ベイズ推論

3.4

3.4. 多次元ガウス分布の学習と予測.

◎ D次元ガウス分布を使うと、共分散行列 Σ によって
データの次元間にまたがるような相関を捉えることができる

精度行列 : $\Lambda = \Sigma^{-1}$

3.4.1. 平均が未知.

<u>状況</u>

$x \in \mathbb{R}^D$ に対して、

- 平均パラメータ $\mu \in \mathbb{R}^D$ のみ未知.
- 精度行列 $\Lambda \in \mathbb{R}^{D\times D}$ は Given.

⇝ 観測モデルは、

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

with Λ : 正定値 $\Rightarrow$ 半正定 $\Leftarrow$ $\exists B$ s.t. $\Lambda = B^T B$ (3.96)

$\Lambda^T = (B^T B)^T = B^T B = \Lambda$

μ の事前分布 : ガウス分布 ⇝ 解析的な計算ができる.

$$p(\mu) = \mathcal{N}(\mu | m, \Lambda_\mu^{-1}) \qquad (3.97)$$

with hyper parameters $m \in \mathbb{R}^D$ and $\Lambda_\mu \in \mathbb{R}^{D\times D}$.

$X = \{x_1, \ldots, x_N\}$: データが与えられたとする.

事後分布は、

$$p(\mu|X) \propto p(X|\mu)\, p(\mu)$$

$$= \left(\prod_{i=1}^{N} p(x_i|\mu)\right) p(\mu)$$

$$= \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \Lambda^{-1})\, \mathcal{N}(\mu|m, \Lambda_\mu^{-1}) \qquad (3.98)$$

対数をとり、μに関して整理する。

$$\ln p(\mu \mid \mathbf{X}) = \ln\left(\prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \Lambda^{-1})\right) \ln \mathcal{N}(\mu \mid m, \Lambda_\mu^{-1}) + \text{const.}$$

$$= \sum_{i=1}^{N}\left[\frac{1}{2}\left(\ln|\Lambda| - D\ln 2\pi\right) - \frac{1}{2}(x_i-\mu)^T\Lambda(x_i-\mu)\right]$$

$$+ \frac{1}{2}\left(\ln|\Lambda_\mu| - D\ln 2\pi\right) - \frac{1}{2}(\mu-m)^T\Lambda_\mu(\mu-m)$$

$$+ \text{const.}$$

ここで、

$$(x_i-\mu)^T\Lambda(x_i-\mu)$$
$$= x_i^T\Lambda x_i - \mu^T\Lambda x_i - x_i^T\Lambda\mu + \mu^T\Lambda\mu$$

より、

$$\sum_{i=1}^{N} -\frac{1}{2}(x_i-\mu)^T\Lambda(x_i-\mu)$$
$$= \mu^T\Lambda\sum_{i=1}^{N}x_i - \frac{1}{2}N\mu^T\Lambda\mu + \text{const.}$$

また、

$$-\frac{1}{2}(\mu-m)^T\Lambda_\mu(\mu-m)$$
$$= -\frac{1}{2}\mu^T\Lambda_\mu\mu + \mu^T\Lambda_\mu m + \text{const.}$$

よって、

$$\ln p(\mu \mid \mathbf{X}) = -\frac{1}{2}\left[\mu^T(N\Lambda + \Lambda_\mu)\mu - 2\mu^T\left(\Lambda\sum_{i=1}^{N}x_i + \Lambda_\mu m\right)\right]$$

$$+ \text{const.}$$

∴ 上に凸の2次関数。 (3.99)

ここで、D次元正規分布

$$\mathcal{N}(\mu \mid \hat{m}, \hat{\Lambda}_\mu^{-1}) = \sqrt{\frac{|\hat{\Lambda}_\mu|}{(2\pi)^D}} \exp\left(-\frac{1}{2}(\mu - \hat{m})^T \hat{\Lambda}_\mu (\mu - \hat{m})\right)$$

の対数は、

$$\ln \mathcal{N}(\mu \mid \hat{m}, \hat{\Lambda}_\mu^{-1})$$

$$= \frac{1}{2}\left[(\ln|\hat{\Lambda}_\mu| - D \ln 2\pi) - (\mu - \hat{m})^T \hat{\Lambda}_\mu (\mu - \hat{m})\right]$$

$$= -\frac{1}{2}\left(\mu^T \underset{\sim}{\hat{\Lambda}_\mu} \mu - 2\mu^T \underset{\sim\sim\sim}{\hat{\Lambda}_\mu \hat{m}}\right) + \text{const.} \qquad (3.101)$$

であるので、 $\hat{\Lambda}_\mu \hat{m} = \Lambda \sum_{i=1}^{N} x_i + \Lambda_\mu m$

(3.99), (3.101) を比較すると、事後分布 $p(\mu \mid X)$ が

$$p(\mu \mid X) = \mathcal{N}(\mu \mid \hat{m}, \hat{\Lambda}_\mu) \qquad (3.100)$$

where

$$\hat{\Lambda}_\mu = N\Lambda + \Lambda_\mu \qquad (3.102)$$

$$\hat{m} = \hat{\Lambda}_\mu^{-1}\left(\Lambda \sum_{i=1}^{N} x_i + \Lambda_\mu m\right) \qquad (3.103)$$

と求まる。

次に、観測されていないデータ点 $x_* \in \mathbb{R}^D$ に関する予測分布を求める。

⇝ 対数計算のみで推論を完結させる。

$$p(\mu \mid x_*) = \frac{p(x_* \mid \mu)\, p(\mu)}{p(x_*)} \qquad (\because \text{Bayes})$$

より、対数をとって x_* に関して整理、

$$\ln p(x_*) = \ln p(x_* \mid \mu) - \ln p(\mu \mid x_*) + \text{const.} \qquad (3.104)$$

$p(\mu \mid x_*)$ は、1点 x_* のみを観測したときの事後分布と考えられる。

$$\rightsquigarrow p(\mu \mid x_*) = \mathcal{N}\left(\mu \mid m(x_*), (\Lambda + \Lambda_\mu)^{-1}\right) \quad (3.105)$$

$$\text{where} \quad m(x_*) = (\Lambda + \Lambda_\mu)^{-1}(\Lambda x_* + \Lambda_\mu m) \quad (3.106)$$

$$(\because (3.100), (3.102), (3.103))$$

よって、

$\ln p(x_*)$

$$= \ln \mathcal{N}(x_* \mid \mu, \Lambda^{-1}) - \ln \mathcal{N}\left(\mu \mid m(x_*), (\Lambda + \Lambda_\mu)^{-1}\right) + \text{const.}$$

$$= \frac{1}{2}(\ln|\Lambda| - D\ln 2\pi) - \frac{1}{2}(x_* - \mu)^T \Lambda (x_* - \mu)$$
$$- \left[\frac{1}{2}(\ln|\Lambda + \Lambda_\mu| - D\ln 2\pi) - \frac{1}{2}(\mu - m(x_*))^T(\Lambda + \Lambda_\mu)(\mu - m(x_*))\right]$$
$$+ \text{const.}$$

$$= -\frac{1}{2}x_*^T \Lambda x_* + x_*^T \Lambda \mu - \mu^T(\Lambda + \Lambda_\mu) m(x_*)$$
$$+ \frac{1}{2} m(x_*)^T (\Lambda + \Lambda_\mu) m(x_*) + \text{const.}$$

ここで、

$$\mu^T(\Lambda + \Lambda_\mu) m(x_*)$$
$$= \mu^T(\Lambda + \Lambda_\mu)(\Lambda + \Lambda_\mu)^{-1}(\Lambda x_* + \Lambda_\mu m)$$
$$= \mu^T \Lambda x_* + \text{const.}$$

であり、

$$m(x_*)^T(\Lambda+\Lambda_\mu)m(x_*)$$

$$=(\Lambda x_*+\Lambda_\mu m)^T\left[(\Lambda+\Lambda_\mu)^{-1}\right]^T(\Lambda+\Lambda_\mu)(\Lambda+\Lambda_\mu)^{-1}(\Lambda x_*+\Lambda_\mu m)$$

$$=(\Lambda x_*+\Lambda_\mu m)^T\left[(\Lambda+\Lambda_\mu)^{-1}\right](\Lambda x_*+\Lambda_\mu m)$$

$$=x_*^T\Lambda^T\left[(\Lambda+\Lambda_\mu)^{-1}\right]^T\Lambda x_*+2x_*^T\Lambda^T\left[(\Lambda+\Lambda_\mu)^{-1}\right]^T\Lambda_\mu m$$
$$+\mathrm{const.}$$

である。

よって、

$$\ln p(x_*)$$
$$=-\frac{1}{2}x_*^T\Lambda x_*+x_*^T\Lambda\mu-\mu^T\Lambda x_*$$
$$+\frac{1}{2}x_*^T\Lambda^T\left[(\Lambda+\Lambda_\mu)^{-1}\right]^T\Lambda x_*+x_*^T\Lambda^T\left[(\Lambda+\Lambda_\mu)^{-1}\right]^T\Lambda_\mu m$$

$$=-\frac{1}{2}\Big[x_*^T\left(\Lambda-\Lambda^T\left[(\Lambda+\Lambda_\mu)^T\right]^{-1}\Lambda\right)x_*$$
$$-2x_*^T\Lambda^T\left[(\Lambda+\Lambda_\mu)^T\right]^{-1}\Lambda_\mu m\Big]+\mathrm{const.}$$

$$=-\frac{1}{2}\Big[x_*^T\left(\Lambda-\Lambda(\Lambda+\Lambda_\mu)^{-1}\Lambda\right)x_*$$
$$-2x_*^T\Lambda(\Lambda+\Lambda_\mu)^{-1}\Lambda_\mu m\Big]+\mathrm{const.} \qquad (3.107)$$

($\because$ Λ, Λ_μ : 対称)

よって、

$$p(x_*) = \mathcal{N}(x_* \mid \mu_*, \Lambda_*^{-1}) \qquad (3.108)$$

where

$$\Lambda_* = \Lambda - \Lambda(\Lambda + \Lambda_\mu)^{-1}\Lambda$$

$$= (\Lambda^{-1} + \Lambda_\mu^{-1})^{-1} \qquad (3.109)$$

$(\because$ Woodbury, $A = \Lambda^{-1}$, $U = V = I$, $B = \Lambda_\mu^{-1})$

$$\mu_* = \Lambda_*^{-1}\Lambda(\Lambda + \Lambda_\mu)^{-1}\Lambda_\mu m$$

$$= (\Lambda^{-1} + \Lambda_\mu^{-1})\Lambda\underline{(\Lambda + \Lambda_\mu)^{-1}}\Lambda_\mu m$$

$$= (\Lambda^{-1} + \Lambda_\mu^{-1})\Lambda\left[\underline{\Lambda^{-1} - \Lambda^{-1}(\Lambda_\mu^{-1} + \Lambda^{-1})^{-1}\Lambda^{-1}}\right]\Lambda_\mu m$$

$(\because$ Woodbury, $A = \Lambda$, $U = V = I$, $B = \Lambda_\mu)$

$$= (\Lambda^{-1} + \Lambda_\mu^{-1})(I - (\Lambda^{-1} + \Lambda_\mu^{-1})^{-1}\Lambda^{-1})\Lambda_\mu m$$

$$= (\Lambda^{-1} + \Lambda_\mu^{-1} - \Lambda^{-1})\Lambda_\mu m$$

$$= m \qquad (3.110)$$

☑ Woodbury formula

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

3.4.2. 精度が未知の場合

状況

$x \in \mathbb{R}^D$ に対して

- 平均パラメータ $\mu \in \mathbb{R}^D$ が既知
- 精度行列 $\Lambda \in \mathbb{R}^{D \times D}$ が未知

⇝ 観測モデルは、

$$p(x \mid \Lambda) = \mathcal{N}(x \mid \mu, \Lambda^{-1}) \qquad (3.111)$$

$$\left(= \sqrt{\frac{|\Lambda|}{(2\pi)^D}} \exp\left(-\frac{1}{2}(x-\mu)^T \Lambda (x-\mu)\right) \right)$$

Λ (正定値) を生成するための確率分布 ⇝ ウィシャート事前分布.

$$p(\Lambda) = \mathcal{W}(\Lambda \mid \nu, W) \left(= C_W(\nu, W) |\Lambda|^{\frac{\nu-D-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(W^{-1}\Lambda)\right) \right) \qquad (3.112)$$

where $W \in \mathbb{R}^{D \times D}$: 正定値, $\nu > D-1$ are hyper parameters.

データ $\mathbb{X} = \{x_1, \dots, x_N\}$ を観測したとする. 事後分布は,

$$p(\Lambda \mid \mathbb{X}) \propto p(\mathbb{X} \mid \Lambda) p(\Lambda)$$

$$= \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \Lambda^{-1}) \mathcal{W}(\Lambda \mid \nu, W)$$

$\mathrm{tr}(ABC)$
$= \mathrm{tr}(CAB)$

より、対数をとると、

$$\ln p(\Lambda \mid \mathbb{X}) = \sum_{i=1}^{N} \frac{1}{2}\left[(\ln|\Lambda| - D\ln 2\pi) - \overbrace{(x_i-\mu)^T \Lambda (x_i-\mu)}^{\text{スカラー}}\right]$$

$$+ \ln C_W(\nu, W) + \frac{\nu-D-1}{2}\ln|\Lambda| - \frac{1}{2}\mathrm{tr}(W^{-1}\Lambda) + \text{const.}$$

$$= \frac{N+\nu-D-1}{2}\ln|\Lambda| - \frac{1}{2}\mathrm{tr}\left[\sum_{i=1}^{N}(x_i-\mu)^T\Lambda(x_i-\mu) + (W^{-1}\Lambda)\right] + \text{const.}$$

$$= \frac{N+\nu-D-1}{2}\ln|\Lambda| - \frac{1}{2}\mathrm{tr}\left[\left(\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T + W^{-1}\right)\Lambda\right] + \text{const.} \qquad (3.114)$$

ここで、
$\mathcal{W}(\Lambda | \hat{\nu}, \hat{W})$ の対数が

$$\ln \mathcal{W}(\Lambda | \hat{\nu}, \hat{W}) = \frac{\hat{\nu} - D - 1}{2} \ln |\Lambda| - \frac{1}{2} \mathrm{tr}\left[\hat{W}^{-1} \Lambda \right] + \mathrm{const.}$$

で与えられることより、事後分布は、

$$p(\Lambda | X) = \mathcal{W}(\Lambda | \hat{\nu}, \hat{W}) \qquad (3.115)$$

where

$$\begin{cases} \hat{W}^{-1} = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T + W^{-1} \\ \hat{\nu} = \nu + N \end{cases} \qquad (3.116)$$

次に、予測分布の計算を行う。
これまでと同様に、ベイズの定理より、

$$p(\Lambda | x_*) = \frac{p(x_* | \Lambda) p(\Lambda)}{p(x_*)}$$

$$\therefore \ln p(x_*) = \ln p(x_* | \Lambda) - \ln p(\Lambda | x_*) + \mathrm{const.} \qquad (3.117)$$

$p(\Lambda | x_*)$ について、1点 x_* を観測したと考えると、

$$p(\Lambda | x_*) = \mathcal{W}(\Lambda | 1 + \nu, W(x_*)) \qquad (3.118)$$

where

$$W(x_*)^{-1} = (x_* - \mu)(x_* - \mu)^T + W^{-1} \qquad (3.119)$$

(3.117)を実際に計算する。

$$\ln p(x_*)$$

$$= \underline{\ln \mathcal{N}(x_* | \mu, \Lambda^{-1})} - \underline{\ln \mathcal{W}(\Lambda | 1+\nu, W(x_*))} + \text{const}.$$

$$= \underline{-\frac{1}{2}(x_* - \mu)^T \Lambda (x_* - \mu)}$$

$$\underline{-\left[-\frac{1}{2} \mathrm{tr}\left[W(x_*)^{-1} \Lambda \right] + \ln C_W(1+\nu, W(x_*)) \right]} + \text{const}.$$

ここで

$$-\frac{1}{2}(x_* - \mu)^T \Lambda (x_* - \mu) + \frac{1}{2} \mathrm{tr}\left(W(x_*)^{-1} \Lambda \right)$$

$$= -\frac{1}{2} \mathrm{tr}\left[\left((x_* - \mu)(x_* - \mu)^T - W(x_*)^{-1} \right) \Lambda \right]$$

$$= -\frac{1}{2} \mathrm{tr}\left[\left((x_* - \mu)(x_* - \mu)^T - (x_* - \mu)(x_* - \mu)^T - W^{-1} \right) \Lambda \right]$$

$$= \text{const}.$$

であり、

$$\ln C_W(1+\nu, W(x_*))$$

$$= -\frac{1+\nu}{2} \ln \left| W(x_*)^{-1} \right|^{-1} + \text{const}.$$

$$= \frac{1+\nu}{2} \ln \left| (x_* - \mu)(x_* - \mu)^T + W^{-1} \right| + \text{const}.$$

$$= \frac{1+\nu}{2} \ln \left| \underbrace{W(x_* - \mu)}_{C} \underbrace{(x_* - \mu)^T}_{D} + I \right| + \text{const}.$$

$$(\because \det(AB) = \det A \det B)$$

$$= \frac{1+\nu}{2} \ln \left\{ 1 + (x_* - \mu)^T W (x_* - \mu) \right\} + \text{const}.$$

$$(\because \det(I_N + CD^T) = \det(I_M + C^T D),\ W = W^T)$$

よって、

$$\ln p(x_*) = -\frac{1+\nu}{2}\ln\left\{1+(x_*-\mu)^T W (x_*-\mu)\right\} + \text{const.} \tag{3.120}$$

これは、スチューデントのt分布のp.d.f.の対数になっている。

$$\begin{aligned} &\mathrm{St}(x \mid \mu_s, \Lambda_s, \nu_s) \\ &= \frac{\Gamma\left(\frac{\nu_s+D}{2}\right)}{\Gamma\left(\frac{\nu_s}{2}\right)} \frac{|\Lambda_s|^{\frac{1}{2}}}{(\pi \nu_s)^{D/2}} \left\{1+\frac{1}{\nu_s}(x-\mu_s)^T\Lambda_s(x-\mu_s)\right\}^{-\frac{\nu_s+D}{2}} \end{aligned} \tag{3.121}$$

where

$\mu_s \in \mathbb{R}^D$, $\Lambda_s \in \mathbb{R}^{D\times D}$：正定値，$\nu_s \in \mathbb{R}$

$$\leadsto \ln \mathrm{St}(x \mid \mu_s, \Lambda_s, \nu_s) = -\frac{\nu_s+D}{2}\left[1+\frac{1}{\nu_s}(x-\mu_s)^T\Lambda_s(x-\mu_s)\right] + \text{const.} \tag{3.122}$$

よって、予測分布は、

$$p(x_*) = \mathrm{St}(x_* \mid \mu_s, \Lambda_s, \nu_s) \tag{3.123}$$

where
$$\begin{cases} \mu_s = \mu \\ \Lambda_s = (1-D+\nu) W \\ \nu_s = 1-D+\nu \end{cases} \tag{3.124}$$

3.4.3. 平均, 精度が未知の場合.
観測モデル

$$p(x \mid \mu, \Lambda) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$$

☆
$$p(X, \mu, \Lambda) = p(X \mid \mu, \Lambda)\, p(\mu, \Lambda)$$

共役事前分布は, ガウス・ウィシャート分布

$$p(\mu, \Lambda) = NW(\mu, \Lambda \mid m, \beta, \nu, W)$$
$$= \mathcal{N}(\mu \mid m, (\beta\Lambda)^{-1})\, W(\Lambda \mid \nu, W) \quad (3.125)$$

データ $X = \{x_1, \dots, x_N\}$ を観測したとする.

☒ 事後分布

$$p(\mu, \Lambda \mid X) = \frac{p(X, \mu, \Lambda)}{p(X)} \quad (\because \text{Bayes})$$
$$= \frac{p(X \mid \mu, \Lambda)\, p(\mu, \Lambda)}{p(X)} \quad (3.126)$$

ここで,

$$p(\mu, \Lambda \mid X) = \frac{p(X, \mu, \Lambda)}{p(X)} = \frac{p(X, \mu, \Lambda)}{p(\Lambda, X)} \cdot \frac{p(\Lambda, X)}{p(X)}$$
$$= \underbrace{p(\mu \mid \Lambda, X)}_{①}\, \underbrace{p(\Lambda \mid X)}_{②} \quad (3.127)$$

より,

① μ の事後分布を求める.

② Λ の事後分布を求める.

と分解できる.

① μの事後分布 $p(\mu|\Lambda, X)$

$$p(\mu|\Lambda, X) \propto p(X|\mu, \Lambda)\, p(\mu)$$
$$= \underbrace{\prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \Lambda^{-1})\, \mathcal{N}(\mu|m, (\beta\Lambda)^{-1})}$$

(3.98) $p(\mu|X) \propto \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \Lambda^{-1})\, \mathcal{N}(\mu|m, \Lambda_\mu^{-1})$

で $\Lambda_\mu \leftarrow \beta\Lambda$ とした形。

よって、μの事後分布は、

$$p(\mu|\Lambda, X) = \mathcal{N}(\mu|\hat{m}, (\hat{\beta}\Lambda)^{-1}) \quad (3.128)$$

where)
$$\hat{\beta} = N + \beta$$
$$\hat{m} = \frac{1}{\hat{\beta}}\Lambda^{-1}\left(\Lambda\sum_{i=1}^{N} x_i + \beta\Lambda m\right)$$
$$= \frac{1}{\hat{\beta}}\left(\sum_{i=1}^{N} x_i + \beta m\right) \quad (3.129)$$

② Λの事後分布 $p(\Lambda|X)$

(3.126), (3.127) より、

$$p(\Lambda|X) = \frac{p(\mu, \Lambda|X)}{p(\mu|\Lambda, X)} \quad (\because (3.127))$$
$$= \frac{p(X|\mu, \Lambda)\, p(\mu, \Lambda)}{p(X)\, p(\mu|\Lambda, X)} \quad (\because (3.126))$$

を用いる。対数をとってΛについて整理すると、

$$\ln p(\Lambda|X)$$
$$= \ln p(X|\mu, \Lambda) + \ln p(\mu, \Lambda) - \ln p(\mu|\Lambda, X) + \text{const.} \quad (3.130)$$

代入して，整理する．

$$\ln p(\Lambda \mid \mathbf{X})$$

$$= \underline{\ln p(\mathbf{X} \mid \mu, \Lambda)} + \underline{\ln p(\mu, \Lambda)} - \ln p(\mu \mid \Lambda, \mathbf{X}) + \text{const.}$$

$$= \underline{\sum_{i=1}^{N} \ln \mathcal{N}(x_i \mid \mu, \Lambda^{-1})} + \underbrace{\ln \mathcal{N}(\mu \mid m, (\beta\Lambda)^{-1}) + \ln \mathcal{W}(\Lambda \mid \nu, W)}_{(3.125)}$$

$$- \underbrace{\ln \mathcal{N}(\mu \mid \hat{m}, (\hat{\beta}\Lambda)^{-1})}_{(3.128)} + \text{const.}$$

$\mathrm{tr}[(x_i-\mu)^T\Lambda(x_i-\mu)] = \mathrm{tr}[(x_i-\mu)(x_i-\mu)^T\Lambda]$

$$= \sum_{i=1}^{N} \frac{1}{2}\left[(\ln|\Lambda| - D\ln 2\pi) - (x_i-\mu)^T\Lambda(x_i-\mu)\right]$$

$\beta\,\mathrm{tr}[(\mu-m)(\mu-m)^T\Lambda]$

ln|Λ|で打ち消し

$$+ \frac{1}{2}\left[(\ln|\beta\Lambda| - D\ln 2\pi) - (\mu-m)^T\beta\Lambda(\mu-m)\right]$$

$\hat{\beta}\,\mathrm{tr}[(\mu-\hat{m})(\mu-\hat{m})^T\Lambda]$

$$+ \frac{\nu-D-1}{2}\ln|\Lambda| - \frac{1}{2}\mathrm{tr}(W^{-1}\Lambda) + \ln C_{\mathcal{W}}(\nu, W)$$

$$- \frac{1}{2}\left[(\ln|\hat{\beta}\Lambda| - D\ln 2\pi) - (\mu-\hat{m})^T\hat{\beta}\Lambda(\mu-\hat{m})\right]$$

$$+ \text{const.}$$

$$= \frac{N+\nu-D-1}{2}\ln|\Lambda|$$

$$-\frac{1}{2}\mathrm{tr}\Big[\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T + \beta(\mu-m)(\mu-m)^T$$

$$- \hat{\beta}(\mu-\hat{m})(\mu-\hat{m})^T + W^{-1}\Big] + \text{const.}$$

ここで.

$$\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$$

$$= \sum_{i=1}^{N} x_i x_i^T - \cancel{2\sum_{i=1}^{N} x_i \mu^T} + \cancel{\overbrace{\sum_{i=1}^{N}\mu\mu^T}^{N\mu\mu^T}}$$

$$\beta(\mu-m)(\mu-m)^T$$

$$= \cancel{\beta\mu\mu^T} \cancel{-2\beta m\mu^T} + \beta m m^T$$

$$-\hat{\beta}(\mu-\hat{m})(\mu-\hat{m})^T$$

$$= -\hat{\beta}\mu\mu^T + 2\hat{\beta}\hat{m}\mu^T \quad \hat{\beta}\hat{m}\hat{m}^T$$

$$= -\hat{\beta}\mu\mu^T + 2\hat{\beta}\cdot\frac{1}{\hat{\beta}}\left(\sum_{i=1}^{N}x_i + \beta m\right)\mu^T - \hat{\beta}\hat{m}\hat{m}^T$$

$$= -\cancel{(N+\beta)\mu\mu^T} + \cancel{2\sum_{i=1}^{N}x_i\mu^T} \cancel{+2\beta m\mu^T} - \hat{\beta}\hat{m}\hat{m}^T$$

より,

$$\ln(\Lambda|X) = \frac{N+\nu-D-1}{2}\ln|\Lambda| - \frac{1}{2}\mathrm{tr}\left[\left\{\sum_{i=1}^{N}x_i x_i^T + \beta m m^T - \hat{\beta}\hat{m}\hat{m}^T + W^{-1}\right\}\Lambda\right] + \mathrm{const}. \quad (3.131)$$

となる。これはウィシャート分布の対数と同じ形.

$$\left(\ln \mathcal{W}(\Lambda|\bar{\nu},\bar{W}) = \frac{\bar{\nu}-D-1}{2}\ln|\Lambda| - \frac{1}{2}\mathrm{tr}\left[\bar{W}^{-1}\Lambda\right] + \mathrm{const.}\right)$$

$$\therefore p(\Lambda|X) = \mathcal{W}(\Lambda|\hat{\nu},\hat{W}) \quad (3.132)$$

where

$$\begin{cases} \hat{W}^{-1} = \sum_{i=1}^{N} x_i x_i^T + \beta m m^T - \hat{\beta}\hat{m}\hat{m}^T + W^{-1} \\ \hat{\nu} = N+\nu \end{cases} \quad (3.133)$$

③ 予測分布.

新しいデータ点 $x_* \in D$ に関する予測分布は、

$$p(x_*) = \iint p(x_* \mid \mu, \Lambda)\, p(\mu, \Lambda)\, d\mu\, d\Lambda \qquad (3.134)$$

~) 対数を使って積分せずに求めたい.

$$p(\mu, \Lambda \mid x_*) = \frac{p(x_* \mid \mu, \Lambda)\, p(\mu, \Lambda)}{p(x_*)} \qquad (\because \text{Bayes})$$

対数をとって整理する.

$$\ln p(x_*) = \ln p(x_* \mid \mu, \Lambda) - \ln p(\mu, \Lambda \mid x_*) + \text{const.} \qquad (3.135)$$

ここで, $p(\mu, \Lambda \mid x_*)$ を(点 x_* を観測した後の事後分布と考える.

$$p(\mu, \Lambda \mid x_*) = \underbrace{p(\mu \mid \Lambda, x_*)}_{(3.128)\text{で } X \leftarrow x_*}\ \underbrace{p(\Lambda \mid x_*)}_{(3.132)\text{で } X \leftarrow x_*}$$

$$= \mathcal{N}\left(\mu \,\middle|\, m(x_*), ((1+\beta)\Lambda)^{-1}\right) \mathcal{W}\left(1+\nu, W(x_*)\right) \qquad (3.136)$$

where

$$\hat{m} = \frac{x_* + \beta m}{1+\beta} =: m(x_*)$$

$$\hat{W}^{-1} = x_* x_*^T + \beta m m^T - (1+\beta) \cdot \frac{x_* + \beta m}{1+\beta} \cdot \frac{x_*^T + \beta m^T}{1+\beta} + W^{-1}$$

$$= \frac{(1+\beta) - 1}{1+\beta} x_* x_*^T + \frac{\beta(1+\beta) - \beta^2}{1+\beta} m m^T - \frac{2\beta x_* m^T}{1+\beta} + W^{-1} \qquad (3.137)$$

$$= \frac{\beta}{1+\beta} (x_* - m)(x_* - m)^T + W^{-1} =: W(x_*)^{-1}$$

よって、

$\ln p(x_*)$

$$= \ln \mathcal{N}(x_* | \mu, \Lambda^{-1})$$

$$- \left(\ln \mathcal{N}(\mu | m(x_*), ((1+\beta)\Lambda)^{-1}) + \ln \mathcal{W}(\Lambda | 1+\nu, W(x_*)) \right)$$

$$= -\frac{1}{2}(x_* - \mu)^T \Lambda (x_* - \mu) \quad Ⓐ$$

$$+ \frac{1}{2}(\mu - m(x_*))^T (1+\beta)\Lambda(\mu - m(x_*)) \quad Ⓑ$$

$$- \left[\underbrace{-\frac{1}{2}\mathrm{tr}(W(x_*)^{-1}\Lambda)}_{Ⓒ} \underbrace{- \frac{1}{2}\ln|W(x_*)|}_{Ⓓ} \right] + \mathrm{const}.$$

ここで、

$$Ⓐ = \frac{1}{2}\mathrm{tr}\left(-(x_* - \mu)^T \Lambda (x_* - \mu)\right) = \frac{1}{2}\mathrm{tr}\Big(\underbrace{-(x_* - \mu)(x_* - \mu)^T}_{Ⓔ}\Lambda\Big)$$

$$Ⓔ = \boxed{-x_* x_*^T} + \cancel{2 x_* \mu^T} + \mathrm{const}.$$

$$Ⓑ = \frac{1}{2}\mathrm{tr}\Big(\underbrace{-(1+\beta)(\mu - m(x_*))(\mu - m(x_*))^T}_{Ⓕ}\Lambda\Big)$$

$$Ⓕ = (1+\beta)\left(\mu - \frac{1}{1+\beta}(x_* + \beta m)\right)\left(\mu - \frac{1}{1+\beta}(x_* + \beta m)\right)^T$$

$$= (1+\beta)\mu\mu^T - 2(x_* + \beta m)\mu^T + \frac{1}{1+\beta}(x_* + \beta m)(x_* + \beta m)^T$$

$$= -2x_*\mu^T + \frac{1}{1+\beta}(x_* x_*^T + 2\beta x_* m^T) + \mathrm{const}.$$

$$= \cancel{-2x_*\mu^T} + \boxed{\frac{1}{1+\beta}x_* x_*^T} + \cancel{\frac{2\beta}{1+\beta}} x_* m^T + \mathrm{const}.$$

$$Ⓒ = \frac{1}{2}\mathrm{tr}\Big(\Big(\underbrace{\frac{\beta}{1+\beta}(x_* - m)(x_* - m)^T}_{Ⓖ} + W^{-1}\Big)\Lambda\Big)$$

$$Ⓖ = \frac{\beta}{1+\beta}(x_* x_*^T - 2x_* m^T) + \mathrm{const}.$$

$$= \boxed{\frac{\beta}{1+\beta}x_* x_*^T} - \cancel{\frac{2\beta}{1+\beta}x_* m^T} + \mathrm{const}.$$

よって、

$$Ⓐ + Ⓑ + Ⓒ = \text{const}.$$

$$Ⓓ = \frac{1}{2} \ln \left| \left(W(x_*)^{-1} \right)^{-1} \right| = -\frac{1}{2} \ln \left| W(x_*)^{-1} \right|$$

$$= -\frac{1}{2} \ln \left[\frac{\beta}{1+\beta} (x_* - m)(x_* - m)^T + W^{-1} \right]$$

$$= -\frac{1}{2} \ln \left[\frac{\beta}{1+\beta} W (x_* - m)(x_* - m)^T + I \right] + \text{const}.$$

$$= -\frac{1}{2} \ln \left\{ 1 + \frac{\beta}{1+\beta} (x_* - m)^T W (x_* - m) \right\} + \text{const}.$$

$$\left(\because \det(I_N + C D^T) = \det(I_M + C^T D),\ W = W^T \right)$$

以上より、

$$\ln p(x_*) = -\frac{1}{2} \ln \left\{ 1 + \frac{\beta}{1+\beta} (x_* - m)^T W (x_* - m) \right\} + \text{const}. \quad (3.138)$$

これはスチューデントのt分布の対数表現と一致。

$$p(x_*) = \mathrm{St}(x_* \mid \mu_s, \Lambda_s, \nu_s) \quad (3.139)$$

where
$$\mu_s = m$$
$$\Lambda_s = \frac{(1 - D + \nu)\beta}{1+\beta} W \quad (3.140)$$
$$\nu_s = 1 - D + \nu$$