PRML の 10 章の数式の補足

サイボウズ・ラボ 光成滋生

2011 年 7 月 26 日

1 概要

この文章は『パターン認識と機械学習』（以下 PRML）の 10 章の式変形を一部埋めたものです. 間違い, 質問などございましたら herumi@nifty.com または twitterID:herumi までご連絡ください.

2 この章でよく使われる公式

9 章と同じようによく使う公式を列挙しておく.

2.1 ガンマ関数

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\,dt, \quad \Gamma(x+1) = x\Gamma(x).$$

ディガンマ関数

$$\phi(x) = \frac{\partial}{\partial x} \log \Gamma(x).$$

2.2 ディリクレ分布

$0 \le \mu_k \le 1$, $\sum_k \mu_k = 1$, $\hat{\alpha} = \sum_k \alpha_k$ として

$$\mathrm{Dir}(\mu|\alpha) = C(\alpha) \prod_{k=1}^K \mu_k^{\alpha_k - 1}, \quad C(\alpha) = \frac{\Gamma(\hat{\alpha})}{\prod_k \Gamma(\alpha_k)}.$$

$$E[\mu_k] = \frac{\alpha_k}{\hat{\alpha}}.$$

2.3 ガンマ分布

$$\mathrm{Gam}(\tau|a,b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}.$$

$$E[\tau] = \frac{a}{b}, \quad \mathrm{var}[\tau] = \frac{a}{b^2}, \quad E[\log \tau] = \phi(a) - \log b.$$

2.4　正規分布

$$\mathcal{N} = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu)^{\boldsymbol{T}}\Sigma^{-1}(x-\mu)).$$

$$E[x] = \mu, \quad \mathrm{cov}[x] = \Sigma, \quad , E[xx^{\boldsymbol{T}}] = \mu\mu^{\boldsymbol{T}} + \Sigma, \quad E[x^{\boldsymbol{T}}x] = \mu^{\boldsymbol{T}}\mu + \mathrm{tr}(\Sigma).$$

$p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})$, $p(y|x) = \mathcal{N}(x|Ax+b, L^{-1})$ のとき

$$p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^{\boldsymbol{T}}).$$

2.5　スチューデントの t 分布

$$\mathrm{St}(x|\mu, \Lambda, \nu) = \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})}\frac{|\Lambda|^{1/2}}{(\pi\nu)^{1/2}}(1 + \frac{\triangle^2}{\nu})^{-\frac{\nu+1}{2}}, \quad \triangle^2 = (x-\mu)^{\boldsymbol{T}}\Lambda(x-\mu).$$

$$E[x] = \mu.$$

2.6　ウィシャート分布

$$\mathcal{W}(\Lambda, W, \nu) = B(W, \nu)|\Lambda|^{\frac{\nu-D-1}{2}}\exp(-\frac{1}{2}\mathrm{tr}(W^{-1}\Lambda).$$

$$B(W, \nu) = |W|^{\nu/2}(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma(\frac{\nu+1-i}{2}))^{-1}.$$

$$E[\Lambda] = \nu W.$$

$$E[\log|\Lambda|] = \sum_{i=1}^{D}\phi(\frac{\nu+1-i}{2}) + D\log 2 + \log|W|.$$

$$H[\Lambda] = -\log B(W, \nu) - \frac{\nu-D-1}{2}E[\log|\Lambda| + \frac{\nu D}{2}.$$

2.7　行列の公式

$$x^{\boldsymbol{T}}Ax = \mathrm{tr}(Axx^{\boldsymbol{T}}).$$

$$\frac{\partial}{\partial A}\log|A| = (A^{-1})^{\boldsymbol{T}}.$$

$$\frac{\partial}{\partial x}\log|A| = \mathrm{tr}(A^{-1}\frac{\partial}{\partial x}A).$$

$$\frac{\partial}{\partial A}\mathrm{tr}(A^{-1}B) = -(A^{-1}BA^{-1})^{\boldsymbol{T}}.$$

$$|I + ab^{\boldsymbol{T}}| = 1 + a^{\boldsymbol{T}}b.$$

2.8 カルバック距離

$$\mathrm{KL}(q||p) = -\int q(Z)\log\frac{p(Z|X)}{q(Z)}\,dZ \geq 0.$$

3 分解による近似の持つ性質

ここで Λ_{ij} はスカラーで $\Lambda_{12} = \Lambda_{21}$. $E[z_1] = m_1$, $E[z_2] = m_2$ より

$$\begin{aligned} m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) \\ &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) - \mu_2) \\ &= \mu_1 + \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^2(m_1 - \mu_1). \end{aligned}$$

よって

$$(m_1 - \mu_1)(\Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^2 - 1) = 0.$$

分布が非特異なら $|\Lambda| = \Lambda_{11}\Lambda_{22} - \Lambda_{12}^2 \neq 0$ より $m_1 = \mu_1$. 同様に $m_2 = \mu_2$.

変数 $Z = \{z_1, \ldots, z_N\}$ に対する分布 $q(Z)$ が

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

と複数のグループの関数の積としてかけていると仮定する. ここで $\{Z_i\}$ は Z の disjoint-union である.

（PRML p.182）$\mathrm{KL}(p||q)$ を Z_j について最小化する問題を考える（以下, 対象変数以外の項をまとめて C と略記する）.

$$\begin{aligned} \mathrm{KL}(p||q) &= -\int p(Z)(\sum_i \log q_i(Z_i))\,dZ + C \\ &= -\int (p(Z)\log q_j(Z_j) + p(Z)\sum_{i\neq j}\log q_i(Z_i))\,dZ + C \\ &= -\int p(Z)\log q_j(Z_j)\,dZ + C \\ &= -\int \log q_j(Z_j)(\int p(Z)\prod_{i\neq j} dZ_i)dZ_j. \end{aligned}$$

$$F_j(Z_j) = \int p(Z)\prod_{i\neq j} dZ_i$$

と置くと

$$\mathrm{KL}(p||q) = \int F_j(Z_j)\log q_j(Z_j)\,dZ_j.$$

$$\int q_j(Z_j)\,dZ_j = 1$$

の条件の元で

$$X = -\int F_j(Z_j)\log q_j(Z_j)\,dZ_j + \lambda(\int q_j(Z_j)\,dZ_j - 1)$$

を最小化する．

$$\frac{\partial}{\partial q_j}X = -\int F_j(Z_j)\log(q_j+\delta q_j)\,dZ_j + \lambda(\int (q_j+\delta q_j)\,dZ_j - 1)$$

$$= (-\int F_j(Z_j)\log q_j\,dZ_j + \lambda(\int q_j\,dZ_j - 1)) - (\int F_j(Z_j)/q_j\,dZ_j - \lambda)\delta q_j = 0.$$

$$F_j/q_j - \lambda = 0.$$

よって $F_j = \lambda q_j$．積分して

$$\int F_j\,dZ_j = \int \lambda q_j\,dZ_j = \lambda = 1.$$

よって

$$q_j^*(Z_j) = q_j = F_j = \int p(Z)\prod_{i\neq j} dZ_i.$$

4 α ダイバージェンス

α を実数として

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2}\,dx)$$

を α ダイバージェンスという．$\alpha \to 1$ のとき $\mathrm{KL}(p||q)$, $\alpha \to -1$ のとき $\mathrm{KL}(q||p)$ になる．

（証明）$\alpha = 1 - 2\epsilon$ と置く．$\alpha \to 1$ で $\epsilon \to 0$ となる．

$$(q/p)^\epsilon = \exp(\epsilon\log(q/p)) \approx 1 + \epsilon\log(q/p)$$

より

$$\begin{aligned} D_\alpha(p||q) &= \frac{1}{\epsilon(1-\epsilon)}(1 - \int p(q/p)^\epsilon\,dx) \\ &\to \frac{1}{\epsilon}(1 - \int p(1+\epsilon\log\frac{q}{p})\,dx) \\ &= \frac{1}{\epsilon}(-\epsilon\int p\log\frac{q}{p}\,dx) = \mathrm{KL}(p||q). \end{aligned}$$

$\alpha \to -1$ も同様.

5 例：一変数ガウス分布

ガウス分布から独立に発生した観測値 x のデータセットを $\mathcal{D} = \{x_1, \ldots, x_N\}$ とする．もとのガウス分布の平均 μ と精度 τ の事後分布をもとめる．

$$p(D|\mu,\tau) = (\frac{\tau}{2\pi})^{N/2}\exp(-\frac{\tau}{2}\sum_n (x_n-\mu)^2).$$

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1}), \quad p(\tau) = \mathrm{Gam}(\tau|a_0,b_0).$$

この問題は厳密にもとめられるがここでは事後分布が次のように分解できると仮定したときの変分近似を考える．

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau).$$

まず μ について

$$
\begin{aligned}
\log q_\mu^*(\mu) &= E_\tau[\log p(D,\mu|\tau)] = E_\tau[\log p(D|\mu,\tau) + \log p(\mu|\tau)] + \mu\text{に依存しない部分（以下略）}C \\
&= \frac{E[\tau}{2}(\sum_n (x_n-\mu)^2) + E_\tau[-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2] + C \\
&= -(E[\tau]/2)(\lambda_0(\mu-\mu_0)^2 + \sum_n (x_n-\mu)^2) + C.
\end{aligned}
$$

μ について平行完成すると

$$
\begin{aligned}
& -(E[\tau]/2)((\lambda_0+N)\mu^2 - 2\mu(\lambda_0\mu_0 + \sum_n x_n) + \lambda_0\mu_0^2 + \sum_n x_n^2) + C \\
& \sum_n x_n = N\bar{x}\text{より} \\
&= -\frac{E[\tau](\lambda_0+N)}{2}(\mu - \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0+N})^2 + \cdots.
\end{aligned}
$$

よってこの分布はガウス分布であることが分かり,

$$
\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0+N}, \quad ,\lambda_N = (\lambda_0+N)[E\tau]
$$

と置くと $\mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$ となることが分かる．$N\to\infty$ のとき $\mu_N \to \bar{x}$ で分散は 0（精度は ∞）．

τ について

$$
\begin{aligned}
\log q_\tau^*(\tau) &= E_\mu[\log p(D,\tau|\mu)] = E_\mu[\log p(D|\mu,\tau) + \log p(\mu|\tau)] + \log p(\tau) \\
&= E_\mu[(N/2)\log\tau - (\tau/2)\sum_n (x_n-\mu)^2] \\
&+ E_\mu[(1/2)\log(\lambda_0\pi) - (\lambda_0\pi/2)(\mu-\mu_0)^2] \\
&+ E_\mu[(a_0-1)\log\tau - b_0\tau - \log\Gamma(a_0) + a_0\log b_0] + C \\
&= (a_0-1)\log\tau - b_0\tau + (N+1)/2\log\tau - (\tau/2)E_\mu[\sum_n (x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2] + C.
\end{aligned}
$$

よって $q_\tau(\tau)$ はガンマ分布となり

$$
a_N = a_0 + \frac{N+1}{2}, \quad b_N = b_0 + \frac{1}{2}E_\mu[\sum_n (x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2]
$$

と置くと $q_\tau(\tau) = \mathrm{Gam}(\tau|a_N, b_N)$．$q_\mu(\mu)$, $q_\tau(\tau)$ の関数の形に何も仮定を置いていないのに, 尤度関数と共役事前分布の構造から決まったことに注意．$N\to\infty$ で

$$
E[\mathrm{Gam}(\tau|a_N,b_N)] = \frac{a_N}{b_N} \to 1/E_\mu[(1/N\sum_n (x_n-\mu)^2] \to 1/\text{分散}.
$$

$$
\sigma[\mathrm{Gam}] = a_N/b_N^2 \to 0.
$$

$\mu_0 = a_0 = b_0 = \lambda_0 = 0$ という無情報事前分布を入れてみる．

$$
a_N = \frac{N+1}{2}, \quad b_N = \frac{1}{2}E_\mu[\sum_n (x_n-\mu)^2].
$$

よって

$$
E[\tau]^{-1} = \frac{b_N}{a_N} = E[\frac{1}{N+1}\sum_n (x_n-\mu)^2] = \frac{N}{N+1}(\overline{x^2} - 2\bar{x}E[\mu] + E[\mu^2]).
$$

$$\mu_N = \frac{0 + N\bar{x}}{0 + N} = \bar{x}, \quad \lambda_N = NE[\tau].$$

よって

$$E[\mu] = \bar{x}, \quad E[\mu^2] = E[\mu]^2 + \frac{1}{\lambda_N} = \bar{x}^2 + \frac{1}{NE[\tau]}.$$

$$\frac{1}{E[\tau]} = \frac{N}{N+1}(\bar{x}^2 - 2\bar{x}^2 + \bar{x}^2 + \frac{1}{NE[\tau]}).$$

$$\frac{N}{N+1}(\overline{x^2} - \bar{x}^2) = 1/E[\tau] - (1/(N+1))(1/E[\tau]) = \frac{N}{N+1}\frac{1}{E[\tau]}.$$

よって

$$\frac{1}{E[\tau]} = \overline{x^2} - \bar{x}^2 = \frac{1}{N}\sum_n (x_n - \bar{x})^2.$$

6 モデル比較

事前確率 $p(m)$ を持つ複数のモデルの比較. 観測データ X の下で事後確率 $p(m|X)$ を近似したい.

$$q(Z, m) = q(Z|m)q(m), \quad p(X, Z, m) = p(X)p(Z, m|X).$$

$\sum_{m,Z} q(Z, m) = 1$ に注意して

$$\begin{aligned}
\log p(X) &= \sum_{m,Z} q(Z, m) \log p(X) \\
&= \sum_{m,Z} q(Z, m) \log \frac{p(X, Z, m)}{q(Z, m)} \frac{q(Z, m)}{p(Z, m|X)} \\
&\mathcal{L} = \sum_{m,Z} q(Z, m) \log \frac{p(Z, X, m)}{q(Z, m)} \text{と置くと} \\
&= \mathcal{L} - \sum_{m,Z} q(Z, m) \log \frac{q(Z, m|X)}{q(Z, m)} \\
&= \mathcal{L} - \sum_{m,Z} q(Z|m)q(m) \log \frac{p(Z, m|X)}{q(Z|m)q(m)}.
\end{aligned}$$

$\mathcal{L}$ を $q(m)$ について最大化する. $\sum_Z q(Z|m) = 1$ に注意して

$$\begin{aligned}
\mathcal{L} &= \sum_{m,Z} q(Z|m)q(m)(\log p(Z, X|m) + \log p(m) - \log q(Z|m) - \log q(m)) \\
&= \sum_m q(m) \left\{ (\log p(m) - \log q(m)) + \sum_Z q(Z|m) \log \frac{p(Z, X|m)}{q(Z|m)} \right\}, \\
&\mathcal{L}_m = \sum_Z q(Z|m) \log \frac{p(Z, X|m)}{p(Z|m)} \text{とくと} \\
&= \sum_m q(m) \log \frac{p(m) \exp \mathcal{L}_m}{q(m)}.
\end{aligned}$$

よって $q(m) \propto \exp \mathcal{L}_m$ のとき $\mathcal{L}$ は最大値をとる.

6.1 変分混合ガウス分布

ガウス混合モデルに変分推論法を適用してみる．x_n に対応する潜在変数 z_n．z_n は K 個の要素 z_{nk} からなる．$z_{nk}=0$ または 1 で $\sum_k z_{nk}=1$.

$X=\{x_1,\ldots,x_N\}$, $Z=\{z_1,\ldots,z_N\}$, 混合比は $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_K)$.

$$p(z_n)=\prod_k \pi_k^{z_{nk}},\quad p(x_n|z_n)=\prod_k \mathcal{N}(x_n|\mu_k,\Sigma_k)^{z_{nk}}.$$

$$p(X|Z,\mu,\Lambda)=\prod_{n,k}\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}.$$

$\boldsymbol{\pi}$ の事前分布はディリクレ分布とする．

$$p(\boldsymbol{\pi})=\mathrm{Dir}(\boldsymbol{\pi}|\alpha_0)=C(\alpha_0)\prod_k \pi_k^{\alpha_0-1}.$$

混合要素の持つガウス分布の事前分布はガウス・ウィシャート分布とする．

$$p(\mu,\Lambda)=p(\mu|\Lambda)p(\Lambda)=\prod_k \mathcal{N}(\mu_k|m_0,(\beta_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k,|W_0,\nu_0).$$

6.2 変分事後分布

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

$q(Z,\boldsymbol{\pi},\mu,\Lambda)=q(Z)q(\boldsymbol{\pi},\mu,\Lambda)$ という変分近似を考える．

Z について（以後対象としている変数以外の項は無視する）

$$\begin{aligned}\log q^*(Z)&=E_{\boldsymbol{\pi},\mu,\Lambda}[\log p(X,Z,\boldsymbol{\pi},\mu,\Lambda)]\\&=E_{\boldsymbol{\pi}}[\log p(Z|\boldsymbol{\pi})]+E_{\mu,\Lambda}[\log p(X|Z,\mu,\Lambda)]\\&=\sum_{n,k} z_{nk}E_{\boldsymbol{\pi}}[\log\pi_k]+\sum_{n,k} z_{nk}E_{\mu,\Lambda}[\frac{1}{2}\log|\Lambda_k|-\frac{1}{2}(x_n-\mu_n)^{\boldsymbol{T}}\Lambda_k(x_n-\mu_n)-\frac{D}{2}\log(2\pi)]\\&=\sum_{n,k} z_{nk}\log\rho_{nk}.\end{aligned}$$

ただし

$$\log\rho_{nk}=E_\pi[\log\pi_k]+\frac{1}{2}E[\log|\Lambda_k|]-\frac{D}{2}\log(2\pi)-\frac{1}{2}E_{\mu_k,\Lambda_k}[(x_n-\mu_k)^{\boldsymbol{T}}\Lambda_k(x_n-\mu_k)].$$

よって

$$q^*(Z)\propto\prod_{n,k}\rho_{nk}^{z_{nk}}.$$

Z について総和をとると $\sum_k z_{nk}=1$ より

$$\sum_Z\prod_{n,k}\rho_{nk}^{z_{nk}}=\prod_n(\sum_k \rho_{nk}).$$

よって

$$r_{nk} = \rho_{nk}/(\sum_k \rho_{nk})$$

と置くと

$$q^*(Z) = \prod_{n,k} r_{nk}^{z_{nk}}$$

とできる．$q(Z)$ の最適解は事前分布 $p(Z|\boldsymbol{\pi})$ と同じ形．$\rho_{nk} = e^?$ の形なので $\rho_{nk} \geq 0$．つまり $r_{nk} \geq 0$．各 n について $\sum_k r_{nk} = 1$

$$E[z_{nk}] = r_{nk}.$$

次の値を定義する：

$$N_k = \sum_n r_{nk}, \quad \bar{x}_k = \frac{1}{N_k}\sum_n r_{nk}x_n, \quad S_k = \frac{1}{N_k}\sum_n r_{nk}(x_n - \bar{x}_k)(x_n - \bar{x}_k)^{\boldsymbol{T}}.$$

$q(\boldsymbol{\pi}, \mu, \Lambda)$ について考える．

$$\begin{aligned}\log q^*(\boldsymbol{\pi}, \mu, \Lambda) &= E_Z[\log p(X, Z, \boldsymbol{\pi}, \mu, \Lambda)] \\ &= \log p(\boldsymbol{\pi}) + \sum_k \log p(\mu_k, \Lambda_k) + E_Z[\log p(Z|\boldsymbol{\pi})] + \sum_{n,k} E[z_{nk}] \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}].\end{aligned}$$

この式は $\boldsymbol{\pi}$ だけを含む項とそれ以外の項に分かれている．更に μ_k, Λ_k の積にもなっている．つまり $q(\boldsymbol{\pi}, \mu, \Lambda) = q(\boldsymbol{\pi}) \prod_k q(\mu_k, \Lambda_k)$ という形になっている．

$\boldsymbol{\pi}$ に依存する部分を見る．

$$\begin{aligned}\log q^*(\boldsymbol{\pi}) &= \log \mathrm{Dir}(\boldsymbol{\pi}|\alpha_0) + E_Z[\sum_{n,k} z_{nk} \log \pi_k] \\ &= (\alpha_0 - 1)\sum_k \log \pi_k + \sum_{n,k} r_{nk} \log \pi_k \\ &= \sum_k (\alpha_0 - 1 + \sum_n r_{nk}) \log \pi_k.\end{aligned}$$

よって $q^*(\boldsymbol{\pi})$ はディリクレ分布となる．その係数は $\alpha_k = \alpha_0 + N_k$ とおいて $\alpha = (\alpha_k)$ とすると $q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\alpha)$.

$q^*(\mu_k, \Lambda_k) = q^*(\mu_k|\Lambda_k)q^*(\Lambda_k)$ を考える．まず

$$\begin{aligned}\log q^*(\mu_k, \Lambda_k) &= \log \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1}) + \log \mathcal{W}(\Lambda_k|W_0, \nu_0) + \sum_n r_{nk} \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) \\ &\quad \mu_k, \Lambda_k\text{の依存部分だけとりだして} \\ &= \frac{1}{2}\log|\beta_0\Lambda_k| - \frac{1}{2}(\mu_k - m_0)^{\boldsymbol{T}}\beta_0\Lambda_k(\mu_k - m_0) + \frac{\nu_0 - D - 1}{2}\log|\Lambda_k| \\ &\quad -\frac{1}{2}\mathrm{tr}(W_0^{-1}\Lambda_k) + \sum_n r_{nk}(\frac{1}{2}\log|\Lambda_k| - \frac{1}{2}(x_n - \mu_k)^{\boldsymbol{T}}\Lambda_k(x_n - \mu_k)).\end{aligned}$$

このうち更に μ_k に依存する部分をみる：

$$\begin{aligned}\log q^*(\mu_k|\Lambda_k) &= -\frac{1}{2}\mu_k^{\boldsymbol{T}}(\beta_0 + \sum_n r_{nk})\Lambda_k\mu_k + \mu_k^{\boldsymbol{T}}\Lambda_k(\beta_0 m_0 + \sum_n r_{nk}x_n) \\ &\quad \beta_k = \beta_0 + N_k, \quad m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k\bar{x}_k) \text{ と置くと} \\ &= -\frac{1}{2}\mu_k^{\boldsymbol{T}}(\beta_k\Lambda_k)\mu_k + \mu_k^{\boldsymbol{T}}(\beta_k\Lambda_k)m_k.\end{aligned}$$

よって

$$q^*(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k\Lambda_k)^{-1}).$$

残りを考える.

$$\begin{aligned}
\log q^*(\Lambda_k) &= \log q^*(\mu_k, \Lambda_k) - \log q^*(\mu_k|\Lambda_k) \\
&= \frac{1}{2}\log|\beta_0\Lambda_k| - \frac{1}{2}(\mu_k - m_0)^{\boldsymbol{T}}(\beta_0\Lambda_k)(\mu_k - m_0) + \frac{\nu_0 - D - 1}{2}\log|\Lambda_k| \\
&\quad -\frac{1}{2}\operatorname{tr}(W_0^{-1}\Lambda_k) + \frac{1}{2}N_k\log|\Lambda_k| - \frac{1}{2}\sum_n r_{nk}(x_n - \mu_k)^{\boldsymbol{T}}\Lambda_k(x_n - \mu_k) \\
&\quad -\frac{1}{2}\log|\beta_0\Lambda_k| + \frac{1}{2}(\mu_k - m_k)^{\boldsymbol{T}}(\beta_0\Lambda_k)(\mu_k - m_k) \\
&= \frac{\nu_k - D - 1}{2} - \frac{1}{2}\log|\Lambda_k| - \frac{1}{2}\operatorname{tr}((\beta_0\Lambda_k)(\mu_k - m_0)(\mu_k - m_0)^{\boldsymbol{T}} \\
&\quad +\sum_n r_{nk}\Lambda_k(x_n - \mu_k)(x_n - \mu_k)^{\boldsymbol{T}} - \beta_k\Lambda_k(\mu_k - m_k)(\mu_k - m_k)^{\boldsymbol{T}}) - \frac{1}{2}(\Lambda_k W_0^{-1}) \\
&\quad v_k = v_0 + N_k \text{と置く} \\
&= \frac{\nu_k - D - 1}{2} - \frac{1}{2}\operatorname{tr}(\Lambda_k(W_0^{-1} + \beta_0(\mu_k - m_0)(\mu_k - m_0)^{\boldsymbol{T}} \\
&\quad +\sum_n r_{nk}(x_n - \mu_k)(x_n - \mu_k)^{\boldsymbol{T}} - \beta_k(\mu_k - m_k)(\mu_k - m_k)^{\boldsymbol{T}}) \\
&= \frac{\nu_k - D - 1}{2} - \frac{1}{2}\operatorname{tr}(\Lambda_k W_k^{-1}) \text{ と置く.}
\end{aligned}$$

W_k を求めよう. まず

$$\begin{aligned}
\sum_n r_{nk}x_n x_n{}^{\boldsymbol{T}} &= \sum_n r_{nk}((x_n - \bar{x}_k)(x_n - \bar{x}_k)^{\boldsymbol{T}} + 2x_n\bar{x}_k - \bar{x}\bar{x}^{\boldsymbol{T}}) \\
&= N_k S_k + 2_N k + 2N_k\bar{x}_k\bar{x}_k^{\boldsymbol{T}} - N_k\bar{x}_k\bar{x}_k^{\boldsymbol{T}} \\
&= N_k S_k + N_k\bar{x}_k\bar{x}_k^{\boldsymbol{T}}.
\end{aligned}$$

よって

$$\begin{aligned}
W_k^{-1} &= W_0^{-1} + \beta_0(\mu_k\mu_k{}^{\boldsymbol{T}} - 2\mu_k m_0^{\boldsymbol{T}} + m_0 m_0{}^{\boldsymbol{T}}) + N_k S_k + N_k\bar{x}_k\bar{x}_k^{\boldsymbol{T}} - 2\sum_n r_{nk}x_n\mu_k^{\boldsymbol{T}} \\
&\quad +\sum_n r_{nk}\mu_k\mu_k{}^{\boldsymbol{T}} - (\beta_0 + N_k)(\mu_k\mu_k{}^{\boldsymbol{T}} - 2\mu_k\frac{1}{\beta_k}(\beta_0 m_0 + N_k\bar{x}_k)^{\boldsymbol{T}} \\
&\quad +\frac{1}{\beta_k^2}(\beta_0 m_0 + N_k\bar{x}_k)(\beta_0 m_0 + N_k\bar{x}_k)^{\boldsymbol{T}} \\
&\quad \sum_n r_{nk} = N_k \text{に注意して} \\
&= W_0^{-1} + N_k S_k + \beta_{m_0 m_0{}^{\boldsymbol{T}}} + N_k\bar{x}_k\bar{x}_k^{\boldsymbol{T}} \\
&\quad -\frac{1}{\beta_k}(\beta_0^2 m_0 m_0{}^{\boldsymbol{T}} + 2\beta_0 N_k m_0\bar{x}_k^{\boldsymbol{T}} + N_k^2\bar{x}_k\bar{x}_k^{\boldsymbol{T}}) \\
&= W_0^{-1} + N_S S_k + \frac{(\beta_0 + N_k)\beta_0 - \beta_0^2}{\beta_k}m_0 m_0{}^{\boldsymbol{T}} + \frac{(\beta_0 + N_k)N_k - N_k^2}{\beta_k}\bar{x}_k\bar{x}_k^{\boldsymbol{T}} - \frac{2\beta_0 N_k}{\beta_k}m_0\bar{x}_k^{\boldsymbol{T}} \\
&= W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_k}(m_0 - \bar{x}_k)(m_0 - \bar{x}_k)^{\boldsymbol{T}}.
\end{aligned}$$

よって

$$q^*(\Lambda_k) = \mathcal{W}(\Lambda_k|W_k, \nu_k), \quad q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|W_k, \nu_k).$$

$\mathcal{N}(x|\mu, \Lambda^{-1})$ について $E[xx^{\boldsymbol{T}}] = \mu\mu^{\boldsymbol{T}} + \Lambda^{-1}$, $\mathcal{W}(\Lambda_k|W_k, \nu_k)$ について $E[\Lambda_k] = \nu_k W_k$ なので

$$\begin{aligned}
E_{\mu_k,\Lambda_k}[(x_n - \mu_k)^{\boldsymbol{T}}\Lambda_k(x_n - \mu_k)] &= \mathrm{tr}(E[\Lambda_k x_n {x_n}^{\boldsymbol{T}}] - 2E[\Lambda_k x_n \mu_k^{\boldsymbol{T}}] + E[\Lambda_k \mu_k {\mu_k}^{\boldsymbol{T}}]) \\
&= \mathrm{tr}\, E[\nu_k W_k x_n {x_n}^{\boldsymbol{T}}] - 2\,\mathrm{tr}\, E[\nu_k W_k x_n m_k^{\boldsymbol{T}}] \\
&\quad + \mathrm{tr}\, E[\Lambda_k(m_k {m_k}^{\boldsymbol{T}} + (\beta_k\Lambda_k)^{-1}] \\
&= \nu_k\,\mathrm{tr}(W_k x_n {x_n}^{\boldsymbol{T}}) - 2\nu_k\,\mathrm{tr}(W_k x_n m_k^{\boldsymbol{T}}) + \mathrm{tr}(\nu_k W_k m_k {m_k}^{\boldsymbol{T}}) + D\beta_k^{-1} \\
&= D\beta_k^{-1} + \nu_k(x_n - m_k)^{\boldsymbol{T}} W_k(x_n - m_k). \qquad (1)
\end{aligned}$$

ウィシャート分布の公式から

$$\log\tilde{\Lambda}_k := E[\log|\Lambda_k|] = \sum_i \phi(\frac{\nu_k + 1 - i}{2}) + D\log 2 + \log|W_k|.$$

ディリクレ分布の公式から

$$\log\tilde{\pi}_k := E[\log\pi_k] = \phi(\alpha_k) - \phi(\hat{\alpha}), \quad \hat{\alpha} = \sum_k \alpha_k.$$

$$\begin{aligned}
\log\rho_{nk} &= E[\log\pi_k] + \frac{1}{2}E[\log|\Lambda_k|] - \frac{D}{2}\log(2\pi) - \frac{1}{2}E[(x_n - \mu_k)^{\boldsymbol{T}}\Lambda_k(x_n - \mu_k)] \\
&= \log\tilde{\pi}_k + \frac{1}{2}\log\tilde{\Lambda}_k - \frac{D}{2}\log(2\pi) - \frac{1}{2}(D\beta_k^{-1} + \nu_k(x_n - m_k)^{\boldsymbol{T}} W_k(x_n - m_k)).
\end{aligned}$$

よって

$$r_{nk} \propto \rho_{nk} \propto \tilde{\pi}_k\Lambda_k^{1/2}\exp(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(x_n - m_k)^{\boldsymbol{T}} W_k(x_n - m_k)).$$

混合ガウス分布の EM アルゴリズムでの負担率は $\gamma(z_{nk}) \propto \pi_k\mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})$ だったので

$$r_{nk} \propto \pi_k|\lambda_k|^{1/2}\exp(-\frac{1}{2}(x_n - \mu_k)^{\boldsymbol{T}}\Lambda_k(x_n - \mu_k)).$$

これは上の式とよく似ている.

ディリクレ分布の平均の式 $E[\mu_k] = \alpha_k/\hat{\alpha}$ より

$$E[\pi_k] = \frac{\alpha_0 + N_k}{\sum_k \alpha_k} = \frac{\alpha_0 + N_k}{K\alpha_0 + \sum_k N_k} = \frac{\alpha_0 + N_k}{K\alpha_0 + N}.$$

ある混合要素 k について $r_{nk} \simeq 0$ なら $N_k \simeq 0$（PRML p.193 はかつとなってるけど片方の条件から出る）．このとき $\alpha_k \simeq \alpha_0$ となる．PRML10 章では分布が幅広いという状態を「なだらか」と表記しているようだ．ちょっとニュアンスが違う気もするけど．

事前分布で $\alpha_0 \to 0$ とすると $E[\pi_k] \to 0$. $\alpha_0 \to \infty$ なら $E[\pi_k] \to 1/K$.

7 変分下限

$q(Z, \boldsymbol{\pi}, \mu, \Lambda) = q(Z)q(\boldsymbol{\pi}, \mu, \Lambda)$ と分解できると仮定すると

$$\begin{aligned}
\mathcal{L}(q) &= \int q(Z)\log\frac{p(X,Z)}{q(Z)}\,dZ \\
&= \sum\int q(Z, \boldsymbol{\pi}, \mu, \Lambda)\log\frac{p(X, Z, \boldsymbol{\pi}, \mu, \Lambda)}{q(Z, \boldsymbol{\pi}, \mu, \Lambda)}\,d\boldsymbol{\pi}d\mu d\Lambda \\
&= E[\log p(X, Z, \boldsymbol{\pi}, \mu, \Lambda)] - E[\log q(Z, \boldsymbol{\pi}, \mu, \Lambda)] \\
&= E[\log p(X|Z, \mu, \Lambda)] + E[\log p(Z|\boldsymbol{\pi})] + E[\log p(\boldsymbol{\pi})] + E[\log p(\mu, \Lambda)] \\
&\quad - E[\log q(Z)] - E[\log q(\boldsymbol{\pi})] - E[\log q(\mu, \Lambda)].
\end{aligned}$$

以下, ひたすら計算する.

$$\begin{aligned}
E[\log p(X|Z,\mu,\Lambda)] &= E[\sum_{n,k} z_{nk} \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})] \\
&= \frac{1}{2} E[\sum_{n,k} z_{nk}(-D\log(2\pi) + \log|\Lambda_k| - (x_n-\mu_k)^{\boldsymbol{T}}\Lambda_k(x_n-\mu_k))] \\
&= \frac{1}{2}\sum_k E[-N_k D\log(2\pi) + N_k \log|\Lambda_k| - \sum_n z_{nk}(x_n-\mu_k)^{\boldsymbol{T}}\Lambda_k(x_n-\mu_k)] \\
&= \frac{1}{2}\sum_k N_k(\log\tilde{\Lambda}_k - D\log(2\pi)) - \frac{1}{2}\underbrace{\sum_{n,k} r_{nk}(D\beta_k^{-1} + \nu_k(x_n-m_k)^{\boldsymbol{T}}W_k(x_n-m_k))}_{X\text{ とおく}}.
\end{aligned}$$

$$X = \sum_k N_k D\beta_k^{-1} + \sum_k \nu_k(\underbrace{\sum_n r_{nk}(x_n-m_k)^{\boldsymbol{T}}W_k(x_n-m_k)}_{Y\text{ とおく}}).$$

$$\sum_n r_{nk}x_n x_n{}^{\boldsymbol{T}} = N_k S_k + N_k \bar{x}_k \bar{x}_k^{\boldsymbol{T}}$$

より

$$\begin{aligned}
Y &= \mathrm{tr}(W_k(\sum_n r_{nk}x_n x_n{}^{\boldsymbol{T}} - 2\sum_n r_{nk}x_n m_k^{\boldsymbol{T}} + \sum_n r_{nk}m_k m_k{}^{\boldsymbol{T}})) \\
&= \mathrm{tr}(W_k(N_k S_k + N_k\bar{x}_k\bar{x}_k^{\boldsymbol{T}} - 2N_k\bar{x}_k m_k^{\boldsymbol{T}} + N_k m_k m_k{}^{\boldsymbol{T}})) \\
&= N_k\,\mathrm{tr}(W_k(S_k + (\bar{x}_k-m_k)(\bar{x}_k-m_k)^{\boldsymbol{T}})) \\
&= N_k(\mathrm{tr}(S_k W_k) + (\bar{x}_k-m_k)^{\boldsymbol{T}}W_k(\bar{x}_k-m_k)).
\end{aligned}$$

よって

$$\begin{aligned}
E[\log p(X|Z,\mu,\Lambda)] = \frac{1}{2}\sum_k N_k(&\log\tilde{\Lambda}_k - D\beta_k^{-1} - \nu_k\,\mathrm{tr}(S_k W_k) \\
&-\nu_k(\bar{x}_k-m_k)^{\boldsymbol{T}}W_k(\bar{x}_k-m_k) - D\log(2\pi)).
\end{aligned}$$

$$E[\log p(Z|\boldsymbol{\pi}) = E[\sum_{n,k} z_{nk}\log\pi_k] = \sum_{n,k} r_{nk}\log\tilde{\pi}_k.$$

$$E[\log p(\boldsymbol{\pi})] = E[\log C(\alpha_0) + \sum_k(\alpha_0-1)\log\pi_k] = \log C(\alpha_0) + (\alpha_0-1)\sum_k \log\tilde{\pi}_k.$$

$$E[\log q(Z)] = E[\sum_{n,k} z_{nk}\log r_{nk}] = \sum_{n,k}\log r_{nk}.$$

$$E[\log q(\boldsymbol{\pi})] = E[\log C(\alpha) + \sum_k(\alpha_k-1)\log\pi_k] = \log C(\alpha) + \sum_k(\alpha_k-1)\log\tilde{\pi}_k.$$

$$E[\log q(\mu,\Lambda)] = \sum_k E[\log \mathcal{N}(\mu_k|m_k,(\beta_k\Lambda_k)^{-1}) + \log \mathcal{W}(\lambda_k|W_k,\nu_k)]$$
$$= \sum_k E[-\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\beta_k\Lambda_k| - \frac{1}{2}(\mu_k - m_k)^{\boldsymbol{T}}(\beta_k\Lambda_k)(\mu_k - m_k)] + E[\log W]$$
$$= \sum_k \frac{1}{2}\log\tilde{\Lambda}_k + \frac{D}{2}\log(\frac{\beta_k}{2\pi}) - \frac{1}{2}\underbrace{\mathrm{tr}\,E[(\beta_k\Lambda_k)(\mu_k - m_k)(\mu_k - m_k)^{\boldsymbol{T}}]}_{X\text{ とおく}} + \underbrace{E[\log W]}_{Y\text{ とおく}}$$

$$X = \mathrm{tr}(\beta_k\Lambda_k)(E[\mu_k{\mu_k}^{\boldsymbol{T}}] - 2E[\mu_k]m_k^{\boldsymbol{T}} + m_k{m_k}^{\boldsymbol{T}})$$
$$= \mathrm{tr}(\beta_k\Lambda_k)(m_k{m_k}^{\boldsymbol{T}} + (\beta_k\Lambda_k)^{-1} - m_k{m_k}^{\boldsymbol{T}})$$
$$= \mathrm{tr}\, I = D.$$

$$Y = E[\log W(\Lambda_k|W_k,\nu_k)]$$
$$= \log B(W_k,\nu_k) + \frac{\nu_k - D - 1}{2}E[\log|\Lambda_k|] - \frac{1}{2}\mathrm{tr}(W_k^{-1}E[\Lambda_k])$$
$$= \log B(W_k,\nu_k) + \frac{\nu_k - D - 1}{2}E[\log|\Lambda_k|] - \frac{1}{2}\nu_k D = -H[\Lambda_k].$$

よって

$$E[\log q(\mu,\Lambda)] = \sum_k(\frac{1}{2}\log\tilde{\Lambda_k} + \frac{D}{2}\log(\frac{\beta_k}{2\pi}) - \frac{D}{2} - H[\Lambda_k]).$$

$$E[\log p(\mu,\Lambda)] = \sum_k \underbrace{E[\log\mathcal{N}(\mu_k|m_0,(\beta_0\Lambda_k)^{-1})]}_{A\text{ とおく}} + \underbrace{E[\log\mathcal{W}(\Lambda_k|W_0,\nu_0)]}_{B\text{ とおく}}.$$

$$A = E[-\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\beta_0\Lambda_k| - \frac{1}{2}(\mu_k - m_0)^{\boldsymbol{T}}(\beta_0\Lambda_k)(\mu_k - m_0)]$$
$$= \frac{D}{2}\log(\frac{\beta_0}{2\pi}) + \frac{1}{2}\log\tilde{\Lambda}_k - \frac{1}{2}\beta_0 E[(m_0 - \mu_k)^{\boldsymbol{T}}\Lambda_k(m_0 - \mu_k)]$$

2 ページ前の式（1）で $x_n = m_0$ として使うと

$$= \frac{1}{2}(D\log(\frac{\beta_0}{2\pi}) + \log\tilde{\Lambda}_k - \beta_0(D\beta_k^{-1} + \nu_k(m_0 - \mu_k)^{\boldsymbol{T}}W_k(m_0 - \mu_k))$$
$$= \frac{1}{2}(D\log(\frac{\beta_0}{2\pi}) + \log\tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} - \beta_0\nu_k(m_k - m_0)^{\boldsymbol{T}}W_k(m_k - m_0)).$$

$$B = E[\log B(W_0,\nu_0) + \frac{\nu_0 - D - 1}{2}\log|\Lambda_k| - \frac{1}{2}\mathrm{tr}(W_0^{-1}\Lambda_k)]$$
$$= \log B(W_0,\nu_0) + \frac{\nu_0 - D - 1}{2}\log\tilde{\Lambda})k - \frac{1}{2}\mathrm{tr}(W_0^{-1}\underbrace{E[\Lambda_k]}_{=\nu_k W_k}).$$

よって

$$E[\log p(\mu,\Lambda)] = \frac{1}{2}\sum_k(D\log(\frac{\beta_0}{2\pi} + \log\tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} - \beta_0\nu_k(m_k - m_0)^{\boldsymbol{T}}W_k(m_k - m_0))$$
$$+K\log B(W_0,\nu_0) + \frac{\nu_0 - D - 1}{2}\sum_k\log\tilde{\Lambda}_k - \frac{1}{2}\sum_k\nu_k\,\mathrm{tr}(W_0^{-1}W_k).$$

最後に $\mathcal{L}$ を求めよう.

$$\sum_k N_k = N,$$

$$H[q(\Lambda_k)] = -\log B(W_k, \nu_k) - \frac{\nu_k - D - 1}{2}\log\tilde{\Lambda}_k + \frac{\nu_k D}{2}$$

に注意する.

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\sum_k N_k \log\tilde{\Lambda}_k - \frac{1}{2}\sum_k N_k \frac{D}{\beta_k} - \frac{1}{2}\sum_k N_k\nu_k \operatorname{tr}(S_k W_k) - \frac{1}{2}\sum_k N_k\nu_k(\bar{x}_k - m_k)^{\boldsymbol{T}} W_k(\bar{x}_k - m_k) \\
&\quad -\frac{1}{2}\sum_k N_k D\log(2\pi) + \sum_k N_k \log\tilde{\pi}_k + \log C(\alpha_0) + (\alpha_0 - 1)\sum_k \log\tilde{\pi}_k + \frac{DK}{2}\log(\frac{\beta_0}{2\pi}) \\
&\quad +\frac{1}{2}\sum_k \log\tilde{\Lambda}_k - \frac{1}{2}\sum_k \frac{D\beta_0}{\beta_k} - \frac{1}{2}\sum_k \beta_0\nu_k(m_k - m_0)^{\boldsymbol{T}} W_k(m_k - m_0) + K\log B(W_0, \nu_0) \\
&\quad +\frac{\nu_0 - D - 1}{2}\sum_k \log\tilde{\Lambda}_k - \frac{1}{2}\sum_k \nu_k \operatorname{tr}(W_0^{-1}W_k) - \sum_{n,k} r_{nk}\log r_{nk} \\
&\quad -\sum_k(\alpha_k - 1)\log\tilde{\pi}_k - \log C(\alpha) - \frac{1}{2}\sum_k \log\tilde{\Lambda}_k - \frac{D}{2}\sum_k \log\frac{\beta_k}{2\pi} + \frac{DK}{2} \\
&\quad +\sum_k(-\log B(W_k, \nu_k) - \frac{\nu_k - D - 1}{2}\log\tilde{\Lambda}_k + \frac{\nu_k D}{2}) \\
&= \log\frac{C(\alpha_0)}{C(\alpha)} - \sum_{n,k} r_{nk}\log r_{nk} + \frac{1}{2}\sum_k \log\tilde{\Lambda}(N_k + 1 - \nu_0 - D - 1 - 1 - \nu_k + D + 1) \\
&\quad +\sum_k \log\tilde{\pi}_k(N_k + \alpha_0 - 1 - \alpha_k + 1) + K\log B(W_0, \nu_0) - \sum_k \log B(W_k, \nu_k) - \frac{DN}{2}\log(2\pi) \\
&\quad -\frac{D}{2}\underbrace{\sum_k(\frac{N_k}{\beta_k} + \frac{\beta_0}{\beta_k})}_{=K} + \frac{DK}{2}(\log\beta_0 - \log(2\pi)) - \frac{D}{2}\sum_k \log\beta_k + \frac{DK}{2}\log(2\pi) + \frac{DK}{2} + \frac{D}{2}\sum_k \nu_k \\
&\quad -\frac{1}{2}\sum_k \nu_k \operatorname{tr}(W_k(\underbrace{N_k S_k + N_k(\bar{x}_k - m_k)(\bar{x}_k - m_k)^{\boldsymbol{T}} + \beta_0(m_k - m_0)(m_k - m_0)^{\boldsymbol{T}} + W_0^{-1}}_{=X\text{ とおく}})) \\
&= \log\frac{C(\alpha_0)}{C(\alpha)} - \sum_{n,k} r_{nk}\log r_{nk} + K\log B(W_0, \nu_0) - \sum_k \log B(W_k, \nu_k) + \frac{DK}{2}\log\beta_0 - \frac{D}{2}\sum_k \log\beta_k \\
&\quad -\frac{DN}{2}\log(2\pi) + \frac{D}{2}\sum_k \nu_k - \frac{1}{2}\sum_k \nu_k \operatorname{tr}(W_k X).
\end{aligned}$$

$$\bar{x}_k - m_k = \bar{x}_k - \frac{1}{\beta_k}(\beta_0 m_0 + N_k\bar{x}_k) = \frac{1}{\beta_k}(\beta_k\bar{x}_k - N_k\bar{x}_k - \beta_0 m0) = \frac{\beta_0}{\beta_k}(\bar{x}_k - m_0).$$

$$m_k - m_0 = \frac{1}{\beta_k}(\beta_0 m_0 + N_k\bar{x}_k) - m_0 = \frac{1}{\beta_k}(N_k\bar{x}_k + \beta_0 m_0 - \beta_k m_0) = \frac{N_k}{\beta_k}(\bar{x}_k - m_0).$$

よって

$$\begin{aligned}
N_k(\bar{x}_k - m_k)(\bar{x}_k - m_k)^{\boldsymbol{T}} + \beta_0(m_k - m_0)(m_k - m_0)^{\boldsymbol{T}} &= (\frac{N_k\beta_0^2}{\beta_k^2} + \frac{\beta_0 N_k^2}{\beta_k^2}) \\
&= \frac{\beta_0 N_k}{\beta_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^{\boldsymbol{T}}.
\end{aligned}$$

よって

$$X = W_k^{-1}, \quad \sum_k \nu_k \operatorname{tr}(W_k W_k^{-1}) = D \sum_k \nu_k.$$

$$\mathcal{L} = \log \frac{C(\alpha_0)}{C(\alpha)} - \sum_{n,k} r_{nk} \log r_{nk} + \sum_k \log \frac{B(W_0, \nu_0)}{B(W_k, \nu_k)} + \frac{D}{2} \sum_k \log \frac{\beta_0}{\beta_k} - \frac{DN}{2} \log(2\pi).$$