162) のうち、x に無関係な係数部分を無視して極限を考える。$z = \nu/\Delta^2$ とおくと

$$
\left(1 + \frac{\Delta^2}{\nu}\right)^{-\nu/2-D/2} = \left\{1 + \frac{1}{z}\right\}^{-z\Delta^2/2-D/2}
$$

$$
\begin{aligned}
&= \left\{\left(1+\frac{1}{z}\right)^{z}\right\}^{-\Delta^2/2}\left(1+\frac{1}{z}\right)^{-D/2} \\
&\to \exp\left\{-\frac{\Delta^2}{2}\right\} \qquad (z\to\infty \text{ のとき})
\end{aligned}
$$

となり、これは平均 $\boldsymbol{\mu}$ 、精度 $\boldsymbol{\Lambda}$ の多変量ガウス分布の密度関数に比例していることが判る。すなわち多変量スチューデントの t 分布の $\nu\to\infty$ における極限は平均 $\boldsymbol{\mu}$ 、精度 $\boldsymbol{\Lambda}$ のガウス分布である。

2.51 （基本）三角関数の公式

(2.297) の左辺に (2.296) を適用すると

$$
\begin{aligned}
\exp(iA)\exp(-iA) &= (\cos A + i\sin A)(\cos A - i\sin A) \\
&= \cos^2 A + \sin^2 A
\end{aligned}
$$

したがって (2.177) が示せた。次に、(2.298) の右辺を (2.296) を使って変形する。

$$
\begin{aligned}
\Re\exp\{i(A-B)\} &= \Re\{\exp(iA)\exp(-iB)\} \\
&= \Re\{(\cos A + i\sin A)(\cos B - i\sin B)\} \\
&= \Re\{(\cos A\cos B + \sin A\sin B) + i(\sin A\cos B - \cos A\sin B)\} \\
&= \cos A\cos B + \sin A\sin B
\end{aligned}
$$

したがって (2.178) が示せた。最後に、$\sin(A-B) = \Im\exp\{i(A-B)\}$ の右辺を (2.296) を使って変形する。

$$
\begin{aligned}
\Im\exp\{i(A-B)\} &= \Im\{\exp(iA)\exp(-iB)\} \\
&= \Im\{(\cos A + i\sin A)(\cos B - i\sin B)\} \\
&= \Im\{(\cos A\cos B + \sin A\sin B) + i(\sin A\cos B - \cos A\sin B)\} \\
&= \sin A\cos B - \cos A\sin B
\end{aligned}
$$

したがって (2.183) が示せた。

2.52 （標準）フォン・ミーゼス分布とガウス分布の関係

フォン・ミーゼス分布 (2.179) のうち、θ に無関係な係数部分を除外して考える。テーラ展開 (2.299) と $\xi = m^{1/2}(\theta-\theta_0)$ の置き換えをすると、

$$
\begin{aligned}
p(\theta|\theta_0, m) &\propto \exp\{m\cos(\theta-\theta_0)\} \\
&= \exp\left[m\left\{1-\frac{(\theta-\theta_0)^2}{2}+O((\theta-\theta_0)^4)\right\}\right] \\
&= \exp\left\{m-\frac{\xi^2}{2}+O(m^{-1})\right\} \\
&\propto \exp\left\{-\frac{\xi^2}{2}+O(m^{-1})\right\}
\end{aligned}
$$

ここで $m\to\infty$ とすると

$$
\exp\left\{-\frac{\xi^2}{2}+O(m^{-1})\right\} \to \exp\left\{-\frac{\xi^2}{2}\right\}
$$

となるので、$m\to\infty$ で $\exp\{-\xi^2/2\}$、つまり $\exp\{-m(\theta-\theta_0)^2/2\}$ に近似できる。すなわち、フォン・ミーゼス分布は $m\to\infty$ で平均 θ_0、精度 m のガウス分布に近似できる。

2.53 （基本）フォン・ミーゼス分布の位置パラメータの最尤推定

(2.182) の左辺に (2.183) を適用すると

$$\begin{aligned}\sum_{n=1}^{N}\sin(\theta_n-\theta_0) &= \sum_{n=1}^{N}(\sin\theta_n\cos\theta_0-\cos\theta_n\sin\theta_0)\\ &= \cos\theta_0\sum_{n=1}^{N}\sin\theta_n-\sin\theta_0\sum_{n=1}^{N}\cos\theta_n\end{aligned}$$

θ_0 が最尤推定解 θ_0^{ML} のときはこの値が 0 となるので、次式が得られる。

$$\tan\theta_0^{\mathrm{ML}}=\frac{\sum_n\sin\theta_n}{\sum_n\cos\theta_n}$$

これを θ_0^{ML} について解くと

$$\theta_0^{\mathrm{ML}}=\tan^{-1}\left\{\frac{\sum_n\sin\theta_n}{\sum_n\cos\theta_n}\right\}$$

となり、(2.184) が得られる。

2.54 （基本）フォン・ミーゼス分布を最大・最小にする値

(2.179) を θ で微分する。

$$\frac{\mathrm{d}}{\mathrm{d}\theta}p(\theta|\theta_0,m)=-\frac{m}{2\pi I_0(m)}\sin(\theta-\theta_0)\exp\{m\cos(\theta-\theta_0)\}$$

これを 0 と置くことにより、(2.179) は $\theta=\theta_0+n\pi\ \ (n=0,\pm1,\pm2,\ldots)$ で極値を取ることが判る。次に (2.179) の 2 階導関数を求める。

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}p(\theta|\theta_0,m) &= -\frac{m}{2\pi I_0(m)}[\cos(\theta-\theta_0)\exp\{m\cos(\theta-\theta_0)\}-m\sin^2(\theta-\theta_0)\exp\{m\cos(\theta-\theta_0)\}]\\ &= -\frac{m}{2\pi I_0(m)}[\cos(\theta-\theta_0)-m\sin^2(\theta-\theta_0)]\exp\{m\cos(\theta-\theta_0)\}\end{aligned}$$

これに $\theta=\theta_0$ および $\theta=\theta_0+\pi\ (\mathrm{mod}\ 2\pi)$ を適用すると、

$$\begin{aligned}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}p(\theta|\theta_0,m)\right]_{\theta=\theta_0} &= -\frac{m}{2\pi I_0(m)}e^m<0\\ \left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}p(\theta|\theta_0,m)\right]_{\theta=\theta_0+\pi} &= \frac{m}{2\pi I_0(m)}e^{-m}>0\end{aligned}$$

したがって、フォン・ミーゼス分布は $\theta=\theta_0$ で最大になり、$\theta=\theta_0+\pi\ (\mathrm{mod}\ 2\pi)$ で最小になる。

2.55 （基本）フォン・ミーゼス分布の集中度の最尤推定

(2.169) と (2.184) から、$\theta_0^{\mathrm{ML}}=\overline{\theta}$ であるので、(2.185) にこの関係と (2.178) を適用すると

$$\begin{aligned}A(m_{\mathrm{ML}}) &= \frac{1}{N}\sum_{n=1}^{N}\cos(\theta_n-\theta_0^{\mathrm{ML}})\\ &= \frac{1}{N}\sum_{n=1}^{N}\cos(\theta_n-\overline{\theta})\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{n=1}^{N}(\cos\theta_n\cos\overline{\theta}+\sin\theta_n\sin\overline{\theta}) \\
&= \frac{1}{N}\cos\overline{\theta}\sum_{n=1}^{N}\cos\theta_n+\frac{1}{N}\sin\overline{\theta}\sum_{n=1}^{N}\sin\theta_n
\end{aligned}
$$

ここで、(2.168) を適用すると

$$
\begin{aligned}
A(m_{\mathrm{ML}}) &= \frac{1}{N}\cos\overline{\theta}\sum_{n=1}^{N}\cos\theta_n+\frac{1}{N}\sin\overline{\theta}\sum_{n=1}^{N}\sin\theta_n \\
&= \overline{r}\cos^2\overline{\theta}+\overline{r}\sin^2\overline{\theta} \\
&= \overline{r}
\end{aligned}
$$

2.56 （標準）指数型分布族

まず、ベータ分布 (2.13) について考える。(2.13) の対数の指数を取ると

$$
\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\{(a-1)\ln\mu+(b-1)\ln(1-\mu)\}
\end{aligned}
$$

この式と (2.149) を比較すると以下のようになっている。

$$
\begin{aligned}
\boldsymbol{\eta} &= \begin{pmatrix} a-1 \\ b-1 \end{pmatrix} \\
\mathbf{u}(\mu) &= \begin{pmatrix} \ln\mu \\ \ln(1-\mu) \end{pmatrix} \\
h(\mu) &= 1 \\
g(\boldsymbol{\eta}) &= \frac{\Gamma(\eta_1+\eta_2+2)}{\Gamma(\eta_1+1)\Gamma(\eta_2+1)}
\end{aligned}
$$

ガンマ分布 (2.146) についても同様に

$$
\begin{aligned}
\mathrm{Gam}(\lambda|a,b) &= \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda) \\
&= \frac{1}{\Gamma(a)}b^a\exp\{(a-1)\ln\lambda-b\lambda\}
\end{aligned}
$$

この式と (2.149) を比較すると以下のようになっている。

$$
\begin{aligned}
\boldsymbol{\eta} &= \begin{pmatrix} -b \\ a-1 \end{pmatrix} \\
\mathbf{u}(\lambda) &= \begin{pmatrix} \ln\lambda \\ \lambda \end{pmatrix} \\
h(\lambda) &= 1 \\
g(\boldsymbol{\eta}) &= \frac{(-\eta_1)^{\eta_2+1}}{\Gamma(\eta_2+1)}
\end{aligned}
$$

最後に、フォン・ミーゼス分布 (2.179) については

$$
\begin{aligned}
p(\theta|\theta_0,m) &= \frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta-\theta_0)\} \\
&= \frac{1}{2\pi I_0(m)}\exp\{m(\cos\theta\cos\theta_0+\sin\theta\sin\theta_0)\}
\end{aligned}
$$

この式と (2.149) を比較すると以下のようになっている。

$$\begin{aligned}\boldsymbol{\eta} &= \begin{pmatrix} m\cos\theta_0 \\ m\sin\theta_0 \end{pmatrix} \\ \mathbf{u}(\theta) &= \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \\ h(\lambda) &= 1 \\ g(\boldsymbol{\eta}) &= \frac{1}{2\pi I_0((\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta})^{1/2})}\end{aligned}$$

2.57 （基本）多変量ガウス分布と指数型分布族

多変量ガウス分布 (2.43) は以下のように書ける。

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right\} \\ &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}-2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu})\right\} \\ &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu}\right\}\exp\left\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}\right\}\end{aligned}$$

ここで、$D\times D$ 次元の行列を D^2 次元のベクトルにする以下の演算子 $\mathrm{vec}(\cdot)$ を導入する。（$\mathbf{a}_n$ は D 次元の列ベクトルとする。）

$$\mathrm{vec}\left(\begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_D\end{pmatrix}\right) \equiv \begin{pmatrix}\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_2^{\mathrm{T}} & \cdots & \mathbf{a}_D^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$$

すると、二次形式の部分は以下のように書ける。

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x} &= \sum_{i,j}\lambda_{ij}x_ix_j \\ &= \sum_{i,j}\lambda_{ij}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]_{ij} \\ &= \mathrm{vec}(\boldsymbol{\Lambda})^{\mathrm{T}}\mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})\end{aligned}$$

この表現を利用すると、

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu}\right\}\exp\left\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}\right\} \\ &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu}\right\}\exp\left\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}-\frac{1}{2}\mathrm{vec}(\boldsymbol{\Lambda})^{\mathrm{T}}\mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})\right\}\end{aligned}$$

この式と (2.149) を比較すると以下のようになっている。

$$\begin{aligned}\boldsymbol{\eta} &= \begin{pmatrix}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Lambda})^{\mathrm{T}}\end{pmatrix} \\ \mathbf{u}(\mathbf{x}) &= \begin{pmatrix}\mathbf{x} \\ \mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})\end{pmatrix} \\ h(\lambda) &= (2\pi)^{-D/2} \\ g(\boldsymbol{\eta}) &= |-2\,\mathrm{Mat}(\boldsymbol{\eta}_2)|^{1/2}\exp\left\{\frac{1}{4}\boldsymbol{\eta}_1\mathrm{Mat}(\boldsymbol{\eta}_2)^{-1}\boldsymbol{\eta}_1^{\mathrm{T}}\right\}\end{aligned}$$

ただし、演算子 $\mathrm{Mat}(\cdot)$ は演算子 $\mathrm{vec}(\cdot)$ の逆演算、すなわち、ベクトルを行列に戻す演算を行う演算子である。なお、ここで $\boldsymbol{\eta}_1$ は D 次元の行ベクトル、$\boldsymbol{\eta}_2$ は D^2 次元の行ベクトルとなっていることに注意。$\mathbf{u}(\mathbf{x})$ の各要素は次元は $\boldsymbol{\eta}$ と同じだがこちらは列ベクトルである。

2.58 （基本）指数分布族の自然パラメータについての 2 階微分

(2.194) の $\boldsymbol{\eta}$ について勾配を計算した (2.225) から以下の式が得られる。

$$-\nabla \ln g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{49}$$

この式の右辺に対して、もう一度 $\boldsymbol{\eta}$ の勾配を計算すると

$$\begin{aligned}
&\nabla \left(g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right) \\
&= \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})^{\mathrm{T}} \, \mathrm{d}\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^{\mathrm{T}} \, \mathrm{d}\mathbf{x} \\
&= \nabla g(\boldsymbol{\eta}) \frac{\mathbb{E}[\mathbf{u}(\mathbf{x})^{\mathrm{T}}]}{g(\boldsymbol{\eta})} + \mathbb{E}[\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^{\mathrm{T}}] \\
&= \nabla \ln g(\boldsymbol{\eta}) \mathbb{E}[\mathbf{u}(\mathbf{x})^{\mathrm{T}}] + \mathbb{E}[\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^{\mathrm{T}}] \\
&= -\mathbb{E}[\mathbf{u}(\mathbf{x})] \mathbb{E}[\mathbf{u}(\mathbf{x})^{\mathrm{T}}] + \mathbb{E}[\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^{\mathrm{T}}] \\
&= \mathrm{cov}[\mathbf{u}(\mathbf{x})]
\end{aligned}$$

ここで、(49) の左辺の $\boldsymbol{\eta}$ の勾配は $-\nabla\nabla \ln g(\boldsymbol{\eta})$ であるから (2.300) が示された。

2.59 （基本）尺度不変性を持つ分布の正規化条件

(2.236) の積分を $y = x/\sigma$ の変数変換を使い計算すると

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x|\sigma) \, \mathrm{d}x &= \frac{1}{\sigma} \int_{-\infty}^{\infty} f\left(\frac{x}{\sigma}\right) \mathrm{d}x \\
&= \frac{1}{\sigma} \int_{-\infty}^{\infty} \sigma f(y) \, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} f(y) \, \mathrm{d}y
\end{aligned}$$

したがって、$f(x)$ が正規化されていればこの積分は 1 になり、$p(x|\sigma)$ も正規化されている。

2.60 （標準）ヒストグラム密度推定法

n 番目の観測値が領域 $j(n)$ で観測されたとすると、対数尤度は次式で表される。

$$\sum_{n=1}^{N} \ln h_{j(n)}$$

ここで、領域 i には n_i 個の観測値があるので、この式は $\sum_i n_i \ln h_i$ に等しい。また、確率密度 $p(\mathbf{x})$ は領域 i の内部では一定値 h_i であるから、各観測において領域 i で観測される確率は $h_i \Delta_i$ である。したがって正規化条件は $\sum_i h_i \Delta_i = 1$ となる。正規化条件の制約のもとで対数尤度の停留点を求めるラグランジュ方程式は次のようになる。

$$\sum_i n_i \ln h_i + \lambda \left(\sum_i h_i \Delta_i - 1 \right)$$

これを h_i で偏微分しそれを 0 と置くと、以下の式が得られる。

$$\frac{n_i}{h_i} + \lambda \Delta_i = 0 \tag{50}$$

この式の両辺に h_i を掛け、i についての総和を取ると

$$\sum_i (n_i + \lambda h_i \Delta_i) = 0$$

ここで $\sum_i n_i = N$、$\sum_i h_i \Delta_i = 1$ を考慮すると、この式から $\lambda = -N$ が得られる。これを (50) に代入すると以下の最尤推定量が求められる。

$$h_i = \frac{n_i}{N\Delta_i}$$

2.61 （基本）K 近傍法の確率密度

N 個の $\mathbf{x}$ の観測値集合（観測値の次元は D）がある場合の K 近傍密度モデルを考える。観測値 $\mathbf{x}_n$ の近傍の確率密度は

$$p(\mathbf{x}_n) = \frac{K}{NV_n}$$

で表される。ただし、V_n は $\mathbf{x}_n$ を中心にして観測点を K 個含むような球の体積である。ここで、この球の半径を r_n とすると V_n は r_n^D に比例するので

$$p(\mathbf{x}_n) = \frac{K}{NV_n} \propto r_n^{-D}$$

この $p(\mathbf{x}_n)$ を $\mathbf{x}$ で積分するのだが r_n は極座標系なので、直交座標系 $\mathbf{x}$ を $\mathbf{x}_n$ を原点とする極座標系 $(r, \boldsymbol{\theta})$ に変換するヤコビアンを計算する必要がある。ただ、直交座標系 $\mathbf{x}$ の各要素は r の一次式であるので、ヤコビアンは

$$\left| \frac{\partial \mathbf{x}}{\partial (r, \boldsymbol{\theta})} \right| \propto r^{D-1}$$

したがって、観測値 $\mathbf{x}_n$ の近傍の確率密度の積分を計算すると、

$$\begin{aligned}
\int_{\mathcal{D}_n} p(\mathbf{x}_n)\,\mathrm{d}\mathbf{x} &\propto \int_0^{r_n} \int_{\mathrm{def}(\boldsymbol{\theta})} r^{-D} \left| \frac{\partial \mathbf{x}}{\partial (r, \boldsymbol{\theta})} \right| \mathrm{d}\boldsymbol{\theta}\,\mathrm{d}r \\
&\propto \int_0^{r_n} r^{-D} r^{D-1}\,\mathrm{d}r \\
&= \int_0^{r_n} r^{-1}\,\mathrm{d}r \\
&= [\,\ln r\,]_0^{r_n} \\
&= \infty
\end{aligned}$$

となり、発散することが判る。ただし、上式の $\mathcal{D}_n$ は $\mathbf{x}_n$ を中心とする半径 r_n の球内部の領域、$\mathrm{def}(\boldsymbol{\theta})$ は極座標系の定義域である。$p(\mathbf{x})$ の全領域での積分はこの式の総和 $\displaystyle\sum_{n=1}^{N} \int_{\mathcal{D}_n} p(\mathbf{x}_n)\,\mathrm{d}\mathbf{x}$ であるから、やはりこれも発散する。