モデル推定

潜在変数がある場合のモデル推論

EMアルゴリズム

EMの混合正規分布への適用例

変分ベイズ法

EP法

潜在変数を考慮する推論

- prior:**w**（＝潜在変数＋分布のパラメター）と観測データ{**x,t**}(ただし、**x**は入力、**t**は結果)の分布にハイパーパラメターα、βを導入する。
- まず、観測データから事後確率を最大化するpriorをベイズで求める。このためにいろいろな技法あり。（例えば経験ベイズ法）

$$p(\mathbf{w}\,|\,\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}\,|\,\mathbf{x},\mathbf{w},\beta)\,p(\mathbf{w}\,|\,\alpha)$$

- 新規データxに対する予測 t を計算するのは以下の式

$$p(t\,|\,x,\mathbf{x},\mathbf{t},\alpha,\beta) = \int p(t\,|\,x,\mathbf{w},\beta)\,p(\mathbf{w}\,|\,\mathbf{x},\mathbf{t},\alpha,\beta)\mathrm{d}\mathbf{w}$$

モデル推定のための学習法

- 事前知識のない場合はK-meansなど
- EM(Expectation and Maximization)アルゴリズム
  - モデル学習法の古典
- 変分ベイズ法
  - 予測モデルqを既知のモデル（prior)とのKL divergence を最小化する関数として変分法で繰り返し計算で求める。速度は速い。
  - KL(p||q)=Σp log(p/q)
- MCMC(Markov Chain Monte Carlo）
  - モデルのパラメター推定をシミュレーションで解いてしまう方法。速度は遅い。

経験ベイズ法 ：
事前分布のパラメターの初期値の推定

- ベイズモデルのパラメターを恣意的に決めると気持ち悪い
- ベイズモデルのパラメターを観測データに基づいて求める方法
- $\pi(\theta \mid \alpha)$: 事前分布　ただし、α は事前分布πを決めるパラメター　$\alpha_{\max} = \arg\max_{\alpha}$
- 観測データ$\mathbf{x}$ から尤度$p(\mathbf{x} \mid \theta)$が与えられたとき、これを用いて事前分布のパラメターα_{max}を求める。

$$\alpha_{\max} = \arg\max_{\alpha} \int p(\mathbf{x} \mid \theta)\pi(\theta \mid \alpha)\mathrm{d}\theta \qquad -(EB10)$$

経験ベイズ法 ：
事前分布のパラメターの初期値の推定例

$$\alpha_{\max} = \arg\max_{\alpha} \int p(\mathbf{x}\,|\,\theta)\pi(\theta\,|\,\alpha)\mathrm{d}\theta \qquad -(EB10)$$

- 多項分布、ディリクレ分布の例

観測データ　iの出現回数　　m_i: $X = (m_1,...,m_K)$,　$\sum_{i=1}^{K} m_i = M, \sum_{i=1}^{K} \alpha_i = \alpha_0$

$$Dir(\mu\,|\,X,\alpha) \propto Mult(X\,|\,\mu)Dir(\mu\,|\,\alpha) \propto \prod_{i=1}^{K} \mu_i^{\alpha_i + m_i - 1}$$

$$\Rightarrow \quad \alpha_{\max} = \arg\max_{\alpha} Dir(\mu\,|\,X,\alpha)$$

$$= \arg\max_{\alpha} Dir(\mu\,|\,\alpha + X)$$

$$= \arg\max_{\alpha} \frac{\Gamma(\alpha_0 + M)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{i=1}^{K} \mu_i^{\alpha_i + m_i - 1}$$

EMアルゴリズム

- 観測変数 observed variable（顕在変数 manifest variable）
- 隠れ変数 hidden variable（潜在変数 latent variable）
  - 例：混合正規分布のとき、どの正規分布から生成されたデータかを示す変数
- 観測変数のデータを用いて母集団の統計モデルの潜在変数を推定する（最尤推定値）
  - パラメータの値を点で推定
- このための数値解法：EMアルゴリズム Expectation and Maximization Algorithm

- 観測されたデータ(＝観測変数の実測データ):X
- 潜在変数のとる値:Z
- 統計モデルのおける未知のパラメター:θ
- 対数尤度関数(log likelihood func.)

logの中のΣが現れると嫌だ！

$$L(X \mid \theta) = \log p(X \mid \theta) = \log \sum_{Z} p(X, Z \mid \theta)$$

- 未知パラメターθの最尤推定値は

$$\hat{\theta} = \arg\max_{\theta} L(X \mid \theta)$$

EMアルゴリズムの導出　その1

$$P(X,Z|\theta)=P(Z|X,\theta)P(X|\theta)$$　を用いると

$$\log P(X,Z|\theta)=\log P(Z|X,\theta)+\log P(X|\theta)$$

ここでZに関する確率分布$q(Z)$を用いると上の式は以下のようになる

$$\sum_Z q(Z)\log\frac{P(X,Z|\theta)}{q(Z)}=\sum_Z q(Z)\log\frac{P(Z|X,\theta)}{q(Z)}+\log P(X|\theta)$$

この式は以下のようにみなせる

$$\sum_Z q(Z)\log\frac{P(X,Z|\theta)}{q(Z)}=-KL(q(Z)\|P(Z|X,\theta))+\log P(X|\theta)$$

すなわち、

$$\log P(X|\theta)=KL(q(Z)\|P(Z|X,\theta))+\sum_Z q(Z)\log\frac{P(X,Z|\theta)}{q(Z)}$$

➢EMアルゴリズムの導出　その2

$$\log P(X \mid \theta) = KL(q(Z) \,\|\, P(Z \mid X, \theta)) + \sum_Z q(Z) \log \frac{P(X, Z \mid \theta)}{q(Z)}$$

ここでZに対する分布の推定は、θの現在の推定値θ^{old}を用いると

$P(Z \mid X, \theta^{old})$なので、これを$q(Z)$とする。すなわち

$$\log P(X \mid \theta) = KL(P(Z \mid X, \theta^{old}) \,\|\, P(Z \mid X, \theta)) + \sum_Z P(Z \mid X, \theta^{old}) \log \frac{P(X, Z \mid \theta)}{P(Z \mid X, \theta^{old})}$$

さて、θ^{pld}を更新したθ^{new}を、$\log P(X \mid \theta)$を改善（より大きくする）ように

選びたい。そこで、$\log P(X \mid \theta^{new}) - \log P(X \mid \theta^{old})$を評価してみよう。

$$\log P(X \mid \theta^{new}) - \log P(X \mid \theta^{old})$$

$$= KL(P(Z \mid X, \theta^{old}) \,\|\, P(Z \mid X, \theta^{new})) - KL(P(Z \mid X, \theta^{old}) \,\|\, P(Z \mid X, \theta^{old}))$$

＝0

$$+ \sum_Z P(Z \mid X, \theta^{old}) \log \frac{P(X, Z \mid \theta^{new})}{P(Z \mid X, \theta^{old})} - \sum_Z P(Z \mid X, \theta^{old}) \log \frac{P(X, Z \mid \theta^{old})}{P(Z \mid X, \theta^{old})}$$

➢EMアルゴリズムの導出　その3

$$\log P(X \mid \theta^{new}) - \log P(X \mid \theta^{old})$$
$$= KL(P(Z \mid X, \theta^{old}) \,\|\, P(Z \mid X, \theta^{new}))$$
$$+ \sum_Z P(Z \mid X, \theta^{old}) \log P(X, Z \mid \theta^{new}) - \sum_Z P(Z \mid X, \theta^{old}) \log P(X, Z \mid \theta^{old})$$

$KL(P(Z \mid X, \theta^{old}) \,\|\, P(Z \mid X, \theta^{new}))$は定義より非負。

また、$\sum_Z P(Z \mid X, \theta^{old}) \log P(X, Z \mid \theta^{old})$は$\theta^{new}$には関係ないので、

結局、できるだけ真のθに近いθ^{new}は

$\Rightarrow$　$\sum_Z P(Z \mid X, \theta^{old}) \log P(X, Z \mid \theta)$を最大化するような$\theta$ である

$$\Rightarrow \quad \theta^{new} = \arg\max_\theta \sum_Z P(Z \mid X, \theta^{old}) \log P(X, Z \mid \theta)$$

なお、以後　$Q(\theta \mid \theta^{old}) = \sum_Z P(Z \mid X, \theta^{old}) \log P(X, Z \mid \theta)$　と書く。

$$= \mathrm{E}_{Z(\theta=\theta^{old}\text{に固定})}[\log P(X, Z \mid \theta)]$$

EMアルゴリズムの詳細

$$Q(\theta \mid \theta^{old}) = \sum_{Z} P(Z \mid X, \theta^{old}) \log P(X, Z \mid \theta) = \mathrm{E}_{Z,\theta^{old}}\left[\log P(X, Z \mid \theta)\right]$$

$$P(Z \mid X, \theta^{old}) = \frac{P(Z, X, \theta^{old})}{\sum_{Z} P(Z, X, \theta^{old})}$$

- 初期化 : θ^{old}を適当に決める
- 以下のEstep, Mstepを収束するまで繰り返す
  - E step: $P(Z|X, \theta^{old})$を計算
  - M step: $\theta^{new} = \underset{\theta}{argmax}\, Q(\theta|\, \theta^{old})$ とし、θ^{old} をθ^{new}に更新

- しかし、上記の導出から分かるように、$\log P(X \mid \theta)$ をθを動かしながら最大化しているので、局所解に陥る可能性あり。

EMとQ関数の再考

- Qを$E_{Z(\theta old)}[\log P(Z,X|\theta)]$で書き直すと
- 以下のEstep, Mstepを収束するまで繰り返す
  - E step: $P(Z|X, \theta^{old})$を計算
  - M step: $\theta^{new} = argmax_{\theta} E_{Z(\theta old)}[\log P(Z,X|\theta)]$とし、$\theta^{old}$を$\theta^{new}$に更新
- つまり、$P(Z,X|\theta)$をθ^{old}を固定してZで期待値をとることでθに関する情報を教師データZから集めてにθ反映させることを繰り返しての良い推定値を求めている。
    - K-meansに似ている
- θはベクトル。よって、Mstepでは、θの全要素を一度に更新する式を求めている点に注意。

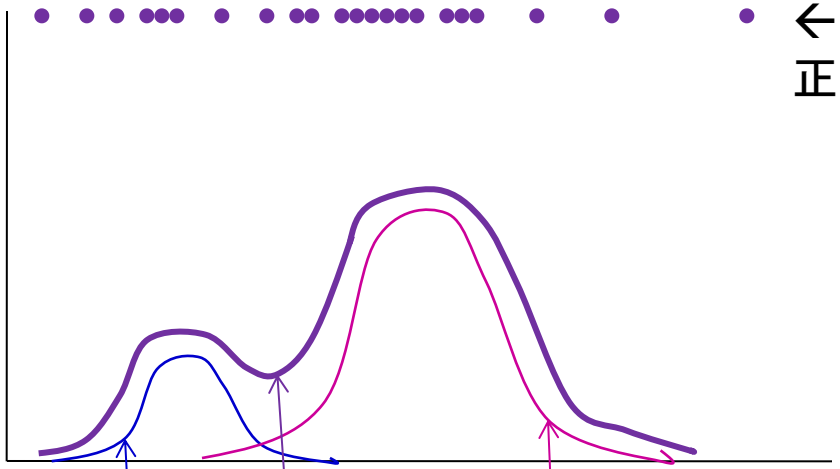
楽屋裏の話：なぜ$Q(\theta|\ \theta^{(t)})$か？

- なぜ
$$L(X\,|\,\theta)=\log p(X\,|\,\theta)=\log\sum_{Z}p(X,Z\,|\,\theta)$$
を直接最適化せずに$Q(\theta|\ \theta^{old})$か？

- $Q(\theta|\ \theta^{old})=\Sigma P(Z|X,\theta^{old})\log P(Z,X\,|\theta^{old})$
である。すなわち、Lは$\log\Sigma$の形で解析的に扱いにくい。Qは$\Sigma\log$の形で解析的に扱いやすい

- では、そもそもなぜ尤度ではなく対数尤度なのか？
- 多くの分布は $exp(f(x))$ だったり（ex 正規分布）、べき乗の形であるから、logをとると扱いやすい。
  - なぜ、expやべき乗なのか？
  - 複数の確率変数の共起は、各々の確率の積だから、という説明も可能
  - 理論的な背景から見れば「指数分布族:exponential family」であることが効果を発揮している。

EMの適用例：1変数正規分布



←このような観測データから、混合正規分布の各パラメタを推定する

混合正規分布　$\pi_1 N(x \mid \mu_1, \sigma_1) + \pi_2 N(x \mid \mu_2, \sigma_2)$　*where*　$\pi_1 + \pi_2 = 1$

要素の正規分布　$N(x \mid \mu_1, \sigma_1)$　$N(x \mid \mu_2, \sigma_2)$

問題の定式化

推定するパラメター：$\pi_1, \mu_1, \sigma_1, \pi_2, \mu_2, \sigma_2$

$$p(x) = \pi_1 N(x \mid \mu_1, \sigma_1) + \pi_2 N(x \mid \mu_2, \sigma_2) \qquad \pi_1 + \pi_2 = 1$$

潜在変数$z_k \in \{0,1\}$ $\qquad p(z_1 = 1) = \pi_1 \qquad p(z_2 = 1) = \pi_2$

$$p(\mathbf{z}) = \pi_1^{z_1} \cdot \pi_2^{z_2} \qquad (GM10)$$

$$p(x \mid \mathbf{z}) = N(x \mid \mu_1, \sigma_1)^{z_1} N(x \mid \mu_2, \sigma_2)^{z_2} \qquad (GM11)$$

$$\Rightarrow p(x) = \sum_{k=1,2} p(z_k) p(x \mid z_k) = \pi_1 N(x \mid \mu_1, \sigma_1) + \pi_2 N(x \mid \mu_2, \sigma_2) \qquad (GM20)$$

ここで次のように定義される$\gamma(z_k)$を導入し、ベイズの定理で書きなおす

$$\gamma(z_k) \equiv p(z_k = 1 \mid x) = \frac{p(z_k = 1) p(x \mid z_k = 1)}{\sum_{k=1,2} p(z_k = 1) p(x \mid z_k = 1)}$$

$$= \frac{\pi_k N(x \mid \mu_k, \sigma_k)}{\sum_{j=1,2} \pi_j N(x \mid \mu_j, \sigma_j)} \qquad (GM30)$$

いよいよEMアルゴリズムの適用

- 次のパラメターに適当な初期値を設定：$\pi_k, \mu_k, \sigma^2{}_k$
- E step: $P(Z|X, \theta^{(t)})$を計算
  - ただし、観測されたデータはN個あるとする。
  - 実際には、$P(Z|X, \theta^{(t)})$ではなく **Zの期待値** を求めておくことにする

$(GM10)$と$(GM11)$より　　$p(Z \mid X, \mu, \sigma^2, \pi) \propto \prod_{n=1}^{N} \left[\pi_1 N(x_n \mid \mu_1, \sigma_1^2)\right]^{z_{n1}} \left[\pi_2 N(x_n \mid \mu_2, \sigma_2^2)\right]^{z_{n2}}$

ここでZ（すなわちz_{n1}, z_{n2}　まとめてz_{nk}　*where*　$k = 1,2$)を評価する。

$$\mathrm{E}[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k N(x_n \mid \mu_k, \sigma_k{}^2)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j N(x_n \mid \mu_j, \sigma_j{}^2)\right]^{z_{nj}}} = \frac{\pi_k N(x_n \mid \mu_k, \sigma_k{}^2)}{\sum_{j=1,2} \pi_j N(x_n \mid \mu_j, \sigma_j{}^2)} = \gamma(z_{nk}) \qquad (GM40)$$

これはデータx_nのz_{nk}(=どちらの正規分布が選ばれるか）への寄与を表す

note!　定義より

$$\sum_{n=1}^{N}\sum_{k=1}^{2} \gamma(z_{nk}) = \sum_{n=1}^{N}\sum_{k=1}^{2} \mathrm{E}[z_{nk}] = N$$

$$\sum_{n=1}^{N} \gamma(z_{nk}) = \sum_{n=1}^{N} \mathrm{E}[z_{nk}] = N_k$$

π_k, μ_k, σ_k の現在の値から計算した更新値

➢ Mstep

$(GM10)(GM11)$より

$$p(X,Z\mid\mu,\sigma,\pi)=p(X\mid Z)p(Z)=\prod_{n=1}^{N}\left[\pi_1 N(x_n\mid\mu_1,\sigma_1)\right]^{z_{n1}}\left[\pi_2 N(x_n\mid\mu_2,\sigma_2)\right]^{z_{n2}}$$

$$\Rightarrow \log p(X,Z\mid\mu,\sigma,\pi)=\sum_{n=1}^{N}\sum_{k=1}^{2} z_{nk}\left\{\log\pi_k+\log N(x_n\mid\mu_k,\sigma_k)\right\}$$

$(GM40)$より

$$\mathrm{E}_Z\left[\log p(X,Z\mid\mu,\sigma^2,\pi)\right]=\sum_{n=1}^{N}\sum_{k=1}^{2}\mathrm{E}[z_{nk}]\left\{\log\pi_k+\log N(x_n\mid\mu_k,{\sigma_k}^2)\right\}$$

$$=\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\log\pi_k+\log N(x_n\mid\mu_k,{\sigma_k}^2)\right\}\qquad(GM50)$$

次に$\gamma(z_{nk})$を固定した上で$(GM50)$において

$\pi_k,\mu_k,{\sigma_k}^2$を最大化してこの*step*での最適値を求める。

μ_kの最適化しμ_k^{new}に更新する

$(GM50)$をμ_kで微分してゼロとおく。

$$\frac{\partial}{\partial \mu_k}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\{\log \pi_k + \log N(x_n \mid \mu_k, \Sigma_k)\}$$

$$= \frac{\partial}{\partial \mu_k}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\log \pi_k - \frac{(x_n-\mu_k)^2}{2\sigma_k^2} + const\right\}$$

$$\propto \sum_{n=1}^{N}\gamma(z_{nk})\frac{(x_n-\mu_k)}{\sigma_k^2} = 0$$

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} x_n\gamma(z_{nk})}{\sum_{n=1}^{N}\gamma(z_{nk})} = \frac{1}{N_k}\sum_{n=1}^{N} x_n\gamma(z_{nk}) \qquad (GM70)$$

$\sigma_k{}^2$の最適化し、$\sigma_k{}^{2^{new}}$に更新する

$(GM50)$をσ_kで微分してゼロとおく。

$$\frac{\partial}{\partial \sigma_k{}^2}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\log \pi_k + \log N(x_n \mid \mu_k, \sigma_k{}^2)\right\}$$

$$= \frac{\partial}{\partial \sigma_k{}^2}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{-\frac{1}{2}\log \sigma_k{}^2 - \frac{(x_n-\mu_k)^2}{2\sigma_k{}^2} + const\right\}$$

$$= \sum_{n=1}^{N}\gamma(z_{nk})\left\{-\frac{1}{2\sigma_k{}^2} + \frac{(x_n-\mu_k)^2}{2\left(\sigma_k{}^2\right)^2}\right\} = 0 \qquad (GM801)$$

$(GM801)$より　$\sigma_k{}^2 \cdot \sum_{n=1}^{N}\gamma(z_{nk}) = \sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)^2$

$$\Rightarrow \quad \sigma_k{}^{2^{new}} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)^2}{N_k} \qquad (GM80)$$

π_kの最適化

$$L=\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\{\log\pi_k+\log N(x_n\mid\mu_k,\sigma_k{}^2)\}+\lambda\left(\sum_{k=1}^{2}\pi_k-1\right)$$

$$\frac{\partial L}{\partial\pi_k}=\sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k}+\lambda=\sum_{n=1}^{N}\gamma(z_{nk})+\lambda\pi_k=N_k{}^{new}+\lambda\pi_k=0$$

$$\text{一方}\sum_{k=1}^{2}\left\{\sum_{n=1}^{N}\gamma(z_{nk})+\lambda\pi_k\right\}=\quad N+\lambda=0$$

以上の2式より

$$\pi_k=\frac{N_k}{N}\qquad(GM90)$$

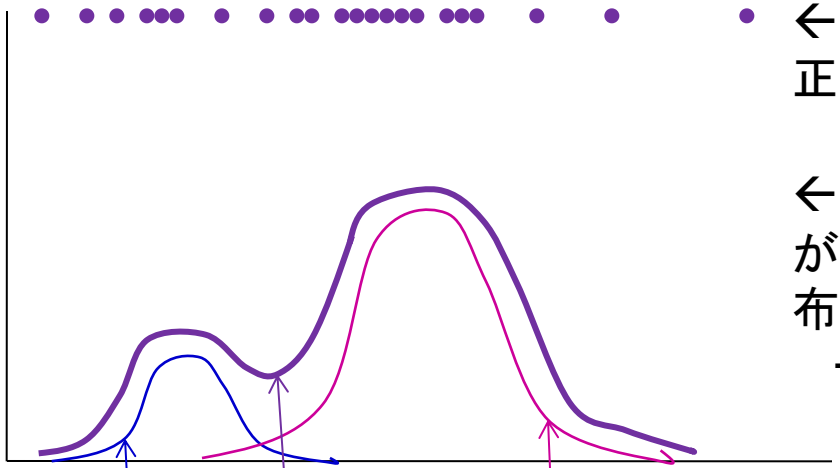
- ここでは、$\gamma(z_k)$ が古い $\pi_k,\mu_k,\sigma_k{}^2$ を反映して計算された値であった。それを固定して、loglikelihoodを最大化する新たな $\pi_k,\mu_k,\sigma_k{}^2$ を求めているわけだ。

- 以上の(GM70)(GM80)(GM90)によって $\pi_k,\mu_k,\sigma_k{}^2$ の更新式が求められた。

- log likelihood (GM50)が収束しなければEstepに戻る

EMの適用例：混合多変数正規分布

1変数の場合と似ているが、少し難しいところあり



←このような観測データから、混合正規分布の各パラメタを推定する

←この例では1変数の正規分布だが以下の導出は多変数の正規分布を仮定している。

→ x, μ_1, μ_2はベクトル,

$\Sigma_1(=\Lambda_1^{-1}), \Sigma_2(=\Lambda_2^{-1})$,は行列

混合正規分布 $\pi_1 N(x \mid \mu_1, \Sigma_1) + \pi_2 N(x \mid \mu_2, \Sigma_2)$ *where* $\pi_1 + \pi_2 = 1$

要素の正規分布 $N(x \mid \mu_1, \Sigma_1)$ $N(x \mid \mu_2, \Sigma_2)$

Σは共分散行列、Λは精度行列であることに注意

下の式はk変数の正規分布においてN個のデータがある場合

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} x_{i1}-\mu_1 \\ \vdots \\ x_{ik}-\mu_k \end{bmatrix}\left[x_{i1}-\mu_1,\ldots,x_{ik}-\mu_k\right]$$

$$= \begin{bmatrix} \sigma_{x1}^2 & \cdots & \sigma_{x1,x2} \\ \vdots & \ddots & \vdots \\ \sigma_{x1,x2} & \cdots & \sigma_{xk}^2 \end{bmatrix} = \Lambda^{-1}$$

問題の定式化

推定するパラメター：$\pi_1, \mu_1, \Sigma_1, \pi_2, \mu_2, \Sigma_2$

$$p(x) = \pi_1 N(x \mid \mu_1, \Sigma_1) + \pi_2 N(x \mid \mu_2, \Sigma_2) \qquad \pi_1 + \pi_2 = 1$$

潜在変数$z_k \in \{0,1\}$ $\qquad p(z_1 = 1) = \pi_1 \qquad p(z_2 = 1) = \pi_2$

$$p(\mathbf{z}) = \pi_1^{z_1} \cdot \pi_2^{z_2} \qquad (GM10)$$

$$p(x \mid \mathbf{z}) = N(x \mid \mu_1, \Sigma_1)^{z_1} N(x \mid \mu_2, \Sigma_2)^{z_2} \qquad (GM11)$$

$$\Rightarrow p(x) = \sum_{k=1,2} p(z_k)\, p(x \mid z_k) = \pi_1 N(x \mid \mu_1, \Sigma_1) + \pi_2 N(x \mid \mu_2, \Sigma_2) \qquad (GM20)$$

ここで次のように定義される$\gamma(z_k)$を導入し、ベイズの定理で書きなおす

$$\gamma(z_k) \equiv p(z_k = 1 \mid x) = \frac{p(z_k = 1)\, p(x \mid z_k = 1)}{\sum_{k=1,2} p(z_k = 1)\, p(x \mid z_k = 1)}$$

$$= \frac{\pi_k N(x \mid \mu_k, \Sigma_k)}{\sum_{j=1,2} \pi_j N(x \mid \mu_j, \Sigma_j)} \qquad (GM30)$$

いよいよEMアルゴリズムの適用

- 次のパラメターに適当な初期値を設定: π_k, μ_k, Σ_k
- E step: $P(Z|X, \theta^{(t)})$を計算
  - ただし、観測されたデータはN個あるとする。
  - 実際には、$P(Z|X, \theta^{(t)})$ではなく **Zの期待値** を求めておくことにする。

$(GM10)$と$(GM11)$より $\quad p(Z \mid X, \mu, \Sigma, \pi) \propto \prod_{n=1}^{N} \left[\pi_1 N(x_n \mid \mu_1, \Sigma_1)\right]^{z_{n1}} \left[\pi_2 N(x_n \mid \mu_2, \Sigma_2)\right]^{z_{n2}}$

ここでZ（すなわちz_{n1}, z_{n2}　まとめてz_{nk}　*where*　$k = 1,2$)を評価する。

$$\mathrm{E}[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k N(x_n \mid \mu_k, \Sigma_k)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j N(x_n \mid \mu_j, \Sigma_j)\right]^{z_{nj}}} = \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1,2} \pi_j N(x_n \mid \mu_j, \Sigma_j)} = \gamma(z_{nk}) \qquad (GM40)$$

これはデータx_nのz_{nk}(=どちらの正規分布が選ばれるか）への寄与を表す

note!　定義より

$$\sum_{n=1}^{N}\sum_{k=1}^{2} \gamma(z_{nk}) = \sum_{n=1}^{N}\sum_{k=1}^{2} \mathrm{E}[z_{nk}] = N$$

$$\sum_{n=1}^{N} \gamma(z_{nk}) = \sum_{n=1}^{N} \mathrm{E}[z_{nk}] = N_k$$

π_k, μ_k, Σ_kの現在の値から計算した更新値

➢ Mstep

$(GM10)(GM11)$より

$$p(X,Z\mid\mu,\Sigma,\pi)=p(X\mid Z)\,p(Z)=\prod_{n=1}^{N}\left[\pi_1 N(x_n\mid\mu_1,\Sigma_1)\right]^{z_{n1}}\left[\pi_2 N(x_n\mid\mu_2,\Sigma_2)\right]^{z_{n2}}$$

$$\Rightarrow \log p(X,Z\mid\mu,\Sigma,\pi)=\sum_{n=1}^{N}\sum_{k=1}^{2} z_{nk}\left\{\log\pi_k+\log N(x_n\mid\mu_k,\Sigma_k)\right\}$$

$(GM40)$より

$$\mathrm{E}_Z\left[\log p(X,Z\mid\mu,\Sigma,\pi)\right]=\sum_{n=1}^{N}\sum_{k=1}^{2}\mathrm{E}\left[z_{nk}\right]\left\{\log\pi_k+\log N(x_n\mid\mu_k,\Sigma_k)\right\}$$

$$=\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\log\pi_k+\log N(x_n\mid\mu_k,\Sigma_k)\right\}\qquad(GM50)$$

$$=\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\log\pi_k+\log N(x_n\mid\mu_k,\Lambda_k^{-1})\right\}$$

次に$\gamma(z_{nk})$を固定した上で$(GM50)$において

π_k,μ_k,Σ_kを最適化してこの*step*での最適値を求める。

μ_kの最適化しμ_k^{new}に更新する

$(GM50)$をμ_kで微分してゼロとおく。

$$\frac{\partial}{\partial \mu_k}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\{\log \pi_k + \log N(x_n \mid \mu_k, \Sigma_k)\}$$

$$= \frac{\partial}{\partial \mu_k}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\log \pi_k + \frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k) + const\right\}$$

$$\propto \sum_{n=1}^{N}\gamma(z_{nk})\Sigma_k^{-1}(x_n - \mu_k) = 0$$

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} x_n \gamma(z_{nk})}{\sum_{n=1}^{N}\gamma(z_{nk})} = \frac{1}{N_k}\sum_{n=1}^{N} x_n \gamma(z_{nk}) \qquad (GM70)$$

Σ_kの最適化し、Σ_k^{new}に更新する

ここが多変数だと難しくなる部分

$\Rightarrow \Lambda_k(=\Sigma_k^{-1})$の最適化し、$\Lambda_k^{new}(=\Sigma_k^{-1^{new}})$に更新する

$(GM50)$をΛ_kで微分してゼロとおく。

$$\frac{\partial}{\partial \Lambda_k}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\log \pi_k + \log N(x_n \mid \mu_k, \Lambda^{-1}{}_k)\right\}$$

$$= \frac{\partial}{\partial \Lambda_k}\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\left\{\frac{1}{2}\log|\Lambda_k| - \frac{1}{2}(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k) + const\right\}$$

$$= \sum_{n=1}^{N}\gamma(z_{nk})\left\{\frac{1}{2}\frac{\partial \log|\Lambda_k|}{\partial \Lambda_k} - \frac{1}{2}\frac{\partial (x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)}{\partial \Lambda_k}\right\} = 0 \qquad (GM801)$$

$$\sum_{n=1}^{N}\gamma(z_{nk})\left\{\frac{1}{2}\frac{\partial\log|\Lambda_k|}{\partial\Lambda_k}-\frac{1}{2}\frac{\partial(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)}{\partial\Lambda_k}\right\}=0 \qquad (GM801)$$

$(GM801)$のおのおのの項の微分を計算する

$$\frac{\partial\log|\Lambda_k|}{\partial\Lambda_k}=\left(\Lambda_k^{-1}\right)^T=\Lambda_k^{-1} \qquad (GM802)$$

公式　$\mathbf{x}^T A\mathbf{x}=trace(A\mathbf{x}\mathbf{x}^T)$　　$\frac{\partial trace(\mathbf{AB})}{\partial\mathbf{A}}=\mathbf{B}^T$　　より

$$\frac{\partial\log|\Lambda_k|}{\partial\Lambda_k}-\frac{\partial}{\partial\Lambda_k}(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)=\Lambda_k^{-1}-\frac{\partial}{\partial\Lambda_k}trace\left(\Lambda_k(x_n-\mu_k)(x_n-\mu_k)^T\right)$$

$$=\Lambda_k^{-1}-\left((x_n-\mu_k)(x_n-\mu_k)^T\right)^T=\Lambda_k^{-1}-(x_n-\mu_k)(x_n-\mu_k)^T=0 \qquad (GM803)$$

$(GM801)(GM802)(GM803)$より $\sum_{n=1}^{N}\gamma(z_{nk})\Lambda_k^{-1}=\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T$

$$\sum_{n=1}^{N}\gamma(z_{nk})\Lambda_k^{-1}=\Lambda_k^{-1}\sum_{n=1}^{N}\gamma(z_{nk})=\Lambda_k^{-1}N_k$$

$$\Rightarrow \quad \Sigma_k^{new}=\Lambda_k^{-1new}=\frac{\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T}{N_k} \qquad (GM80)$$

π_kの最適化

$$L=\sum_{n=1}^{N}\sum_{k=1}^{2}\gamma(z_{nk})\{\log\pi_k+\log N(x_n\mid\mu_k,\Sigma_k)\}+\lambda\left(\sum_{k=1}^{2}\pi_k-1\right)$$

$$\frac{\partial L}{\partial\pi_k}=\sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k}+\lambda=\sum_{n=1}^{N}\gamma(z_{nk})+\lambda\pi_k=N_k^{new}+\lambda\pi_k=0$$

$$\text{一方}\sum_{k=1}^{2}\left\{\sum_{n=1}^{N}\gamma(z_{nk})+\lambda\pi_k\right\}=\quad N+\lambda=0$$

以上の2式より

$$\pi_k=\frac{N_k}{N}\qquad(GM90)$$

- ここでは、$\gamma(z_k)$ が古い π_k,μ_k,Σ_k を反映して計算された値であった。それを固定して、loglikelihoodを最大化する新たな π_k,μ_k,Σ_k を求めているわけだ。

- 以上の(GM70)(GM80)(GM90)によって π_k,μ_k,Σ_k の更新式が求められた。

- log likelihood (GM50)が収束しなければEstepに戻る

EM法の応用：不完全観測データの場合のモデル推定：多項分布の場合

- 観測値が不完全な場合としては、複数の確率変数があるのに観測されるのは、K個の和だけなどという場合。
- 例題：母集団が次の多項分布である場合にN個の観測値からパラメターを推定する問題を考える。

$$p(x_1, x_2, x_3, x_4, x_5) = \frac{N!}{x_1! x_2! x_3! x_4! x_5!} \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3} \theta_4^{x_4} \theta_5^{x_5}$$

$$\text{ただし} (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = \left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4} \right) \quad (1)$$

- 観測値としては、x_1,x_2,x_3,x_4,x_5ではなく、x_1+x_2に対応するyとx_3,x_4,x_5が得られていたとする。
- このため、観測値から直接にパラメター θを求められない。

- そこで、以下のstep1, step2を繰り返して θ を近似する。ただし、θの初期値を $\theta(0)$とする。また、以下はk+1回目の繰り返しとする。

step1　既に求まっている$\theta(\mathrm{k})$を用いて x_1,x_2,x_3,x_4,x_5の推定値を求める。

θの近似値として$\theta(k)$が与えられていたときの対数尤度は次式

$$\log\frac{N!}{x_1!x_2!x_3!x_4!x_5!}-x_1\log 2-(x_2+x_3+x_4+x_5)\log 4+(x_2+x_5)\log\theta+(x_3+x_4)\log(1-\theta)$$

$$=(x_2+x_5)\log\theta+(x_3+x_4)\log(1-\theta)+const$$

x_1,x_2,x_3,x_4,x_5のなす多項分布は次式

$$\frac{N!}{x_1!x_2!x_3!x_4!x_5!}\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y}\left(\frac{1-\theta}{4}\right)^{x_3}\left(\frac{1-\theta}{4}\right)^{x_4}\left(\frac{\theta}{4}\right)^{x_5}$$

$\Rightarrow$ ただし、$y=x_1+x_2$なので、x_1,x_2の分布は次式

$$\frac{y!}{x_1!x_2!}\left(\frac{\frac{1}{2}}{\frac{1}{2}+\frac{\theta_{old}}{4}}\right)^{x_1}\left(\frac{\frac{\theta_{old}}{4}}{\frac{1}{2}+\frac{\theta_{old}}{4}}\right)^{x_2}=\frac{y!}{x_1!x_2!}\left(\frac{2}{2+\theta_{old}}\right)^{x_1}\left(\frac{\theta_{old}}{2+\theta_{old}}\right)^{x_2}$$

step2 step1の結果を用いてQ関数

$$Q\left(\theta \mid \theta^{old}\right) = E_{x_1,x_2,x_3,x_4,x_5,\theta=\theta^{old}}\left[p(x_1,x_2,x_3,x_4,x_5,\theta)\right]$$

θの新しい近似値$\theta(k+1)$を次式で求める

$$\theta(k+1) = \arg\max_{\theta} E_{x_1,x_2,x_3,x_4,x_5,\theta=\theta^{old}}\left[p(x_1,x_2,x_3,x_4,x_5,\theta)\right]$$

具体的には

$$Q\left(\theta \mid \theta^{old}\right) = E_{x_1,x_2,x_3,x_4,x_5,\theta=\theta^{old}}\left[p(x_1,x_2,x_3,x_4,x_5,\theta)\right]$$

$$= E_{\theta^{old}}\left[(x_2+x_5)\log\theta + (x_3+x_4)\log(1-\theta) + const\right] \quad (*)$$

yは$x1, x2$が各々確率$\dfrac{\frac{1}{2}}{\frac{1}{2}+\frac{\theta_{old}}{4}}, \dfrac{\frac{\theta_{old}}{4}}{\frac{1}{2}+\frac{\theta_{old}}{4}}$である2項分布で$y = x1 + x2$

2項分布$\binom{n}{k}p^k(1-p)^{n-k}$の期待値は$np$

この場合には　$n=y, p=\dfrac{\theta_{old}}{2+\theta_{old}} \Rightarrow E[x_2]=y\dfrac{\theta_{old}}{2+\theta_{old}}$

$$(*)=\left(y\frac{\theta_{old}}{2+\theta_{old}}+x_5\right)\log\theta+(x_3+x_4)\log(1-\theta)+const$$

$$\frac{\partial Q(\theta\mid\theta^{old})}{\partial\theta}=\frac{\left(y\dfrac{\theta_{old}}{2+\theta_{old}}+x_5\right)}{\theta}-\frac{x_3+x_4}{1-\theta}=0$$

$$\Rightarrow \theta^{new}=\arg\max_{\theta} Q(\theta\mid\theta^{old})=\frac{y\dfrac{\theta_{old}}{2+\theta_{old}}+x_5}{y\dfrac{\theta_{old}}{2+\theta_{old}}+x_3+x_4+x_5}$$

準備：KL divergence

- 相対エントロピー　or Kullback-Leibler divergence or KL divergence: KL(P||Q)：分布PとQの類似性を測る尺度

$$KL(P||Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

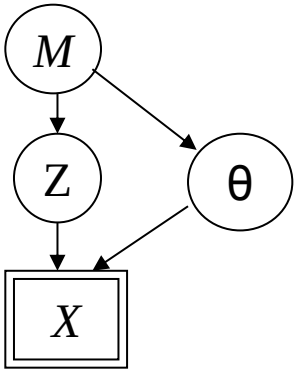
  - KL(P||P)=0
  - KL(P||Q)≠KL(Q||P)
    - 非対称なので擬距離
    - 対称性を持たせるために SymmetricKL(P||Q)=(KL(P||Q)+KL(Q||P))/2　という尺度もある
- 相互情報量：

$$I(x,y) = KL(P(x,y) \| P(x)P(y)) = \sum P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

この部分をpointwise mutual informationとして使うこともある

変分ベイズ法（Variational Bayse:VB)

- 観測データからのベイズ推定
- 観測データ：X、未知パラメター：θ、
  モデル構造：M、潜在変数集合：Z



$$p(X,Z,\theta,M) = p(X,Z \mid \theta,M)\,p(\theta \mid M)\,p(M)$$

$$p(\theta,Z,M \mid X) = \frac{p(X,Z \mid \theta,M)\,p(\theta \mid M)\,p(M)}{p(X)} = \frac{p(X,Z \mid \theta,M)\,p(\theta \mid M)\,p(M)}{\sum_{M}\sum_{Z}\int p(X,Z,\theta,M)\mathrm{d}\theta}$$

$$p(\theta \mid X) = \sum_{M}\sum_{Z} p(\theta,Z,M \mid X)$$

この積分の計算が
困難

- 新たなデータxの事後予測分布

$$p(x \mid X) = \int p(x \mid \theta)\,p(\theta \mid X)\mathrm{d}\theta$$

計算の困難さの問題点を詳しくいうと

- ベイズ推定は、最尤推定と異なり、未知データの予測値ではなく予測分布を求める
- 教師データが少ない場合でも、汎化能力の高い予測器が作れる
- ただし、$P(Z|\mathbf{x},\theta,M)$、$\mathbf{x}$は1個の観測データ（L次元）で、ベイズの定理で次のように変形するが

$$P(Z \mid \mathbf{x},\theta,M) = \frac{P(\mathbf{x},Z \mid \theta,M)}{\sum_{Z} P(\mathbf{x},Z \mid \theta,M)}$$

- 右辺 $P(\mathbf{x},Z|\theta,M)$は
  - $\mathbf{x}$はZの成分$(z_1,z_2,\ldots,z_K)$に組み合わせで依存しているため、次式のように分解できない。

$$P(\mathbf{x},Z \mid \theta,M) = \prod_{k=1}^{K}\prod_{l=1}^{L} P(x_l, z_k \mid \theta,M)$$

  - よって、Kが大きくなるとZの成分の組み合わせの数が膨大になり計算が困難

変分ベイズ法の考え方

- 問題であった期待値の計算を近似的に確率的シミュレーションで解くMCMC法が有力。
- ただし、MCMCは計算が膨大。数理モデルを工夫し計算を効率化する方法として変分ベイズ法
- EM法では、$Q(\theta)$を最大にするθを求めた。
- VB法では、θの値を最大化の対象にするのではなく、θの分布の形そのものを求める→変分法

変分ベイズ法のトリック

$$L(X) = \log P(X) = \log \sum_M \sum_Z \int p(X, Z, \theta, M) \mathrm{d}\theta$$

$$= \log \sum_M \sum_Z \int q(Z, \theta, M) \frac{p(X, Z, \theta, M)}{q(Z, \theta, M)} \mathrm{d}\theta$$

*Jensen*の不等式 $\log(E[x]) \geq E[\log(x)]$ より

$$\geq \sum_M \sum_Z \int q(Z, \theta, M) \log \frac{p(X, Z, \theta, M)}{q(Z, \theta, M)} \mathrm{d}\theta \equiv \mathcal{F}(q)$$

$\sum_M \sum_Z \int q(Z, \theta, M) \mathrm{d}\theta = 1$ に注意。

すなわち、$\log P(X)$はM, θ, Zに対して

周辺化しているので、M, θ, Zに依存しない。

$$L(X) - \mathcal{F}(q)$$

$$= \sum_M \sum_Z \int q(Z,\theta,M) \log p(X) \mathrm{d}\theta - \sum_M \sum_Z \int q(Z,\theta,M) \log \frac{p(X,Z,\theta,M)}{q(Z,\theta,M)} \mathrm{d}\theta = (1)$$

$$\frac{p(X,Z,\theta,M)}{p(X)} = p(Z,\theta,M \mid X) \quad \text{だから}$$

$$(1) = -\sum_M \sum_Z \int q(Z,\theta,M) \log \frac{p(Z,\theta,M \mid X)}{q(Z,\theta,M)} \mathrm{d}\theta = KL(q \| p)$$

よって $L(X) = \mathcal{F}(q) + KL(q \| p)$

- $L(X)$はqに依存しないから、$\mathcal{F}(q)$を最大化することは$KL(q||p)$を最小化、すなわちpに最も似たqを求めることになる
- EMでは、qを$P(X,Z \mid \theta^{(t)})$ と決め打ちしていたが、VBでは、より柔軟な決め方を行っている。

- 上記のqを求めるプロセスをいきなり計算することは困難なので、パラメターの事前分布とq関して以下の仮定をおく。

- 因子化の仮定

$$p(\theta \mid M) = \prod_{i=1}^{I} p(\theta_i \mid M)$$

$$q(Z, \theta, M) = q(M) q(Z \mid M) \prod_{i=1}^{I} q(\theta_i \mid M)$$

- この仮定の下で、変分法を使えば、次に述べる変分ベイズ法のアルゴリズムが導ける

- しかし、この仮定が成立しないこともあるので、適用する問題によっては注意が必要。

因子化の仮定下でのVBの導出　その1

$$\mathcal{F}(q)=\sum_{M}\sum_{Z}\int q(M)q(Z\mid M)\prod_{i=1}^{I}q(\theta_i\mid M)\log\frac{p(X,Z\mid\theta,M)p(\theta\mid M)p(M)}{q(Z\mid M)\left(\prod_{i=1}^{I}q(\theta_i\mid M)\right)q(M)}\mathrm{d}\theta$$

$\mathrm{d}\theta=\mathrm{d}\theta_1\cdots\mathrm{d}\theta_I$　　$p(\theta\mid M)=\prod_{i=1}^{I}p(\theta_i\mid M)$　に注意すると

$$\mathcal{F}(q)=\sum_{M}q(M)\left\{\begin{array}{l}\displaystyle\sum_{Z}\int q(Z\mid M)\prod_{i=1}^{I}q(\theta_i\mid M)\log\frac{p(X,Z\mid\theta,M)}{q(Z\mid M)}\mathrm{d}\theta\\ \displaystyle+\sum_{i=1}^{I}\int q(\theta_i\mid M)\log\frac{p(\theta_i\mid M)}{q(\theta_i\mid M)}\mathrm{d}\theta_i\\ \displaystyle+\log\frac{p(M)}{q(M)}\end{array}\right.$$

$$\sum_{Z}q(Z\mid M)=1$$

$$\int q(\theta_j\mid M)\mathrm{d}\theta_j=1$$

$$\sum_{Z}q(Z\mid M)\int\prod_{i=1}^{I}q(\theta_i\mid M)\mathrm{d}\theta_1\cdots\mathrm{d}\theta_I=1$$

因子化の仮定下でのVBの導出　その2

モデルMが与えられたときのZの最適な分布$q(Z\mid M)$は

$\sum_{Z} q(Z\mid M)=1$　という条件下で$\mathcal{F}(q)$を$q(Z\mid M)$に対して最大化

すなわち、次の$J[q(Z\mid M)]$の極値問題を解く。

$$J[q(Z\mid M)]=\sum_{M} q(M)\sum_{Z}\int q(Z\mid M)\prod_{i=1}^{I} q(\theta_i\mid M)\log\frac{p(X,Z\mid\theta,M)}{q(Z\mid M)}\mathrm{d}\theta$$

$$+\lambda\left(\sum_{Z} q(Z\mid M)-1\right)$$

$J[q(Z\mid M)]$が$q(Z\mid M)$の微分を含まないので

$$\frac{\partial J[q(Z\mid M)]}{\partial q(Z\mid M)}=0 \qquad \frac{\partial J[q(Z\mid M)]}{\partial\lambda}=0$$

ただし、$q(Z\mid M)$はθと無関係なので、$J[q(Z\mid M)]$の右辺$\int$ の前に出、

次のようになる。

因子化の仮定下でのVBの導出　その3

$$\frac{\partial J[q(Z\mid M)]}{\partial q(Z\mid M)}=$$

$$\frac{\partial}{\partial q(Z\mid M)}\left[\sum_M q(M)\left\{\begin{array}{c}\sum_Z q(Z\mid M)\int\prod_{i=1}^{I} q(\theta_i\mid M)\log p(X,Z\mid\theta,M)\mathrm{d}\theta \\ -\sum_Z q(Z\mid M)\log q(Z\mid M)\int\prod_{i=1}^{I} q(\theta_i\mid M)\mathrm{d}\theta \end{array}\right\}+\lambda\left(\sum_Z q(Z\mid M)-1\right)\right]$$

$$=\sum_M q(M)\left\{\begin{array}{l}\int\prod_{i=1}^{I} q(\theta_i\mid M)\log p(X,Z\mid\theta,M)\mathrm{d}\theta \\ -\log q(Z\mid M)\int\prod_{i=1}^{I} q(\theta_i\mid M)\mathrm{d}\theta-\int\prod_{i=1}^{I} q(\theta_i\mid M)\mathrm{d}\theta+\lambda\end{array}\right\}=0$$

←なぜなら$\sum_M q(M)=1$だから$\lambda\sum_M q(M)=\lambda$とおけるので

因子化の仮定下でのVBの導出　その4

さらに$\int \prod_{i=1}^{I} q(\theta_i \mid M)\mathrm{d}\theta = 1$だから、結局

$$\int \prod_{i=1}^{I} q(\theta_i \mid M)\log p(X, Z \mid \theta, M)\mathrm{d}\theta - 1 + \lambda = \log q(Z \mid M)$$

$$\therefore \quad q(Z \mid M) = C\exp\left\{\int \prod_{i=1}^{I} q(\theta_i \mid M)\log p(X, Z \mid \theta, M)\mathrm{d}\theta\right\}$$

もし、θが確率変数でなく確定していると、

$q(\theta_i \mid M) = \delta(\theta_i = \theta)$　　θはあるiに対するθ_iの値（スカラー）$-(*)$

よって、$q(Z \mid M) = Cp(X, Z \mid \theta, M)$　　$C = \dfrac{1}{\sum_Z p(X, Z \mid \theta, M)}$

この式のθは$(*)$のθからなるベクトル

因子化の仮定下でのVBの導出　その5

モデルMが与えられたときの　Z　の最適な分布　$q(\theta_i \mid M)$　は

次の$J[q(\theta_i \mid M)]$(=下の式の[ ]の内部）の極値問題を解く。

$$\frac{\partial}{\partial q(\theta_i \mid M)}\left[\sum_M q(M)\left\{\begin{array}{l}\sum_Z \int q(Z \mid M)\prod_{i=1}^{I} q(\theta_i \mid M)\left(\log p(X,Z \mid \theta, M) - \log q(Z,M)\right)\mathrm{d}\theta \\ \quad + \sum_{i=1}^{I}\int q(\theta_i \mid M)\log\frac{p(\theta_i \mid M)}{q(\theta_i \mid M)}\mathrm{d}\theta_i \\ \quad + \lambda\left(\int q(\theta_i \mid M)\mathrm{d}\theta_i - 1\right)\end{array}\right\}\right]$$

$$= \sum_M q(M)\left\{\begin{array}{l}\int\left[\sum_Z q(Z \mid M)\int\prod_{i\neq j} q(\theta_j \mid M)\log p(X,Z \mid \theta, M)\mathrm{d}\theta_{-i}\right]\mathrm{d}\theta_i \\ \quad + \int\log\frac{p(\theta_i \mid M)}{q(\theta_i \mid M)} - 1\,\mathrm{d}\theta_i \qquad + \int\lambda\mathrm{d}\theta_i\end{array}\right\} = 0$$

$\rightarrow$　　$\int f(\theta_i)\mathrm{d}\theta_i = 0$　だから　$f(\theta_i) = 0$

よって　$\sum_Z q(Z \mid M)\int\prod_{i\neq j} q(\theta_j \mid M)\log p(X,Z \mid \theta, M)\mathrm{d}\theta_{-i} = \log\frac{q(\theta_i \mid M)}{p(\theta_i \mid M)} + C$

$\rightarrow$　$$q(\theta_i \mid M) = C\, p(\theta_i \mid M)\exp\left\{\sum_Z q(Z \mid M)\int\prod_{i\neq j} q(\theta_j \mid M)\log p(X,Z \mid \theta, M)\mathrm{d}\theta_{-i}\right\}$$

変分ベイズ法のアルゴリズム

- 初期化として、以下の初期分布を設定

$$\{q(\theta_i \mid M)^{old}\}, \{p(\theta_i \mid M)^{old}\} \qquad i = 1, \ldots, I$$

- 反復計算　以下を収束するまで繰り返す。
  - VB-E step

$$q(Z \mid M)^{new} = C \exp\left(\sum_M \sum_Z \int q(\theta \mid M)^{old} \log p(X, Z \mid \theta, M) \mathrm{d}\theta\right) = C \exp\left(E_{M,Z,\theta}[\log p(X, Z \mid \theta, M)]\right)$$

  - VB-M step

$$q(\theta_i \mid M)^{new}$$

$$= C' p(\theta_i \mid M) \exp\left(\sum_M \sum_Z \int q(Z \mid M)^{new} q(\theta \setminus \{\theta_i\} \mid M)^{old} \log p(D, Z \mid \theta, M) dZ d\theta \setminus \{\theta_i\}\right)$$

$$= C' p(\theta_i \mid M) \exp\left(E_{M^{new}, Z, \theta \setminus \{\theta_i\}^{old}}[\log p(X, Z \mid \theta, M)]\right)$$

変数^{new}を変数^{old}とする。

- $\theta \setminus \{\theta_i\}$ はθの構成要素からθ_iを除いた残りを意味する

変分ベイズ法再考

- EMの再考を思い出して比較してみる。

$$VB-Estep: \quad q(Z\mid M)^{new} = C\exp(E_{M,Z,\theta}[\log p(X,Z\mid\theta,M)])$$

$$VB-Mstep: \quad q(\theta_i\mid M)^{new} = C' p(\theta_i\mid M)\exp(E_{M^{new},Z,\theta\setminus\{\theta_i\}^{old}}[\log p(X,Z\mid\theta,M)])$$

変数^{new}を変数^{old}とする。

- $P(Z,X|\ \theta,M)$をθ^{old}を固定してZ, θ,Mで期待値をとることによって、Z,θ,Mに関する情報を教師データZから集めて再度推定することを繰り返しての良い推定値を求めている。
- ただし、因子化仮定によってθ_iを別々に更新している。だから解析的に更新式が求まる場合もあるわけだ。

変分ベイズ法の例題

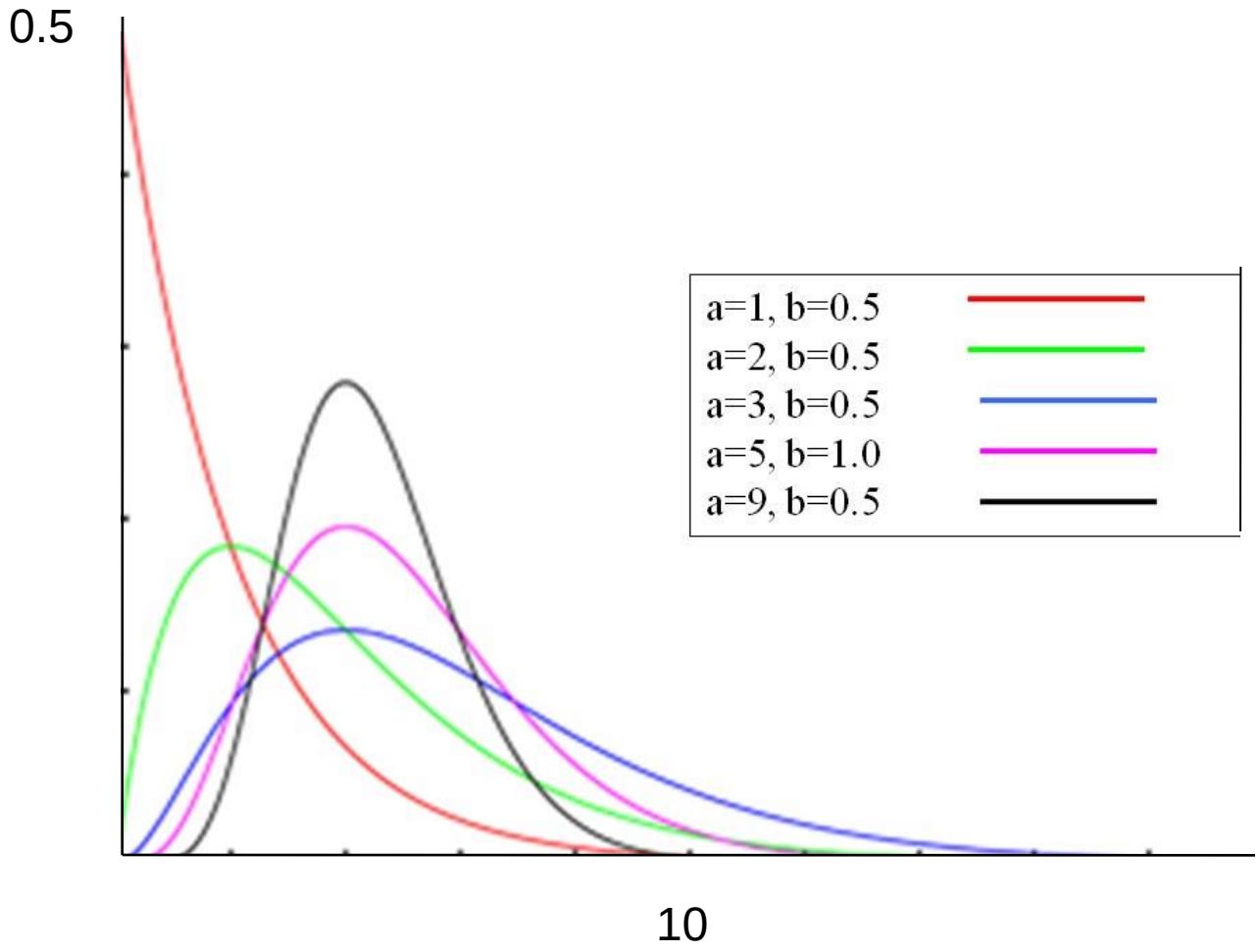
- 1変数正規分布$p(X|\mu,\tau)$のパラメターを観測データから推定。ただし、平均値μ（←これを潜在変数Zとみなす） 精度（＝分散の逆数）τ （←これをパラメターθとみなす）の事前分布は以下のように与えられているとする。
- $p(\tau|a,b)$を定義するガンマ関数は、パラメターa,bによっていろいろな分布形になるので、a,bを変化させて適した分布を得る目的でVBの事前分布として使うことが多い。

$$p(X|\mu,\tau)=\left(\frac{\tau}{2\pi}\right)^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}$$

$$p(\mu|\tau)=N(\mu|\mu_{0,}(\lambda_0\tau)^{-1})$$

$$p(\tau|a_0,b_0)=Gamma(\tau|a_0,b_0)=\frac{{b_0}^{a_0}\tau^{a_0-1}}{\Gamma(a_0)}\exp(-b_0\tau)$$

ガンマ分布



- factorizedな変分近似の事後分布を

  $q(\mu,\tau)=q_{\mu}(\mu)q_{\tau}(\tau)$とすると、以下のようにVB-Eステップ、VB-Mステップの計算ができる。左辺のqは更新した結果とする。

- VB-E：ここでは、内部変数μ,λを更新する。

$$q^{*}{}_{\mu}(\mu)=C_0\exp\left\{E_{\tau}[\log p(X\mid\mu,\tau)+\log p(\mu\mid\tau)]\right\}$$

$$=C_1\exp\left\{-\frac{E[\tau]}{2}\left(\sum_{i=1}^{N}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\right)\right\}$$

$E[\log(\tau/2\pi)^{N/2}]$などはμには関係しないので定数とみなす

$$=C_1\exp\left\{-\frac{E[\tau]}{2}(\lambda_0+N)\left(\mu^2-2\mu\frac{\lambda_0\mu_0+\sum_{i=1}^{N}x_i}{\lambda_0+N}+\frac{\lambda_0{\mu_0}^2+\sum_{i=1}^{N}{x_i}^2}{\lambda_0+N}\right)\right\}$$

よって

$$q_{\mu}(\mu)=N(\mu\mid\mu_N,\lambda_N^{-1})$$

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

- VB-Mステップ

$$q_\tau^*(\tau) = C_0 p(\tau) \exp\left\{E_\mu[\log p(X \mid \mu, (\lambda\tau)^{-1}) + \log p(\mu \mid \tau)]\right\}$$

$$= C_1 \exp\left\{\begin{array}{l}(a_0 - 1)\log\tau - b_0\tau + \frac{N+1}{2}\log\tau \\ -\frac{\tau}{2}E_\mu[\sum_{i=1}^{N}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2]\end{array}\right\}$$

- Gamma分布が指数分布族（事前分布と事後分布が同じタイプだから、以下のように推論できる。

この結果を$Gamma(\tau \mid a_N, b_N)$に対応させると以下の更新式が得られる

$$a_N = a_0 + \frac{N+1}{2}$$

$$b_N = b_0 + \frac{1}{2}E_\mu[\sum_{i=1}^{N}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2]$$

- こうして　$p(\tau)$を定義するGamma分布のパラメターが更新された。
- 以下、同様にVB-E,VB-Mを収束するまで繰り返すことになる。

Expectation Propagation

- 変分ベイズ法では$KL(q||p)$を最小化した。しかし、pに近いqを求めればよいのだから$KL(p||q)$を最小化する方法もありうる。
- これがExpectation Propagation:EP法。

EP法の背景

- KL(p||q)を最小化するqをpから求める。
- qはexponential family

$$q(\mathbf{z}) = h(\mathbf{z})\exp(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{z}) - a(\boldsymbol{\eta}))$$

$$KL(p \,\|\, q) = a(\boldsymbol{\eta}) - \boldsymbol{\eta}^{\mathrm{T}} E_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})] + const$$

$$then\ to\ \mathrm{minimize} \quad \mathrm{KL}(p \,\|\, q) \quad \Rightarrow \quad \frac{\partial a(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = E_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})]$$

- 一方、exponential familyであるqにおいて

qの積分が１だから　$\int h(\mathbf{z})\exp\left(\eta^{\mathrm{T}} u(\mathbf{z}) - a(\boldsymbol{\eta})\right)\mathrm{d}\mathbf{z} = 1$

より　$-\frac{\partial a(\eta)}{\partial \eta}\int h(\mathbf{x})\exp(\eta^T u(\mathbf{x}) - a(\eta))d\mathbf{x} + \int h(\mathbf{x})\exp(\eta^T u(\mathbf{x}) - a(\eta))u(\mathbf{x})d\mathbf{x} = 0$

ゆえに　$\frac{\partial a(\eta)}{\partial \eta} = E_{q(z)}[u(\mathbf{z})]$

あわせると　$E_{q(z)}[u(\mathbf{z})] = E_{p(z)}[u(\mathbf{z})]$　となりmomentが保存されている。

EP法

- 確率変数θをパラメターとするとき、第i番目のデータの出現確率が$f_i(\theta)$とする。
- このとき、観測データDの結合確率は $p(D,\theta)=\prod_i f_i(\theta)$

- そこで、狙いは事後分布$p(\theta \mid D)$を近似する分布qを以下のように求めること。

$$q(\theta)=\frac{1}{Z}\prod_i \tilde{f}_i(\theta)$$

EP法の処理手順

1. $\tilde{f}_i(\theta)$ を全て初期化。
2. 事後分布の近似を以下のように設定 $q(\theta) \propto \prod_i \tilde{f}_i(\theta)$
3. 収束するまで以下を繰り返す

   (a)改善すべき $\tilde{f}_j(\theta)$ を決める

   (b)これを次のようにして取り除く $q^{\setminus j}(\theta) = \frac{q(\theta)}{\tilde{f}_j(\theta)}$

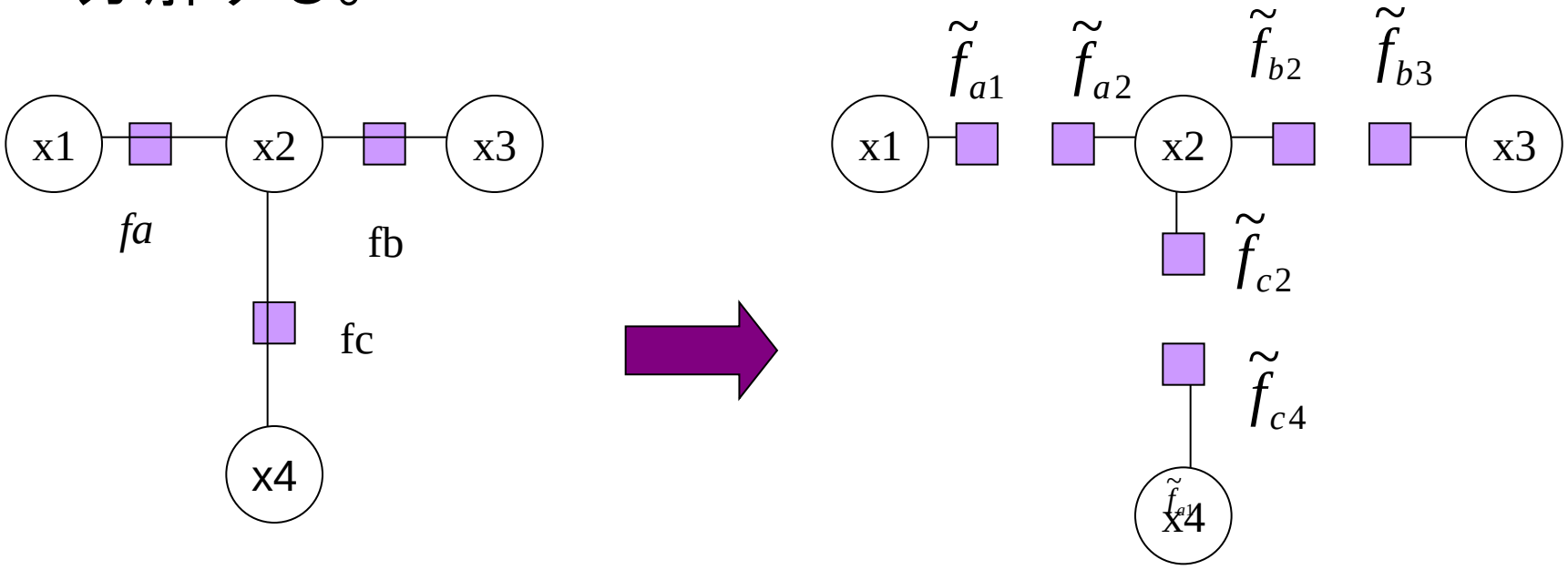
   (c)良い十分統計量（exモメント）が保存されるような新しい事後分布 $q^{new}(\theta) = q^{\setminus j}(\theta) f_j(\theta)$ を求め、正規化定数も次のように求める。 $Z_j = \int q^{\setminus j}(\theta) f_j(\theta) \mathrm{d}\theta$

   （この計算を解析的に行うので、けっこう面倒である。）

   (d)以下の更新を行う。 $\tilde{f}_j(\theta) = Z_j \frac{q^{new}(\theta)}{q^{\setminus j}(\theta)}$
4. 得られたモデルの近似度を評価する。 $p(D) \cong \int \prod_i \tilde{f}_i(\theta) \mathrm{d}\theta$

EP法の例:グラフィカルモデル

- 左側の構造のグラフィカルモデルを、右側のように分解する。



$$p(\mathbf{x}) = f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_3, x_4)$$

$$q(\mathbf{x}) \propto \tilde{f}_a(x_1, x_2) \tilde{f}_b(x_2, x_3) \tilde{f}_c(x_3, x_4)$$

ここで、前頁の右のグラフのように分解してみると

$$q(\mathbf{x}) \propto \tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{b2}(x_2) \tilde{f}_{b3}(x_3) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4)$$

ここでＥＰアルゴリズムにおいて$\tilde{f}_b(x_2, x_3) = \tilde{f}_{b2}(x_2) \tilde{f}_{b3}(x_3)$を選ぶ。

すると

$$q^{\setminus b}(\mathbf{x}) = \tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4)$$

そして、更新したf_bをかけて

$$\hat{p}(\mathbf{x}) = q^{\setminus b}(\mathbf{x}) f_b(x_2, x_3) = \tilde{f}_{a1}(x_1) \tilde{f}_{a2}(x_2) f_b(x_2, x_3) \tilde{f}_{c2}(x_2) \tilde{f}_{c4}(x_4)$$

ここで、$q^{new}(\mathbf{x})$は、$KL(\hat{p} \| q^{new})$を最小化するものとして得る。

$KL(\hat{p} \| q^{new})$を最小化するものは以下のようにして求める。

KLを最小化するには

$$KL(p\|q) = -\int p(\mathbf{Z})\left[\sum_{i=1}^{M}\log q_i(Z_i)\right]\mathrm{d}\mathbf{Z}$$

$+const$(じつは$\int p(\mathbf{Z})\left[\sum_{i=1}^{M}\log p_i(Z_i)\right]\mathrm{d}\mathbf{Z}$すなわち、$q$に無関係な$p$のエントロピー)

これをq_iについて最小化するのだが、*Lagrangue*未定乗数法により

$$h = KL(p\|q) + \lambda\left(\sum_{i=1}^{M} q_i - 1\right)$$

$$0 = \frac{\partial h}{\partial q_j} = -\int p(\mathbf{Z})\prod_{i\neq j}\mathrm{d}\mathbf{Z}_i\,\frac{1}{q_i(Z_i)} + \lambda$$

$$\int p(\mathbf{Z})\prod_{i\neq j}\mathrm{d}\mathbf{Z}_i = p(Z_j) = \lambda q_i(Z_i)$$

ここで$\sum p(Z_j) = 1 = \lambda\sum q_i(Z_i)$かつ$\sum q_i(Z_i) = 1$より$\lambda = 1$

ゆえに最適なqすなわち$q^*(Z_i) = \int p(\mathbf{Z})\prod_{i\neq j}\mathrm{d}\mathbf{Z}_i$　　周辺化したp

よってq^{new}は更新したpを周辺化すればよい

関連する周辺化された　p　は以下の通り

$$\hat{p}(x_1) \propto \tilde{f}_{a1}(x_1)$$

$$\hat{p}(x_2) \propto \tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2)\sum_{x_3} f_b(x_2, x_3)$$

$$\hat{p}(x_3) \propto \sum_{x_2}\left\{\tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2) f_b(x_2, x_3)\right\}$$

$$\hat{p}(x_4) \propto \tilde{f}_{c4}(x_4)$$

よって、q^{new}において更新された$\tilde{f}_{b2}, \tilde{f}_{b3}$は以下のとおり。

$$\tilde{f}_{b2}(x_2) \propto \sum_{x_3} f_b(x_2, x_3)$$

$$\tilde{f}_{b3}(x_3) \propto \sum_{x_2}\left\{\tilde{f}_{a2}(x_2)\tilde{f}_{c2}(x_2) f_b(x_2, x_3)\right\}$$

以下、同様にf_a, f_b, f_cについてこれらの操作を収束するまで繰り返す。