論　文

変分法的ベイズ推定による混合主成分分析

大羽　成征[†]　　石井　信[†,††]　　佐藤　雅昭[††,†††]

Variational Bayes Method for Mixture of Principal Component Analyzers

Shigeyuki OBA[†], Shin ISHII[†,††], and Masa-aki SATO[††,†††]

あらまし　高次元データを扱う際には、データの主要な特徴のみを抽出して適切な特徴空間を構成することが重要である。混合主成分分析はデータ空間のクラスタリングと主成分分析とを併用することによって特徴空間構成を行う特徴抽出器であり、その確率モデルと最尤推定アルゴリズムが Tipping と Bishop(1999) によって提案されている。しかし特徴抽出を確率モデルで表現するメリットは最尤推定ではなく、ベイズ推定を使用して初めて十分に享受できる。ベイズ推定には実行困難な積分計算が伴うが、近年提案された変分法的ベイズ推定 (variational Bayes:VB) 法によって、現実的に使用可能になってきた。本論文では混合主成分分析の変分法的ベイズ推定のアルゴリズムを導出する。また、手書き数字の認識問題に応用した結果についても述べる。

キーワード　変分法的ベイズ推定, 混合主成分分析, 特徴抽出, ベイズ推定, モデル選択

1. 導　　入

高次元のデータを扱うためには低次元の特徴空間へ変換する手続きが重要である。混合主成分分析はクラスタリングと主成分分析の組合せによって効果的な特徴空間を構成する手法であり、Tipping と Bishop [1] によって確率的生成モデルと、その最尤推定問題として定式化された。混合主成分分析には、モデル構造を決める定数として、ユニット (クラスタ) の個数と、各ユニットの有効主成分次元数との 2 種類があり、それらは階層的である。しかし、彼らが示したのはモデルパラメータの決定法であって、モデル構造の決定法については余り述べていない。

最尤推定法は観測データを最も良く説明するモデルパラメータの点推定を行う。一方ベイズ推定法は様々なモデル構造やモデルパラメータを持つモデルの集合 (アンサンブル) を考え、データを観測した後での各モデル構造やモデルパラメータの確からしさ (事後確率分布) を推定する。またモデル構造の良し悪しはモデル事後確率に基いて定量的に判定することができる。特に階層的なモデル構造を持つ混合主成分分析においてベイズ推定の必要性は高い。ところがベイズ推定には一般に困難な積分計算が伴い、混合主成分分析も例外では無い。パラメータの事後分布をガウス近似するラプラス近似法 [2] や、積分をモンテカルロ法で行うマルコフ連鎖モンテカルロ法 [3] などが、ベイズ推定の近似法として提案されてきたが、前者は近似精度、後者は計算時間の点で問題がある。

近年このような問題点を改善する手法として、変分法的ベイズ推定 (variational Bayes:VB) 法 [4]〜[9] が研究されてきた。VB 法では、事後分布を近似するための試験事後分布を用意し、事後分布推定の問題を試験事後分布に関する (変分法的) 自由エネルギー最大化の問題として定式化する。自由エネルギーを最大にする試験事後分布は真の事後分布に等しくなる。VB 法はさらに試験事後分布において確率変数間の因子化仮定を行う。とくに隠れ変数を持つ確率モデルに対しては、隠れ変数とモデルパラメータとが条件付き独立であるという独立因子化仮定を行なう。この仮定の下で自由エネルギー最大化問題を解くと最尤推定法における EM 法 [11] に類似した VB アルゴリズム [4]〜[8] が得られる。この VB アルゴリズムを用いることによりラプラス近似よりも高精度な近似が、マルコフ連鎖

[†] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Takayama 8916–5, Ikoma-shi, 630–0101, Japan
[††] 科学技術振興事業団 CREST
CREST, Japan Science and Technology Corporation
[†††] 国際電気通信基礎技術研究所
Advanced Telecommunication Research Institute International, Soraku-gun, Kyoto, Japan

モンテカルロ法よりも高速に計算できる。

この VB 法を、Bishop [7] は確率的主成分分析に対して、Ghahramani ら [6] は混合因子分析に対して適用し、適切なモデル構造の推定が行なえることを示した。しかしながらこれらの研究 [6], [7] では VB アルゴリズムを導くために、各モデルパラメータの間の条件付き独立性も仮定している。この仮定の下で得られる VB アルゴリズムは各モデルパラメータ事後分布に関して自由エネルギーを逐次的に最大化する多段ステップアルゴリズムになる。VB 法は解の局所最適性しか保証していない [4], [5] ため、各パラメータ事後分布に関する最大化を行う順序が異なれば得られる解が異なってくる可能性があり、結果がアルゴリズム依存になってしまうという問題点がある。本論文ではこの点を改善するために、混合主成分分析に対して各モデルパラメータ間の条件付き依存性を考慮した VB アルゴリズムを導く。すなわち隠れ変数とモデルパラメータ間の条件付き独立性のみを仮定し、各パラメータ間の条件付き相互依存性に関しては何の仮定もしない。この結果得られる VB アルゴリズムは、EM 法のように隠れ変数事後分布とパラメータ事後分布に関する最大化を交互に繰り返す 2 ステップアルゴリズム (階層的事前分布を用いる場合は 3 ステップアルゴリズム) になり、アルゴリズム的にも簡潔になる。また各パラメータ間の条件付き独立性を仮定する方法に比べて近似精度が向上することが期待できる。本研究では提案手法の基本的性能を調べるために、人工問題と実データを用いた手書き数字認識の問題に応用したのでその結果について述べる。

2. ベイズ推定と変分法的近似

2.1 ベイズ推定とアンサンブル学習

確率変数 $\mathbf{y}$ に対する確率モデルが確率分布 $p(\mathbf{y}|\theta)$ で定義されているものとする。θ はモデルパラメータである。データ $Y = \{\mathbf{y}(t)|t = 1, \ldots, T\}$ を観測した時、$p(Y|\theta) = \prod_{t=1}^{T} p(\mathbf{y}(t)|\theta)$ を θ の尤度関数と呼ぶ。

最尤推定は尤度関数の最大化により θ の値を一つ定める点推定であるが、ベイズ推定はパラメータ θ を事後分布 $p(\theta|Y)$ の形で求める。事後分布はパラメータ θ の値の集合 (アンサンブル) を考えた時に、データ Y の下で各 θ の値がどれくらい確からしいかを表わす。推定結果に基づき予測を行う際にも、以下のように事後分布を用いてアンサンブルに関する積分を行う。

$$\mathbf{y}_{\text{predict}} = \int \mathbf{y} p(\mathbf{y}|\theta) p(\theta|Y) d\theta \qquad (1)$$

パラメータアンサンブルによる予測は一般に点推定によるものよりも汎化性能が高くなることが知られている。

事後分布 $p(\theta|Y)$ はベイズの定理から以下で定義される。

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \qquad (2)$$

ここで右辺分子の $p(\theta)$ はパラメータの値に関する事前知識を反映する事前分布である。また右辺分母は、事後分布が確率分布になるための条件 $\int p(\theta|Y)d\theta = 1$ を満たすことを保証する正規化係数であり、

$$p(Y) = \int p(Y|\theta)p(\theta)d\theta \qquad (3)$$

によって定義される。

2.2 ベイズ推定とモデル構造

ガウス混合分布におけるクラスタ数など、それによって確率分布関数のパラメータ数や関数空間が変化するような変数をモデル構造変数と呼び、モデルパラメータと区別する。ここまでに現れた確率分布のそれぞれについて、$p(\mathbf{y}|\theta, M), p(Y|\theta, M), p(\theta|M), p(\theta|Y, M), p(Y|M)$ のようにモデル構造 M を明示的に書くと、モデル構造の事後分布 $p(M|Y)$ も、ベイズの定理を用いて次のように求めることができる。

$$p(M|Y) \propto p(Y|M)p(M) \qquad (4)$$

$$p(Y|M) = \int p(Y|\theta, M)p(\theta|M)d\theta \qquad (5)$$

モデル構造の確からしさについて特に事前知識が無ければ、全候補が等確率をとるようにするのが自然であり、$p(M)$ は定数となる。するとモデル構造の確からしさは式 (5) で決まる。この量はモデル周辺尤度もしくはエビデンスと呼ばれる。

パラメータ次元数が多く複雑なモデル構造であればあるほど確率分布形状の表現能力が高い。表現能力が高いモデル構造はパラメータの値次第でデータセットに沿うことができるが、そのせいでノイズに引きずられて現実とは合わないモデルを選択しやすくなる。最尤推定のように、尤度 $p(Y|\theta, M)$ を θ と M の両方について最大化させようとすると、際限無く複雑なモデルが選択されてしまう。これはニューラルネットにお

ける過学習と同様で、汎化性能の低下を招く。ベイズ推定を用いると、例えばモデル構造の事後確率最大化 (MAP) 推定を周辺尤度 $p(Y|M)$ 最大化基準に基いて行うことで、この困難を解消できる。

以下では簡潔性のために特に必要な場合を除き、モデル構造を明示しない。

2.3 変分法的近似

変分法的ベイズ推定[4]〜[9]では、モデル周辺尤度 (5) における積分を直接実行する代りに変分法を用いてこれを求める。特に隠れ変数 $\mathbf{x}$ を持つ確率モデル $p(\mathbf{y}|\theta) = \int d\mathbf{x} p(\mathbf{y}, \mathbf{x}|\theta)$ に対して有効である。観測データ Y に対応する隠れ変数を $X = \{\mathbf{x}(t)|t = 1, \ldots, T\}$ とすると、隠れ変数 X とパラメータ θ の同時事後確率分布は次式で与えられる。

$$p(X, \theta|Y) = \frac{P(Y, X|\theta)p(\theta)}{p(Y)} \tag{6}$$

これを用いて対数周辺尤度に対する以下の関係式が導ける。

$$\begin{aligned}
\ln p(Y) &= \int dX d\theta Q(X, \theta) \ln p(Y) \\
&= \int dX d\theta Q(X, \theta) \ln \frac{p(Y, X|\theta)p(\theta)}{p(X, \theta|Y)} \\
&= \int dX d\theta Q(X, \theta) \Big[\ln \frac{p(Y, X|\theta)p(\theta)}{Q(X, \theta)} \\
&\qquad + \ln \frac{Q(X, \theta)}{p(X, \theta|Y)}\Big] \\
&= \mathcal{F}[Q|Y] + \mathrm{KL}\left[Q \,||\, p(X, \theta|Y)\right] \\
&\geq \mathcal{F}[Q|Y]
\end{aligned} \tag{7}$$

ここで試験事後分布 $Q(X, \theta)$ は事後分布を近似するために導入されたものであり、確率分布条件 $\int dX d\theta Q(X, \theta) = 1$ を満たす任意の確率分布である。$\mathrm{KL}\left[Q \,||\, p\right]$ は Kullback-Leibler(KL) 距離と呼ばれる非負値の擬距離関数である。任意の試験事後分布 $Q(X, \theta)$ について上の不等式が成り立ち、$p(X, \theta|Y)$ と $Q(X, \theta)$ が確率分布として一致するときに限り等号が成立する。$\mathcal{F}[Q|Y]$ は (変分法的) 自由エネルギーと呼ばれ、次式で定義される。

$$\mathcal{F}[Q|Y] = \int dX d\theta Q(X, \theta) \ln \frac{p(Y, X|\theta)p(\theta)}{Q(X, \theta)} \tag{8}$$

KL 距離は非負値であるから、自由エネルギーを試験事後分布に関して最大化すると、試験事後分布は真の事後分布に等しくなり、同時に自由エネルギーは対数周辺尤度 $\ln p(Y)$ に一致する。

変分法的近似では、試験事後分布に対して確率変数間に因子化仮定 (因子化近似) を導入する。因子化仮定

$$Q(X, \theta) = Q_X(X) Q_\theta(\theta)$$

により、計算困難な周辺化積分が

$$\begin{aligned}
\langle f(X) g(\theta) \rangle_{X,\theta} &\stackrel{\mathrm{def}}{=} \int dX d\theta Q(X, \theta) f(X) g(\theta) \\
&= \int dX Q_X(X) f(X) \int d\theta Q_\theta(\theta) g(\theta) \\
&= \langle f(X) \rangle_X \langle g(\theta) \rangle_\theta
\end{aligned}$$

のように分離されてそれぞれ計算可能な項の積になる。これにより自由エネルギーは

$$\begin{aligned}
\mathcal{F}[Q|Y] = &\langle \ln p(Y, X|\theta) \rangle_{X,\theta} + \langle \ln p(\theta) \rangle_\theta \\
&- \langle \ln Q_X(X) \rangle_X - \langle \ln Q_\theta(\theta) \rangle_\theta
\end{aligned} \tag{9}$$

のように書ける。

2.4 自由エネルギーの最大化アルゴリズム

(9) 式の自由エネルギーを $Q(X, \theta)$ に関して最大化するために、$Q_X(X)$ に関する最大化と $Q_\theta(\theta)$ に関する最大化を交互に繰り返す逐次的アルゴリズムを用いる。

$Q_X(X)$ に関する最大化は、$Q_\theta(\theta)$ を固定しつつ、$\frac{\delta \mathcal{F}}{\delta Q_X(X)} = 0$ より、

$$\ln Q_X(X) = \langle \ln p(Y, X|\theta, M) \rangle_\theta + \mathrm{const.} \tag{10}$$

のように得られる。また $Q_\theta(\theta)$ に関する最大化は、$Q_X(X)$ を固定しつつ、$\frac{\delta \mathcal{F}}{\delta Q_\theta(\theta)} = 0$ より、

$$\ln Q_\theta(\theta) = \langle \ln p(Y, X|\theta, M) \rangle_X + \ln p(\theta) + \mathrm{const.} \tag{11}$$

のように得られる。ここで const. は変数に関わらない定数であり、確率分布条件 $\int dX Q_X(X) = \int d\theta Q_\theta(\theta) = 1$ により決められる。

2.5 事前分布

以下で考える混合主成分分析は隠れ変数を持つ指数型分布族

$$p(\mathbf{y}, \mathbf{x}|\theta) = \exp\big[\boldsymbol{\phi}(\theta) \cdot \boldsymbol{r}(\mathbf{y}, \mathbf{x}) - \Psi(\theta)\big] \tag{12}$$

の特別な場合になっている[5]。ここで $\boldsymbol{\phi}(\theta)$ は自然パラメータ、$\boldsymbol{r}(\mathbf{y}, \mathbf{x})$ は十分統計量、$\Psi(\theta)$ は正規化項である。この時、隠れ変数を含めた完全データ尤度は、

以下のように表現される。

$$p(Y,X|\theta) = \exp\big[\boldsymbol{\phi}(\theta)\cdot\sum_{t=1}^{T}\boldsymbol{r}(\mathbf{y}(t),\mathbf{x}(t)) - T\Psi(\theta)\big] \tag{13}$$

この完全データ尤度をパラメータ θ の確率分布として見た場合、次式で与えられる共役分布の形をしている。

$$p(\theta) = \exp\big[\gamma\{\boldsymbol{\phi}(\theta)\cdot\boldsymbol{h} - \Psi(\theta)\} - \Phi(\boldsymbol{h},\gamma)\big] \tag{14}$$

ここで、$\boldsymbol{h},\gamma$ は共役分布のハイパーパラメータ、$\Phi(\boldsymbol{h},\gamma)$ は正規化項である。事前分布として共役分布を用いれば事後分布もまた同様の共役形となり、計算が簡単になる。

ハイパーパラメータ $\boldsymbol{h}$ を定数ではなく確率変数として扱うとき、パラメータ θ の事前分布 $p(\theta|\boldsymbol{h})p(\boldsymbol{h})$ を階層的事前分布と呼ぶ。この場合、対数周辺尤度を求める際にこのハイパーパラメータ $\boldsymbol{h}$ に関しても積分 (周辺化) を行なう。Neal [3] と MacKay [2] は多層パーセプトロンの枠組みの中で結合重みに階層的事前分布を与えることで、中間層ニューロン個数を抑圧した。これは ARD (automatic relevance determination) と呼ばれる。Bishop [10] はこれを確率的主成分分析に適用し、自動的に有効な主成分次元を決定した。

3. 混合主成分分析の確率モデル

3.1 確率的主成分分析

確率的主成分分析 (PPCA) [1] は、隠れ変数を含んだ確率モデルを用いて主成分分析を再定式化したものである。

d 次元観測データベクトル $\mathbf{y}$ が、q 次元隠れ変数ベクトル $\mathbf{x}$ の線形変換とノイズから次式のように生成されるものとする。

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{15}$$

ここで、$\mathbf{W}$ は $(d\times q)$ 行列、$\boldsymbol{\mu}$ は d 次元ベクトルであり、合せて線形変換を定義している。また隠れ変数ベクトル $\mathbf{x}$ は平均 $\mathbf{0}$、分散 1 の q 次元正規分布 $\mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{I}_q)$ に従って生起し、d 次元ノイズベクトル $\boldsymbol{\epsilon}$ は分散 $1/\tau$ の等分散正規分布 $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},(1/\tau)\mathbf{I}_d)$ に従って乗るものとする。またここで $\mathbf{I}_q,\mathbf{I}_d$ はそれぞれ q 次元、d 次元の単位行列を表す。ここで $\mathcal{N}(\mathbf{x}|\mathbf{m},\boldsymbol{\Sigma})$ は確率変数ベクトル $\mathbf{x}$ に対する平均 $\mathbf{m}$、共分散 $\boldsymbol{\Sigma}$ を持つ多次元正規分布の確率密度関数である。

以上から、隠れ変数 $\mathbf{x}$ と観測ベクトル $\mathbf{y}$ の同時確率分布は以下のように表せる。

$$\begin{aligned} p(\mathbf{y},\mathbf{x}|\theta) &= \exp\big[-\frac{1}{2}\tau||\mathbf{y}-\mathbf{W}\mathbf{x}-\boldsymbol{\mu}||^2 - \frac{1}{2}||\mathbf{x}||^2 \\ &\quad + \frac{d}{2}\ln\tau - \frac{d+q}{2}\ln 2\pi\big] \\ &= \mathcal{N}(\mathbf{x}\mid \tau\boldsymbol{R}_{\mathbf{x}}{}^{-1}\mathbf{W}'(\mathbf{y}-\boldsymbol{\mu}), \boldsymbol{R}_{\mathbf{x}}{}^{-1}) \\ &\quad \times\mathcal{N}(\mathbf{y}\mid \boldsymbol{\mu},\ \frac{1}{\tau}\boldsymbol{R}_{\mathbf{y}}) \end{aligned} \tag{16}$$

ここで以下の表記を使用した。

$$\tau \stackrel{\text{def}}{=} 1/\sigma^2, \quad \boldsymbol{R}_{\mathbf{x}} \stackrel{\text{def}}{=} \tau\mathbf{W}'\mathbf{W} + \mathbf{I}_q, \quad \boldsymbol{R}_{\mathbf{y}} \stackrel{\text{def}}{=} \tau\mathbf{W}\mathbf{W}' + \mathbf{I}_d$$

また、パラメータは $\theta = (\mathbf{W},\boldsymbol{\mu},\tau)$ である。

するとデータ Y のもとでパラメータ θ について対数尤度が得られる。

$$\begin{aligned} \ln p(Y|\theta) &= \sum_{t=1}^{T}\ln\int p(\mathbf{y}(t),\mathbf{x}(t)|\theta)d\mathbf{x}(t) \\ &= -\frac{T}{2}\tau\,\mathrm{tr}\boldsymbol{R}_{\mathbf{y}}{}^{-1}\mathbf{C} \\ &\quad -\frac{T}{2}(-d\ln(\tau/2\pi) + \ln|\boldsymbol{R}_{\mathbf{y}}|) \end{aligned} \tag{17}$$

ここで $\mathbf{C}$ はデータ Y の共分散行列である。$\mathrm{tr}(\cdot)$ は対角和である。式 (17) の対数尤度に基く最尤推定によって得られる行列 $\mathbf{W}$ は通常の主成分分析における主成分行列と同じものになり、主方向と直交する分散成分は、分散 $1/\tau$ として近似される [1]。

3.2 混合主成分分析の確率モデル

データ $\mathbf{y}$ に対して m 個の確率的データ生成モデル $M \equiv \{M_i|i=1,\ldots,m\}$ があるとき、それらの混合モデルは以下のように与えられる。

$$p(\mathbf{y}|\mathbf{g},M) = \sum_{i=1}^{m} p(\mathbf{y}|M_i)g_i \tag{18}$$

ここで $\mathbf{g} = (g_1,\cdots,g_m)$ は混合比であり、$\sum_{i=1}^{m} g_i = 1$ を満たす。

便宜上データ $\mathbf{y}$ がどのモデル M_i から生起したかを示す確率変数 $\mathbf{z}$ を導入する。$\mathbf{z} \equiv (z_1,\cdots,z_m)$ の各成分 z_i は 0 もしくは 1 の値をとり、$\sum_{i=1}^{m} z_i = 1$ を満たす。$z_i = 1$ はデータ $\mathbf{y}$ が確率モデル M_i から生起したことを意味する。すると、$\mathbf{y}$ と $\mathbf{z}$ の同時確率分布は次のように書ける。

$$p(\mathbf{y},\mathbf{z}|\mathbf{g},M) = \prod_{i=1}^{m} p(\mathbf{y}|M_i)^{z_i} g_i^{z_i} \tag{19}$$

この特別な場合として、確率的主成分分析の混合モデルにおける完全データ $(\mathbf{y}, \mathbf{x}, \mathbf{z})$ の同時確率分布は以下のように定義される。

$$p(\mathbf{y}, \mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(\mathbf{y}, \mathbf{x}_i|\theta_i)^{z_i} g_i^{z_i}$$
$$= \exp[\sum_{i=1}^{m} z_i(\ln p(\mathbf{y}, \mathbf{x}_i|\theta_i) + \ln g_i)] \qquad (20)$$

ここで $p(\mathbf{y}, \mathbf{x}_i|\theta_i)$ は式 (16) で与えられる。$\boldsymbol{\theta} = \{\mathbf{g}, \theta_i | i = 1, \ldots, m\}$、$\theta_i = \{\mathbf{W}_i, \boldsymbol{\mu}_i, \tau_i\}$ である。m 個の確率モデルを以下ではユニットと呼ぶ。主成分を表す隠れ変数ベクトル $\mathbf{x}$ は各ユニットごとに独立に定義されており、$\mathbf{x} = \{\mathbf{x}_i | i = 1, \ldots, m\}$ と表現される。各ユニットごとに主成分ベクトル $\mathbf{x}_i$ の次元 q_i が異っていることが許されるから、生成モデルの構造にはユニット数 m と、各ユニットの主成分次元 q_i との二種類の自由度があり、階層化されていることに注意する。

3.3 混合主成分分析と正規混合分布の関係

確率的主成分分析の確率モデルは d 次元多変量正規分布 $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ における共分散行列 $\boldsymbol{\Sigma}$ について $\mathbf{W}\mathbf{W}' + \tau^{-1}\mathbf{I}_d$ という形の制約を入れたものに対応する。したがって確率的主成分分析と正規分布とは良く似た表現能力を持つが、パラメータの自由度が異なる。

多変量正規分布では、共分散行列 $\boldsymbol{\Sigma}$ は $d(d+1)/2$ 次元の自由パラメータを持ち、最も一般的な形をしている。この他にも共分散行列として、$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_d$ のようにスカラー分散 1 つだけの自由度を持つもの、$\boldsymbol{\Sigma}$ を対角行列に制限して d 個の自由度で表わすものなどが考えられる。これに対して確率的主成分分析の共分散行列は、$d \times q$ 主軸行列 $\mathbf{W}$ とスカラー逆分散 τ と併せて $dq + 1$ 個の自由度を持っている。また因子分析は $d \times q$ 因子行列と対角分散成分を併せて $d \times (q+1)$ 個の自由度を持っている。$q \ll d$ のとき、確率的主成分分析は、一般共分散とスカラー分散や対角共分散との中間に位置することになる。

Moerland [12] は、正規混合分布における共分散行列表現の様々なもの (球状 (スカラー分散)、対角行列、完全行列) と、混合主成分分析、混合因子分析の最尤推定をさまざまな応用問題において比較し、多くの場合に混合主成分分析や混合因子分析が有利であることを示した。

4. 変分法的ベイズ推定による混合主成分分析

この節では、混合主成分分析の変分法的ベイズ推定アルゴリズムを導出する。

4.1 事前分布

混合主成分分析のパラメータ $\boldsymbol{\theta} = \{(\mathbf{g}, \tau_i, \mathbf{W}_i, \boldsymbol{\mu}_i) | i = 1, \ldots, m\}$ の事前分布として、2.5 節で述べた一般論に基いて共役分布を与える。また、とくに主軸行列 $\mathbf{W}_i$ の事前分布のハイパーパラメータ $\boldsymbol{\alpha}$ に関して階層的事前分布を与える。

混合比 $\mathbf{g}$ に関して以下の Dirichlet 分布を事前分布とする。

$$\ln p(\mathbf{g}) = \sum_{i=1}^{m} (\gamma_{0i} \ln g_i - \ln\Gamma(\gamma_{0i} + 1)) + \ln\Gamma\left(\sum_{i=1}^{m} \gamma_{0i} + m\right) \qquad (21)$$

ただし、γ_{0i} はハイパーパラメータ、Γ はガンマ関数 $\Gamma(\gamma) = \int_0^\infty dt e^{-t} t^{\gamma-1}$

以下では簡潔性のため添字 i を省略して書く。

ユニット中心 $\boldsymbol{\mu}$ に関して以下の正規分布を事前分布とする。

$$\ln p(\boldsymbol{\mu}|\tau) = -\frac{1}{2}\gamma_{\mu_0}\tau(\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}_0)^2 + \frac{d}{2}\ln(\gamma_{\mu_0}\tau/2\pi) \qquad (22)$$

ただし、$\gamma_{\mu_0}, \overline{\boldsymbol{\mu}}_0$ はハイパーパラメータ。主軸行列 $\mathbf{W}$ に関しても以下の正規分布を事前分布とする。

$$\ln p(\mathbf{W}|\tau, \boldsymbol{\alpha}) = -\frac{1}{2}\tau \mathrm{tr}\mathbf{W}'\mathbf{W}\boldsymbol{A} + \frac{d}{2}\ln|\boldsymbol{A}| + \frac{dq}{2}\ln(\tau/2\pi) \qquad (23)$$

ただし $\boldsymbol{A}$ はその対角成分が $\boldsymbol{\alpha}$ であるような $(q \times q)$ 対角行列。

等方逆分散 τ は以下のガンマ分布を事前分布とする。

$$\ln p(\tau) = -\gamma_{\tau_0}\bar{\tau}_0^{-1}\tau + (\gamma_{\tau_0} - 1)\ln\tau - \ln\Gamma(\gamma_{\tau_0}) + \gamma_{\tau_0}\ln(\gamma_{\tau_0}\bar{\tau}_0^{-1}) \qquad (24)$$

ただし、$\gamma_{\tau_0}, \bar{\tau}_0$ はハイパーパラメータ。

ハイパーパラメータ $\boldsymbol{\alpha}$ の事前分布も以下のガンマ分布を事前分布とする。

$$\ln p(\boldsymbol{\alpha}) = \sum_{j=1}^{q}[-\gamma_{\alpha\,0}\bar{\alpha}_{0j}^{-1}\alpha_j + (\gamma_{\alpha\,0}-1)\ln\alpha_j - \ln\Gamma(\gamma_{\alpha\,0}) + \gamma_{\alpha\,0}\ln(\gamma_{\alpha\,0}\bar{\alpha}_{0j}^{-1})] \quad (25)$$

ただし、$\gamma_{\alpha\,0}, \bar{\alpha}_{0j}$ は分布 $p(\boldsymbol{\alpha})$ のハイパーパラメータ。

4.2 アルゴリズム

混合主成分分析は隠れ変数 $(X, Z) = \{\mathbf{x}_i(t), z_i(t) | t = 1, \ldots, T\ , i = 1, \ldots, m\ \}$ とパラメータ $\boldsymbol{\theta} = \{\mathbf{g}, \mathbf{W}_i, \boldsymbol{\mu}_i, \tau_i | i = 1, \ldots, m\ \}$、及びハイパーパラメータ $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_i | i = 1, \ldots, m\ \}$ を持つ。これらの未知量に対する試験事後分布に対して以下のような因子化仮定を行う。

$$Q(X, Z, \boldsymbol{\theta}, \boldsymbol{\alpha}) \ = \ Q(X, Z)Q(\boldsymbol{\theta})Q(\boldsymbol{\alpha}) \quad (26)$$

この因子化仮定は文献 [6], [7] で使われた各パラメータごとの因子化仮定 $Q(\boldsymbol{\theta}) = Q(\mathbf{g})Q(\mathbf{W})Q(\boldsymbol{\mu})Q(\tau)$ よりも弱く、近似精度の向上が期待でき、アルゴリズムも簡潔になる。この因子化仮定の下で自由エネルギーは、以下のように書ける。

$$\begin{aligned}\mathcal{F}[Q, Y] = &\langle \ln p(Y, X, Z|\boldsymbol{\theta}, \boldsymbol{\alpha})\rangle_{X,Z,\boldsymbol{\theta},\boldsymbol{\alpha}} \\ &- \langle \ln Q(X, Z)\rangle_{X,Z} - \langle \ln Q(\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}} - \langle \ln Q(\boldsymbol{\alpha})\rangle_{\boldsymbol{\alpha}} \\ &+ \langle \ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})\rangle_{\boldsymbol{\theta},\boldsymbol{\alpha}} + \langle \ln p(\boldsymbol{\alpha})\rangle_{\boldsymbol{\alpha}} \quad (27)\end{aligned}$$

変分法的ベイズ推定のアルゴリズムは、試験分布の各因子に関して他を固定しながら、式 (27) の自由エネルギーを最大化することによって以下のように得られる。

- 他を固定しつつ、$Q(X, Z)$ を更新

$$\begin{aligned}\ln Q(X, Z) &= \langle \ln p(Y, X, Z|\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}} + \text{const.} \\ &= \sum_{t=1}^{T}\langle \ln p(\mathbf{y}(t), \mathbf{x}(t), \mathbf{z}(t)|\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}} + \text{const.} \quad (28)\end{aligned}$$

より、

$$\begin{aligned}Q(X, Z) &= \prod_{t=1}^{T} Q(\mathbf{x}(t), \mathbf{z}(t)|\mathbf{y}(t)) \\ \ln Q(\mathbf{x}(t), \mathbf{z}(t)|\mathbf{y}(t)) &= \langle \ln p(\mathbf{y}(t), \mathbf{x}(t), \mathbf{z}(t)|\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}} \\ &\quad + \text{const.} \quad (29)\end{aligned}$$

- 他を固定しつつ、$Q(\boldsymbol{\theta})$ を更新

$$\ln Q(\boldsymbol{\theta}) = \langle \ln p(Y, X, Z|\boldsymbol{\theta})\rangle_{X,Z} + \langle \ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})\rangle_{\boldsymbol{\alpha}} + \text{const.} \quad (30)$$

- 他を固定しつつ、$Q(\boldsymbol{\alpha})$ を更新

$$\ln Q(\boldsymbol{\alpha}) = \langle \ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})\rangle_{\boldsymbol{\theta}} + \ln p(\boldsymbol{\alpha}) + \text{const.} \quad (31)$$

- 以上を収束するまで繰り返す

また、各 const. は確率分布条件により定まる。具体的なアルゴリズムは付録 1. を参照されたい。

4.3 事前分布のハイパーパラメータ

各パラメータの値に関する事前知識 (先験的偏見) が存在しないことを表現するためには、無情報事前分布を与えるのが理想的である。しかしパラメータ空間が無限大の体積を持つ連続パラメータに対する無情報事前分布は、積分が発散するため正規化できないという問題点がある。一方で、2.5 節で述べたように指数型分布族に対する変分法的ベイズ推定の定式化においては共役形の事前分布を与えることで計算が簡単になる。事前分布のハイパーパラメータを適切に設定することによって、共役事前分布は近似的に無情報事前分布と見做すことができる。上に示した $\boldsymbol{\mu}, \tau, \mathbf{g}$ の事前分布に含まれるハイパーパラメータ $\gamma_0, \gamma_{\boldsymbol{\mu}\,0}, \gamma_{\tau\,0}$ は事前分布に含まれる情報の多さ (実効的なデータ数) を表現している。そこでこれらを小さな値に調整することで、事前分布の影響力を弱めることができる。

4.4 自由エネルギー計算

自由エネルギーの値は、対数尤度項とモデル複雑度項により $\mathcal{F} = \sum_{t=1}^{T} L(\mathbf{y}(t)) - H$ のように表せる。モデル複雑度項 H は以下で定義される。

$$\begin{aligned}H = &\langle \ln Q(\theta)\rangle_{\theta} - \langle \ln p(\theta|\boldsymbol{\alpha})\rangle_{\theta,\boldsymbol{\alpha}} \\ &+ \langle \ln Q(\boldsymbol{\alpha})\rangle_{\boldsymbol{\alpha}} - \langle \ln p(\boldsymbol{\alpha})\rangle_{\boldsymbol{\alpha}}\end{aligned}$$

これはパラメータ事後分布のエントロピー項 (あるいは事前と事後の KL 距離) である。この項はモデルが複雑な程大きくなるので、複雑なモデルに対するペナルティー項として働く。対数尤度が寄与する分は $Q(X, Z)$ を更新した直後には以下のように表せる。

$$\begin{aligned}&\langle \ln p(Y, X, Z|\boldsymbol{\theta})\rangle_{X,Z,\boldsymbol{\theta}} - \langle \ln Q(X, Z)\rangle_{X,Z} \\ &= \sum_{t=1}^{T}\int d\mathbf{x}d\mathbf{z}Q(\mathbf{x}, \mathbf{z}|\mathbf{y}(t))\big[\langle \ln p(\mathbf{y}(t), \mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}} \\ &\quad - \ln Q(\mathbf{x}, \mathbf{z}|\mathbf{y}(t))\big] \\ &= \sum_{t=1}^{T}\int d\mathbf{x}d\mathbf{z}Q(\mathbf{x}, \mathbf{z}|\mathbf{y}(t))\big[\ln\sum_{i=1}^{m} U_i(\mathbf{y}(t))\big] \\ &= \sum_{t=1}^{T}\ln\left[\sum_{i=1}^{m} U_i(\mathbf{y}(t))\right] = \sum_{t=1}^{T} L(\mathbf{y}(t))\end{aligned}$$

これを、対数尤度項と呼び、データと試験事後分布の合致度を表す。ただし $U_i(\mathbf{y}(t))$ の定義は付録 1. を参照のこと。

5. モデル選択とユニット操作

ベイズ的モデル推定において、MAP 推定によるモデル選択とは周辺尤度 $p(Y|M)$ を最大にするようなモデル構造 M を求めることである。混合主成分分析において、モデル構造を決定するのはユニット (クラスタ) の個数 m 及び、各ユニット i に関する主成分の次元 q_i である。このうち、各ユニットの主成分次元 q_i は主軸行列 $\mathbf{W}$ に階層型事前分布を適用することによって、ほぼ自動的に求めることもできる。一方で、ユニットの個数を適切に決定するのには工夫が必要である。

直接選択法

ユニットの個数 m について事前にいくつか用意して、それぞれに対して試験事後分布を収束させる試行を行い、自由エネルギーの収束値を比較する。自由エネルギーを最大にする m の値を最適値として決定する。

しかしこれには計算量的な無駄が多く、またオンライン的状況に対応できない。そこで変分法的ベイズ推定アルゴリズムの途中でユニット操作を行うことでモデル選択の効率化を図る。

ユニット操作

クラスタリングの確率モデル構造を扱うユニット操作としては、これまでに、削除、挿入、分離、結合などが提案されている [5], [6], [8]。本研究では特に高次元データに対して有効な以下の削除と挿入操作を用いる。

（1） 各ユニットが担当するデータの個数 $TE[z_i]$ が 1 を下回るユニットを削除する。($E[z_i]$ の定義は付録 1. 参照。)

（2） 既存のユニットがうまく説明できていないデータ、すなわち $L(\mathbf{y}(t))$ の値が小さいデータを一定個数抽出して、部分データを用意する。次に部分データを 2-3 個のユニット数で混合主成分分析にかける。またこの部分データを取り去った残りのデータによって、既存ユニットを訓練する。その後で新規ユニットを既存ユニットに追加する。

ユニット操作はそれを行った後にパラメータを収束させ、自由エネルギー値の改善が見られたときにのみ受理されるものとした。

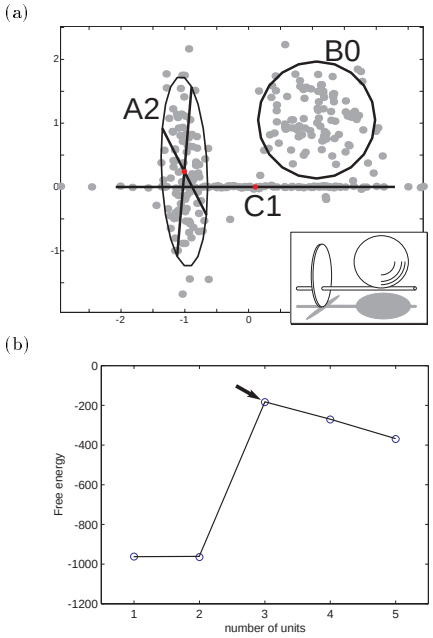


図 1 3 次元データの学習結果。(a) 使用した 3 次元データの 2 次元プロファイルと、学習結果のクラスタプロット。(b) ユニット数ごとの自由エネルギーの収束値。

6. 実　　験

6.1 人工データを用いた実験

まず混合主成分分析の能力を示すための簡単な実験を示す。図 1(a) 中の点は実験に用いた人工データ点である。3 次元空間内の 200 個のデータ点は、3 つのクラスタから成り、各クラスタは図の右下の模式図に示すように、それぞれ球状、円板状、葉巻状の分布を成している。これに混合主成分分析を適用した。混合数を 1 から 5 までのそれぞれに対して 1000 エポックの繰り返し計算による自由エネルギーの収束値を調べた結果を図 1(b) に示す。直接選択法により、 混合数 3 が最大事後確率を与える混合数となる。図 1(a) に、求められた 3 つのクラスタと各クラスタの有効次元を示す。アルファベットの横の数字は各クラスタの有効次元、太線は各クラスタの主軸を表す。有効次元を 0 と推定することは、特定の一方向を主方向として取り出せなかったことを意味し、球状 (等方等分散) で分布するデータに対して、正しい結果である。

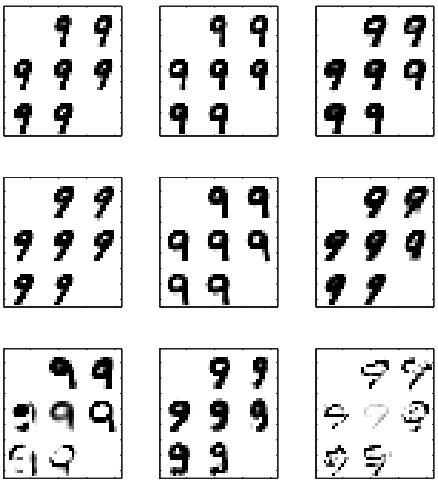
6.2 手書き数字画像認識の実験

混合主成分分析を手書き数字画像の認識に応用した。実験に使用したデータは、AT& T 社 MNIST デー

タベースの手書き数字画像(注1) である。学習データは 28×28 ピクセル 256 階調グレースケールの数字画像が '0' から '9' までの各数字について約 6000 個、合計で 60000 個用意され、試験データは同様のものが合計 10000 個用意されている。各データはそれがどの数字を表すものであるかを示すラベルがつけられている。実験では各数字画像を 14×14 ピクセルの 196 次元ベクトルデータにサンプリングしなおしたものについて、60000 個の学習データで学習後、10000 個の試験データで性能を調べた。

学習において、'0' から '9' までの各数字画像それぞれに対して、混合主成分分析モデル $M_l, (l = 0, \ldots, 9)$ を求めた。またこの際、削除と挿入のユニット操作を用いた。最大主成分次元 q は $q = 1, 5, 10$ の三種類を用意して比較した。さらに $q = 5, 10$ のものについては ARD の階層型事前分布を使うもの、使わないものとを比較した。

(a)



(b)

	$\boldsymbol{\mu}_i + \mathbf{w}_{i,1}$	$\boldsymbol{\mu}_i + \mathbf{w}_{i,3}$
$\boldsymbol{\mu}_i - \mathbf{w}_{i,2}$	$\boldsymbol{\mu}_i$	$\boldsymbol{\mu}_i + \mathbf{w}_{i,2}$
$\boldsymbol{\mu}_i - \mathbf{w}_{i,3}$	$\boldsymbol{\mu}_i - \mathbf{w}_{i,1}$	

図 2　数字画像に対する混合主成分分析結果の例

混合主成分分析による特徴抽出の例を図 2 に示す。数字 '9' を学習したモデル中の 9 つのユニットを 2(a) に示す。各ユニットを、3×3 の領域に分けて表示しており、各小領域の意味を図 2(b) の凡例で説明している。すなわち、第 i ユニット中心 $\boldsymbol{\mu}_i$ ベクトル及び、そこから第 j 主成分ベクトル $\mathbf{w}_{i,j}$ 分のずれを加えたものを、主成分次元 3 まで表示している。

学習によって得られた確率モデルを用いて、各試験データ $\mathbf{y}_{\text{test}}$ がどの数字画像であるかを認識するには、各モデルに基づく尤度 $p(\mathbf{y}_{\text{test}}|M_l)$ を計算してやり、これを最大にする $l = \arg\max_l p(\mathbf{y}_{\text{test}}|M_l)$ を認識結果として出力すれば良い。10000 点の試験データに対して混合主成分分析による認識結果と正解とを比べて誤答率を求め、性能評価値とした。MNIST データベースで評価を行った、他の文字認識アルゴリズムと本アルゴリズム(注2) とを比較したのが表 1 である。本手法 (VBMPCA) において、ARD とあるのが ARD を用いたもので、ないのは用いていないものである。比較的パラメータ数の多い階層型ニューラルネット (NN) や、RBF などよりも良い結果が得られていることが分かる。

METHOD	test error (%)
VBMPCA $q = 1$ 14x14	6.05
VBMPCA $q = 5$ 14x14	3.42
VBMPCA $q = 5$ ARD 14x14	2.92
VBMPCA $q = 10$ 14x14	2.99
VBMPCA $q = 10$ ARD 14x14	3.14
linear classifier (1-layer NN)	12.0
pairwise linear classifier	7.6
K-nearest-neighbors, Euclidean	5.0
40 PCA + quadratic classifier	3.3
1000 RBF + linear classifier	3.6
K-NN, Tangent Distance, 16x16	1.1
SVM deg 4 polynomial	1.1
Reduced Set SVM deg 5 polynomial	1.0
Virtual SVM deg 9 poly [distortions]	0.8
2-layer NN, 1000 hidden units	4.5
3-layer NN, 500+150 hidden units	2.95
Boosted LeNet-4, [distortions]	0.7

表 1　MNIST 数字認識課題の性能比較

q	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	mean
1	42	35	30	36	26	22	46	48	35	29	34.9
5	22	20	20	20	19	19	22	22	16	24	20.4
5 ARD	24	20	23	18	24	19	24	23	25	24	22.4
10	14	16	12	13	12	11	18	13	14	12	13.5
10 ARD	15	15	20	14	15	20	18	16	15	21	16.9

表 2　各数字データの分布を表現するユニット数

また学習結果として得られたユニット数を表 2 に示す。ここに示されたユニット数は、逐次的なユニット

(注1)：http://www.research.att.com/~yann/exdb/mnist/

(注2)：http://www.research.att.com/~yann/exdb/mnist/ に基く

挿入・削除操作と、パラメータ収束後の、自由エネルギー最大基準から決定されたものである。パラメータ収束値が大域的最適値である保証は無いため、決定されたユニット数もまた大域的最適である保証は無い。しかし同じデータに関して、q が小さいモデルのユニット数が、q が大きいモデルのユニット数よりも大きくなっていることは自然な結果である。ベイズ推定においては、データの持つ情報量とモデルの複雑度がバランスすることが要求される。q が大きいユニットは q が小さいユニットに比べてモデル複雑度が高く、ユニット数を増やすことに対するペナルティも高くなるために、ユニット数が少くなると考えられる。

7. ま と め

混合主成分分析の変分法的ベイズ推定アルゴリズムを、不必要な因子化仮定を置かずに導出した。これを高次元の実問題である手書き数字認識課題に応用したところ高いパフォーマンスを示した。

因子化仮定の置きかたによるパフォーマンスの違いを示すことや、オンライン計算への適用は今後の課題である。

文　　献

[1] M. E. Tipping, C. M. Bishop, “Mixtures of Probabilistic Principal Component Analyzers,” Advances in Neural Information Processing Systems, vol. 11, pp. 443–482, 1999.

[2] D. J. C. MacKay, “Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks,” Network: Computation in Neural Systems, Vol. 6(3), pp. 469–505, 1995

[3] R. M. Neal, Bayesian learning for neural networks, Springer Verlag. ,1996.

[4] H. Attias, “Learning parameters and structure of latent variable models by variational Bayes.” in Proceedings of Uncertainty in Artificial Intelligence, 1999.

[5] M. Sato, “On-Line Model Selection Based on the Variational Bayes,” ATR-Technical Report TR-H-282, 2000.

[6] Z. Ghahramani and M. J. Beal. “Variational Inference for Bayesian Mixture of Factor Analysers,” Advances in Neural Information Processing Systems, vol. 12, pp.449–455, 2000.

[7] C. M. Bishop, “Variational Principal Components,” In IEE Conference Publication on Artificial Neural Networks, pp. 509–514, 1999.

[8] 上田修功, ” 局所解の回避と最適モデル探索を同時実現する Variational Bayes 学習,” 信学技報, vol.NC99-132, pp.113-120, 2000

[9] S. Waterhouse, D. Mackay, T. Robinson, “Bayesian methods for mixture of experts,” Advances in Neural Information Processing Systems, vol. 8, pp. 351–357, 1996

[10] C. M. Bishop, “Bayesian PCA,” Advances in Neural Information Processing Systems, vol. 11, pp. 509–514, 1999.

[11] R. M. Neal, G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models, 355–368 (Jordan, M.I. et al, Eds). Kluwer Academic Press, Norwell, MA.

[12] P. Moerland, “A comparison of mixture models for density estimation,” in Proceedings of the International Conference on Artificial Neural Networks (ICANN’99), pp. 15–30, 1999,

付　　録

1. 変分法的ベイズ推定による混合主成分分析の試験分布更新則

1.1 隠れ変数試験事後分布の更新

(29) 式より、あるデータ $\mathbf{y} = \mathbf{y}(t)$ に対応する隠れ変数 $(\mathbf{x}, \mathbf{z})$ の試験分布 $Q(\mathbf{x}, \mathbf{z}|\mathbf{y})$ は次式で与えられる。

$$\ln Q(\mathbf{x},\mathbf{z}|\mathbf{y}) = \langle \ln p(\mathbf{y},\mathbf{x},\mathbf{z}|\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}} + \text{const.}$$
$$= \sum_{i=1}^{m} z_i \{ \langle \ln p(\mathbf{y},\mathbf{x}_i|\theta_i)\rangle_{\theta_i} + \langle \ln g_i \rangle_{\mathbf{g}} \} + \text{const.}$$

これはこの時点でのパラメータアンサンブル平均 (次節参照)

$$\langle \tau_i \rangle = \bar{\tau}_i, \quad \langle \tau_i \boldsymbol{\mu}_i \rangle = \bar{\tau}_i \overline{\boldsymbol{\mu}}_i, \quad \langle \tau_i \boldsymbol{\mu}_i^2 \rangle, \langle \tau_i \mathbf{W}_i' \boldsymbol{\mu}_i \rangle$$
$$\langle \tau_i \mathbf{W}_i \rangle = \bar{\tau}_i \overline{\mathbf{W}}_i,$$
$$\langle \tau_i \mathbf{W}_i' \mathbf{W}_i \rangle = \bar{\tau}_i \overline{\mathbf{W}}_i' \overline{\mathbf{W}}_i + d\Delta_{\mathbf{W}i}^{-1}$$
$$\langle \ln \tau_i \rangle, \quad \langle \ln g_i \rangle$$

を用いて以下のように求まる。

$$Q(\mathbf{x},\mathbf{z}|\mathbf{y}) = \frac{\prod_{i=1}^{m} \left[U_i(\mathbf{y}) \mathcal{N}(\mathbf{x}_i|\bar{\mathbf{x}}_i, \bar{\boldsymbol{R}}_{\mathbf{x}i}^{-1}) \right]^{z_i}}{\sum_{i=1}^{m} U_i(\mathbf{y})} \quad \text{(A·1)}$$

ここで U_i は以下で与えられる。

$$U_i(\mathbf{y}) \stackrel{\text{def}}{=} \int d\mathbf{x}_i \exp\Big[\langle \ln g_i \rangle + \langle \ln p(\mathbf{y},\mathbf{x}_i|\theta_i)\rangle \Big]$$
$$= \exp\Big[-\frac{\bar{\tau}_i}{2}(\mathbf{y}-\overline{\boldsymbol{\mu}}_i)^2 + \frac{1}{2}(\bar{\tau}_i \overline{\mathbf{W}}_i' \mathbf{y} - \langle \tau_i \mathbf{W}_i' \boldsymbol{\mu}_i \rangle)'$$
$$\times \bar{\boldsymbol{R}}_{\mathbf{x}i}^{-1} (\bar{\tau}_i \overline{\mathbf{W}}_i' \mathbf{y} - \langle \tau_i \mathbf{W}_i' \boldsymbol{\mu}_i \rangle) - \frac{1}{2}(\langle \tau \boldsymbol{\mu}_i^2 \rangle - \bar{\tau}_i \overline{\boldsymbol{\mu}}_i^2)$$
$$- \frac{1}{2} \ln |\bar{\boldsymbol{R}}_{\mathbf{x}i}| + \langle \ln g_i \rangle + \frac{d}{2} \langle \ln \tau_i \rangle - \frac{d}{2} \ln 2\pi \Big] \quad \text{(A·2)}$$

ただし、

$$\bar{\boldsymbol{R}}_{\mathbf{x}i} = \bar{\tau}_i \overline{\mathbf{W}}_i' \overline{\mathbf{W}}_i + d\Delta_{\mathbf{W}i}^{-1} + \mathbf{I}_q,$$
$$\bar{\mathbf{x}}_i = \bar{\boldsymbol{R}}_{\mathbf{x}i}^{-1} (\bar{\tau}_i \overline{\mathbf{W}}_i' \mathbf{y} - \langle \tau_i \mathbf{W}_i' \boldsymbol{\mu}_i \rangle)$$

この試験事後分布を用いて隠れ変数のアンサンブル平均は、

$$\langle z_i(t) \rangle = U_i(\mathbf{y}(t)) / \textstyle\sum_{i'=1}^{m} U_{i'}(\mathbf{y}(t)),$$
$$\langle z_i(t) \mathbf{x}_i(t) \rangle = \langle z_i(t) \rangle \bar{\mathbf{x}}_i,$$
$$\langle z_i(t) \mathbf{x}_i(t) \mathbf{x}_i'(t) \rangle = \langle z_i(t) \rangle (\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' + \bar{\boldsymbol{R}}_{\mathbf{x}i}^{-1}).$$

のようになり、これを用いて十分統計量の期待値は

$$E[z_i] = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \langle z_i(t) \rangle$$
$$E[z_i \mathbf{x}_i] = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \langle z_i(t) \mathbf{x}_i(t) \rangle$$
$$E[z_i \mathbf{x}_i \mathbf{x}_i'] = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \langle z_i(t) \mathbf{x}_i(t) \mathbf{x}_i'(t) \rangle$$
$$E[z_i \mathbf{y} \mathbf{x}_i'] = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \mathbf{y}(t) \langle z_i(t) \mathbf{x}_i(t) \rangle'$$
$$E[z_i \mathbf{y}] = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \langle z_i(t) \rangle \mathbf{y}(t)$$
$$E[z_i \mathbf{y}' \mathbf{y}] = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \langle z_i(t) \rangle \mathbf{y}'(t) \mathbf{y}(t)$$

のように計算できる。後の便利のため、拡張ベクトル $\mathbf{x}_e{}' = [\mathbf{x}'\ 1]$ を定義する。これを用いると十分統計量が以下のように表せる。

$$E[z_i \mathbf{x}_{ei} \mathbf{x}_{ei}'] = \begin{bmatrix} E[z_i \mathbf{x}_i \mathbf{x}_i'] & E[z_i \mathbf{x}_i] \\ E[z_i \mathbf{x}_i]' & E[z_i] \end{bmatrix} \quad \text{(A·3)}$$
$$E[z_i \mathbf{y} \mathbf{x}_{ei}'] = \begin{bmatrix} E[z_i \mathbf{y} \mathbf{x}_i'] & E[z_i \mathbf{y}] \end{bmatrix} \quad \text{(A·4)}$$

またパラメータに関しても、ユニット中心と主軸行列を拡張主軸行列 $\boldsymbol{W}_e = [\mathbf{W}\ \boldsymbol{\mu}]$ として併せて扱う。これに対する事前分布は以下のような正規分布で書ける。

$$\ln p(\boldsymbol{W}_e|\tau,\boldsymbol{\alpha}_e) = -\frac{1}{2}\tau \mathrm{tr}(\boldsymbol{W}_e - \overline{\boldsymbol{W}}_{e0})' (\boldsymbol{W}_e - \overline{\boldsymbol{W}}_{e0}) \boldsymbol{A}_e$$
$$+ \frac{d}{2} \ln |\boldsymbol{A}_e| + \frac{d(q+1)}{2} \ln(\tau/2\pi) \quad \text{(A·5)}$$

ただし、$\boldsymbol{A}_e$ はその対角成分が $\boldsymbol{\alpha}_e = [\boldsymbol{\alpha}\ \gamma_{\mu 0}]$ であるような対角行列であり、$\overline{\boldsymbol{W}}_{e0} = [\mathbf{0}\ \overline{\boldsymbol{\mu}}_0]$ は $d \times (q+1)$ 行列である。

1.2 パラメータ試験事後分布の更新

自由エネルギーの $Q(\boldsymbol{\theta})$ に関する最大化条件から、(30) 式が得られる:

$$\ln Q(\boldsymbol{\theta}) = \quad \text{(A·6)}$$
$$\langle \ln p(Y,Z,X|\boldsymbol{\theta})\rangle_{Z,X} + \langle \ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})\rangle_{\boldsymbol{\alpha}} + \text{const.}$$

この右辺を計算すると $Q(\boldsymbol{\theta})$ が以下の共役形で書けることがわかる。

$$Q(\boldsymbol{\theta}) = Q(\mathbf{g}) \prod_{i=1}^{m} Q(\boldsymbol{W}_{ei}|\tau_i) Q(\tau_i) \quad \text{(A·7)}$$

$$\ln Q(\mathbf{g}) = \sum_{i=1}^{m} \gamma_i \ln g_i - \sum_{i=1}^{m} \ln \Gamma(\gamma_i + 1)$$
$$+ \ln \Gamma\left(\sum_{i=1}^{m} \gamma_i + m \right) \quad \text{(A·8)}$$

$$\ln Q(\boldsymbol{W}_{ei}|\tau_i) =$$
$$- \frac{\tau_i}{2} \mathrm{tr}(\boldsymbol{W}_{ei} - \overline{\boldsymbol{W}}_{ei})' (\boldsymbol{W}_{ei} - \overline{\boldsymbol{W}}_{ei}) \tilde{\Delta}_{\boldsymbol{W}i}$$
$$+ \frac{d}{2} \ln |\tilde{\Delta}_{\boldsymbol{W}i}| + \frac{d(q+1)}{2} \ln(\tau_i/2\pi) \quad \text{(A·9)}$$

$$\ln Q(\tau_i) = -\gamma_{\tau i} \bar{\tau}_i^{-1} \tau_i + (\gamma_{\tau i} - 1) \ln \tau_i$$
$$- \ln \Gamma(\gamma_{\tau i}) + \gamma_{\tau i} \ln(\gamma_{\tau i} \bar{\tau}_i^{-1}) \quad \text{(A·10)}$$

ここで試験事後分布のハイパーパラメータとして、新たに

$$\gamma_i,\quad \overline{\boldsymbol{W}}_{\boldsymbol{e}\,i} = [\overline{\mathbf{W}}_i\ \overline{\boldsymbol{\mu}}_i],\quad \gamma_{\tau\,i},\quad \bar{\tau}_i,$$
$$\tilde{\Delta}_{\boldsymbol{W}\,i}^{-1} = \begin{bmatrix} \Delta_{\mathbf{W}_i}^{-1} & \mu_{W_i} \\ \mu'_{W_i} & \delta_{\mathbf{W}_i} \end{bmatrix}$$

を導入した。これらのハイパーパラメータは以下のように求まる。

1.2.1 混　合　比

$$\gamma_i = TE[z_i] + \gamma_{0i} \tag{A·11}$$
$$\langle \ln g_i \rangle = \psi(\gamma_i + 1) - \psi\left(\textstyle\sum_{i=1}^{m} \gamma_i + m\right) \tag{A·12}$$

ここで ψ は $\psi(\gamma) = \frac{d}{d\gamma}\ln\Gamma(\gamma)$ で定義される digamma 関数である。以下では簡潔性のために添字 i を省略する。

1.2.2 拡張主軸行列

$$\tilde{\Delta}_{\boldsymbol{W}} = TE[z\mathbf{x}_e\mathbf{x}_e{}'] + \langle \boldsymbol{A}_{\boldsymbol{e}} \rangle \tag{A·13}$$
$$\overline{\boldsymbol{W}}_{\boldsymbol{e}} = (TE[z\mathbf{y}\mathbf{x}_e{}'] + \overline{\boldsymbol{W}}_{\boldsymbol{e}\,0}\langle \boldsymbol{A}_{\boldsymbol{e}} \rangle)\tilde{\Delta}_{\boldsymbol{W}}{}^{-1} \tag{A·14}$$

1.2.3 誤差逆分散

$$\gamma_\tau = \frac{d}{2}TE[z] + \gamma_{\tau\,0} \tag{A·15}$$
$$\begin{aligned} 2\gamma_\tau\bar{\tau}^{-1} &= TE[z\mathbf{y}^2] - \mathrm{tr}\overline{\boldsymbol{W}}_{\boldsymbol{e}}\tilde{\Delta}_{\boldsymbol{W}}\overline{\boldsymbol{W}}_{\boldsymbol{e}}{}' \\ &\quad + \mathrm{tr}\overline{\boldsymbol{W}}_{\boldsymbol{e}\,0}\langle \boldsymbol{A}_e \rangle \overline{\boldsymbol{W}}_{\boldsymbol{e}\,0}{}' + 2\gamma_{\tau\,0}\bar{\tau}_0 \\ &= TE[z(\mathbf{y} - \overline{\boldsymbol{W}}_{\boldsymbol{e}}\mathbf{x}_e)^2] \\ &\quad + \mathrm{tr}(\overline{\boldsymbol{W}}_{\boldsymbol{e}} - \overline{\boldsymbol{W}}_{\boldsymbol{e}\,0})\langle \boldsymbol{A}_{\boldsymbol{e}} \rangle (\overline{\boldsymbol{W}}_{\boldsymbol{e}} - \overline{\boldsymbol{W}}_{\boldsymbol{e}\,0})' + 2\gamma_{\tau\,0}\bar{\tau}_0 \end{aligned} \tag{A·16}$$

1.2.4 パラメータアンサンブル平均

以上に基き、各パラメータに関するアンサンブル平均は以下のように求められる。

$$\langle \tau \rangle = \bar{\tau},\quad \langle \ln\tau \rangle = \ln\bar{\tau} + \psi(\gamma_\tau) - \ln(\gamma_\tau)$$
$$\langle \boldsymbol{W}_{\boldsymbol{e}} \rangle = \overline{\boldsymbol{W}}_{\boldsymbol{e}},\quad \langle \tau \boldsymbol{W}_{\boldsymbol{e}} \rangle = \bar{\tau}\overline{\boldsymbol{W}}_{\boldsymbol{e}}$$
$$\langle \tau \boldsymbol{W}_{\boldsymbol{e}}{}' \boldsymbol{W}_{\boldsymbol{e}} \rangle = \bar{\tau}\overline{\boldsymbol{W}}_{\boldsymbol{e}}{}'\overline{\boldsymbol{W}}_{\boldsymbol{e}} + d\tilde{\Delta}_{\boldsymbol{W}}{}^{-1}$$
$$= \begin{bmatrix} \langle \tau\mathbf{W}'\mathbf{W} \rangle & \langle \tau\mathbf{W}'\boldsymbol{\mu} \rangle \\ \langle \tau\mathbf{W}'\boldsymbol{\mu} \rangle' & \langle \tau\boldsymbol{\mu}^2 \rangle \end{bmatrix}$$

1.3 ハイパーパラメータの試験事後分布更新

自由エネルギーの $Q(\boldsymbol{\alpha})$ に関する最大化条件 (31) 式から次式が得られる。

$$\begin{aligned} \ln Q(\boldsymbol{\alpha}) &= \langle \ln p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \rangle_{\boldsymbol{\alpha}} + \ln p(\boldsymbol{\alpha}) + \mathrm{const.} \\ &= -\frac{1}{2}\mathrm{tr}\langle \tau\mathbf{W}'\mathbf{W} \rangle \boldsymbol{A} + \frac{d}{2}\ln|\boldsymbol{A}| \\ &\quad + \sum_{j=1}^{q}[-\gamma_{\alpha\,0}\alpha_{0j}^{-1}\alpha_j + (\gamma_{\alpha\,0} - 1)\ln\alpha_j] + \mathrm{const.} \\ &= \sum_{j=1}^{q}[-\gamma_\alpha\bar{\alpha}_j^{-1}\alpha_j + (\gamma_\alpha - 1)\ln\alpha_j \\ &\quad - \ln\Gamma(\gamma_\alpha) + \gamma_\alpha\ln(\gamma_\alpha\bar{\alpha}_j^{-1})] \end{aligned} \tag{A·17}$$

ここで試験事後分布のハイパーパラメータ $\gamma_\alpha, \bar{\alpha}$ は次のように更新される。

$$\gamma_\alpha = \gamma_{\alpha\,0} + d/2 \tag{A·18}$$
$$\gamma_\alpha\bar{\alpha}_j^{-1} = \frac{1}{2}\langle \tau\mathbf{W}'\mathbf{W} \rangle_{j,j} + \gamma_{\alpha\,0}\boldsymbol{\alpha}_{0j}^{-1} \tag{A·19}$$

また $\boldsymbol{\alpha}$ のアンサンブル平均は以下のように与えられる。

$$\langle \alpha_j \rangle = \bar{\alpha}_j \tag{A·20}$$

（平成 x 年 xx 月 xx 日受付）