10.3.3 変分下界

これまでの議論で、ベイズ線形回帰における事後分布と予測分布を変分推論で求めることができた。混合ガウスモデルと同様に、ベイズ線形回帰においても共役事前分布を設定することで、事前分布と同一の分布が更新式で得られることを確かめた。

この節では、次元数の異なる線形回帰モデルを比較するために、**変分下界**（エビデンス下界）を求めることとする。変分下界の導出を行い、これを実際のデータに当てはめてみた場合に、複雑なモデルが選択されなくなるケースを確かめる。

変分下界の導出に入る前に、変分下界を求めるモチベーションについて改めて整理したい。混合ガウス分布で変分下界を求めた節（10.2.2節）では、事後分布の収束判定のために変分下界を求めていた。今回のベイズ線形回帰でも同じく収束判定のために利用できるが、ここでは特にモデル比較のために利用する。

「10.1.4 モデル比較」では、下記の式より、モデルの事後分布をモデルの事前分布と変分下界で表すことができた。

$$q(m) \propto p(m)\exp\{\mathcal{L}_m\}$$

ただし $\mathcal{L}_m = \sum_{\mathbf{Z}} q(\mathbf{Z}|m)\ln\left\{\frac{p(\mathbf{Z},\mathbf{X}|m)}{q(\mathbf{Z}|m)}\right\}$

モデル m における変分下界

モデルの事前分布に一様な分布を仮定すると、モデルの事後分布 $q(m)$ は $\mathcal{L}_m$ のみに比例することとなり、事後分布の大小は変分下界のみに依存する。事後分布最大のモデルを選択するには、次元数の異なる各のモデルで変分下界を求め、最大のモデルを選択すれば良い。

次に変分下界の導出に入る。変分下界はこれまで得られた分解の式を使って、次のように表現できる。

$$\begin{aligned}
\mathcal{L}(q) &= \iint q(\mathbf{w},\alpha)\ln\left\{\frac{p(\mathbf{w},\alpha,\mathbf{t})}{q(\mathbf{w},\alpha)}\right\} d\mathbf{w}\,d\alpha \\
&= \iint q(\mathbf{w},\alpha)\ln p(\mathbf{w},\alpha,\mathbf{t})\,d\mathbf{w}\,d\alpha - \iint q(\mathbf{w},\alpha)\ln q(\mathbf{w},\alpha)\,d\mathbf{w}\,d\alpha \\
&= E[\ln p(\mathbf{w},\alpha,\mathbf{t})] - E[\ln q(\mathbf{w},\alpha)] \\
&= E[\ln p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)] - E[\ln q(\mathbf{w})q(\alpha)] \\
&= E_{\mathbf{w}}[\ln p(\mathbf{t}|\mathbf{w})] + E_{\mathbf{w}}[\ln p(\mathbf{w}|\alpha)] + E_{\alpha}[\ln p(\alpha)] \\
&\quad - E_{\mathbf{w}}[\ln q(\mathbf{w})] - E_{\mathbf{w}}[\ln q(\alpha)]
\end{aligned}$$

ここで分解された期待値は分布の性質を使って次のように求めることができる。

$$
\begin{aligned}
E_w[\ln p(t|w)] &= E_w\Big[\sum_{n}^{N}\ln \mathcal{N}(t_n|w^T\phi_n,\beta^{-1})\Big] \\
&= \frac{N}{2}\ln\Big(\frac{\beta}{2\pi}\Big) - \frac{\beta}{2}E_w\big[(t-\Phi w)^T(t-\Phi w)\big] \\
&= \frac{N}{2}\ln\Big(\frac{\beta}{2\pi}\Big) - \frac{\beta}{2}t^Tt + \beta E_w[w^T\Phi^T]t - \frac{\beta}{2}E_w[w^T\Phi^T\Phi w] \\
&= \frac{N}{2}\ln\Big(\frac{\beta}{2\pi}\Big) - \frac{\beta}{2}t^Tt + \beta m_N^T\Phi^Tt - \frac{\beta}{2}E_w[Tr(\Phi^T\Phi ww^T)] \\
&= \frac{N}{2}\ln\Big(\frac{\beta}{2\pi}\Big) - \frac{\beta}{2}t^Tt + \beta m_N^T\Phi^Tt - \frac{\beta}{2}Tr(\Phi^T\Phi(m_Nm_N^T+S_N))
\end{aligned}
$$

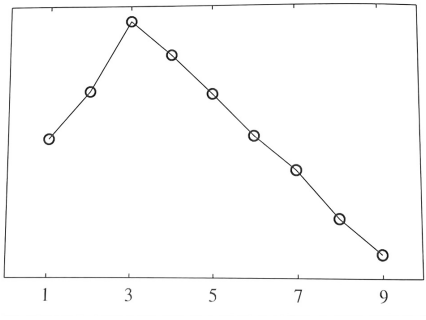
$$
\begin{aligned}
E_{w,\alpha}[\ln p(w|\alpha)] &= E_{w,\alpha}\Big[\ln \mathcal{N}(w|0,\alpha^{-1}I)\Big] \\
&= E_{w,\alpha}\Big[-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\alpha^{-1}I| - \frac{\alpha}{2}w^Tw\Big] \\
&= -\frac{M}{2}\ln(2\pi) - \frac{1}{2}E\big[M\ln(\alpha^{-1})\big] - \frac{1}{2}E[\alpha]E[w^Tw] \\
&\quad \text{分けていいんだっけ？} \\
&= -\frac{M}{2}\ln(2\pi) + \frac{M}{2}E[\ln\alpha] - \frac{a_N}{2b_N}E\big[Tr(ww^T)\big] \qquad (\because (10.102)) \\
&= -\frac{M}{2}\ln(2\pi) + \frac{M}{2}\big(\psi(a_N)-\ln b_N\big) - \frac{a_N}{2b_N}Tr\big[m_Nm_N^T+S_N\big] \\
&\qquad \Big(\because \begin{matrix}(10.103)\\(B.30)\end{matrix}\Big) \\
&= -\frac{M}{2}\ln(2\pi) + \frac{M}{2}\big(\psi(a_N)-\ln b_N\big) - \frac{a_N}{2b_N}\big[m_N^Tm_N + Tr(S_N)\big]
\end{aligned}
$$

$$
\begin{aligned}
E[\ln p(\alpha)] &= E\big[\ln \mathrm{Gam}(\alpha|a_0,b_0)\big] \\
&= E\Big[\ln\Big(\frac{1}{\Gamma(a_0)}b_0^{a_0}\alpha^{a_0-1}e^{-b_0\alpha}\Big)\Big] \\
&= -\ln\Gamma(a_0) + a_0\ln b_0 + (a_0-1)E[\ln\alpha] - b_0E[\alpha] \\
&= -\ln\Gamma(a_0) + a_0\ln b_0 + (a_0-1)\big[\psi(a_N)-\ln b_N\big] - b_0\frac{a_N}{b_N} \qquad (10.110)
\end{aligned}
$$

$$
\begin{aligned}
E_{w,\alpha}[\ln q_*(w)] &= E_{w,\alpha}\left[\ln \mathcal{N}(w \mid m_N, S_N)\right] \\
&= E_{w,\alpha}\left[-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|S_N| - \frac{1}{2}(w-m_N)^\top S_N^{-1}(w-m_N)\right] \\
&= -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|S_N| - \frac{1}{2}E\left[\mathrm{Tr}\left\{(w-m_N)(w-m_N)^\top S_N^{-1}\right\}\right] \\
&= -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|S_N| - \frac{1}{2}\mathrm{Tr}\left[E\left\{(w-m_N)(w-m_N)^\top\right\}S_N^{-1}\right] \\
&= -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|S_N| - \frac{1}{2}\mathrm{Tr}\left[S_N S_N^{-1}\right] \\
&= -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|S_N| - \frac{1}{2}\mathrm{Tr}[I] \\
&= -\frac{1}{2}\ln|S_N| - \frac{M}{2}\ln(2\pi) - \frac{M}{2} \\
&= -\frac{1}{2}\ln|S_N| - \frac{M}{2}\left\{1+\ln(2\pi)\right\} \qquad (10.111)
\end{aligned}
$$

$$
\begin{aligned}
E[\ln q(\alpha)] &= E\left[\ln \mathrm{Gam}(\alpha \mid a_N, b_N)\right] \\
&= E\left[\ln \frac{1}{\Gamma(a_N)} b_N^{a_N} \alpha^{a_N-1} e^{-b_N\alpha}\right] \\
&= -\ln\Gamma(a_N) + a_N \ln b_N + (a_N-1)E[\ln\alpha] - b_N E(\alpha) \\
&= -\ln\Gamma(a_N) + a_N \ln b_N + (a_N-1)\left\{\psi(a_N) - \ln b_N\right\} - b_N\frac{a_N}{b_N} \\
&= -\ln\Gamma(a_N) + (a_N-1)\psi(a_N) + \ln b_N - a_N \qquad (10.112)
\end{aligned}
$$

以上より、あるモデルにおける変分下界を求めることができた。下図は、人口データについて次数の異なる線形モデルの変分下界をプロットしたものである。図より、次数３において変分下界が最も大きくなり、モデルの事後分布も次数３で最大確率が割り当てられていると推察できる。これは複雑なモデルが選択されにくい性質を示している。



10.4 指数型分布族

混合ガウス分布の例とベイズ線形回帰の例を通じて、我々は次の2点の洞察を得た。
1. 共役事前分布を設定したパラメータは、変分推論後の近似分布も同一の分布となる
2. 変分推論における独立性の仮定は、潜在変数とパラメータについての独立のみで良い
実はこれらの性質は指数型分布族についてより一般的に成り立つ。

本節では、一般的な性質として次のことを示す。「潜在変数が指数型分布で、パラメータがこれの共役事前分布を持つ場合、潜在変数とパラメータの独立性を仮定した変分推論により共役な事後分布が得られる」ここで、パラメータは**内包的変数**、潜在変数は**外延的変数**と呼ばれる。

さらに後半では、一般的なグラフィカルモデルの事後分布が局所的なメッセージパッシングを繰り返すことにより求まることを示す。これに必要な条件は、前半で示すように全ての確率変数が共役事前分布をもつ指数型分布族であることである。

まずは、一般的な指数型分布族についての変分推論により、共役な事後分布が得られることを示す。前提として、観測変数 $\mathbf{X}$ と潜在変数 $\mathbf{Z}$ 、パラメータ η は次のような分布をもつとする。

$$p(\mathbf{X}, \mathbf{Z} \mid \eta) = \prod_n h(\mathbf{x}_n, \mathbf{z}_n) g(\eta) \exp\left\{ \eta^\top u(\mathbf{x}_n, \mathbf{z}_n) \right\}$$

$$p(\eta \mid \nu_0, \chi_0) = f(\nu_0, \chi_0) g(\eta)^{\nu_0} \exp\left\{ \nu_0 \eta^\top \chi_0 \right\}$$

同時分布 $p(\mathbf{X}, \mathbf{Z} \mid \eta)$ は指数型分布族であり、パラメータの事前分布 $p(\eta \mid \nu_0, \chi_0)$ は共役となるような指数型分布族として表している。(ref. 上巻 114頁 2.4.2節)

変分推論を行うため、$q(\mathbf{z}, \eta) = q(\mathbf{z}) q(\eta)$ という独立性を仮定する。この仮定により、一般的な変分の更新式 (10.9) を適用することができる。

潜在変数の更新式は次のようになる。

$$\begin{aligned}
\ln q^\star(\mathbf{Z}) &= E_\eta \left[\ln \left\{ p(\mathbf{X}, \mathbf{Z} \mid \eta) p(\eta) \right\} \right] + \text{const.} \\
&= E_\eta \left[\ln p(\mathbf{X}, \mathbf{Z} \mid \eta) \right] + \text{const.} \\
&= \sum_n E_\eta \left[\ln h(\mathbf{x}_n, \mathbf{z}_n) + \ln g(\eta) + \eta^\top u(\mathbf{x}_n, \mathbf{z}_n) \right] + \text{const.} \\
&= \sum_n \left\{ \ln h(\mathbf{x}_n, \mathbf{z}_n) + E[\eta^\top] u(\mathbf{x}_n, \mathbf{z}_n) \right\} + \text{const.}
\end{aligned}$$

両辺に指数をとると、指数型分布族として更新式を得る。

$$\begin{aligned}
q^\star(\mathbf{Z}) &\propto \prod_n h(\mathbf{x}_n, \mathbf{z}_n) \exp\left\{ E[\eta^\top] u(\mathbf{x}_n, \mathbf{z}_n) \right\} \\
\therefore q^\star(\mathbf{Z}) &= \prod_n h(\mathbf{x}_n, \mathbf{z}_n) g(E[\eta]) \exp\left\{ E[\eta^\top] u(\mathbf{x}_n, \mathbf{z}_n) \right\}
\end{aligned}$$

指数型分布の一般的な形が求まる。

次にパラメータの更新式は次のように求まる。

$$
\begin{aligned}
\ln q^*(\eta) &= E_z\left[\ln p(X, Z|\eta) + \ln p(\eta|\nu_0, \chi_0)\right] + \text{const.} \\
&= \sum_n^N \left\{\ln E(h(x_n, z_n)) + \ln g(\eta) + \eta^T E(u(x_n, z_n))\right\} \\
&\quad + \ln f(\nu_0, \chi_0) + \nu_0 \ln g(\eta) + \nu_0 \eta^T \chi_0 + \text{const.} \\
&\qquad \text{const.} \\
&= N \ln g(\eta) + \eta^T \sum_n^N E(u(x_n, z_n)) + \nu_0 \ln g(\eta) + \nu_0 \eta^T \chi_0 + \text{const.} \\
&= (\nu_0 + N) \ln g(\eta) + \eta^T \left[\nu_0 \chi_0 + \sum_n^N E\{u(x_n, z_n)\}\right] + \text{const.} \\
&= \nu_N \ln g(\eta) + \nu_N \eta^T \chi_N + \text{const.}
\end{aligned}
$$

$$
\left(\text{ここで、} \quad \nu_N = \nu_0 + N, \quad \nu_N \chi_N = \nu_0 \chi_0 + \sum_n E\{u(x_n, z_n)\}\right)
$$

両辺に指数をとり正規化すると、次の指数型分布族が得られる。

$$
q^*(\eta) = f(\nu_N, \chi_N)\, g(\eta)^{\nu_N} \exp\left\{\nu_N \eta^T \chi_N\right\}
$$

以上で、より一般的な仮定の元で変分推論の更新式を得られることを確かめた。また、得られた二つの更新式は互いに依存関係があるため、これまで見てきた例と同様に繰り返し法によって分布を更新していく必要がある。

10.4.1 変分メッセージパッシング

10.4 指数型分布族の節では、 $\mathbf{z}$ と η の2変数間のみの依存関係について考えてきた。より多くの変数の依存関係を表現する場合有向グラフで表すことができる。

この節では有向グラフで表現された確率分布について変分推論を実施できることを示す。特に、変分推論の一般的な結果 (10.9) を利用してノード $\mathbf{x}_j$ の更新式を得る場合、ノード $\mathbf{x}_j$ のマルコフブランケット内の期待値に関する局所的な計算のみが必要なことを示す。

まず、有向グラフにおける同時分布は次のように分解できる。

$$p(\mathbf{x}) = \prod_i p(\mathbf{x}_i \mid pa_i)$$

また、変分近似による分布は次のように分解できるとする。

$$q(\mathbf{x}) = \prod_i q_i(\mathbf{x}_i)$$

ノード $\mathbf{x}_j$ の変分推論による更新式は、(10.9) より次のように求まる。

$$\begin{aligned}
\ln q_j^*(\mathbf{x}_j) &= E_{i \neq j}\left[\ln p(\mathbf{x})\right] + \text{const.} \\
&= E_{i \neq j}\left[\sum_i \ln p(\mathbf{x}_i \mid pa_i)\right] + \text{const.} \\
&= E_{i \neq j}\left[\ln p(\mathbf{x}_j \mid pa_j) + \sum_{i \in ch_j} \ln p(\mathbf{x}_i \mid pa_i)\right] + \text{const.} \\
&= E_{i \in (MB_j \setminus j)}\left[\ln p(\mathbf{x}_j \mid pa_j) + \sum_{i \in ch_j} \ln p(\mathbf{x}_i \mid pa_i)\right] + \text{const.}
\end{aligned}$$

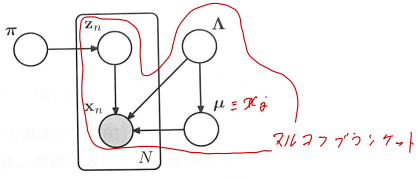
$\mathbf{x}_i$ が pa_i どうしがに $\mathbf{x}_j$ がある頂のみ残る。

ch_j の集合

$\mathbf{x}_j$ のマルコフブランケットから $\mathbf{x}_j$ を除いた集合

この更新式から、ノード　　の変分事後分布は、ノード　　のマルコフブランケットに閉じた局所的な計算によって得られることがわかる。

（感想）ベイズ混合ガウス分布に当てはめて考えるとわかりやすかった。(10.55) から π と Λ，μ をそれぞれ別々に更新式を得ているのは、それぞれの変数がマルコフブランケットの外側にあること成り立っていると理解できる。



10.5 局所変分推論法

（ここでやろうとしていることをほぼ理解できなかったので、読み合わせとせさてください）

この節では特に、ロジスティックシグモイド関数の下界を新たなパラメータを用いて表現する。ここで求めた下界は、10.6.1 節で分布関数の近似として代用され、事後分布が求まるらしい（10.152 - 10.156 あたり）

また、今回の内容の中では、209頁最後の 10.131 の後あたりから何が言いたいのかわからなくなった。図10.10 の解釈などがわからなかったので相談したい。