混合回帰モデルのためのSMEMアルゴリズム

鈴木　敏[†]　　上田　修功[†]

SMEM algorithm for mixture regression models

Satoshi SUZUKI[†] and Naonori UEDA[†]

あらまし　混合モデルにおけるモデル配置の不均衡を解決する最適化手法として提案された併合分割操作付きEM(Split and Merge Expectation Maximization: SMEM)アルゴリズムの回帰問題への適用を検討する．回帰問題への適用に際して，同時分布を仮定した正規化ガウス関数ネットワーク(Normalized Gaussian network: NGnet)を混合回帰モデルとして利用する．本論文では，SMEMアルゴリズムのNGnetへの適用のために必要な，併合分割操作に加える変更点を明らかにし，人工データ及び実データを用いた実験により，SMEMアルゴリズムの回帰問題に対する有効性を検証する．**また，同時分布の仮定が回帰に与える影響についても検討を行う．**

キーワード　EMアルゴリズム，回帰問題，NGnet，混合モデル，最尤推定，同時分布

1. はじめに

筆者の一人は，混合モデルにおけるEMアルゴリズム[1]の局所最適性の問題を解決するための一手法としてSMEM(Split and Merge EM)アルゴリズムを先に提案し，混合潜在変数モデルを含む混合分布推定問題に対しその有効性を示した[2], [3]．SMEMアルゴリズムは，EMアルゴリズムが陥る局所最適解からの脱出を，要素モデルの併合分割操作により行い，より良い解へと誘導する手法である．混合正規分布推定問題及び次元圧縮問題に対して有効な手段であることが，実験的にも示されている．

しかしながら，これらの課題は全て混合分布の推定問題であり，回帰問題に対するSMEMアルゴリズムの有効性は自明ではない．本論文ではSMEMアルゴリズムの回帰問題における有効性を示すことを主な目的とする．

一般に，混合回帰モデルとしては，Jordanらの提唱したMixture Expert network (MEnet)が有名であり[4]，EMアルゴリズムによる最適化学習手法も提案されている[5]．ところが，SMEMアルゴリズムでは，Q関数の要素モデル毎の直和分解を必要とするため，一般のMEnetへは適用できない．

ところで，Xuらは，EMアルゴリズムの最適化計算を効率化する手法として，MEnetのgating networkをガウス関数に置き換えた学習モデルを提案し，確率モデルの変更により，収束の速い学習が可能であることを示した[6]．この時，同時分布を仮定することにより，Q関数の直和分解を可能とし，パラメタを要素モデル毎に独立に計算できることも示している．このモデルは，NGnet(Normalized Gaussian network)とも呼ばれており[7]，Q関数の直和分解が可能であるためSMEMアルゴリズムを適用できる．また同時に，SMEMアルゴリズムはEMアルゴリズムを繰り返し利用する手法であるため，同様に収束の速い学習が期待できる．

以上より，本論文では，NGnetを対象に，SMEMアルゴリズムの混合回帰問題への有効性を検証する．また，SMEMアルゴリズムのNGnetへの適用に際し，アルゴリズムの中核部分である併合分割操作について，回帰問題に適した形式への**拡張を行う．**

以下，2.ではNGnetを概説し，3.において，SMEMアルゴリズムについて述べる．ここでは特に，併合分割操作について詳細に記す．4.の計算機実験では，人工データによる詳細な検討に加えて高次元実データへの適用も示す．**5.では，同時分布の仮定が回帰に与える影響について検討する．**最後に，6.において本論文のまとめを記す．

[†] NTTコミュニケーション科学基礎研究所，京都府
NTT Communication Science Laboratories, Kyoto, 619-0237 Japan

2. NGnet と EM アルゴリズム

NGnet は MEnet と同様のモジュール構造を持つネットワークであり，d_x 次元の入力 $\boldsymbol{x} \in R^{d_x}$ から d_y 次元の出力 $\boldsymbol{y} \in R^{d_y}$ への変換は，

$$\boldsymbol{y} = \sum_{i=1}^{M} \frac{G_i(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i)}{\sum_{j=1}^{M} G_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} f_i(\boldsymbol{x}; W_i) \qquad (1)$$

により与えられる．但し，$G_i(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i)$ は d_x 次元正規分布：

$$G_i(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i) \equiv (2\pi)^{-d_x/2} |\Sigma_i|^{-1/2} \times \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}$$

である．ここに，M はモデル要素数，N は観測データ数，i はモデル指標であり，$\boldsymbol{\mu}_i, \Sigma_i, W_i$ はそれぞれ要素モデル i の入力分布の平均ベクトル，共分散行列，及び変換行列である．

MEnet と NGnet は形式的には gating network の関数の違いであると見なせるが，これらの本質的な違いは入力変数を確率変数として取り扱っているか否かにある．即ち，観測データ $D = \{(\boldsymbol{x}_n, \boldsymbol{y}_n), n = 1, \cdots, N\}$ に関して，MEnet では出力 $\boldsymbol{y}$ のみを確率変数として扱うのに対し，NGnet では入力 $\boldsymbol{x}$ も確率変数として扱う．即ち，NGnet では観測データ D を同時分布と見なす．

要素モデル i に入力 $\boldsymbol{x}$ が与えられた下での出力 $\boldsymbol{y}$ の分布は，通常の回帰モデルと同様，平均 $f_i(\boldsymbol{x}_i; W_i)$，共分散行列 S_i のガウス分布：

$$p(\boldsymbol{y}|\boldsymbol{x}, i, \boldsymbol{\mu}_i, \Sigma_i) = G_i(\boldsymbol{y}|f_i(\boldsymbol{x}; W_i), S_i)$$

を仮定する．以上より，推定すべき未知パラメタは

$$\Theta = \{(\boldsymbol{\mu}_i, \Sigma_i, W_i, S_i)\}_{i=1}^{M}$$

となる．この時，完全データ $(\boldsymbol{x}_n, \boldsymbol{y}_n)$に対する 対数尤度関数は

$$L(\Theta|\boldsymbol{x}_n, \boldsymbol{y}_n, i) \equiv \log p(\boldsymbol{x}_n, \boldsymbol{y}_n, i|\Theta) = \log\{p(\boldsymbol{y}_n|\boldsymbol{x}_n, i, W_i, S_i) p(\boldsymbol{x}_n|i, \boldsymbol{\mu}_i, \Sigma_i) P(i)\} \qquad \mathbf{(2)}$$

により表される．但し，$P(i)$ はモデルの prior であり，ここでは一様分布 $P(i) = 1/M$ とする．従って，NGnet に対する EM アルゴリズムの Q 関数は，ステップ t でのパラメタ推定値を $\Theta^{(t)}$ として，

$$Q(\Theta|\Theta^{(t)}) = \sum_{n=1}^{N}\sum_{i=1}^{M}\Big[P(i|\boldsymbol{x}_n, \boldsymbol{y}_n, \Theta^{(t)}) \times \log\{p(\boldsymbol{y}_n|\boldsymbol{x}_n, i, W_i, S_i) p(\boldsymbol{x}_n|i, \boldsymbol{\mu}_i, \Sigma_i) P(i)\}\Big] \qquad \mathbf{(3)}$$

となる．式 (3) 中の $p(\boldsymbol{x}|i, \boldsymbol{\mu}_i, \Sigma_i)$ は要素モデル i の入力分布で，$p(\boldsymbol{x}|i, \boldsymbol{\mu}_i, \Sigma_i) = G_i(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i)$ である．

この時，$p(\boldsymbol{y}|\boldsymbol{x}, i, W_i, S_i), p(\boldsymbol{x}|i, \boldsymbol{\mu}_i, \Sigma_i)$ は要素モデル毎に独立であるため，式 (3) は更に

$$Q(\Theta|\Theta^{(t)}) = \sum_{i=1}^{M} q_i(\boldsymbol{\mu}_i, \Sigma_i, W_i, S_i|\Theta^{(t)}) \qquad (4)$$

と書ける．この時 $\Theta^{(t)}$ が定数であることに注意すると，式 (4) は，Q がモデル i のパラメタのみに依存する q_i に直和分解でき，SMEM アルゴリズムが適用可能となることがわかる．また，入出力にガウス分布を仮定している為，$Q(\Theta|\Theta^{(t)})$ は Θ の 2 次形式で与えられるという利点がある．

一方，MEnet では入力 $\boldsymbol{x}$ を確率変数として取り扱っていないので，EM アルゴリズムの Q 関数は

$$Q(\Theta|\Theta^{(t)}) = \sum_{n=1}^{N}\sum_{i=1}^{M}\Big[P(i|\boldsymbol{x}_n, \boldsymbol{y}_n, \Theta^{(t)}) \times \log\{p(\boldsymbol{y}_n|\boldsymbol{x}_n, i, W_i, S_i) P(i|\boldsymbol{x}_n, \Theta)\}\Big] \qquad (5)$$

により与えられる．このとき，$P(i|\boldsymbol{x}_n, \Theta)$ は，gating network 出力層の要素モデル i に対応するユニットの出力を $s_i(\boldsymbol{x}, \boldsymbol{\mu}_i, \Sigma_i)$ として，

$$P(i|\boldsymbol{x}_n, \Theta) = \frac{\exp\left(s_i(\boldsymbol{x}_n, \boldsymbol{\mu}_i, \Sigma_i)\right)}{\sum_{j=1}^{M} \exp\left(s_j(\boldsymbol{x}_n, \boldsymbol{\mu}_i, \Sigma_i)\right)} \qquad (6)$$

により与えられる．よって，MEnet では $P(i|\boldsymbol{x}_n, \Theta)$ の中に全てのパラメタ $\{\boldsymbol{\mu}_i, \Sigma_i, W_i, S_i | i = 1, \cdots, M\}$ が含まれるため，要素モデル毎の計算ができない．従って，一般に MEnet へ EM アルゴリズムを適用した場合，M ステップにおける Q 関数の最大化の際，更なる反復計算が必要となる [6].

NGnet では，各要素モデルでの $f_i(\boldsymbol{x}; W_i)$ による変換を線形変換：

$$f_i(\boldsymbol{x}; W_i) = W_i \boldsymbol{x}^*,$$

但し，$\boldsymbol{x}^* = [\boldsymbol{x}^T, 1]^T \in R^{d_x+1}$

とし，観測データ D を同時分布と見なせば，EMアルゴリズムのMステップにおける最適解の計算が解析解として求められる．即ち，Mステップでの各パラメタの更新式は以下で与えられる [6].

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{\sum_{n=1}^{N} P\left(i|\boldsymbol{x}_n, \boldsymbol{y}_n, \Theta^{(t)}\right) \boldsymbol{x}_n}{\sum_{l=1}^{N} P\left(i|\boldsymbol{x}_l, \boldsymbol{y}_l, \Theta^{(t)}\right)} \tag{7}$$

$$\Sigma_i^{(t+1)} = \frac{\left[\sum_{n=1}^{N} P\left(i|\boldsymbol{x}_n, \boldsymbol{y}_n, \Theta^{(t)}\right)}{\sum_{l=1}^{N} P\left(i|\boldsymbol{x}_l, \boldsymbol{y}_l, \Theta^{(t)}\right)} \times\left(\boldsymbol{x}_n - \boldsymbol{\mu}_i^{(t+1)}\right)\left(\boldsymbol{x}_n - \boldsymbol{\mu}_i^{(t+1)}\right)^T\right] \tag{8}$$

$$W_i^{(t+1)} = \left[\sum_{n=1}^{N} P\left(i|\boldsymbol{x}_n, \boldsymbol{y}_n, \Theta^{(t)}\right) \boldsymbol{y}_n \boldsymbol{x}^{*T}{}_n\right] \times\left[\sum_{n=1}^{N} P\left(i|\boldsymbol{x}_n, \boldsymbol{y}_n, \Theta^{(t)}\right) \boldsymbol{x}^*{}_n \boldsymbol{x}^{*T}{}_n\right]^{-1} \tag{9}$$

$$S_i^{(t+1)} = \frac{\left[\sum_{n=1}^{N} P\left(i|\boldsymbol{x}_n, \boldsymbol{y}_n, \Theta^{(t)}\right)}{\sum_{l=1}^{N} P\left(i|\boldsymbol{x}_l, \boldsymbol{y}_l, \Theta^{(t)}\right)} \times\left(\boldsymbol{y}_n - W_i^{(t+1)} \boldsymbol{x}^*{}_n\right)\left(\boldsymbol{y}_n - W_i^{(t+1)} \boldsymbol{x}^*{}_n\right)^T\right] \tag{10}$$

3. SMEMアルゴリズム

3.1 アルゴリズムの概要

SMEMアルゴリズムは，EMアルゴリズムの学習過程において，モデル要素の併合分割操作による局所最適解からの脱出を図るアルゴリズムであり，より良い解へ誘導する機能を備えている [2].

SMEMアルゴリズムの概要は以下の通りである．

step 1. 通常のEMアルゴリズムを実行．$(\Theta^*, Q^*$ を収束後のパラメタ推定値および Q 関数値とする)

step 2. Θ^* に基づき併合分割基準を計算し (3.2.1及び3.2.2参照)，併合分割候補 $\{i,j,k\}_c$ の優先順位 c を定める (3.2.3参照).

step 3. 優先順位の最も高い候補 $\{i,j,k\}_c$ を取り出し，併合分割を実行する．新しいモデルに初期値を与え，再学習を行う (3.2.4参照)．$(\Theta^{**}, Q^{**}$ を収束値とする)

step 4. $Q^{**} > Q^*$ ならば，$Q^* \leftarrow Q^{**}, \Theta^* \leftarrow \Theta^{**}$ とし，step 2. へ．それ以外はstep 3. に戻り，次の併合分割候補を選ぶ．候補がなくなればstep 5. へ．

step 5. Θ^* を最終結果として終了．

上記アルゴリズムでは，併合及び分割を同時に行っており，モデルの複雑さ，即ち，モデル数を一定に保っている．最尤推定の場合，一般に，パラメタ数の増加に伴い尤度関数値が単調増加するため，パラメタ数を変化させると尤度関数値によりパラメタの良さを評価することが困難になる．そこでSMEMアルゴリズムでは，局所解からの脱出を目的として，混合数 (パラメタ数) を固定している．

3.2 回帰問題への適用

3.2.1 併合基準

併合基準は2つのモデルに同程度の帰属確率を有するデータが他に比べて多く存在する場合に，その2つのモデルを優先的に併合するための基準である．

回帰問題では，観測データ D を同時分布とするため，分布推定問題の併合基準で用いたモデル i への帰属確率 $P(i|\boldsymbol{x},\Theta)$ を $P(i|\boldsymbol{x},\boldsymbol{y},\Theta)$ として置き換える．従って，各データのモデル i への帰属確率 $P(i|\boldsymbol{x},\boldsymbol{y},\Theta)$ から成るベクトル $\boldsymbol{P}_i \in R^N$ を

$$\boldsymbol{P}_i = \left(P\left(i|\boldsymbol{x}_1, \boldsymbol{y}_1, \Theta\right), \cdots, P\left(i|\boldsymbol{x}_N, \boldsymbol{y}_N, \Theta\right)\right)^T$$

として，併合基準 $J_{merge}(i,j)$ を

$$J_{merge}(i,j) = \frac{\boldsymbol{P}_i^T \boldsymbol{P}_j}{\|\boldsymbol{P}_i\| \|\boldsymbol{P}_j\|} \tag{11}$$

とする．J_{merge} の大きい組が高い優先順位を持つ．

3.2.2 分割基準

分割基準はデータの当てはまりの悪いモデルを優先的に分割するための基準であり，以下に定義する局所Kullback-Leibler divergence(局所KL情報量) により与えられる．

分割基準でも併合基準と同様の理由で，分布推定問題の分割基準で用いたモデル k による推定分布 $p(\boldsymbol{x}|k,\Theta)$ および事後確率で重み付けした局所経験分布 $f_k(\boldsymbol{x})$ を**修正する必要がある．**回帰問題では，推定分布を $p(\boldsymbol{x},\boldsymbol{y}|k,\Theta)$，事後確率で重み付けした局所経験分布を $f_k(\boldsymbol{x},\boldsymbol{y})$ とする．従って，分割基準 $J_{split}(k)$ は，入力の局所経験分布 $f_k(\boldsymbol{x},\boldsymbol{y})$ と推定分布 $p(\boldsymbol{x},\boldsymbol{y}|k,\Theta)$ との局所KL情報量：

$$J_{split}(k)$$

$$= \sum_{n=1}^{N} f_k(\boldsymbol{x}_n, \boldsymbol{y}_n) \log \frac{f_k(\boldsymbol{x}_n, \boldsymbol{y}_n)}{p(\boldsymbol{x}_n, \boldsymbol{y}_n | k, \Theta)} \qquad (12)$$

とする．但し，局所経験分布 $f_k(\boldsymbol{x}, \boldsymbol{y})$ は

$$f_k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n)\delta(\boldsymbol{y} - \boldsymbol{y}_n) \frac{P(k|\boldsymbol{x}, \boldsymbol{y}, \Theta)}{\sum_{l=1}^{N} P(k|\boldsymbol{x}_l, \boldsymbol{y}_l, \Theta)} \qquad (13)$$

である．J_{split} の大きいモデルが高い優先順位を持つ．

3.2.3 併合分割候補の順序付け

併合分割のための優先順位は，SMEM アルゴリズムの計算量に大きな影響を与える要素であり，その決定方法の選択は重要である．$J_{merge}(i,j)$ と $J_{split}(k)$ を用いた優先順位の決定方法として，次の 2 種類が考えられる．即ち，M 個のモデルの中から併合の候補 $\{i,j\}_c$ を順序付けし，それぞれの候補 $\{i,j\}_c$ に対して残りの $(M-2)$ 個のモデルを分割候補 $\{k\}_c$ として順序付けする併合優先の手法と，M 個のモデルを分割の候補 $\{k\}_c$ として順序付けし，それぞれの候補 $\{k\}_c$ に対して残りの $(M-1)$ 個のモデルから併合候補 $\{i,j\}_c$ を順序付けする分割優先の手法である．本論文では後者の手法を用いる．(注 1)

3.2.4 併合分割の実行と再学習

併合分割は併合分割候補 $\{i,j,k\}_c$ から，モデル i,j を併合してモデル i' を生成し，モデル k を分割してモデル j',k' を作る処理である．このとき，各モデルのパラメタの初期値はモデル i' については，式 (7)〜(10) に対して，事後確率

$$P\left(i'|\boldsymbol{x}, \boldsymbol{y}, \Theta\right) = P\left(i|\boldsymbol{x}, \boldsymbol{y}, \Theta\right) + P\left(j|\boldsymbol{x}, \boldsymbol{y}, \Theta\right)$$

を適用することにより得られる．

一方，モデル j',k' については，モデル k の各パラメタに微少の摂動 $\boldsymbol{\epsilon}$ を加えた値を用いる．即ち，$m = j',k'$ として，

$$\boldsymbol{\mu}_m^{(t+1)} = \boldsymbol{\mu}_k^{(t)} + \boldsymbol{\epsilon}_m^{\boldsymbol{\mu}} \qquad (14)$$

$$\Sigma_m{}^{(t+1)} = \Sigma_k{}^{(t)} + \epsilon_m^{\Sigma} I \qquad (15)$$

$$W_m^{(t+1)} = W_k^{(t)} + \epsilon_m^{W} \qquad (16)$$

$$\sigma_m^{(t+1)} = \sigma_k^{(t)} + \epsilon_m^{\sigma} \qquad (17)$$

(注 1)：併合分割候補の順序付けは，従来手法 [2] では前者の手法を用いていたが，後者の手法によると greedy 探索における解の発見までの計算量が節約できることが実験により示されている (4.参照)．

とする．以上により，初期値が与えられる．

次に，これらの初期値から再学習を行う．再学習は 2 段階の学習に分けて行われる．まずはじめに，partial-EM ステップと呼ばれる学習を行う．partial-EM ステップでは，併合分割により新しくできた 3 つのモデル $m = i',j',k'$ のパラメタを，他のモデルに影響を与えないように更新する．具体的には，EM アルゴリズムにおいて，M ステップで式 (7)〜(10) の事後確率を

$$P(m|\boldsymbol{x}, \boldsymbol{y}, \Theta) = \frac{P(\boldsymbol{x}, \boldsymbol{y}, m|\Theta)}{\sum_{l=i',j',k'} P(\boldsymbol{x}, \boldsymbol{y}, l|\Theta)} \times \sum_{r=i,j,k} P(\boldsymbol{x}, \boldsymbol{y}, r|\Theta) \qquad (18)$$

として学習する．但し，$m = i',j',k'$ である．

partial-EM ステップの学習が収束した後，後処理として，全モデルでの EM アルゴリズムによる学習を行う．この過程は，partial-EM ステップに対して full-EM ステップと呼ばれている．学習は通常の EM アルゴリズムと同じである．

4. 計算機実験

計算機実験では人工データによる 3 次元曲面の推定問題を用いて，本アルゴリズムの有効性を詳細に検討した．また，高次元の実データについても，時系列データによる予測問題を用いて有効性を調べた．

4.1 人工データによる検討

まず，効果を可視化するため，2 次元入力，1 次元出力の人工データによる実験結果を以下に示す．

図 1にモデル数を 5 とした場合の実験における学習過程の例を示す．(a) はターゲット関数で，2 次元入力から 1 次元出力を与える関数である．学習には入力空間から任意に選ばれた 1000 点を用い，出力範囲の 5% を標準偏差とするガウスノイズを出力に対して加えた．また，テストにおいては入力範囲から等間隔に 21×21 点取り出し，入力とした．

(b) は学習の初期状態を表している．図 (b-1) は初期値として任意に与えられたパラメタにより推定された関数である．図 (b-2) では各モデルで推定された入力分布が太線で表されており，背後の細線はターゲット関数の等高線を示している．

(c) は通常の EM アルゴリズムによる学習結果である．この例では，EM ステップを 172 回繰り返すことにより学習が収束している．推定された関数はターゲット関数に近づいているが，十分とは言えない (c-1)．ま

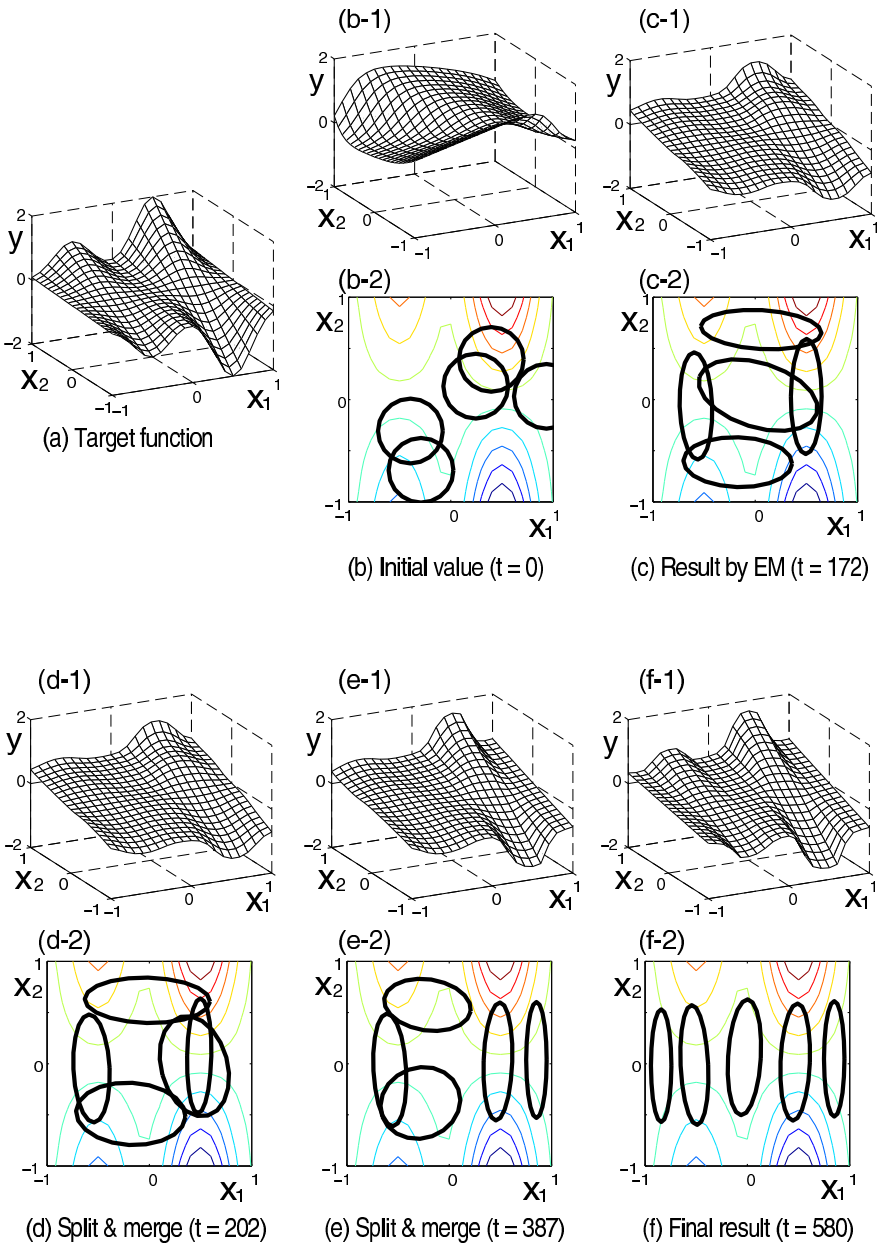

図 1　学習過程の例．(a) ターゲット関数，(b) 初期値による推定結果，(c)EM アルゴリズムによる推定結果，(d)SMEM アルゴリズムの学習過程 (ステップ数 202) での推定状況，(e) 学習過程 (ステップ数 387) での推定状況，(f) 最終推定結果．

Fig. 1　An example of the learning processes. (a)Target function, (b)estimation result from initial values, (c)result by the EM algorithm, (d),(e) results in process of SMEM algorithm, (f)final result by the SMEM algorithm.

表 1　アルゴリズムの比較
Table 1　Comparison between two algorithms.

	対数尤度			計算量	
	初期値	EM	SMEM	EM	SMEM
平均	-131819	-1739	-1426	52	662
標準偏差	21320	182	124	35	477
最大値	-78206	-1372	-1226	156	2561
最小値	-191712	-2625	-1749	6	27

表 2　同等の計算量における尤度の比較
Table 2　Comparison in equal calculation volume.

	EM	SMEM
平均	-1749	-1582
標準偏差	159	129

た，入力分布の推定も，変化の大きい領域を入力領域とするモデルがあり，最適な入力分布とはいえない (c-2).

(d) および (e) は SMEM アルゴリズムによる学習過程の途中結果である．それぞれに，一つの SM(Split & Merge) ループが終了した時点 (3.1のアルゴリズム step 4. が終了した時点) での結果が示されている．学習が進むにつれて，ターゲット関数により近い結果が得られていることが分かる．

(f) は SMEM アルゴリズムによる最終結果で，受理されなかった EM ステップも含め，合計 580 回の EM ステップを要した．(c),(d),(e) に比べ，より良い解が得られていることが分かる (f-1)．モデル毎の入力分布の推定も，平面に近い部分が入力領域となるように学習され，妥当な推定結果である (f-2).

次に，上記の実験を異なる初期値から 100 回行った場合の，通常の EM アルゴリズムと SMEM アルゴリズムとの比較を表 1に示す．この表は，それぞれのアルゴリズムによる学習結果の対数尤度および計算量 (総 EM ステップ数) の平均，標準偏差，最大値，最小値を比較したものである．表からは，SMEM アルゴリズムによる尤度の改善が認められる．但し，計算量は，平均で約 13 倍を要している．

上記の結果から，SMEM アルゴリズムによる計算は EM アルゴリズムを 13 回行う場合と，計算量がほぼ等しいことが分かった．従って，初期値を任意に変えた 13 回の EM アルゴリズムによる学習結果から最適なものを選ぶ場合と，1 回の SMEM アルゴリズムによる学習結果とを比較することにより，同等の計算量においての比較ができる．表 2は，13 回の EM アルゴリズムによる学習の中から最適値を取り出す作業を 1 セットとして 30 セットの学習結果および，SMEM アルゴリズムによる 30 回の学習結果を尤度の平均と標準偏差として示し

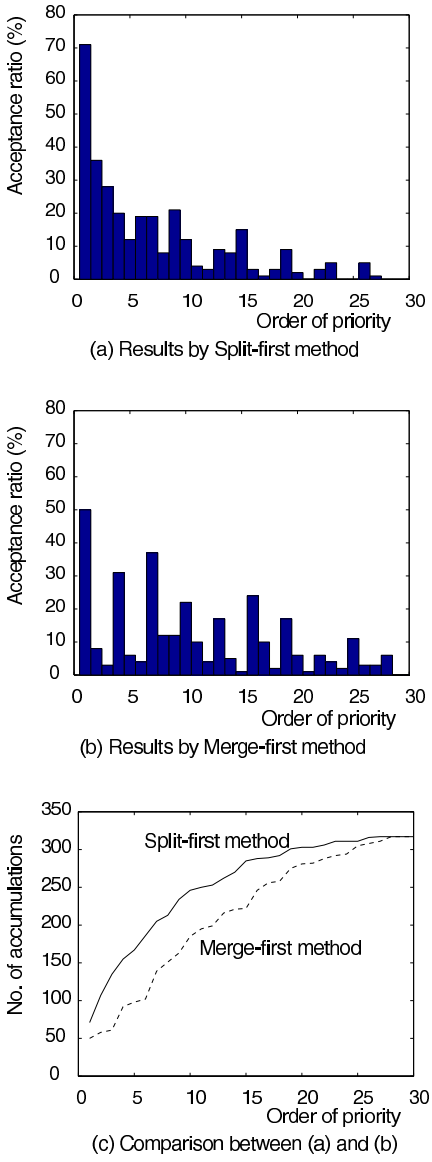

図 2　併合分割候補の計算方法の比較．(a) 分割優先の場合の優先順位に対するアクセプト率，(b) 併合優先の場合の優先順位に対するアクセプト率，(c) 累積による比較.
Fig. 2　Comparison between two types of computations determining the order of priority of SM candidates. (a)Acceptance ratios of each order by Split-first method, (b)acceptance ratios of each order by Merge-first method, (c)comparison of accumulations between the two methods.

ている．表より，SMEMアルゴリズムによる尤度の平均値はEMアルゴリズムによる尤度の平均値よりも大きく，同等の計算量においてもSMEMアルゴリズムがEMアルゴリズムに比べ，より良い解が得られることを確認した．

併合分割候補の順序付けについての考察

併合分割のための優先順位の決定方法として，併合優先の手法と分割優先の手法が考えられる (3.2.3参照)．SMEMアルゴリズムで用いている greedy 探索を効率的に行うために，これら2種類の優先順位決定法の比較を行った．

上記の実験において，それぞれの手法で求めた優先順位の中から，併合分割後の学習により $Q^{**} > Q^{*}$ を満たした順位 c を全て取り出し，統計を比較した．図2は100回の試行による結果である．(a) は分割優先の場合，(b) は併合優先の場合を示している．(a),(b) はそれぞれ横軸に優先順位，縦軸に100回の試行の中で各順位が $Q^{**} > Q^{*}$ を満たした比率を表している．換言すれば，それぞれの順位で $Q^{**} > Q^{*}$ を満たす確率が表されている．分割優先の場合には順位が高いほど $Q^{**} > Q^{*}$ を満たす確率が高くなることが図から明らかであるが (a)，併合優先の場合には，その傾向は見られるものの，分割優先の場合ほど明確ではない (b)．これらは，(c) の図からも明らかである．(c) は (a),(b) のヒストグラムから，それぞれで順位の若い順に $Q^{**} > Q^{*}$ を満たした回数の累積を取り，グラフ化したものであり，分割優先の方が $Q^{**} > Q^{*}$ を満たす確率が高いことが示されている．

$Q^{**} > Q^{*}$ を満たす併合分割候補を早く見つけられないと，それだけ無駄なSMループの計算が増えることになるので，上記の結果は分割優先が望ましいことを示している．この結果は，併合分割操作による尤度の変化は，併合よりも分割による尤度上昇の影響が大きいと考えられる点からも妥当と言える．

4.2 実データによる検討

次に，高次元実データによる実験結果を示す．利用したデータは，遠赤外線レーザーから観測されたカオス的挙動を示す1次元の時系列物理量である [8](注2)．この1次元時系列データにおいて，時間的に連続な25点から次の1点を推定することをネットワークに学習させる．即ち，

(注2)：Santa Fe Institute Time Series Prediction and Analysis Competition で用いられたデータである．

表3　実データによる実験結果
Table 3　Results by real data.

モデル数		EM	SMEM
10	平均	-39274	-38571
	標準偏差	52	621
	最大	-39194	-36988
	最小	-39355	-39001
50	平均	-39112	-32159
	標準偏差	52	21
	最大	-39040	-32095
	最小	-39214	-32167

表4　他手法との比較
Table 4　Comparison with other methods.

	平均	標準偏差
SMEM(10)	0.0233	0.79e-3
SMEM(50)	0.0135	0.39e-5
MSE	0.0396	0.018
MDL	0.0123	0.43e-4

$$\boldsymbol{x}^{(t)} = F\left(\boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(t-2)}, \cdots, \boldsymbol{x}^{(t-25)}; \Theta\right) \quad (19)$$

を考える．

高次元データの学習では，データが粗に分布する為，式 (8) による共分散行列 Σ_i が，計算機上で非正則になる場合がある．本実験では，微少量 ϵ に対し，$|\Sigma_i| < \epsilon$ を満たす場合のみ，$\Sigma_i = \Sigma_i + \epsilon' I$ として，この問題に対処した．但し，ϵ' は微少量である．

図3(a) に学習で用いたデータを示す．学習時には連続した1000点を用い，テストではそれに続く100点を推定する．表3はモデル数を10及び50としたときの，10回の学習に対する対数尤度の統計量を比較したものである．いずれの場合もSMEMアルゴリズムによる学習結果がEMアルゴリズムの結果と比べて優れていることが確認できる．

モデル数を50とした時の推定結果を図3(b) に示す．実線は真の値，太い破線は真の入力値を与えたときの推定値，細い破線は最初の25点から順次推定を繰り返した結果である．真の入力値を与えた場合は，一部で大きな誤差を示しているが，全体的には良い結果を与えている．

他手法との二乗誤差による比較を表4に示す．比較対象としたのは10の中間層を持つ三層MLP(Multi-Layer Perceptron) を用いた二乗誤差最小化 (MSE) 法とMDL正則化法 [9] である．モデル数50のSMEMアルゴリズムによる推定結果は，誤差，標準偏差ともにMDL正則化法と大差はない．即ち，推定精度も安定性もかなり高いことが示された．

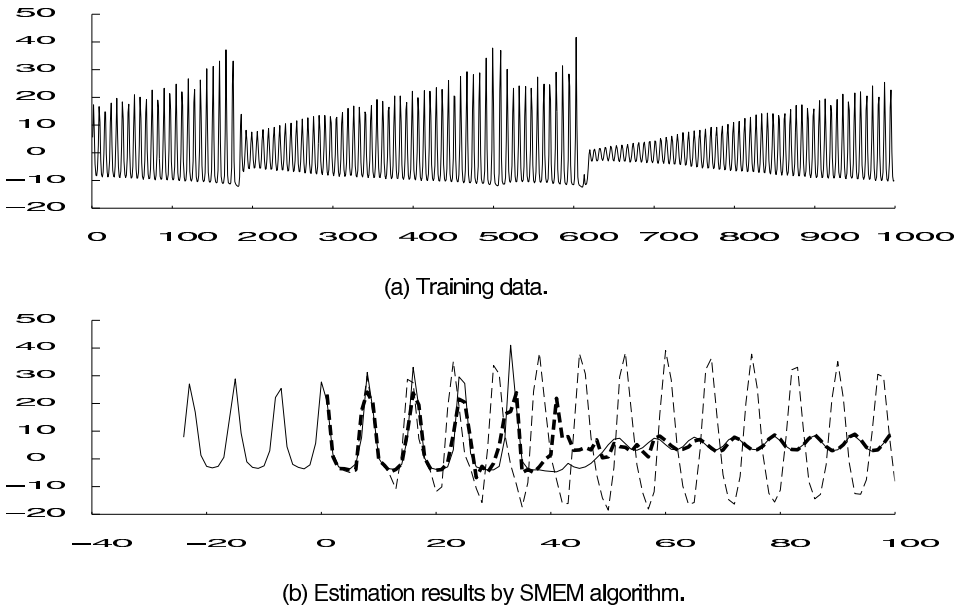

図3 実データによる実験．(a) 学習データ，(b) テストデータ (実線) 及び推定値 (破線)．太い破線は実データからの推定，細い破線は時系列推定．
Fig. 3 An experiment using real data. (a)Training data, (b)test data (solid line) and estimation results by the SMEM algorithm (broken lines).

一方，繰り返し推定する手法での結果は50モデルの場合で，平均0.20，標準偏差0.234e-4となり，安定はしているが誤差は非常に大きいという結果となった．これは，NGnetによるモデルの適用限界を示しており，非定常のダイナミクスに対するモデルとしてはふさわしくないと推察される．

5. 考　　察

5.1 同時分布における分布推定と回帰

同時分布を仮定した回帰と分布推定との差異は以下に示す通りである．

混合正規分布による分布推定では，完全データの尤度関数は $\boldsymbol{z}_n = (\boldsymbol{x}_n, \boldsymbol{y}_n)$ として，

$$p(\boldsymbol{z}_n, i|\Theta) = p(\boldsymbol{z}_n|i, \boldsymbol{\mu}_i, \Sigma_i)P(i)$$

により表される．この時，未知パラメタはモデル毎の $\boldsymbol{z}$ の平均 $\boldsymbol{\mu}_i \in R^{d_x+d_y}$ と共分散行列 $\Sigma_i \in R^{(d_x+d_y)\times(d_x+d_y)}$ (i はモデル指標) の二種類である．

一方，回帰では，2.に記したように，完全データの尤度は，

$$p(\boldsymbol{x}_n, \boldsymbol{y}_n, i|\Theta)$$

$$= p(\boldsymbol{y}_n|\boldsymbol{x}_n, i, W_i, S_i)p(\boldsymbol{x}_n|i, \boldsymbol{\mu}_i, \Sigma_i)P(i)$$

となり，形式的には $p(\boldsymbol{y}_n|\boldsymbol{x}_n, i, W_i, S_i)$ が加わっただけであるが，未知パラメタが $\boldsymbol{\mu}_i, \Sigma_i, W_i, S_i$ と増加する．即ち，入力から出力への変換 W_i と出力の共分散行列 S_i の推定を行うか否かの違いである．

5.2 同時分布による回帰の長所と短所

一般に，同時分布を仮定した場合，出力 $\boldsymbol{y}$ のみを確率変数とする場合と比べて，入力 $\boldsymbol{x}$ の分布推定が加わる分，パラメタが多く必要になる．このため，過少数のデータに対する推定精度の悪化は避けられない．

これに関し，表4は一つの目安を与えている．実験結果はデータ数1000の分布に対し，パラメタ数の異なるモデルによる学習の比較であるが，パラメタ数271のMSEに対しパラメタ数18850のSMEM(モデル数50)の方が推定誤差が小さい．即ち、この程度のデータ数／パラメタ数の比であれば十分な推定が可能であることが示されている．

ところで，NGnetを前提とする場合には，パラメタ数は同時分布の仮定の有無に関わらず同じである．なぜなら，$\boldsymbol{y}$ のみを分布と見なす場合でも，ガウス関数を用いたモデル選択を行うために，平均 $\boldsymbol{\mu}_i$ および共分散 Σ_i

表5　同時分布の仮定の有無による比較
Table 5　Effects of assumption of joint distribution.

		同時分布の仮定	
		有り	無し
計算時間	平均	0.407	61.298
	標準偏差	0.041	74.3
	最大	0.50	202.83
	最小	0.37	0.48
二乗誤差	平均	0.011	0.188
	標準偏差	0.054	0.395
	最大	0.062	1.363
	最小	0.0034	0.0073

が必要になるためである．この場合は，同時分布とした方がMステップの学習が高速化できる点で有利だと考えられる．

この点を明らかにするために，同時分布を仮定する場合としない場合との性能比較を行った．モデル数3の混合モデルにEMアルゴリズムを適用し，回帰問題で同時分布の仮定の有無による違いを計算時間(cpu time)と二乗誤差により比較した．回帰関数は$y = 10 \times \tanh(x), \{-10 < x < 10\}$を用いた．

初期値の異なる10回の実験結果の統計を表5に示す．計算時間は同時分布を仮定すると，平均で約25倍速くなっている．これは，Mステップでの繰り返し計算を解析的に解くことができるためである．二乗誤差についても，同時分布を仮定した方が小さく，良い結果を示している．二乗誤差の最小値はどちらの場合も大きな違いはないが，同時分布を仮定しない場合には標準偏差が大きく，初期値の影響を強く受けることがわかる．以上の結果から，同時分布を仮定すると学習が高速になるのみならず，初期値への依存の少ない安定した結果を得られることがわかった．

6. むすび

本論文ではNGnetによる同時分布モデルを用いて，SMEMアルゴリズムの混合回帰問題への適用を行った．この際，併合分割操作へ加えるべき変更点を明らかにし，併合分割候補の順序付け方法についても検証を行った．

人工データ及び実データを用いた実験では，従来のEMアルゴリズムによる回帰問題の推定に比べ，推定精度は確実に向上することが示され，回帰問題でのSMEMアルゴリズムの有効性を確認した．さらに，同時分布の仮定による計算時間の短縮及び，初期値依存度の小さい安定した解が得られることを確認した．

文　献

[1] A.P.Dempster, N.M.Laird and D.B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of Royal Statistical Society B, vol.39, pp.271-182, 1977.

[2] 上田, 中野, "混合モデルのための併合分割操作付きEMアルゴリズム," 信学論D-II, vol.J82-S-II, no.5, pp.930-940, 1999.

[3] N.Ueda, R.Nakano, Z.Gharamani and G.E.Hinton, "SMEM algorithm for mixture models," to appear in Neural Computation.

[4] R.A.Jacobs, M.I.Jordan, S.J.Nowlan and G.E.Hinton, "Adaptive mixtures of local experts," Neural computation, vol.3, pp.79-87, 1991.

[5] M.I.Jordan and R.A.Jacobs, "Hierarchical mixtures of experts and the EM algorithm," Neural Computation, vol.6, pp.181-214, 1994.

[6] L.Xu, M.I.Jordan and G.E.Hinton, "An aleternative model for mixtures of experts," NIPS 7, The MIT Press, pp633-640, 1995.

[7] 石井, 佐藤, "正規化ガウス関数ネットワーク、Mixture of experts とEMアルゴリズム," 日本神経回路学会誌, vol.6, No.1, pp30-40, 1999.

[8] A.Weigend and N.Gershenfeld, "Time series prediction: forcasting future and understanding the past," Addison-Wesley, 1993.

[9] 斎藤, 中野, "MDL原理に基づく新正則化法," 人工知能学会誌, vol.13, No.1, pp123-130, 1998.

(平成x年xx月xx日受付)

鈴木　敏　(正員)

平2東大・教養・基礎科学科卒．同年NTT入社．平4ATR人間情報通信研究所出向．平9よりNTT復帰．現在，コミュニケーション科学基礎研究所所属．物体認識に関わる計算モデルの研究に従事．平6日本神経回路学会研究賞受賞．日本神経回路学会会員．

上田　修功　(正員)

昭57阪大・工・通信卒．昭59同大大学院修士課程了．同年NTT電気通信研究所入所．以来，画像処理，パターン認識・学習，ニューラルネットワーク，統計的学習理論の研究に従事．現在，NTTコミュニケーション科学基礎研究所知識処理研究部 学習理論研究グループ 主幹研究員(特別研究員)，奈良先端大客員助教授．平5～6米国Purdue大学客員研究員．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(テレコムシステム技術賞)受賞，工博．日本神経回路学会，日本統計学会，AVIRG, IEEE各会員．