変分ベイズ法とは

正田 備也

masada@rikkyo.ac.jp

Contents

ベイズ的モデリングにおける変分法

- 観測データを表す確率変数を $\mathcal{X} \equiv \{x_1, \ldots, x_N\}$ とする
- データモデルのパラメータを $\mathbf{\Theta}$ とする
- ベイズ的なモデリングでは、$\mathcal{X}$ だけでなく $\mathbf{\Theta}$ も確率変数
- 知りたいのは事後分布 $p(\mathbf{\Theta}|\mathcal{X})$

$$p(\mathbf{\Theta}|\mathcal{X}) = \frac{p(\mathcal{X}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathcal{X})} \tag{1}$$

- 変分ベイズ法は $p(\mathbf{\Theta}|\mathcal{X})$ を近似する分布 $q(\mathbf{\Theta})$ を求める
  - $q(\mathbf{\Theta})$ を変分法 (variational methods) で求める（後述）
  - $q(\mathbf{\Theta})$ を変分事後分布 (variational posterior distribution) と呼ぶ

前回のEMアルゴリズムでの議論のパターン

- 潜在変数 $\mathcal{Z} = \{z_1, \dots, z_N\}$ を含むモデリングを行いたい
- 確率モデルを指定することで同時分布 $p(\mathcal{X}, \mathcal{Z}) = p(\mathcal{Z})p(\mathcal{X}|\mathcal{Z}) = \prod_{i=1}^{N} p(z_i)p(x_i|z_i)$ が得られる
- 潜在変数 $\mathcal{Z}$ の周辺化 $\sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z})$ により観測データの尤度 $p(\mathcal{X})$ は得られるのだが、大抵この尤度は計算できない
- Jensen の不等式を使い、対数尤度 $\ln p(\mathcal{X})$ の下界を得る

$$\ln p(\mathcal{X}) \geq \sum_{i=1}^{N} \sum_{z_i=1}^{K} q_{i,z_i} \ln \frac{p(z_i)p(x_i|z_i)}{q_{i,z_i}}$$

- この下界を最大化することで、様々な未知量を推定する

この議論のパターンを事後分布の推論へ適用

- 潜在変数 $\mathbf{\Theta}$ を含むモデリングを行いたい
- 確率モデルを指定することで観測データと潜在変数の同時分布 $p(\mathcal{X}, \mathbf{\Theta}) = p(\mathbf{\Theta})p(\mathcal{X}|\mathbf{\Theta}) = p(\mathbf{\Theta}) \prod_{i=1}^{N} p(x_i|\mathbf{\Theta})$ が得られる
- 潜在変数 $\mathbf{\Theta}$ の周辺化 $\int p(\mathcal{X}, \mathbf{\Theta}) d\mathbf{\Theta}$ により観測データの周辺尤度 $p(\mathcal{X})$ は得られるのだが、大抵この尤度は計算できない
- Jensen の不等式を使い、対数周辺尤度 $\ln p(\mathcal{X})$ の下界を得る

$$\ln p(\mathcal{X}) \geq \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta})p(\mathcal{X}|\mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta}$$

- この下界を最大化することで、様々な未知量を推定する
  - この下界を ELBO(Evidence Lower BOund; 変分下限) と呼ぶ

変分ベイズ法 (variational Bayesian methods) とは

- Jensen の不等式を適用することで、ELBO を次のように得た
$$\ln p(\mathcal{X}) \geq \int q(\mathbf{\Theta}) \ln \frac{p(\mathbf{\Theta})p(\mathcal{X}|\mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta}$$
- 実は、ELBO を大きくすればするほど、$\mathbf{\Theta}$ が従う確率分布である $q(\mathbf{\Theta})$ が、事後分布 $p(\mathbf{\Theta}|\mathcal{X})$ に近くなっていく
- つまり、この $q(\mathbf{\Theta})$ は、事後分布を近似する分布とみなせるような分布になっている
- $q(\mathbf{\Theta})$ は変分法 (variational method) で求められるので、変分事後分布 (variational posterior) と呼ばれる

「変分 (variational)」の意味

- ELBO の最大化は、$q(\boldsymbol{\Theta})$ を変化させることでおこなう
- $q(\boldsymbol{\Theta})$ の密度関数がどんなかたちを持つかに制約を設けない
- 逆に言うと、$q(\boldsymbol{\Theta})$ の密度関数が特定のかたちを持つと仮定した上で、その関数のパラメータだけを動かすのではない
  - パラメータについて微分することで最大化問題を解くのではなく、いわば "関数について微分する" ことで最大化問題を解いている
- とても直感的に言うと、関数のかたちを決めてそのパラメータを動かすのではなく、関数のかたち自体を動かすことで問題を解く方法を、変分法と呼ぶ（cf. 汎関数微分）

ELBO を最大化する根拠

- Jensen の不等式の左辺から右辺を引いたものを求めてみる

$$
\begin{aligned}
&\ln p(\mathcal{X}) - \int q(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{\Theta}|\mathcal{X})p(\mathcal{X})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} \\
&= \ln p(\mathcal{X}) - \int q(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{\Theta}|\mathcal{X})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} - \int q(\boldsymbol{\Theta}) \ln p(\mathcal{X}) d\boldsymbol{\Theta} \\
&= \ln p(\mathcal{X}) - \int q(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{\Theta}|\mathcal{X})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} - \ln p(\mathcal{X}) \int q(\boldsymbol{\Theta}) d\boldsymbol{\Theta} \\
&= \int q(\boldsymbol{\Theta}) \ln \frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta}|\mathcal{X})} d\boldsymbol{\Theta} = D_{\mathsf{KL}}(q(\boldsymbol{\Theta}) \parallel p(\boldsymbol{\Theta}|\mathcal{X}))
\end{aligned} \tag{2}
$$

$\therefore$ ELBO を $\ln p(\mathcal{X})$ に近づける $\Leftrightarrow$ $q(\boldsymbol{\Theta})$ を $p(\boldsymbol{\Theta}|\mathcal{X})$ に近づける (3)

変分事後分布に関する factorization の仮定

- モデルパラメータ $\boldsymbol{\Theta}$ について、以下のような排他的なグループ分けを行い・・・

$$\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \cup \ldots \cup \boldsymbol{\Theta}_m \tag{4}$$

- 変分事後分布が以下のように分解される、と仮定したりする

$$q(\boldsymbol{\Theta}) = q(\boldsymbol{\Theta}_1) \cdots q(\boldsymbol{\Theta}_m) \tag{5}$$

  - 「posterior factorizes」でググってみよう
- このような仮定をすることで、ELBO の最大化問題が簡単になることがある

平均場近似(mean-field approximation)

- 最も極端な場合、個々のパラメータが従う確率分布の積へ分解されると仮定することも、わりとある
- $\mathbf{\Theta}$ が r 個のパラメータ $\theta_1, \ldots, \theta_r$ からなるとすると、以下のような積へ分解されると仮定する

$$q(\mathbf{\Theta}) = q(\theta_1) \cdots q(\theta_r) \tag{6}$$

- これを平均場近似(mean-field approximation)と呼ぶ

より実際的な変分ベイズ法

- 上述の factorization の仮定をおくと、それだけで、変分事後分布の密度関数のかたちが決まってしまうこともある
  - この後に示す例が、そうなっている
- しかし実際には、$q(\boldsymbol{\Theta})$ の密度関数が特定のかたちを持つと仮定してしまった上で、その関数のパラメータを動かすことによって、ELBO を最大化することも多い
  - 例えば、$q(\boldsymbol{\Theta})$ が多変量正規分布だと仮定して、ELBO を最大化するような平均パラメータと共分散行列パラメータを求める、など
  - 変分オートエンコーダでは、$q(\boldsymbol{\Theta})$ が多変量正規分布だと仮定し、さらにその共分散行列が対角行列だと仮定する

Contents

変分ベイズ法によるデータモデリングの手順

- データモデル $p(\mathcal{X}|\mathbf{\Theta})$ とモデルパラメータの事前分布 $p(\mathbf{\Theta};\mathbf{\Xi})$ を指定する
  - $\mathbf{\Xi}$ は事前分布のパラメータ、つまり、ハイパーパラメータ
- 同時分布 $p(\mathcal{X},\mathbf{\Theta};\mathbf{\Xi})$ を書き下す
- ELBO を書き下す
- 変分事後分布 $q(\mathbf{\Theta};\mathbf{\Psi})$ を扱いやすくするための仮定を行う
  - factorization の仮定、既知の確率分布であるという仮定、等
- その仮定を利用して、変分事後分布のパラメータ $\mathbf{\Psi}$ を求めるための式（多くの場合、更新式）を得る
- この式を実装して計算機で動かす

例：メッセージ受信数の変化点の検知

- この授業の最初に採り上げた例
  - 参考: http://machine-learning.hatenablog.com/entry/2017/08/19/200841

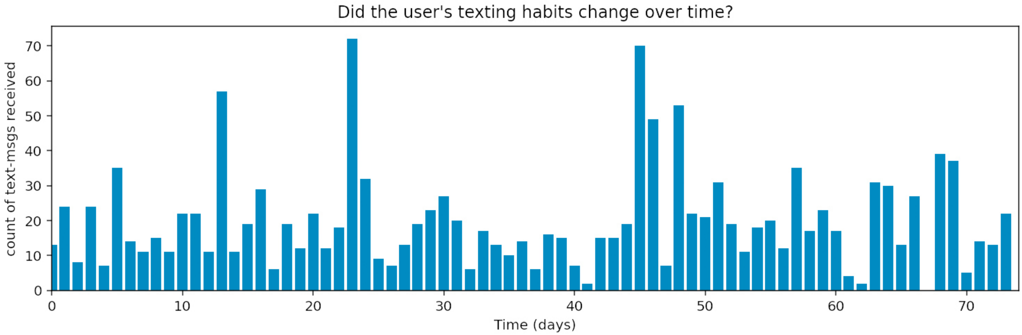

Figure: メッセージの受信数

モデルを指定する

- c_n を n 日目の受信数、τ を受信数の変化点とする
- λ_1 は $n < \tau$ の場合のポアソン分布のパラメータ
- λ_2 は $n \geq \tau$ の場合のアソン分布のパラメータ

$$
\begin{aligned}
\tau &\sim \mathtt{Uniform}(1, N) \\
\lambda_1 &\sim \mathtt{Gam}(a, b) \\
\lambda_2 &\sim \mathtt{Gam}(a, b) \\
c_n &\sim \mathtt{Poi}(\lambda_1) \quad \text{for } n < \tau \\
c_n &\sim \mathtt{Poi}(\lambda_2) \quad \text{for } n \geq \tau
\end{aligned}
$$

同時分布を書き下す

同時分布は、観測データを $\boldsymbol{c} = \{c_1, \ldots, c_N\}$ とすると

$$\begin{aligned} p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau) &= p(\boldsymbol{c}|\lambda_1, \lambda_2, \tau)p(\lambda_1; a, b)p(\lambda_2; a, b)p(\tau) \\ &= p(\lambda_1; a, b)p(\lambda_2; a, b)p(\tau) \prod_{n=1}^{N} p(c_n|\lambda_1)^{\delta(n<\tau)} p(c_n|\lambda_2)^{\delta(n\geq\tau)} \end{aligned} \tag{7}$$

- $p(\lambda_i; a, b) \equiv \frac{b^a}{\Gamma(a)} \lambda_i^{a-1} e^{-b\lambda_i}$ for $i = 1, 2$
- $p(\tau) \equiv \frac{1}{N}$
- $\delta(\cdot)$ は、カッコ内の命題が真ならば 1、偽ならば 0
- $p(c_n|\lambda_i) \equiv \frac{\lambda_i^{c_n} e^{-\lambda_i}}{c_n!}$ for $i = 1, 2$

ELBOを書き下す

$$
\begin{aligned}
\ln p(\boldsymbol{c}) &= \ln \int \sum_{\tau=1}^{N} p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau) d\lambda_1 d\lambda_2 \\
&\geq \int \sum_{\tau=1}^{N} q(\lambda_1, \lambda_2, \tau) \ln \frac{p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau)}{q(\lambda_1, \lambda_2, \tau)} d\lambda_1 d\lambda_2 \qquad (8)
\end{aligned}
$$

- このままではこれ以上議論を進められない
- 変分事後分布 $q(\lambda_1, \lambda_2, \tau)$ について、それを扱いやすくするような、何らかの仮定をおこなう

factorization の仮定

ここでは、変分事後分布 $q(\lambda_1, \lambda_2, \tau)$ が $q(\lambda_1, \lambda_2, \tau) = q(\lambda_1, \lambda_2)q(\tau)$ と factorize することを仮定する

$$\begin{aligned} \ln p(\boldsymbol{c}) &\geq \int \sum_{\tau=1}^{N} q(\lambda_1, \lambda_2, \tau) \ln \frac{p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau)}{q(\lambda_1, \lambda_2, \tau)} d\lambda_1 d\lambda_2 \\ &= \int \sum_{\tau=1}^{N} q(\lambda_1, \lambda_2) q(\tau) \ln \frac{p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau)}{q(\lambda_1, \lambda_2) q(\tau)} d\lambda_1 d\lambda_2 \end{aligned} \tag{9}$$

- 同時分布の式 (7) を使って、ELBO をさらに詳しく書き下す

λ_1, λ_2 が従うガンマ事前分布のハイパーパラメータ a, b は省略する。

$$\ln p(\boldsymbol{c}) \geq \int \sum_{\tau=1}^{N} q(\lambda_1, \lambda_2) q(\tau) \ln \frac{p(\boldsymbol{c}, \lambda_1, \lambda_2, \tau)}{q(\lambda_1, \lambda_2) q(\tau)} d\lambda_1 d\lambda_2$$

$$= \int \sum_{\tau=1}^{N} q(\lambda_1, \lambda_2) q(\tau) \ln \frac{p(\lambda_1) p(\lambda_2) p(\tau) \prod_{n=1}^{N} p(c_n|\lambda_1)^{\delta(n<\tau)} p(c_n|\lambda_2)^{\delta(n\geq\tau)}}{q(\lambda_1, \lambda_2) q(\tau)} d\lambda_1 d\lambda_2$$

$$= \int q(\lambda_1, \lambda_2) \ln \frac{p(\lambda_1)}{q(\lambda_1)} d\lambda_1 d\lambda_2 + \int q(\lambda_1, \lambda_2) \ln \frac{p(\lambda_2)}{q(\lambda_2)} d\lambda_1 d\lambda_2 + \sum_{\tau=1}^{N} q(\tau) \ln \frac{p(\tau)}{q(\tau)}$$

$$+ \sum_{\tau=1}^{N} \sum_{n=1}^{N} \delta(n<\tau) \int q(\lambda_1, \lambda_2) \ln p(c_n|\lambda_1) d\lambda_1 d\lambda_2 + \sum_{\tau=1}^{N} \sum_{n=1}^{N} \delta(n\geq\tau) \int q(\lambda_1, \lambda_2) \ln p(c_n|\lambda_2) d\lambda_1 d\lambda_2$$

$$= \int q(\lambda_1) \ln \frac{p(\lambda_1)}{q(\lambda_1)} d\lambda_1 + \int q(\lambda_2) \ln \frac{p(\lambda_2)}{q(\lambda_2)} d\lambda_2 + \sum_{\tau=1}^{N} q(\tau) \ln \frac{p(\tau)}{q(\tau)}$$

$$+ \sum_{\tau=1}^{N} \sum_{n=1}^{N} \delta(n<\tau) \int q(\lambda_1) \ln p(c_n|\lambda_1) d\lambda_1 + \sum_{\tau=1}^{N} \sum_{n=1}^{N} \delta(n\geq\tau) \int q(\lambda_2) \ln p(c_n|\lambda_2) d\lambda_2 \tag{10}$$

変分事後分布を求める

- この例の場合は、$q(\lambda_1, \lambda_2, \tau) = q(\lambda_1, \lambda_2)q(\tau)$ と仮定すると、$q(\lambda_1, \lambda_2) = q(\lambda_1)q(\lambda_2)$ と factorize することが得られた
- さらにはこの factorization により、変分事後分布の密度関数のかたちが決まってしまう
- 具体的には、$q(\lambda_1)$ と $q(\lambda_2)$ と $q(\tau)$ のうち 2 つを固定して、残りの 1 つだけを動かすことで、ELBO を最大化する
- すると、変分事後分布の密度関数のかたちが自ずと決まる
  - $q(\lambda_1)$ と $q(\lambda_2)$ の密度関数の式は、ガンマ分布のそれに一致
  - $q(\tau)$ の質量関数の式は、カテゴリカル分布のそれに一致

$q(\lambda_1)$ の密度関数のかたちを求める(1/2)

$q(\lambda_2)$ と $q(\tau)$ を固定し、$D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}))$ を最小にする $q(\lambda_1)$ を求める。
(この KL 情報量の最小化は、式 (3) より、ELBO の最大化と等価であることに注意)

$$
\begin{aligned}
D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})) &= \int \sum_{\tau=1}^{N} q(\lambda_1)q(\lambda_2)q(\tau) \ln \frac{q(\lambda_1)q(\lambda_2)q(\tau)}{p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})} d\lambda_1 d\lambda_2 \\
&= \int q(\lambda_1)\Big\{ \ln q(\lambda_1) - \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2 \Big\} d\lambda_1 + const. \\
&= \int q(\lambda_1) \ln \frac{q(\lambda_1)}{\exp \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2} d\lambda_1 + const. \qquad (11)
\end{aligned}
$$

$\exp \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2$ は、λ_2 については積分消去しており、τ についても総和をとって消去しているので、λ_1 の関数である。そこで、規格化定数 Z を導入することによって、$\exp \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2$ を、λ_1 の単なる関数から、$\frac{1}{Z} \exp \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2$ という密度関数に変える。

$q(\lambda_1)$ の密度関数のかたちを求める (2/2)

すると、

$$
\begin{aligned}
& D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})) \\
&= \int q(\lambda_1) \ln \frac{q(\lambda_1)}{\exp \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2} d\lambda_1 + const. \\
&= \int q(\lambda_1) \ln \frac{q(\lambda_1)}{\frac{1}{Z} \exp \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2} d\lambda_1 - \int q(\lambda_1) \ln Z d\lambda_1 + const.
\end{aligned}
$$

となるが、$\int q(\lambda_1) \ln Z d\lambda_1 = \ln Z \int q(\lambda_1) d\lambda_1 = \ln Z = const.$ なので

$$
\begin{aligned}
& D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})) \\
&= D_{\mathsf{KL}}(q(\lambda_1) \parallel \frac{1}{Z} \exp \int q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2 d\tau) + const. \qquad (12)
\end{aligned}
$$

上の KL 情報量は、$q(\lambda_1) = \frac{1}{Z} \exp \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2$ のとき、最小。
つまり、$\ln q(\lambda_1) = \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_2 - \ln Z$ のとき最小。

$q(\lambda_1)$ のパラメータを求める(1/2)

上述の KL 情報量が最小となるとき、$q(\lambda_1)$ がどのような分布になるかを調べるため、
$\ln q(\lambda_1) = \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln p(\lambda_1, \lambda_2, \tau | \boldsymbol{c}) d\lambda_2 - \ln Z$ の右辺を、計算してみる。

$$
\begin{aligned}
\ln q(\lambda_1) &= \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln \frac{p(\lambda_1, \lambda_2, \tau, \boldsymbol{c})}{p(\boldsymbol{c})} d\lambda_2 - \ln Z \\
&= \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau) \ln \left\{ p(\lambda_1; a, b) p(\lambda_2; a, b) p(\tau) \prod_{n=1}^{N} p(c_n|\lambda_1)^{\delta(n<\tau)} p(c_n|\lambda_2)^{\delta(n\geq\tau)} \right\} d\lambda_2 + const. \\
&= \ln p(\lambda_1; a, b) + \int q(\lambda_2) \ln p(\lambda_2; a, b) d\lambda_2 + \sum_{\tau=1}^{N} q(\tau) \ln p(\tau) \\
&\quad + \sum_{n=1}^{N} \sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \ln p(c_n|\lambda_1) + \sum_{n=1}^{N} \int \sum_{\tau=1}^{N} q(\lambda_2)q(\tau)\delta(n \geq \tau) \ln p(c_n|\lambda_2) d\lambda_2 + const. \\
&= \ln p(\lambda_1; a, b) + \sum_{n=1}^{N} \sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \ln p(c_n|\lambda_1) + const.
\end{aligned}
\tag{13}
$$

$q(\lambda_1)$ のパラメータを求める (2/2)

ここで、事前分布 $p(\lambda_1; a, b)$ にガンマ分布の密度関数を、観測データの尤度 $p(c_n|\lambda_1)$ にポアソン分布の質量関数の式を、それぞれあてはめると、

$$
\begin{aligned}
&\ln q(\lambda_1) \\
&= \ln \frac{b^a}{\Gamma(a)} \lambda_1^{a-1} e^{-b\lambda_1} + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \right) \ln \frac{\lambda_1^{c_n} e^{-\lambda_1}}{c_n!} + const. \\
&= \left(a - 1 + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \right) c_n \right) \ln \lambda_1 - \left(b + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \right) \right) \lambda_1 + const.
\end{aligned}
$$

よって、$q(\lambda_1)$ は、shape パラメータが $a + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \right) c_n$ で、rate パラメータが $b + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \right)$ のガンマ分布となる。

$q(\lambda_2)$ についても同様に計算すると、やはりガンマ分布であることが分かる。その shape パラメータを α_2、rate パラメータを β_2 とすると・・・

$q(\lambda_1; \alpha_1, \beta_1)$ と $q(\lambda_2; \alpha_2, \beta_2)$ の更新式

$$\alpha_1 \leftarrow a + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \right) c_n \tag{14}$$

$$\beta_1 \leftarrow b + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n < \tau) \right) \tag{15}$$

$$\alpha_2 \leftarrow a + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n \geq \tau) \right) c_n \tag{16}$$

$$\beta_2 \leftarrow b + \sum_{n=1}^{N} \left(\sum_{\tau=1}^{N} q(\tau)\delta(n \geq \tau) \right) \tag{17}$$

$q(\tau)$ のかたちを求める

$q(\lambda_1)$ と $q(\lambda_2)$ を固定し、$D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}))$ を最小にする $q(\tau)$ を求める。

$$
\begin{aligned}
& D_{\mathsf{KL}}(q(\lambda_1)q(\lambda_2)q(\tau) \parallel p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})) \\
&= \int \sum_{\tau=1}^{N} q(\lambda_1)q(\lambda_2)q(\tau) \ln \frac{q(\lambda_1)q(\lambda_2)q(\tau)}{p(\lambda_1, \lambda_2, \tau|\boldsymbol{c})} d\lambda_1 d\lambda_2 \\
&= \sum_{\tau=1}^{N} q(\tau) \Big\{ \ln q(\tau) - \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2 \Big\} + const. \\
&= \sum_{\tau=1}^{N} q(\tau) \ln \frac{q(\tau)}{\exp \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2} + const. \\
&= D_{\mathsf{KL}}\Big(q(\tau) \parallel \frac{1}{Z} \exp \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2\Big) + const. \qquad (18)
\end{aligned}
$$

$q(\tau) = \frac{1}{Z} \exp \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2$ のとき、上の KL 情報量は最小。
つまり、$\ln q(\tau) = \int q(\lambda_1)q(\lambda_2) \ln p(\lambda_1, \lambda_2, \tau|\boldsymbol{c}) d\lambda_1 d\lambda_2 - \ln Z$ のとき最小。

上述の KL 情報量が最小となるとき、$q(\tau)$ がどういう分布になるかを調べるため、
$\ln q(\tau) = \int q(\lambda_1)q(\lambda_2)\ln p(\lambda_1,\lambda_2,\tau|\boldsymbol{c})d\lambda_1 d\lambda_2 - \ln Z$ の右辺を計算してみる。

$$\ln q(\tau) = \int q(\lambda_1)q(\lambda_2)\ln\frac{p(\lambda_1,\lambda_2,\tau,\boldsymbol{c})}{p(\boldsymbol{c})}d\lambda_1 d\lambda_2 - \ln Z$$

$$= \int q(\lambda_1)q(\lambda_2)\ln\left\{p(\lambda_1;a,b)p(\lambda_2;a,b)p(\tau)\prod_{n=1}^{N}p(c_n|\lambda_1)^{\delta(n<\tau)}p(c_n|\lambda_2)^{\delta(n\geq\tau)}\right\}d\lambda_1 d\lambda_2 + const.$$

$$= \int q(\lambda_1)\ln p(\lambda_1;a,b)d\lambda_1 + \int q(\lambda_2)\ln p(\lambda_2;a,b)d\lambda_2$$

$$+\sum_{n=1}^{N}\delta(n<\tau)\int q(\lambda_1)\ln p(c_n|\lambda_1)d\lambda_1 + \sum_{n=1}^{N}\delta(n\geq\tau)\int q(\lambda_2)\ln p(c_n|\lambda_2)d\lambda_2 + const.$$

この式は、異なる τ ごとに単に別々の値をとる。よって、$q(\tau)$ はカテゴリカル分布であり、

$$q(\tau) \propto \exp\left[\sum_{n=1}^{N}\delta(n<\tau)\int q(\lambda_1)\ln p(c_n|\lambda_1)d\lambda_1 + \sum_{n=1}^{N}\delta(n\geq\tau)\int q(\lambda_2)\ln p(c_n|\lambda_2)d\lambda_2\right]$$

となる。$q(\lambda_1)$ と $q(\lambda_2)$ がガンマ分布であることを利用し、さらに式変形する。

$$\sum_{n=1}^{N} \delta(n < \tau) \int q(\lambda_1; \alpha_1, \beta_1) \ln p(c_n|\lambda_1) d\lambda_1 = \sum_{n=1}^{N} \delta(n < \tau) \int q(\lambda_1; \alpha_1, \beta_1) \ln \frac{\lambda_1^{c_n} e^{-\lambda_1}}{c_n!} d\lambda_1$$

$$= \{\psi(\alpha_1) - \ln(\beta_1)\} \sum_{n=1}^{N} \delta(n < \tau) c_n - \frac{\alpha_1}{\beta_1} \sum_{n=1}^{N} \delta(n < \tau) - \sum_{n=1}^{N} \delta(n < \tau) \ln c_n! \quad (19)$$

$$\sum_{n=1}^{N} \delta(n \geq \tau) \int q(\lambda_2) \ln p(c_n|\lambda_2) d\lambda_2 = \cdots$$

$$= \{\psi(\alpha_2) - \ln(\beta_2)\} \sum_{n=1}^{N} \delta(n \geq \tau) c_n - \frac{\alpha_2}{\beta_2} \sum_{n=1}^{N} \delta(n \geq \tau) - \sum_{n=1}^{N} \delta(n \geq \tau) \ln c_n! \quad (20)$$

$$\therefore q(\tau) \propto \exp\left[\{\psi(\alpha_1) - \ln(\beta_1)\} \sum_{n=1}^{\tau-1} c_n + \{\psi(\alpha_2) - \ln(\beta_2)\} \sum_{n=\tau}^{N} c_n - \frac{(\tau - 1)\alpha_1}{\beta_1} - \frac{(N - \tau + 1)\alpha_2}{\beta_2} \right] \quad (21)$$

まとめ

- メッセージ受信数の変化点を検知するため、ベイズ的なモデルを立てた
- 事後分布を近似するために、変分ベイズ法を使った
- その際、変分事後分布 $q(\lambda_1, \lambda_2, \tau)$ について、$q(\lambda_1, \lambda_2, \tau) = q(\lambda_1)q(\lambda_2)q(\tau)$ と分解できることを仮定した
- このように仮定すると、$q(\lambda_1)$ と $q(\lambda_2)$ はガンマ分布となり、$q(\tau)$ はカテゴリカル分布となった

今日の課題

- メッセージ受信数の変化点検知の例を考える。
- λ_1 の値が従う変分事後分布 $q(\lambda_1)$ は、ガンマ分布であることが分かった。
- そこで、$q(\lambda_1)$ の shape パラメータを α_1 とし、rate パラメータを β_1 とする。
- このとき、$\int q(\lambda_1; \alpha_1, \beta_1) \ln p(\lambda_1; a, b) d\lambda_1$ を計算せよ。
  - これは ELBO の算出に必要な計算。
  - ヒント１：$q(\lambda_1)$ のパラメータ α_1 と β_1 を使って答えを表す。
  - ヒント２：$p(\lambda_1; a, b)$ がガンマ分布で、shape パラメータは a、rate パラメータは b であることも当然使う。