

最良モデル探索のための変分ベイズ学習

Variational Bayesian Learning for Optimal Model Search

上田 修功
Naonori Ueda

NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
ueda@cslab.kecl.ntt.co.jp, <http://www.kecl.ntt.co.jp/as/members/ueda/index-j.html>

keywords: Bayesian learning, variational approximation, model search, split and merge operations, local optima problem.

Summary

When learning a nonlinear model, we suffer from two difficulties in practice: (1) the local optima, and (2) appropriate model complexity determination problems. As for (1), I recently proposed the split and merge Expectation Maximization (SMEM) algorithm within the framework of the maximum likelihood by simultaneously splitting and merging model components, but the model complexity was fixed there. To overcome these problems, I first formally derive an objective function that can optimize a model over parameter and structure distributions *simultaneously* based on the variational Bayesian approach. Then, I devise a Bayesian SMEM algorithm to efficiently optimize the objective function. With the proposed algorithm, we can find the optimal model structure while avoiding being trapped in poor local maxima. I apply the proposed method to the learning of a mixture of experts model and show the usefulness of the method.

1. はじめに

統計的学習の目的は、観測データの背後にある生成モデルの推定である。近年、構造が複雑なデータを対象に、様々な非線形モデルが適用されている。しかしながら、非線形モデルを学習する際、実用上、(1) 局所最適性 (local optimality)、および、(2) モデルの複雑さ (model complexity) の決定、の問題に悩まされる。

(1) は学習アルゴリズムが所望の大域的最適解に収束せず、初期解の近傍の局所最適解に収束するという問題である。また (2) は、例えば 3 層ニューラルネットワークの場合、中間ユニット数の決定、また、混合モデルでは混合要素数の決定など、非線形モデルの構造決定問題を指す。モデルの複雑さを学習タスクの複雑さに応じて適切に定めないと汎化能力 (未学習データに対する予測能力) の低下を招く。

筆者は先に、(1) の問題に対処すべく、混合モデルを対象に最尤推定法の一般的数値解法である EM アルゴリズム [Dempster 77] に、モデルの同時併合分割による局所解からの脱出とより尤度値の高い解への誘導を図る併合分割操作を導入した併合分割操作付き EM (Split and Merge EM: SMEM) アルゴリズムを考案し [Ueda 99a], 様々な応用での有効性を示した [Ueda 99b, Ueda 00]。

しかしながら、SMEM アルゴリズムは (2) の問題には対処していなかった。その理由は、最尤推定では、一般にモデルパラメータの次元の増加と共に尤度が単調増加す

るため、尤度を基準にして併合分割操作で最適なモデルを探索することができないことに因る。そこで、SMEM アルゴリズムでは、モデルの併合と分割を同時に行い、モデルの複雑さ (混合数) を固定していた。

これに対し、尤度値ではなく AIC, MDL 等の情報量基準 [Akaike 73, Rissanen 87] を用いたモデル探索も考えられるが、多くの非線形モデルの場合、情報量基準の導出で仮定される最尤推定量の漸近正規性が成り立たないため、これら情報量基準に基づくモデル探索は実用上うまく作用しない。

本論文では、SMEM アルゴリズムの“モデルの併合分割操作”の考え方をベイズ学習の近似法である変分ベイズ学習 (variational Bayesian learning) [Waterhouse 95] に導入し、非線形モデルの上記 (1),(2) の問題を同時解決する新たな学習法 (最良モデル探索のための変分ベイズ学習法) を提案し、混合回帰モデルへの適用実験により手法の有効性を示す。

変分ベイズ学習法は後述するように、ベイズ学習の核となる事後分布の一近似法であるが、ラプラス近似法 [MacKay 92] より近似精度が高く、また、マルコフ連鎖モンテカルロ (MCMC) 法 [Geman 97] より遥かに効率的な手法として近年注目されている。

変分ベイズ学習は、当初、汎化性能の観点で最尤法に対する優位性を示していた [Waterhouse 95] が、上記 (1),(2) の問題は全く取り扱われていなかった。最近、Attias は、モデルの複雑さも確率変数として取り扱うこと

により上記 (2) の一解決法を示している [Attias 99] が、(1) の問題は取り扱っていない。また、モデルのパラメータ学習とモデルの複雑さの決定が 2 段階に実行され、候補モデルの中から最良なモデルを選ぶという“モデル選択的”手法であった。

最近、Ghahramani らは混合因子分析モデルを対象として、変分ベイズ学習法の枠組みでモデルの削除と追加に基づくモデル探索法を提案している [Zoubin 00]。しかしながら、そこではモデル探索のアイデアに留まり、その正当性に関する根拠は示されていない。

これに対し本論文では、一般の非線形モデルを対象に、変分ベイズ学習においてモデルのパラメータ学習とモデルの複雑さの決定が、同一の評価関数の最小化問題として定式化できることを示す。これにより、モデル探索学習法という新たな枠組みの正当性を保証する。

以下の本文では、2. 節でベイズ学習の概要を述べ、3. 節でベイズ学習の近似法である変分ベイズ法、およびその問題点について述べる。次いで、4. 節でその問題解決法を提案し、5. 節で適用実験結果を示す。

2. ベイズ学習

2.1 予測事後分布

今、モデルの複雑さの指標 m とモデルパラメータ θ で規定されるパラメトリックな確率分布 (確率モデル) のクラスを $\mathcal{H}_m = \{p(\cdot|\theta, m)\}$ とし、これを仮説空間と呼ぶこととする。

統計的学習のゴールは、観測データ $D = \{d_i\}_{i=1}^N$ に基づいて仮説空間上で真のモデルを最良近似する仮説を“探索”することと言える。最尤学習ではその良さの基準として尤度 (対数尤度) を用いる。即ち、最尤学習での最適仮説のモデルパラメータは次式で与えられる。

$$\theta^* = \arg \max_{\theta} \{\log p(D|\theta, m)\}$$

尚、混合分布モデルのように非観測変数 (Z とする) を取り扱う場合は、 $p(D|\theta, m) = \sum_Z p(D, Z|\theta, m)$ とすれば良い。

一方、ベイズ学習では尤度に加えてパラメータ θ の事前分布 $p(\theta|m)$ をも考慮する。即ち、最尤学習の様に一つの仮説 $p(D|\hat{\theta}, m)$ を求めるのではなく、未知データ d_{N+1} に対し、 D が与えられた下での θ の事後分布 $p(\theta|D, m)$ で仮説 $p(d_{N+1}|\theta, m)$ を重み付き平均した“事後の予測分布” $p(d_{N+1}|D, m)$ を次式で求め d_{N+1} についての確率的な言明を行う。

$$p(d_{N+1}|D, m) = \int p(d_{N+1}|\theta, m)p(\theta|D, m)d\theta \quad (1)$$

従って、一般に、ベイズ学習は最尤学習に比べ過学習が抑制される。更に、ベイズ学習ではモデル指標 m も確率

変数として取り扱える。 m の事前分布 $P(m)$ も考慮すると、式 (1) は次式のように書き換えられる。

$$p(d_{N+1}|D) = \sum_m \int p(d_{N+1}|\theta, m)p(\theta, m|D)d\theta \quad (2)$$

2.2 ベイズ学習の実用上の問題点

上記事後予測分布は特殊な場合を除き解析的に求めることが困難で何らかの近似法を援用して求める。その一近似法としてラプラス近似法 [MacKay 92] がある。ラプラス近似法では事後分布をガウス関数近似し上記積分を解析的に求める手法である。しかしながら、この近似は無数個のデータ数を前提 (漸近正規性) にした近似であり、現実問題での近似精度に問題がある。

より正確な近似解法としてマルコフ連鎖モンテカルロ (MCMC) 法がある。通常モンテカルロ法との相違点は、 x 空間全てを評価するのではなく、 $p(x)$ を近似する有限個の $\{x_t\}$ をサンプリングという形式で“生成”する点にある。サンプリングの具体的手法としてメトロポリス法、Gibbs サンプリング法が著名である [Geman 97]。

尚、従来 MCMC ではモデルパラメータの次元 (モデルの複雑さ) は固定されていたが、最近、“reversible jump MCMC 法”と呼ばれるパラメータ次元を変更しながらサンプリングを実行する手法が提案されている [Richardson 97]。しかしながら、これら MCMC 法はサンプリングに膨大な時間を要し、また、収束判定も一般には容易ではないという問題がある。次節では、ラプラス近似よりも近似精度が高く、MCMC に比べ遥かに効率的な、Bayes 学習の第三のアプローチである変分ベイズ学習法について述べる。

3. 変分ベイズ学習

3.1 変分近似

ベイズ学習では、前述した様に全ての未知量 Z, θ, m を確率変数として取り扱う。そこで、これら全ての未知量を周辺化した次式の周辺尤度を考える。

$$\begin{aligned} \mathcal{L}(D) &= \log p(D) \\ &= \log \sum_m \sum_Z \int p(D, Z, \theta, m)d\theta \end{aligned} \quad (3)$$

ここに、 Z は潜在変数 (非観測データ) を表す。

全ての確率変数の結合分布 $p(D, Z, \theta, m)$ は

$$p(D, Z, \theta, m) = p(D, Z|m)p(\theta|m)P(m) \quad (4)$$

と分解できる。式 (4) の右辺第 1 項はモデル指標が与えられた下での完全データ尤度に、第 2 項はモデル指標が与えられた時のパラメータ θ の事前分布、そして、第 3 項はモデル指標の事前分布に、各々対応している。

ここで、新たな分布 Q を導入し、対数関数に対する Jensen の不等式を適用することにより以下の様に $\mathcal{L}(D)$ の下限値 $\mathcal{F}[Q]$ を得る。

$$\begin{aligned} \mathcal{L}(D) &= \log \sum_m \sum_Z \int Q(Z, \theta, m) \frac{p(D, Z, \theta, m)}{Q(Z, \theta, m|D)} d\theta \\ &= \log \left\langle \frac{p(D, Z, \theta, m)}{Q(Z, \theta, m|D)} \right\rangle_{Q(Z, \theta, m)} \\ &\geq \left\langle \log \frac{p(D, Z, \theta, m)}{Q(Z, \theta, m|D)} \right\rangle_{Q(Z, \theta, m)} \\ &= \sum_m \sum_Z \int Q(Z, \theta, m) \log \frac{p(D, Z, \theta, m)}{Q(Z, \theta, m|D)} d\theta \\ &\equiv \mathcal{F}[Q] \end{aligned} \quad (5)$$

但し、表記 $\langle f(x) \rangle_{p(x)}$ は $f(x)$ の $p(x)$ に関する期待値:

$$\langle f(x) \rangle_{p(x)} = \int f(x)p(x)dx$$

を表すものとする。

$\mathcal{F}[Q]$ は Q を変関数とする汎関数であることに注意。ある関数の下限値 (もしくは上限値) を定数ではなく関数で抑える近似法を一般に変分近似 (variational approximation) と呼ぶ [Jordan 97]。上記では、 $\mathcal{F}[Q]$ が対数周辺尤度 \mathcal{L} の下限値となっている。

\mathcal{L} と \mathcal{F} の間には次式に示す関係式が成り立つ。

$$\mathcal{L}(D) = \mathcal{F}[Q] + \text{KL}(Q(Z, \theta, m|D) \parallel p(Z, \theta|D)) \quad (6)$$

ここに、右辺第 2 項は分布 $Q(Z, \theta, m|D)$ と $p(Z, \theta|D)$ との Kullback-Leibler 距離である。

式 (6) で \mathcal{L} が D のみに依存する定数であることに注意すると、下限値を最大化すべく、 $\mathcal{F}[Q]$ を Q に関して最大化することは、 Q と真の事後分布 $p(\cdot|D)$ との KL 情報量を最小化することと等価である。換言すれば、 \mathcal{F} を最大化する分布 Q は真の事後分布の最良の近似となっている。真の事後分布を変分近似する事後分布であることから、 Q は変分事後分布と呼ばれる*1。

上記は [Waterhouse 95] らの定式化であるが、これは最尤学習における変分近似 [Saul 96] をベイズ拡張したものである。 Q として、通常、次式のように各未知変量毎に分解した形 (factorization form) を仮定するが、各分布族は任意で良い。

$$Q(Z, \theta, m) = Q(Z|m)Q(\theta|m)Q(m) \quad (7)$$

式 (7) の制約された形で真の事後分布を推定するため一般には真の分布に一致しないが、全パラメータの同時事後分布を単一の正規分布で近似するラプラス近似法に比べれば、遥かに近似精度が高いと言える。

3.2 最適事後分布とモデル選択

モデル指標 m が与えられた下での θ の最適変分事後分布 $Q(\theta|m)$ は、制約条件 $\int Q(\theta|m)d\theta = 1$ の下で $\mathcal{F}[Q]$ を Q に関して最大化することにより容易に得られる [Attias 99]。

$$Q(\theta|m) = \frac{1}{C_\theta} \exp \left\{ \left\langle \log p(D, Z|\theta, m) \right\rangle_{Q(Z|m)} + \log p(\theta|m) \right\} \quad (8)$$

但し、 C_θ は $\int Q(\theta|m)d\theta = 1$ となるための規格化定数である。同様に、

$$Q(Z|m) = \frac{1}{C_Z} \exp \left\{ \left\langle \log p(D, Z|\theta, m) \right\rangle_{Q(\theta|m)} \right\} \quad (9)$$

式 (8),(9) より明らかな様に、 $Q(\theta|m)$ と $Q(Z|m)$ は相互に依存関係にあり閉形式で解くことはできず逐次解法により求める。即ち、第 t 反復での事後分布の推定値を各々 $Q(Z|m)^{(t)}$ の $Q(\theta|m)^{(t)}$ とすると、第 $t+1$ 反復での推定値は各々以下で計算すれば良い。

$$Q(Z|m)^{(t+1)} = \frac{1}{C_Z} \exp \left\{ \left\langle \log p(D, Z|\theta, m) \right\rangle_{Q(\theta|m)^{(t)}} \right\} \quad (10)$$

$$Q(\theta|m)^{(t+1)} = \frac{1}{C_\theta} \exp \left\{ \left\langle \log p(D, Z|\theta, m) \right\rangle_{Q(Z|m)^{(t+1)}} + \log p(\theta|\varphi, m) \right\} \quad (11)$$

式 (10),(11) を反復して実行することにより局所最適事後分布 $Q(Z|m)^*$, $Q(\theta|m)^*$ を得る。

但し、上記は非線形最適化法の coordinate ascent 法と同様な反復写像法であるが、変分近似問題に反復写像法を用いた場合、収束性に関する理論的保証はない。しかしながら、文献 [Attias 99][Zoubin 00] および後述する筆者の実験に関する限り、(局所)最適解への収束を確認している。収束性に関する厳密な議論は別途検討課題と言える。

次いで、モデル指標 m の最適事後分布は \mathcal{F} の $Q(m)$ に関する最大化より解析的に求まる。

$$Q(m)^* = \frac{1}{C_m} \exp \left\{ \left\langle \log \frac{p(D, Z|\theta, m)}{Q(Z|m)^*} \right\rangle_{Q(Z, \theta|m)^*} + \left\langle \log \frac{p(\theta|\varphi, m)}{Q(\theta|m)^*} \right\rangle_{Q(\theta|m)^*} + \log P(m) \right\} \quad (12)$$

$Q(m)^*$ を最大にする m が事後分布最大化 (Maximum a posteriori Probability: MAP) の観点で最適なモデル指標となる [Attias 99]。

ベイズ推定の場合、式 (2) に示した様に、本来は全ての可能なモデルのアンサンブルとして予測分布を求める。しかしながら、実用的にはある最良なモデルのみに着目して単一のモデルを選択することも可能である。 $Q(m)$ が単峰でかつ鋭いピークを持つ場合には十分な近似が得

*1 Q は事後分布故、本来は $Q(\cdot|m, D)$ と書くべきであるが表記を簡単にする為、 D を省略している。

られることが期待される。次節で提案するモデル探索法は、この最良単一モデル選択の考え方に基づく。但し、単にモデルを選択するのではなく、その過程で 1. 節で述べた (1) の局所最適性の問題も同時に解決する新たな変分ベイズ学習法の提案である。

4. モデル探索学習

4.1 モデル探索

式 (5) の $\mathcal{F}[Q]$ において、 $Q(m)$ を含まない項をまとめて \mathcal{F}_m と書くと次式を得る。

$$\mathcal{F}[Q] = \langle \mathcal{F}_m \rangle_{Q(m)} - \text{KL}(Q(m) || P(m)) \quad (13)$$

ここで、 \mathcal{F}_m は $Q(m)$ には依存しないが、 m に依存することに注意。 \mathcal{F}_m は次式で与えられる。

$$\mathcal{F}_m = \left\langle \log \frac{p(D, Z | \theta, m) p(\theta | \varphi, m)}{Q(Z | m) Q(\theta | m)} \right\rangle_{Q(Z, \theta | m)} \quad (14)$$

式 (13) の右辺第 1 項の \mathcal{F}_m は、 $Q(Z | m)$ 、 $Q(\theta | m)$ に依存する。従って 3.2 節で述べた $\mathcal{F}[Q]$ の最大化は以下の 2 ステップの最大化と等価である。

[従来の変分ベイズ学習アルゴリズム]

Step 1: 各 $m \in \mathcal{M}$ に対し、 $Q(Z | m)^{(0)}$ 、 $Q(\theta | m)^{(0)}$ を設定し、 $t \leftarrow 0$ とし、以下を収束するまで実行。

$$Q(Z | m)^{(t+1)} = \arg \max_{Q(Z | m)} \mathcal{F}_m [Q(Z | m), Q(\theta | m)^{(t)}]$$

$$Q(\theta | m)^{(t+1)} = \arg \max_{Q(\theta | m)} \mathcal{F}_m [Q(Z | m)^{(t+1)}, Q(\theta | m)]$$

$$t \leftarrow t + 1$$

Step 2: 各 $m \in \mathcal{M}$ に対し、 \mathcal{F}_m を m に関し最大化する。 $(\mathcal{M}$ は候補モデル指標集合を表す。)

ここで、 \mathcal{F}_m^* を Step 1 で得られた \mathcal{F}_m の最適値を表すものとして式 (13) より次式を得る。

$$\mathcal{F}[Q] = \langle \mathcal{F}_m^* \rangle_{Q(m)} - \text{KL}(Q(m) || P(m)) \quad (15)$$

従って、Step 2 は、 $\sum_m Q(m) = 1$ の下で $Q(m)$ に関する式 (15) の最大化により

$$Q(m)^* = \frac{P(m) e^{\mathcal{F}_m^*}}{\sum_{l=1}^M P(l) e^{\mathcal{F}_l^*}} \quad (16)$$

と求まる。式 (16) が式 (12) と等価であることは容易に確認できる。

ここで式 (16) を注意深く見ると、分母は m に依存しないので $Q(m)$ の m に関する最大化は $P(m) e^{\mathcal{F}_m^*}$ の最大化と等価であることが分かる。簡単のため m の事前分

布を一様分布 $P(m) = 1/M$ とすると、 $Q(m)$ の最大化は単純に \mathcal{F}_m の最大化となる。換言すれば、 $Q(m)$ を最大化する m は \mathcal{F}_m を最大化する m に他ならない。

従って、

$$\mathcal{F}_m^{(t)} = \left\langle \log \frac{p(D, Z | \theta, m) p(\theta | \varphi, m)}{Q(Z | m)^{(t)} Q(\theta | m)^{(t)}} \right\rangle_{Q(Z, \theta | m)^{(t)}} \quad (17)$$

および、

$$Q(m)^{(t)} = \frac{P(m) e^{\mathcal{F}_m^{(t)}}}{\sum_{l=1}^M P(l) e^{\mathcal{F}_l^{(t)}}} \quad (18)$$

とし、更に、 $P(m) = 1/M$ とすると、次の単調性が成り立つ。

$$\mathcal{F}_{m'}^{(t)} \geq \mathcal{F}_m^{(t)} \Rightarrow Q(m')^{(t)} \geq Q(m)^{(t)}$$

これは \mathcal{F}_m を $Q(\theta | m)$ 、 $Q(Z | m)$ のみならず m に関しても同時に最大化することにより、式 (18) を計算することなく最適なモデル指標 m が同時に求まることを意味する。

つまり、 $\mathcal{F}[Q]$ ではなく \mathcal{F}_m を目的関数として $Q(Z | m)$ 、 $Q(\theta | m)$ および m に関して同時に最大化することにより事後分布最大化 (Maximum a posteriori Probability: MAP) の観点で最適なモデルパラメータおよび最適なモデル指標が次式の様に得られるわけである。

$$\begin{cases} \theta_{\text{MAP}} = \arg \max_{\theta} \{Q(\theta | m)^*\} \\ m_{\text{MAP}} = \arg \max_m \{Q(m)^*\} \end{cases} \quad (19)$$

θ_{MAP} および m_{MAP} が得られれば、式 (2) に示した未知データ d_{N+1} に対する予測分布は次式の様に近似的に求まる。

$$p(d_{N+1} | D) \simeq p(d_{N+1} | \theta_{\text{MAP}}, m_{\text{MAP}}) \quad (20)$$

4.2 併合分割操作付き変分ベイズ学習

4.1 節では、同一の目的関数で θ および m の最適値が同時に学習可能であることを示した。本節ではその実現アルゴリズムとして併合分割操作付き変分ベイズ学習アルゴリズムを提案する。

今、仮説空間 \mathcal{H}_i が i に関する直和として

$$\mathcal{H} = \cup_{i=1}^m \mathcal{H}_i \quad (21)$$

で与えられる場合を考える。一般の混合モデルでは常に成立する。この場合、局所解の大半はあるデータ領域に過剰数のモデルが割り当てられ、かつ、あるデータ領域に過少数のモデルが割り当てられた状況に相当する。実際、4.1 節の Step 1 の \mathcal{F}_m の最大化 (式 (10), (11) の逐次増大化) では、適切な初期値を設定しない限り、上記のような不均衡なモデル配置 (低品質な局所解) に収束してしまう。

このモデル配置の不均衡を解消し、より良いモデル配置を実現するために、最尤学習の枠組みで提案したモデ

ルの併合分割操作 [Ueda 99a] を変分ベイズ学習に導入する。但し、ここではモデル指標 m も同時に最適化するという点で更に拡張している。

式 (21) が成立する場合、 \mathcal{F}_m は次式の様に、各要素モデルの目的関数の直和で書き表せる。

$$\mathcal{F}_m = \sum_{i=1}^m \mathcal{F}_{(i)} \quad (22)$$

$\mathcal{F}_{(i)}$ は混合モデルの第 i 要素モデルの目的関数に対応する。今、ある m に対し、式 (10),(11) により得た事後分布 (局所最適解) を Q^* 、その時の \mathcal{F}_m の値を \mathcal{F}_m^* と書くこととすると、式 (22) は更に次式の様に書ける。

$$\mathcal{F}_m^* = \mathcal{F}_{(i)}^* + \mathcal{F}_{(j)}^* + \mathcal{F}_{(k)}^* + \sum_{u, u \neq i, j, k} \mathcal{F}_{(u)}^* \quad (23)$$

この時、式 (23) の右辺の最初の 3 項のみに着目し、要素モデル i と要素モデル j とを新たな要素モデル i' として併合し、要素モデル k を二つの要素モデル j' と k' とに分割することにより、 \mathcal{F}_m 値の更なる増大を試みる。

前述した様に、最尤学習では m を増加 (減少) させると一般に尤度が増加 (減少) するので、例えば分割のみを行うと、分割と再学習により局所解から脱出してより良い解に到達して尤度が増加したのか、単に m が増加したことで尤度が増加したかの識別が困難となる。それ故、SMEM アルゴリズムでは m を固定すべく、併合と分割を同時に行っていた。

一方、4.1 節で述べた様に、変分ベイズ学習では目的関数 \mathcal{F}_m を用いてパラメータとモデルの複雑さの最適化が同時実行できる。即ち、 m の増加と共に \mathcal{F}_m 値は単調増加せず、最適な m の値に対し最大値をとる。そこで、同時併合分割操作だけでなく、“併合操作のみ”、あるいは、“分割操作のみ”、も試みる。明らかに“併合 (分割) 操作のみ”は m を 1 だけ増加 (減少) させることを意味する。

従って、これら 3 種類の操作を実行し、 \mathcal{F}_m を増大させることにより、局所最適性の問題と最適なモデルの複雑さの決定の問題が同時解決可能となる。以上が変分ベイズ学習の枠組みに併合分割操作を導入した併合分割操作付き変分ベイズ学習アルゴリズムの概要である。アルゴリズムの詳細は以下の通りである。

[最良モデル探索のための変分ベイズ学習アルゴリズム]

- Step 1: m および事後分布の初期値を設定し、式 (16),(17) に基づく通常の変分ベイズ学習を実行する。収束した時の事後分布の値を $Q(\theta|m)^*$ 、 $Q(Z|m)^*$ とし、 $F^* \leftarrow \mathcal{F}_m^*$ 、 $m^* \leftarrow m$ とする。
- Step 2: 現在の事後分布に基づき併合分割候補 (C 個) をソートする。
- Step 3: 以下の (3-1),(3-2),(3-3) を各々実行する。
 - (3-1): 併合: C 個の併合候補を順に、目的関数の値が F^* より大きくなるまで併合操作のみによる探索を行う。その時の目的関数の値を F_1^{**} とする。

- (3-2): 併合分割: C 個の併合分割候補を順に、目的関数が F^* を上回るまで同時併合分割操作による探索を行う。その時の目的関数の値を F_2^{**} とする。

- (3-3): 分割: C 個の分割候補を順に、目的関数が F^* を上回るまで分割操作のみによる探索を行う。その時の目的関数の値を F_3^{**} とする。

Step 4: Step 3 で F^* を上回る候補がなければアルゴリズムを終了。さもなければ、

$$F^* \leftarrow \max\{F_1^{**}, F_2^{**}, F_3^{**}\}$$

とし、 $F^* = F_1^{**}$ なら (3-1) の探索結果を採用し、 $m^* \leftarrow m^* - 1$ として Step 2 へ。 $F^* = F_2^{**}$ なら (3-2) の探索結果を採用し、Step 2 へ。 $F^* = F_3^{**}$ なら (3-3) の探索結果を採用し、 $m^* \leftarrow m^* + 1$ として Step 2 へ。

上記アルゴリズムの Step 3 の (3-1),(3-2),(3-3) の各々は m を固定した下で、 $Q(Z, \theta|m)$ の局所解からの脱出とより良い解への誘導を行う。そして、最適モデル選択の観点で、この 3 通りのモデルの複雑さから最良のものを Step 4 で選択する。これら一連の処理を反復することにより、局所解を回避しながら最適モデルを探索することができる。

Step 2 での併合分割候補基準、および Step 3 での併合、分割直後の初期化および再学習は SMEM アルゴリズムの時と同様に行えるので省略する (文献 [Ueda 00] 参照)。Step 3 はいわゆる greedy search 故、上記アルゴリズムは \mathcal{F}_m のより良い極大値の探索であり、大域的最大値が得られる理論的保証はない。しかしながら、 \mathcal{F}_m の単調増加性は保証される為、より良い極大値の探索が効率良く実現できる。

5. 混合回帰 (MoE) モデルへの適用

本節では、本学習法の有効性を検証すべく、ニューラルネットの代表モデルの一つである混合回帰モデル (Mixture of Experts: MoE)[Jacobs 91] による回帰問題 (関数近似問題) を対象に、本学習法を適用した結果について述べる。

5.1 MoE の確率モデル

MoE は、 m 個の回帰モデルと入力データを個々の回帰モデルに割り当てる役目を担うゲート関数から成る混合回帰モデルである。個々の回帰モデルは“expert”、混合回帰モデルは“mixture of experts”と通常呼ばれている。直観的には、関数近似問題を単一の関数で回帰するのでなく、データ領域を分割し、単純な関数近似問題に分割して解くモデルである。但し、領域分割は関数近似

と独立して実行されるのではなく、関数近似と同時に実行されるといふ点で柔軟なモデルとなっている。

MoE に対する変分ベイズ学習は既に提案されている [Waterhouse 95] が、そこでは 1. 節の問題 (1) および (2) は全く取り扱われていない。また、本節では、MoE の確率モデルとして入出力の同時分布を用いているため、文献 [Waterhouse 95] と異なり全ての変分事後分布を解析的に導出している。これらの点で本節で述べる MoE の変分ベイズ学習は文献 [Waterhouse 95] の成果をより発展させていると言える。

$x \in \mathcal{R}^d$ を入力、 $f_i(x, \theta_i) \in \mathcal{R}$ を入力 x に対応するモデル i の出力^{*2}を表すものとする、MoE の入力 x に対する出力 y は次式で与えられる。

$$y = \sum_{i=1}^m G_i(x|\Phi) f_i(x, \theta_i) \quad (24)$$

文献 [Waterhouse 95] 同様、各要素回帰モデル (expert) として線形モデルを採用した。

$$f_i(x, w_i) = w_i^T x', \quad i = 1, \dots, m$$

$x' = (x^T 1)^T$, $w_i = (w_{id}, \dots, w_{i1}, w_{i0})^T \in \mathcal{R}^{d+1}$ とし、 w_{i0} はバイアスに相当する。

ここに、 $G_i \in \mathcal{R}$ はゲート関数の第 i 出力で、通常、softmax 関数が通常用いられるが、ここでは、正規化ガウス関数を用いた [Xu 94]。

$$G_i(x|\Phi) = \frac{\varphi_i \mathcal{N}(x|\mu_i, S_i^{-1})}{\sum_{j=1}^m \varphi_j \mathcal{N}(x|\mu_j, S_j^{-1})}$$

但し、 φ_i は混合比 ($\varphi_i \geq 0$ かつ $\sum_{i=1}^m \varphi_i = 1$) で、表記 $\mathcal{N}(x|\mu, S^{-1})$ は平均ベクトル μ 、精度行列 (共分散行列の逆行列) S とする多次元正規分布を表す。

$$\mathcal{N}(x|\mu, S^{-1}) = (2\pi)^{-\frac{1}{2}} |S|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T S(x-\mu)\right\}$$

$|S|$ は行列式を表す。

また、出力のノイズモデルを正規分布と仮定すると、入力が与えられたもとの第 i モデルの出力の条件付き分布は次式となる。

$$p(y|x, i, \theta_i) = \mathcal{N}(y|w_i^T x', \beta_i^{-1})$$

以上より未知パラメータは以下となる。

$$\{(\varphi_i, \mu_i, S_i, w_i, \beta_i), \quad i = 1, \dots, m\}$$

今、 $D = \{(x_n, y_n)\}_{n=1}^N$ を観測学習データ集合とし、 $Z = \{z_i^n\}_{i=1, n=1}^{m, N}$ を潜在変数集合とする。但し、 z_i^n は入力 x_n に対応する出力 y_n が第 i モデルから生成されたと

する時 1 でそれ以外は零をとるものとする。この時、完全データの対数尤度関数は次式となる。

$$p(D, Z|\Phi, \Theta, m) = \prod_{i=1}^m \prod_{n=1}^N \left\{ \varphi_i \mathcal{N}(x_n|\mu_i, S_i^{-1}) \mathcal{N}(y_n|w_i^T x'_n, \beta_i^{-1}) \right\}^{z_i^n}$$

ここで注意すべきは、上記尤度は、文献 [Waterhouse 95] と異なり、 i に関して積の形に分解可能である点である。これにより全ての変分事後分布が各 i 毎に独立に導出できる。一方、文献 [Waterhouse 95] ではゲート関数のパラメータが i に関して分解できない為、近似を余儀なくされていた。

5.2 MoE の変分ベイズ学習

ベイズ学習では、未知パラメータおよびモデル指標 m はある事前分布を持つ確率変数として取り扱われる。各モデルの独立性を仮定すると全変数の結合分布は次式となる。

$$p(D, Z, \Phi, \Theta, m) = p(D, Z|\Phi, \Theta, m) p(\{\varphi_i\}_{i=1}^m | m) \times P(m) \left\{ \prod_{i=1}^m p(\mu_i | S_i) p(\beta_i) p(S_i) \times p(w_i | \{\alpha_{i,j}\}_{j=1}^{d+1}, \beta_i) \prod_{j=1}^{d+1} p(\alpha_{i,j}) \right\} \quad (25)$$

式 (25) 中の未知パラメータの事前分布は自然共役事前分布を用いた。具体的には、 $\{\varphi_i\}_{i=1}^m$ は Dirichlet 分布、 μ_i, w_i は正規分布、 S_i は Wishart 分布とし、さらに $\beta_i, \alpha_{i,j}$ は Gamma 分布を仮定した。また、 m は一様分布とした。

$$p(\{\varphi_i\}_{i=1}^m | m) = \mathcal{D}(\{\varphi_i\}_{i=1}^m | \delta_0) \propto \prod_{i=1}^m \varphi_i^{\delta_0 - 1}$$

$$p(\mu_i | S_i) = \mathcal{N}(\mu_i | \nu_0, (\xi_0 S_i)^{-1})$$

$$p(S_i) = \mathcal{W}(S_i | \eta_0, B_0)$$

$$\propto |S_i|^{\frac{1}{2}(\eta_0 - d - 1)} \exp\left\{-\frac{1}{2}\text{Tr}\{B_0 S_i\}\right\}$$

$$p(w_i | \beta_i, \{\alpha_{i,j}\}_{j=1}^{d+1}) = \mathcal{N}(w_i | \mathbf{0}, (\beta_i \Lambda_i)^{-1})$$

$$p(\beta_i) = \mathcal{G}(\beta_i | \rho_0, \lambda_0) \propto \beta_i^{\rho_0 - 1} e^{-\lambda_0 \beta_i}$$

$$p(\alpha_{i,j}) = \mathcal{G}(\alpha_{i,j} | \kappa_0, \zeta_0)$$

$$P(m) = 1/M_0$$

$\Lambda_i = \text{diag}(\alpha_{i,1}, \dots, \alpha_{i,d+1})$ である。また、添字 '0' のついた変数 (例えば δ_0) はハイパーパラメータ (定数) を表す。 $\alpha_{i,j} \beta_i$ は w_{ij} の精度 (分散の逆数) である。

変分事後分布 Q は以下の分解形を仮定する。

$$Q = Q(m) Q(Z|m) Q(\Phi|m) Q(\Theta|m) \\ = Q(m) Q(Z|m) Q(\{\varphi_i\}_{i=1}^m | m) Q(\mu|m) Q(S|m) \\ \times Q(w|m) Q(\beta|m) Q(\alpha|m)$$

*2 本稿では、 $f_i(\cdot)$ はスカラー関数とするが、本稿の議論は全てベクトル値関数の場合にも容易に拡張可能である。

但し, $\varphi = \{\varphi_i\}_{i=1}^m$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_i\}_{i=1}^m$, $\boldsymbol{S} = \{\boldsymbol{S}_i\}_{i=1}^m$, $\boldsymbol{W} = \{\boldsymbol{w}_i\}_{i=1}^m$, $\boldsymbol{\alpha} = \{\alpha_{ij}\}_{i=1, j=1}^{m, d+1}$, $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^m$ とする.

これらより, 式 (14) の目的関数は以下の様に具体化される.

$$\begin{aligned} \mathcal{F}_m = & \left\langle \log \frac{p(D, Z | \Phi, \Theta, m)}{Q(Z | m)} \right\rangle_{Q(Z, \varphi, \boldsymbol{\mu}, \boldsymbol{S}, \boldsymbol{W}, \boldsymbol{\beta} | m)} \\ & + \left\langle \log \frac{p(\varphi | m)}{Q(\varphi | m)} \right\rangle_{Q(\varphi | m)} + \left\langle \log \frac{p(\boldsymbol{\mu} | \boldsymbol{S}, m)}{Q(\boldsymbol{\mu} | m)} \right\rangle_{Q(\boldsymbol{\mu}, \boldsymbol{S} | m)} \\ & + \left\langle \log \frac{p(\boldsymbol{S} | m)}{Q(\boldsymbol{S} | m)} \right\rangle_{Q(\boldsymbol{S}, \boldsymbol{B} | m)} + \left\langle \log \frac{p(\boldsymbol{W} | m)}{Q(\boldsymbol{W} | m)} \right\rangle_{Q(\boldsymbol{W}, \boldsymbol{\alpha} | m)} \\ & + \left\langle \log \frac{p(\boldsymbol{\alpha} | m)}{Q(\boldsymbol{\alpha} | m)} \right\rangle_{Q(\boldsymbol{\alpha} | m)} + \left\langle \log \frac{p(\boldsymbol{\beta} | m)}{Q(\boldsymbol{\beta} | m)} \right\rangle_{Q(\boldsymbol{\beta} | m)} \end{aligned}$$

5.3 最適変分分布

式 (8), (9) に従って最適変分事後分布を求めると以下を得る. 本論文は, 変分ベイズ学習に基づく最適モデル探索法の一般的枠組みの提案を主眼とし, MoE に対する各変分事後分布の導出の詳細は論文の本質ではないので紙面の都合上省略し, 読者が追試できる様, 結果のみを以下に整理しておく.

$$Q(\{\varphi_i\}_{i=1}^m | m) = \mathcal{D}(\{\varphi_i\}_{i=1}^m | \{\delta_0 + \bar{N}_i\}_{i=1}^m),$$

$$\bar{N}_i = \sum_{n=1}^N \bar{z}_i^n.$$

$$Q(\boldsymbol{\mu}_i | m) = \mathcal{T}(\boldsymbol{\mu}_i | \bar{\boldsymbol{\mu}}_i, \Sigma \boldsymbol{\mu}_i, f \boldsymbol{\mu}_i),$$

$$\bar{\boldsymbol{\mu}}_i = \frac{\bar{N}_i \bar{\boldsymbol{x}}_i + \xi_0 \boldsymbol{\mu}_0}{\bar{N}_i + \xi_0}$$

$$\Sigma \boldsymbol{\mu}_i = \frac{1}{\bar{N}_i + \xi_0} f \boldsymbol{\mu}_i^{-1} \boldsymbol{B}_i$$

$$f \boldsymbol{\mu}_i = \bar{N}_i + \eta_0 + 1 - d$$

$$\boldsymbol{B}_i = \boldsymbol{B}_0 + \sum_{n=1}^N \bar{z}_i^n (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_i)^T$$

$$+ \frac{\bar{N}_i + \xi_0}{\bar{N}_i \xi_0} (\bar{\boldsymbol{x}}_i - \boldsymbol{\nu}_0)(\bar{\boldsymbol{x}}_i - \boldsymbol{\nu}_0)^T$$

$$\bar{\boldsymbol{x}}_i = \frac{1}{\bar{N}_i} \sum_{n=1}^N \bar{z}_i^n \boldsymbol{x}_n$$

ここに, $\mathcal{T}()$ は d 次元 Student's- t 分布で次式で定義される.

$$\begin{aligned} \mathcal{T}(\boldsymbol{x} | \boldsymbol{\mu}, \Sigma, \nu) = & \left(1 + \frac{1}{\nu} \text{Tr}\{\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\} \right)^{-\frac{\nu+d}{2}} \end{aligned}$$

ν は自由度を表す.

$$Q(\boldsymbol{S}_i | m) = \mathcal{W}(\boldsymbol{S}_i | \eta_0 + \bar{N}_i, \boldsymbol{B}_i)$$

$$Q(\beta_i | m) = \mathcal{G}(\beta_i | \rho_0 + \frac{1}{2} \bar{N}_i, \lambda_i),$$

$$\lambda_i = 2\lambda_0 + \sum_{n=1}^N \bar{z}_i^n (y_n - \bar{\boldsymbol{w}}_i^T \boldsymbol{x}'_n)^2 + \bar{\boldsymbol{w}}_i^T \bar{\boldsymbol{\Lambda}}_i \bar{\boldsymbol{w}}_i$$

$$Q(\alpha_{ij} | m) = \mathcal{G}(\alpha_{ij} | \kappa, \zeta),$$

$$\kappa = \kappa_0 + \frac{1}{2}$$

$$\zeta = \zeta_0 + \frac{1}{2} \left\{ (\Sigma \boldsymbol{w}_i)_{jj} + (\bar{\boldsymbol{w}}_{ij})^2 \right\}$$

$$Q(\boldsymbol{w}_i | m) = \mathcal{T}(\boldsymbol{w}_i | \bar{\boldsymbol{w}}_i, \lambda_i f \boldsymbol{w}_i^{-1} \Sigma \boldsymbol{w}_i, f \boldsymbol{w}_i),$$

$$f \boldsymbol{w}_i = 2\rho_0 + \bar{N}_i,$$

$$\bar{\boldsymbol{w}}_i = \Sigma \boldsymbol{w}_i \sum_{n=1}^N \bar{z}_i^n y_n \boldsymbol{x}'_n,$$

$$\Sigma \boldsymbol{w}_i = \left(\sum_{n=1}^N \bar{z}_i^n \boldsymbol{x}'_n \boldsymbol{x}'_n^T + \bar{\boldsymbol{\Lambda}}_i \right)^{-1}$$

$$\bar{\boldsymbol{\Lambda}}_i = \text{diag}(\bar{\alpha}_{i,1}, \dots, \bar{\alpha}_{i,d+1})$$

さらに,

$$\bar{z}_i^n = Q(z_i^n = 1 | m) = \frac{\exp\{\gamma_i^n\}}{\sum_{j=1}^m \exp\{\gamma_j^n\}}$$

$$\gamma_i^n = \Psi(\delta_0 + \bar{N}_i) - \Psi\left(m\delta_0 + \sum_{i=1}^m \bar{N}_i\right)$$

$$+ \frac{1}{2} \sum_{k=1}^d \Psi\left(\eta_0 + \bar{N}_i - \frac{k-1}{2}\right) - \log |\boldsymbol{B}_i|$$

$$- \frac{1}{2} (\bar{N}_i + \eta_0) \left\{ \frac{1}{\bar{N}_i + \xi_0} f \boldsymbol{\mu}_i^{-1} + \text{Tr}\{\boldsymbol{B}_i^{-1} (\boldsymbol{x}_n - \bar{\boldsymbol{\mu}}_i)(\boldsymbol{x}_n - \bar{\boldsymbol{\mu}}_i)^T\} \right\}$$

$$+ \frac{1}{2} \Psi\left(\rho_0 + \frac{\bar{N}_i}{2}\right) - \frac{1}{2} \log \lambda_i$$

$$- \frac{1}{2} \left\{ \frac{1}{\lambda_i} f \boldsymbol{w}_i (y_n - \bar{\boldsymbol{w}}_i^T \boldsymbol{x}'_n)^2 \right.$$

$$\left. + \frac{f \boldsymbol{w}_i}{f \boldsymbol{w}_i - 2} \boldsymbol{x}'_n^T \Sigma \boldsymbol{w}_i \boldsymbol{x}'_n \right\}.$$

但し, $\Psi(x)$ は digamma 関数で次式で定義される.

$$\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$$

ここで, $\Gamma()$ は Gamma 関数を表す. 実際には, 以上を式 (10), (11) に示した様に反復式に書き換えて実行することになる.

5.4 事後予測分布

アルゴリズムが収束し, 最適変分事後分布および得られれば, 新たな入力 \boldsymbol{x}_{N+1} に対する出力 y_{N+1} の事後予測分布は以下で算出される.

$$p(y_{N+1} | \boldsymbol{x}_{N+1}, D) =$$

$$\int \sum_{i=1}^m G_i(\boldsymbol{x}_{N+1} | \Phi) \boldsymbol{w}_i^T \boldsymbol{x}'_{N+1} Q(\Phi, \Theta | m) d\Phi d\Theta$$

(26)

式 (26) は更に次式の様に近似できる .

$$p(y_{N+1} | \mathbf{x}_{N+1}, D) \simeq \sum_{i=1}^m \pi_i \mathcal{T}(y_{N+1} | \bar{\mathbf{w}}_i^T \mathbf{x}'_{N+1}, \lambda_i f \mathbf{w}_i^{-1} V_{i,N+1}^{-1}, f \mathbf{w}_i)$$

即ち, y_{N+1} の事後予測分布は各要素分布が自由度 $f \mathbf{w}_i = 2\rho_0 + \bar{N}_i$ の Student- t 分布から成る混合 Student- t 分布となる . 但し ,

$$\pi_i = G_i(\mathbf{x} | \Phi^{\text{MAP}})$$

$$V_{i,N+1} = 1 - \mathbf{x}'_{N+1}^T \left(\mathbf{x}'_{N+1} \mathbf{x}'_{N+1}^T + \Sigma_{\mathbf{w}_i}^{-1} \right)^{-1} \mathbf{x}'_{N+1}$$

また, Φ^{MAP} は $Q(\Phi | m^*)$ の MAP 推定値を表す . y_{N+1} の平均および分散は以下の様に得られる .

$$\bar{y}_{N+1} = \sum_{i=1}^m \pi_i \bar{\mathbf{w}}_i^T \mathbf{x}'_{N+1}$$

$$\sigma_{y_{N+1}}^2 = \sum_{i=1}^m \frac{\pi_i \lambda_i}{f \mathbf{w}_i - 2} V_{i,N+1}^{-1}$$

5.5 実験結果

[人工データ]

提案アルゴリズムの挙動を可視化すべく, 図 1(a) に示す人工データを用いた (明らかに, 6 個の線形 expert による回帰が最適である.) 図 1(b) に示す混合数 $m = 6$ の初期値に対し, 従来の変分ベイズ学習を実行した結果を図 1(c) に示す . 尚, 図中の直線は各 expert ($f_i, i = 1, \dots, m$) を, 太い曲線は混合モデルによる予測曲線 \bar{y} を, 両側の曲線 (点線) は $\bar{y} \pm \sigma_y$ を, 各々示す . 明らかに低品質の局所最適解に収束している .

一方, 図 1(d) に示す様に, 過少数のモデル $m = 3$ から提案学習法を実行した結果, 混合数が 4,5,6 と変化しながら最終的に最適値 (図 1(h)) に収束した . 但し, 図中のステップ数 t は提案アルゴリズムの Step 4 で採録されなかった探索過程のステップ数は含まれていない事に注意 .

図 1(f) は図 1(d) から (h) に至るまでの \mathcal{F}_m 値, および未学習データ (テストデータ) に対する 1 サンプルあたりの平均自乗誤差 (MSE) の値の推移を示したグラフである . 図 1(e),(f),(g),(h) の順に, \mathcal{F}_m 値は $-41.5, -15.8, -1.6, 1.2$ となり, より良いモデルに近づくにつれて \mathcal{F}_m 値が増加していくこと, 更に, MSE 値がそれに伴い確実に減少していることが確認できる .

尚, 従来の変分ベイズ学習結果の図 1(c) の \mathcal{F}_m 値は -14.6 で $m = 5$ の時の図 1(g) よりも下回っている . これは従来の変分ベイズ学習では局所最適性により最適なモデル探索が困難となり得ることを示している .

[Delve データ]

高次元データ (Delve データ [Rasmussen 96]) 中の “kin-8nm” データへの適用実験を行った . このデータは

Table 1: \mathcal{F}_m and MSE values for each m and by the proposed model search (*).

m	5	6	7	8	9	10	*	
\mathcal{F}_m	min	-3002	-2985	-2911	-2821	-2969	-2927	-2401
	max	-2671	-2590	-2587	-2514	-2567	-2715	-2381
MSE	min	0.481	0.498	0.476	0.465	0.475	0.480	0.457
	max	0.502	0.497	0.481	0.489	0.502	0.531	0.465

8 リンクのロボットアームのフォワードキネマティックスのシミュレーションデータ (8 次元入力, 1 次元出力, 学習, テストデータ数共 256 で, 高い非線形性と中程度のノイズが含まれている) である .

表 1 の最後の欄を除く各欄は, 各 m に対して従来の変分ベイズ学習を異なる 10 通りの初期化で実行して得られた \mathcal{F}_m 値および MSE 値の平均値を示す . 明らかに, 局所最適性のために各 m での \mathcal{F}_m 値は大きくばらついており, \mathcal{F}_m を用いた従来のパッチタイプのモデル選択法では信頼性が低いと言える .

一方, $m = 5, \dots, 10$ の各々を初期モデルとして提案モデル探索学習法を独立に実行した . 各 m について 1 回の実行にも関わらず全ての $m = 5, \dots, 10$ に対し, 同じ $m = 8$ に収束した . その時の \mathcal{F}_m 値および MSE 値の最大値, 最小値を表 1 の最後の欄 (* が付記した欄) に示す . 明らかに最大値と最小値の差が小さく安定した結果が得られていることが分かる . また, 1 回の実行にも関わらず, 提案学習法が最良の結果を得ていることも確認できる .

6. ま と め

本論文では, 変分ベイズ学習の枠組みで, 非線形モデルの学習における局所最適性の問題とモデルの複雑さの決定の問題を同時解決する学習法を提案し, 混合回帰モデルへの適用実験により有効性を確認した, 本論文で提案したアルゴリズムは, 形式的には, 筆者らが先に提案した SMEM アルゴリズムのベイズ拡張と見なせるが, SMEM アルゴリズムと異なり, 最良なモデル探索が実現できるという点で大きく進展していると言える . 現在, モデルの複雑さの場合の数指数オーダーとなる複雑な非線形モデル (隠れマルコフモデル (HMM)) への適用を進めている . これについては別の機会に報告する .

謝 辞

議論して頂いたロンドン大学 (Gatsby Computational Neuroscience Unit) の Dr. Zoubin Ghahramani に感謝します .

◇ 参 考 文 献 ◇

- [Akaike 73] Akaike, H. : A new look at the statistical model identification, *IEEE Trans. Autom. Contor.*, Vol.AC-19, pp.716-723 (1973).

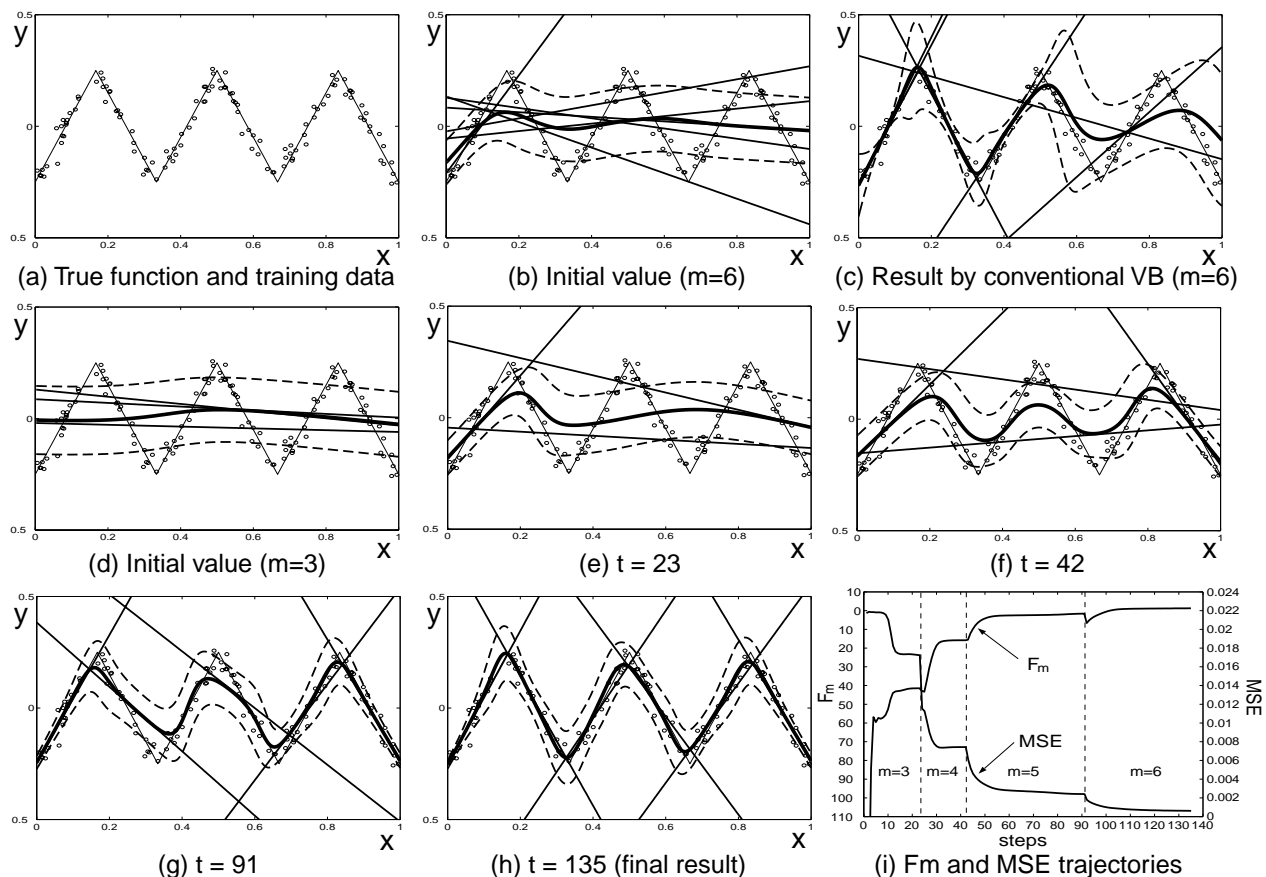


図 1 MoE の学習結果 . (a) の人工データに対し , (b) の様に最適混合数 $m = 6$ で初期化しても従来の変分ベイズ学習では (c) に示す様に局所最適解に収束する . 一方 , 提案学習法では , (d) の過少混合数 $m = 3$ から探索を開始したところ分割操作を繰り返し , 最終的に (h) に示すような所望の解に収束している . (i) は採用された探索過程における F_m , MSE(テストデータに対する平均自乗誤差) 値の推移を示す . F_m の増加に伴い MSE もほぼ単調に減少していることが確認できる .

[Attias 99] Attias, H. : Learning parameters and structure of latent variable models by variational Bayes, In proc. *Uncertainty in Artificial Intelligence*, (1999).
 [Bartlett 96] Bartlett, P. : The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, Technical report, Australian National University (1996).
 [Dempster 77] Dempster, A.P., Laird, N.M., and Rubin, D.B. : Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society B*, Vol.9, pp.1-38 (1977).
 [Gamerman 97] Gamerman, D. : *Markov chain Monte Carlo*, Chapman & Hall (1997).
 [Jacobs 91] Jacobs, R.J., Jordan, M.I. Nowlan, S.J., and Hinton, G.E. : Adaptive mixtures of local experts, *Neural Computation*, Vol.3, pp.79-87 (1991).
 [Jordan 97] Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. : An introduction to variational methods for graphical models, *Machine Learning*, Vol.37, No.2 (1997).
 [Richardson 97] Richardson, S. and Green, P. : On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society B*, Vol.59, pp.731-792 (1997).
 [MacKay 92] MacKay, D. : Bayesian interpolation, *Neural Computation*, Vol.4, pp.405-447 (1992).
 [Rasmussen 96] Rasmussen, C.E., Neal, R.M., Hinton, G.E., van Camp, D., Revow, M., Ghahramani, Z., Kustra, R. and Tibshirani, R. : *The DELVE Manual*,

<http://www.cs.utoronto.ca/~delve/> (1996).
 [Rissanen 87] Rissanen, J. : Stochastic complexity, *Journal of the Royal Statistical Society B*, Vol.49, pp.223-239 (1987).
 [Saul 96] Saul, L.K., Jaakkola, T., and Jordan, M.I. : Mean field theory for sigmoid belief networks, *Journal of Artificial Intelligence Research*, Vol.4, pp.61-76 (1996).
 [Ueda 99a] Ueda, N. and Nakano, R., Ghahramani, Z., and Hinton, G.E. : SMEM algorithm for mixture models, *Advances in Neural Information Processing Systems 11* (1999).
 [Ueda 99b] 上田, 中野 : 確率的混合部分空間法 -混合因子分析を用いたパターン認識法-, *信学論*, Vol.J82-D-II, No.12, pp.2394-2401 (1999).
 [Ueda 00] Ueda, N. and Nakano, R., Ghahramani, Z., and Hinton, G.E. : SMEM algorithm for mixture models, *Neural Computation*, Vol.12, No.9, pp.2109-2128 (2000).
 [Waterhouse 95] Waterhouse, S.R., MacKay, D. and Robinson, A.J. : Bayesian methods for mixture of experts, *Advances in Neural Information Processing Systems 8*, (1995).
 [Xu 94] Xu, L., Jordan, M. I., and Hinton, G.E. : An alternative model for mixtures of experts, *Advances in Neural Information Processing Systems 7*, pp.633-640 (1994).
 [Zoubin 00] Ghahramani, Z. and Beal, M.J. : Variational inference for Bayesian mixture of factor analyzers, *Advances in Neural Information Processing Systems 12*, (2000).

〔担当委員：新田克己〕

2000 年 11 月 1 日 受理

著 者 紹 介



上田 修功

1982 年大阪大学工学部通信工学科卒業。1984 年同大学院修士課程修了。同年 NTT 入社。1993-1994 年米国 Purdue 大学客員研究員。現在、NTT コミュニケーション科学基礎研究所 知能情報研究部創発学習研究グループリーダー、主幹研究員（特別研究員）、奈良先端科学技術大学院大学客員助教授。統計的学習理論の研究に従事。1992 年日本神経回路学会研究奨励賞、1997 年電気通信普及財団賞、2000 年電子情報通信学会論文賞受賞。共著書“わかりやすいパターン認識（オーム社）”など。工学博士。電子情報通信学会、日本神経回路学会、日本統計学会、IEEE 各会員。