概要

前回はEMアルゴリズムの一般的な解法について学んだ。その手続きとは、Eステップで潜在変数の事後分布を求め、Mステップで完全データ対数尤度の期待値を最大化するようにパラメータ更新するというものであった。さらに、混合ガウス分布についてEMアルゴリズムを適用する例を示し、 K-means アルゴリズムがハードな割り当て、EMアルゴリズムがソフトな割り当てという対比関係にあることについて考察した。

今回扱う内容ではさらに別のケースについてEMアルゴリズムを適用する例を学ぶ。前回学んだ通り、EMアルゴリズムは潜在変数を含む場合の最尤推定を行う一般的な手法であり、多くの応用ができることを確かめるのが今回の目的である。

9.3.3 節では混合ベルヌーイ分布を例にEMアルゴリズムを適用する。この例では混合ガウス分布のケースと同様に、どの混合要素に割り当てられるかを表す変数を潜在変数として扱うことができ、負担率の計算と期待値最大化をもってパラメータ更新を行う。さらに、手書き数字データを混合ベルヌーイ分布でモデル化し、正しくクラスタリングできることを確かめる。さらに、手書き数字データを混合ベルヌーイ分布でモデル化し、正しくクラスタリングできることを確かめる。

9.3.4 節ではベイズ線形回帰とRVMのハイパーパラメータ推定をEMアルゴリズムで解く例を与える。これらの問題は以前エビエンス近似（3.5 節）によって解法を得たが、改めてEMアルゴリズムで解いてみようという話である。さらに、EMアルゴリズムにおける収束解とエビデンス近似における収束解が一致することも確かめる。

9.3.3 混合ベルヌーイ分布

EMアルゴリズムを混合ベルヌーイ分布に適用する例を扱う。後の例でみるように、このモデルを扱うことで、2値の画像のクラスタリングなどを実行することができる。

前提条件の細かい部分は読み合わせとし、押さえておきたいポイントのみ記す。観測変数 $\mathbf{x}$ は次のような混合ベルヌーイ分布に従うとする。

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \qquad (9.47)$$

ただし、

$$p(\mathbf{x}|\boldsymbol{\mu}_k) = \underbrace{\prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{(1-x_i)}}_{\text{ベクトル}\,\mathbf{x}\,\text{の各要素が独立にベルヌーイ分布に従っている}} \qquad (9.48)$$

[イメージ]

$$\mathbf{x} = (x_1, x_2, \cdots, x_D)$$

$$\uparrow \quad \uparrow \qquad \uparrow \quad \text{独立にベルヌーイ分布に従って発生する。}$$

$$\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, \cdots, \mu_{kD})$$

このような $\boldsymbol{\mu}_k$ が K 本あり、混合分布を成している。

我々の目的は次の対数尤度関数を最大化するパラメータ $\boldsymbol{\mu}, \boldsymbol{\pi}$ の推定である。

$$\begin{aligned}
\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\pi}) &= \ln \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu},\boldsymbol{\pi}) \\
&= \ln \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k) \\
&= \sum_{n=1}^{N} \ln \Big\{ \underbrace{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}_{\ln\,\text{の中に}\,\sum_{k=1}^{K}\,\text{があるためバラせず困った。}} \Big\}
\end{aligned}$$

これを陽に解くことができないため、潜在変数を導入して完全データ対数尤度の期待値を代わりに最大化するというのが EM アルゴリズムの手続きであった。潜在変数は 1-of-K 符号化法を取った $\mathbf{z}$ とする。

zを導入した後の尤度と事前分布は次のように表せる。

$$p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k} \qquad (9.52)$$

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad (9.53)$$

$\mathbf{x}$ と $\mathbf{z}$ の同時分布を $\mathbf{z}$ について周辺化すると、（9.47）が再び現れるので、このモデルは混合ベルヌーイ分布を正しく表現できている［演習9.14］

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu})p(\mathbf{z}|\boldsymbol{\pi})$$

$$= \sum_{\mathbf{z}} \prod_{k=1}^{K} \left(\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\right)^{z_k}$$

$\mathbf{z} = (1, 0, \cdots, 0)$

$\vdots$ に分解して考える。

$\mathbf{z} = (0, 0, \cdots, 1)$

$=$ <中略>

$$= \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \qquad (9.47)$$

話を戻し、EMアルゴリズムにおけるEステップを考える。潜在変数の事後分布は次のように求まる。

$$p(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\mu},\boldsymbol{\pi}) = \frac{p(\mathbf{x}_n,\mathbf{z}_n|\boldsymbol{\mu},\boldsymbol{\pi})}{\sum_{\mathbf{z}_n} p(\mathbf{x}_n,\mathbf{z}_n|\boldsymbol{\mu},\boldsymbol{\pi})}$$

$$= \frac{p(\mathbf{x}_n|\mathbf{z}_n,\boldsymbol{\mu},\boldsymbol{\pi})p(\mathbf{z}_n|\boldsymbol{\pi})}{\sum_{\mathbf{z}_n} p(\mathbf{x}_n|\mathbf{z}_n,\boldsymbol{\mu},\boldsymbol{\pi})p(\mathbf{z}_n|\boldsymbol{\pi})}$$

$$= \frac{\sum_{k'}^{K} \left\{\pi_{k'} p(\mathbf{x}_n|\boldsymbol{\mu}_k)\right\}^{z_{nk'}}}{\sum_{\mathbf{z}_n}\sum_{k'}^{K} \left\{\pi_{k'} p(\mathbf{x}_n|\boldsymbol{\mu}_k)\right\}^{z_{nk'}}}$$

後の期待値計算に必要な、k要素についての期待値は次のようになる。

$$E[z_{nk}] = \gamma(z_{nk}) = \sum_{z_n} z_{nk}\, p(z_n \mid x_n, \mu, \pi)$$

$$= \frac{\sum_{z_n} z_{nk} \prod_{k'}^{K} \left\{ \pi_{k'} p(x_n \mid \mu_k) \right\}^{z_{nk'}}}{\sum_{z_n} \prod_{k'}^{K} \left\{ \pi_{k'} p(x_n \mid \mu_k) \right\}^{z_{nk'}}}$$

$z_n = (1, 0, \cdots, 0)$
$\vdots$
$z_n = (0, 0, \cdots, 1)$ に分解する

$= \langle$ 中略 $\rangle$

$$= \frac{\pi_k p(x_n \mid \mu_k)}{\sum_{j=1}^{K} \pi_j p(x_n \mid \mu_j)} \qquad (9.56)$$

次に、Mステップについて考える。Mステップは完全データ対数尤度の期待値を算出し、これのパラメータに関する停留点を求めるという手順であった。まず、完全データ対数尤度の期待値は次のようになる。

$$E_z\left[\ln p(X, Z \mid \mu, \pi) \right]$$

$$= E_z\left[\ln p(X \mid Z, \mu) + \ln p(Z \mid \pi) \right]$$

$$= E_z\left[\sum_n^N \ln \underbrace{p(x_n \mid z_n, \mu)}_{\downarrow (9.52)} + \sum_n^N \ln \underbrace{p(z_n \mid \pi)}_{\downarrow (9.53)} \right]$$

$$= E_z\left[\sum_n^N \ln \prod_k^K \underbrace{p(x \mid \mu_k)}_{\downarrow (9.48)}{}^{z_{nk}} + \sum_n^N \ln \prod_k^K \pi_k^{z_{nk}} \right]$$

$$= E_z\left[\sum_n^N \ln \prod_k^K \left\{ \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}} \right\}^{z_{nk}} + \sum_n^N \ln \prod_k^K \pi_k^{z_{nk}} \right]$$

$= \langle$ 中略 $\rangle$

$$= E_z\left[\sum_n^N \sum_k^K z_{nk} \left\{ \ln \pi_k + \sum_i^D \left[x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln(1-\mu_{ki}) \right] \right\} \right]$$

$$= \sum_n^N \sum_k^K \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_i^D \left[x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln(1-\mu_{ki}) \right] \right\}$$

まず、Mステップにおける μ_k の更新式を得る。上で得られた期待値を μ_{ki} で微分し、0とおくと、次の式が更新式が得られる。

$$0 = \frac{dE}{d\mu_{ki}} = \frac{d}{d\mu_{ki}}\left\{\sum_n^N \gamma(z_{nk})\left[x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\right]\right\} + 0$$

$$= \sum_n^N \gamma(z_{nk})\left(\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\right)$$

$$\Leftrightarrow \quad 0 = \sum_n^N \gamma(z_{nk})\left\{x_{ni}(1-\mu_{ki}) - \mu_{ki}(1-x_{ni})\right\}$$

$$\Leftrightarrow \quad 0 = \sum_n^N \gamma(z_{nk})(x_{ni} - \mu_{ki})$$

$$\begin{cases} \overline{x_{ki}} = \frac{1}{N_k}\sum_n^N \gamma(z_{nk})x_{ni} \\ N_k = \sum_n^N \gamma(z_{nk}) \end{cases} \quad \text{とおく}$$

$$\Leftrightarrow \quad 0 = N_k\overline{x_{ki}} - N_k\mu_{ki}$$

$$\Leftrightarrow \quad \mu_{ki} = \overline{x_{ki}}$$

$$\therefore \quad \mu_k = \overline{x_k}$$

次に、Mステップにおける π の更新式を得る。π についての最大化は、制約式 $\sum_k \pi_k = 1$ を考慮したラグランジュ未定乗数法を用いる。ラグランジュ関数を L は次のようになる。

$$L = \sum_n^N \sum_k^K \gamma(z_{nk})\left\{\ln\pi_k + \sum_i^D\left[x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\right]\right\} - \lambda\left(\sum_k^K \pi_k - 1\right)$$

これを π_k で微分すると、次の式を得る。

$$0 = \frac{dL}{d\pi_k} = \sum_n^N \gamma(z_{nk})\frac{1}{\pi_k} - \lambda$$

$$= \frac{N_k}{\pi_k} - \lambda$$

$$\Leftrightarrow \quad \pi_k = \frac{N_k}{\lambda}$$

ここで、両辺に $\sum_k^K$ を与えると、$\sum_k^K \pi_k = 1$、$\sum_k^K N_k = N$ より、$\lambda = N$ を得る。
よって、

$$\pi_k = \frac{N_k}{N}$$

手書き画像における応用については読み合わせとする。

9.3.4 ベイズ線形回帰に関するEMアルゴリズム

これまでは混合分布に対するEMアルゴリズムの適用例を見てきたが、この節ではエビデンス近似をEMアルゴリズムで解く例を扱う。混合分布の場合、Eステップをいわゆる「ソフト割り当て」として解釈できたが、この例ではもはやそのような解釈はできない。一般的なEMアルゴリズムではEステップを「潜在変数の事後分布」としての解釈したように、あくまで定義通りのEMアルゴリズムに則って各ステップを確かめていく。（この節のタイトルは「ベイズ線形回帰に関する」よりも「エビデンス近似に関する」の方が正確な気がした。というのも後半で扱うRVMモデルがベイズ線形回帰であるとはすぐに結びつかないためである。）

前半ではベイズ線形回帰に関するエビデンス近似をEMアルゴリズムで解く手続きを見ていく。エビデンス近似は、モデルパラメータ $\mathbf{w}$ で周辺化した尤度（=モデルエビデンス）を最大にするハイパーパラメータを決定する枠組みであった。以前学んだ際は、エビデンスをハイパーパラメータで微分して0とおくことで更新式を得たが、EMアルゴリズムでは $\mathbf{w}$ を潜在変数とした期待値を微分することで更新式を得る。ここの違いに注目して詳しく見ていく。

後半ではRVMのエビデンス近似にEMアルゴリズムを適用する。ここでも $\mathbf{w}$ を潜在変数としてハイパーパラメータの更新式を得る。計算式の中身は違うものの、おおまかな手続きは共通していることを確かめる。

■ ベイズ線形回帰のEMアルゴリズム

まず、ベイズ線形回帰に関してすでに得られている結論をまとめる。

尤度関数 : $p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$ (3.10)

事前分布 : $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ (3.52)

事後分布 : $p(\mathbf{w}|\mathbf{t},\alpha,\beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ (3.49)

対数エビデンス : $\ln p(\mathbf{t}|\alpha,\beta) = \ln\int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,d\mathbf{w}$ (3.77)

$$= \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{\beta}{2}\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2 - \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi) \quad (3.86)$$

エビデンス近似の枠組みでは、対数エビデンスを最大にする α, β の値を得るために、それぞれのパラメータの微分を0とおくことで更新式を得た。ただし、ここでは陽な解が得られず、有向パラメータ数 $\gamma = \sum(\lambda_i/\alpha+\lambda_i)$ の更新と α, β の更新を繰り返すことにより解を得ていた（See. 上巻 p.168）

一方 EM アルゴリズムでは、エビデンスを最大化するために、 $\mathbf{w}$ を潜在変数とした事後分布推定（Eステップ）と、完全データ対数尤度の $\mathbf{w}$ に関する期待値最大化（Mステップ）を実施する。 $\mathbf{w}$ の事後分布推定は式 (3.49) で既に求まっているので、ここではMステップにのみ着目する。完全データ対数尤度の期待値は次のようになる。

$$
\begin{aligned}
E_{\mathbf{w}}[\ln p(\mathbf{t}, \mathbf{w} \mid \alpha, \beta)]
&= E_{\mathbf{w}}\left[\ln p(\mathbf{t} \mid \mathbf{w}, \beta) + \ln p(\mathbf{w} \mid \alpha)\right] \\
&= E_{\mathbf{w}}\left[\sum_{n=1}^{N} \ln \mathcal{N}(t_n \mid \mathbf{w}^T \boldsymbol{\phi}_n, \beta^{-1}) + \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1} \mathbf{I})\right] \\
&= E_{\mathbf{w}}\left[\sum_{n=1}^{N}\left\{\frac{1}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2\right\} + \frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbf{w}'\mathbf{w}\right] \\
&= \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N} E_{\mathbf{w}}\left[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2\right] + \frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}E_{\mathbf{w}}[\mathbf{w}'\mathbf{w}]
\end{aligned}
\tag{9.62}
$$

これの α に関する微分を0とおくと次の更新式を得る。

$$
\begin{aligned}
0 = \frac{dE}{d\alpha} &= \frac{M}{2}\frac{2\pi}{\alpha}\frac{1}{2\pi} - \frac{1}{2}E_{\mathbf{w}}[\mathbf{w}'\mathbf{w}] \\
&= \frac{M}{2\alpha} - \frac{1}{2}\sum_{i=1}^{M} E_{\mathbf{w}}[w_i^2] \\
&= \frac{M}{2\alpha} - \frac{1}{2}\sum_{i=1}^{M}\left(\mathrm{Cov}[w_i] + E[w_i]^2\right) \\
&= \frac{M}{2\alpha} - \frac{1}{2}\sum_{i=1}^{M}\left(S_{N\,ii} + m_i^2\right) \\
&= \frac{M}{2\alpha} - \frac{1}{2}\left(\mathrm{Tr}(S_N) + \mathbf{m}_N'\mathbf{m}_N\right)
\end{aligned}
$$

$$
\Leftrightarrow \quad \alpha = \frac{M}{\mathbf{m}_N'\mathbf{m}_N + \mathrm{Tr}(S_N)} \tag{9.63}
$$

次にβに関する微分を0と置くと次の更新式を得る。

$$
\begin{aligned}
0 = \frac{dE}{d\beta} &= \frac{N}{2}\,\frac{2\pi}{\beta}\,\frac{1}{2\pi} - \frac{1}{2}\sum_{n=1}^{N} E_w\left[(t_n - w'\phi_n)^2 \right] \\
&= \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N} E_w\left[\left\{ (t_n - m'\phi_n) + (m'\phi_n - w'\phi_n) \right\}^2 \right] \\
&= \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N} \Big\{ E_w\left[(t_n - m'\phi_n)^2 \right] + 2(t_n - m'\phi_n)\underbrace{(m'\phi_n - E[w']\phi_n)}_{= m'\phi_n - m'\phi_n = 0} \\
&\qquad + E_w\left[(m'\phi_n - w'\phi_n)^2 \right] \Big\} \\
&= \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N} \Big\{ (t_n - m'\phi_n)^2 + 0 + \underbrace{E_w\left[\phi_n'(m - w)\cdot(m - w)'\phi_n \right]}_{= \phi_n' E_w\left[(w - m)(w - m)'\right]\phi_n = \phi_n' S_N \phi_n} \Big\} \\
&= \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N} \left\{ (t_n - m'\phi_n)^2 \right\} - \frac{1}{2}\sum_{n=1}^{N} \phi_n' S_N \phi_n \\
&= \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N} \left\{ (t_n - m'\phi_n)^2 \right\} - \frac{1}{2}\mathrm{Tr}(\Phi S_N \Phi^T) \\
&= \frac{N}{2\beta} - \frac{1}{2}(t - \Phi m)'(t - \Phi m) - \frac{1}{2}\mathrm{Tr}(\Phi S_N \Phi^T) \\
&= \frac{N}{2\beta} - \frac{1}{2}\| t - \Phi m \|^2 - \frac{1}{2}\mathrm{Tr}(\Phi S_N \Phi^T)
\end{aligned}
$$

説明できないけど、こうするとRVMのとき教科書と一致する

$$
\Leftrightarrow \quad \beta^{-1} = \frac{\| t - \Phi m \|^2 + \mathrm{Tr}(\Phi S_N \Phi^T)}{N}
$$

以上より、Mステップの更新式を得られた。EMアルゴリズムの手順をまとめると、Eステップにおいて w の事後分布を表す m_N と S_N を求め、Mステップにおいて上記更新式に m_N と S_N を代入し、新たな α と β を更新する。

この手順で得られたハイパーパラメータは、(3.92) で導出した解析計算によるハイパーパラメータと同じ結果に収束する。これは、(3.92) の式を変形することで導かれる。

$$(3.92) \iff \alpha = \frac{\gamma}{m_N' m_N}$$

$$= \frac{1}{m_N' m_N}\left(M - \alpha \sum_i^M \frac{1}{\lambda_i + \alpha}\right) \qquad (\because 3.90)$$

$$= \frac{1}{m_N' m_N}\left\{M - \alpha \, \mathrm{Tr}(A^{-1})\right\}$$

$\left(\because\ \lambda_i + \alpha\right.$ は A の固有値なので、$\mathrm{Tr}(A) = \sum(\lambda_i + \alpha)$ (C.48) $\left.\iff \mathrm{Tr}(A^{-1}) = \sum \frac{1}{\lambda_i + \alpha}\right)$

$$= \frac{1}{m_N' m_N}\left\{M - \alpha \, \mathrm{Tr}(S_N)\right\} \qquad (\because 3.54,\ 3.81)$$

$$\iff \alpha = \frac{M}{m_N' m_N + \mathrm{Tr}(S_N)}$$

(9.63) と一致した

■ RVMのEMアルゴリズム

7.2.1節で扱ったRVMのエビデンス近似を、EMアルゴリズムによって解く手続きを確かめる。まずは、RVMにおいて既に得られている結論をまとめる。

尤度関数 : $p(t \mid \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t \mid \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \beta^{-1})$ (7.76) (7.77)

事前分布 : $p(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i \mid \alpha_i^{-1})$ (7.80)

事後分布 : $p(\mathbf{w} \mid \mathbf{t}, \mathbf{X}, \boldsymbol{\alpha}, \beta) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \boldsymbol{\Sigma})$ (7.81)

ベイズ線形回帰では S_N

対数エビデンス : $\ln p(\mathbf{t} \mid \mathbf{X}, \boldsymbol{\alpha}, \beta) = -\frac{1}{2}\left\{N \ln(2\pi) + \ln|\mathbf{C}| + \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t}\right\}$ (7.85)

ただし、$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi} \mathbf{A}^{-1} \boldsymbol{\Phi}^T$

繰り返しになるが、7.2.1 では周辺化によって得られていた対数エビデンス関数を最大化していたのを、EMアルゴリズムでは期待値を最大化する。Eステップは (7.81) そのものである。Mステップで必要な完全データ対数尤度の期待値は次のようになる。

$$
\begin{aligned}
&E_{\mathbf{w}}\left[\ln p(\mathbf{t}, \mathbf{w} \mid \mathbf{X}, \boldsymbol{\alpha}, \beta)\right] \\
&= E_{\mathbf{w}}\left[\ln p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) + \ln p(\mathbf{w} \mid \boldsymbol{\alpha})\right] \\
&= E_{\mathbf{w}}\left[\sum_{n=1}^{N} \ln \mathcal{N}(t_n \mid \mathbf{w}^T \boldsymbol{\phi}_n, \beta^{-1}) + \sum_{i=1}^{M} \ln \mathcal{N}(w_i \mid 0, \alpha_i^{-1})\right] \\
&= E_{\mathbf{w}}\left[\sum_{n=1}^{N}\left\{\frac{1}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2\right\} + \sum_{i=1}^{M}\left\{\frac{1}{2}\ln\left(\frac{\alpha_i}{2\pi}\right) - \frac{\alpha_i}{2} w_i^2\right\}\right] \\
&= \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2} E_{\mathbf{w}}\left[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2\right] + \frac{1}{2}\sum_{i=1}^{M}\left\{\ln\left(\frac{\alpha_i}{2\pi}\right) - \alpha_i E\left[w_i^2\right]\right\}
\end{aligned}
$$

α_i に関する微分を0と置くと次の更新式を得る。

$$0 = \frac{dE}{d\alpha_i} = \frac{1}{2}\left\{ \frac{2\pi}{\alpha_i} \frac{1}{2\pi} - E[w_i^2] \right\}$$

$$= \frac{1}{2\alpha_i} - \frac{1}{2}\left\{ V(w_i) + E(w_i)^2 \right\}$$

$$= \frac{1}{2\alpha_i} - \frac{1}{2}\left(\Sigma_{ii} + m_i^2 \right)$$

$$\Leftrightarrow \quad \alpha_i = \frac{1}{m_i^2 + \Sigma_{ii}} \qquad (9.67)$$

同様に、 β に関する微分を0と置くと次の更新式を得る。

$$0 = \frac{dE}{d\beta} = \frac{N}{2} \frac{2\pi}{\beta} \frac{1}{2\pi} - \frac{1}{2} E_{w}\left[(t_n - w' \phi_n)^2 \right]$$

= < 8ページと全く同じ式 >

$$= \frac{N}{2\beta} - \frac{1}{2} \| t - \Phi m \|^2 - \frac{1}{2} \mathrm{Tr}(\Phi \Sigma \Phi^T)$$

$$\Leftrightarrow \quad \beta^{-1} = \frac{\| t - \Phi m \|^2 + \mathrm{Tr}(\Phi \Sigma \Phi^T)}{N}$$

$$= \frac{\| t - \Phi m \|^2 + \beta^{-1} \sum_{i=1}^{M} \gamma_i}{N} \qquad (9.68)$$

(ただし、$\gamma_i = 1 - \alpha_i \Sigma_{ii}$ (7.89) である)

(9.68)で用いた $\mathrm{Tr}(\Phi\Sigma\Phi^T) = \beta^{-1}\sum_{i=1}^{m}\gamma_i$ を示す。

w の事後分布の共分散の式 $\Sigma = (A + \beta\Phi^T\Phi)^{-1}$ (7.83) より示せる。

$$
\begin{aligned}
(7.83) \iff\ & \Sigma(A + \beta\Phi^T\Phi) = I \\
\iff\ & \Sigma A + \beta\Sigma\Phi^T\Phi = I \\
\iff\ & \beta\Sigma\Phi^T\Phi = I - \Sigma A \\
\iff\ & \beta\,\mathrm{Tr}(\Sigma\Phi^T\Phi) = \mathrm{Tr}(I - \Sigma A) \\
\iff\ & \mathrm{Tr}(\Phi\Sigma\Phi^T) = \frac{1}{\beta}\sum_{i=1}^{m}\left\{1 - \alpha_i\Sigma_{ii}\right\} \quad (\because \text{トレースの循環性}) \\
\iff\ & \mathrm{Tr}(\Phi\Sigma\Phi^T) = \beta^{-1}\sum_{i=1}^{m}\gamma_i \qquad \square
\end{aligned}
$$

以上より、EMアルゴリズムにおけるエビデンス近似の解法を得た。この更新式の収束先は、7.2.1 での更新式の収束先と同じらしいが、ベイズ線形回帰と同じようなお話なので照明は割愛する（演習9.23）

（考察）

我々はエビデンス近似により最適なハイパーパラメータを更新する方法を2つ手に入れた。当然、どちらの方が優れているかは気になるところである。下巻 p.58 の下の方では、「（EMアルゴリズムよりも）直接最適化を行う手法の方がやや高速である」とあるので、EMアルゴリズムいらないじゃん、と思われるが、実はそうでもなさそう。

3章や7章で扱ったエビデンス近似では、尤度関数について潜在変数 w の周辺化が解析的に実行することができ、エビデンス関数が陽な式で求められているという前提があるのである。一方で、EMアルゴリズムでは潜在変数の周辺化は不要であり、完全データ対数尤度についての期待値さえ求められれば良い。つまり、より緩い条件のもとで実行できるのはEMアルゴリズムなのかもしれない。