計測自動制御学会論文集
Vol.39, No.5, 503/512（2003）

変分法的ベイズ推定法に基づく正規化ガウス関数ネットワークと階層的モデル選択法

吉　本　潤一郎*・石　井　　信**・佐　藤　雅　昭***

Normalized Gaussian Network Based on Variational Bayes Inference and Hierarchical Model Selection

Junichiro Yoshimoto*, Shin Ishii** and Masa-aki Sato***

This paper presents a model selection method for normalized Gaussian network (NGnet). We introduce a hierarchical prior distribution of the model parameters and the NGnet is trained based on the variational Bayes (VB) inference. The free energy calculated in the VB inference is used as a criterion for the model selection. In order to efficiently search for the optimal model structure, we develop a hierarchical model selection method. The performance of our method is evaluated by using function approximation and nonlinear dynamical system identification problems. Our method achieved better performance than existing methods.

Key Words: normalized Gaussian network, variational Bayes inference, hierarchical prior distribution, model selection, function approximation

1. はじめに

正規化ガウス関数ネットワーク (normalized Gaussian network, NGnet) [1),2)] は，正規化されたガウス関数を用いて入力空間を滑らかに分割し，分割された部分空間ごとに線形近似を行うモデルである．空間全体をすべてのユニットを用いて近似する多層パーセプトロンのような大域モデルと異なり，NGnet は各ユニットに割り当てられた部分空間内で近似を行う局所モデルであるため，効率良い学習が可能である．局所モデルを高次元データに適用する場合には『次元の呪い』が問題になるが，現実のデータは高次元空間のある一部分に局在しやすいため，多くの場合は局所モデルで十分に対応可能である．実際に，NGnet は非線形力学システム同定問題や強化学習における関数近似器として用いられ，良い結果が得られている [3),4)]．

NGnet は入出力変数の同時分布を近似する確率モデルとして定式化すると，混合正規分布の特殊な場合となる [1),2)]．これは，最尤推定 (Maximum Likelihood estimation, ML) 法の実現法である Expectation-Maximization (EM) アルゴリズム [5)] を用いて，高速な学習が可能であることを意味している．しかしながら，ML 法はモデルの複雑さや推定の確からしさを考慮しないために，モデル構造(注 1)の決定が困難であり，また，しばしば過学習の問題を引き起こす．我々は以前の研究 [6)] において確率的解釈に基づくユニットの生成・削除機構を導入することによりこの問題の解決を図ったが，ヒューリスティクスであるために，課題に応じてあらかじめ適切なメタパラメータを設定する必要があった．

本論文では，この問題点を解決する手法としてベイズ推定法に基づく NGnet の学習法とモデル選択法を提案する．ベイズ推定法では，モデルパラメータに対する事前知識を事前分布の形で導入することにより，学習データに対する各モデルパラメータの確信度が事後分布として獲得される．推定後には，事後分布に関するアンサンブル平均を用いて予測を行うことができる．これはアンサンブル学習の一種であるために，学習データが少ない場合でも過学習を避けることができ，近似精度が高くなる．さらに，推定の際に計算されるモデル周辺化尤度はモデル構造に対する尤度に相当するために，これを用いてモデル構造の良さを定量的に評価することができる．

一方で，NGnet のような非線形混合モデルに対するベイズ推定は，困難な積分計算を含んでいるために近似計算が必要となる．近年，この近似手法として変分法的ベイズ推定 (Variational Bayes inference, VB) 法が提案された [7)~9)]．VB 法では，ある試験事後分布を用いて真の事後分布の近似を行う．この近似は，両分布間の Kullback-Leibler (KL)

* 科学技術振興事業団　CREST　銅谷プロジェクト
** 奈良先端科学技術大学院大学情報科学研究科
*** ATR 人間情報科学研究所
* CREST Doya Project, Japan Science and Technology Corporation
** Graduate School of Information Science, Nara Institute of Science and Technology
*** ATR Human Information Science Laboratories
（Received July 22, 2002）
（Revised February 7, 2003）

（注 1）NGnet ではユニット数がモデル構造に対応する．

divergence によって定義される自由エネルギーの最大化問題を解くことによって実現される．推定後の自由エネルギーは (対数) モデル周辺化尤度の下界になっているので，最適モデル構造を決定するための評価基準を与える．また，試験事後分布における未知変量の部分独立性を仮定すると，VB 法は ML 法における EM アルゴリズムと類似した効率良い反復アルゴリズムとして実装することができる．

VB 法における反復アルゴリズムは自然勾配法 [10] の一種である [9] ために高速な収束が期待される一方で，アルゴリズムの初期条件に依存して局所最適解に収束する．この局所最適解の良さは，さらにモデル構造にも依存する．したがって，悪い局所解を避けながら良い結果を得るためのモデル構造探索手法は実用上重要な問題であり，本研究では階層モデル選択法を導入することによりこの問題の解決を試みる．

本研究と独立・並行して，Ueda らは一般の混合モデルに対して，VB 法に基づくモデル選択法を提案している [11]．ここでも，その応用例として NGnet が取り上げられているが，本研究と以下の点で異なる．第一にモデルパラメータに関する事前分布が異なる点である．Ueda らの研究では，Automatic Relevance Determination (ARD) [12] と呼ばれる事前分布のパイパーパラメータのみに階層的な事前分布を与えていたが，本研究では入出力分散のハイパーパラメータにも階層的な事前分布を与える．これにより，事前分布による推定のバイアスを極力抑えることができる．第二にモデル構造探索アルゴリズムが異なる点である．Ueda らの手法は，ユニットを局所的に分割・併合することによりユニット配置の不均衡性を解消する局所的なアプローチである．一方で，本研究で提案する手法は，データ空間全体を部分空間に階層的に分割し，階層構造を訪問しながら各部分空間で最適化を行う手続きであるので，大域的なデータ構造も考慮したものになっている．

本論文は以下のように構成される．第 2 節で NGnet とその確率モデルについて概説する．第 3 節では，階層的事前分布を導入し，VB 法に基づく NGnet の学習法を導出する．第 4 節では，VB 法に基づくモデル構造選択法を提案し，その実装法について議論する．第 5 節では，提案手法を関数近似問題と未知の非線形力学システムの同定問題に応用することにより，その性能を示し，第 6 節で本論文をまとめる．

2. NGnet と確率モデル

N 次元入力ベクトル $x \equiv (x_1, \cdots, x_N)'$ を D 次元出力ベクトル $y \equiv (y_1, \cdots, y_D)'$ に変換する NGnet は以下で定義される [1),2)]．

$$y = \sum_{i=1}^{M} \left(\frac{g_i \mathcal{N}_N(x|\mu_i, S_i)}{\sum_{j=1}^{M} g_j \mathcal{N}_N(x|\mu_j, S_j)} \right) W_i \tilde{x} \qquad (1)$$

ここで，$\tilde{x} \equiv (x', 1)'$ であり，プライム記号 ($'$) は転置を表している．M は NGnet を構成するユニット数であり，$i \in \{1, \cdots, M\}$ はユニットインデクスである．$g \equiv (g_1, \cdots, g_M)'$ は混合比であり，$g_i \geq 0\ (i = 1, \cdots, M)$，かつ，$\sum_{i=1}^{M} g_i = 1$ を満たす．$\mathcal{N}_N(x|\mu_i, S_i)$ は N 次元中心ベクトル μ_i と $N \times N$ 次逆共分散行列 S_i をパラメータとして持つガウス関数を表している (付録 A.1 を参照)．$W_i \equiv (w_{i1}, \cdots, w_{iD})'$ は $D \times (N+1)$ 次線形回帰行列を表すパラメータであり，w_{ih} は出力 y_h に対応する $N+1$ 次元線形係数ベクトルである．

NGnet は入力と出力の組 (x, y) を確率事象とする確率モデルとして定式化できる [1),2)]．各事象は 1 つのユニット i から確率的に生成されると仮定し，生成したユニットインデクス i を隠れ変数とみなす．このとき，(x, y, i) の組に対する確率モデルを以下で与える．

$$P(x, y, i|\theta) = g_i \mathcal{N}_N(x|\mu_i, S_i) \mathcal{N}_D(y|W_i\tilde{x}, B_i) \qquad (2)$$

ここで，$B_i \equiv diag(\beta_{i1} \cdots, \beta_{iD})$ は出力逆分散を表すパラメータである．$\theta \equiv \{(g_i, \theta_i)|i = 1, \cdots, M\}$ は NGnet のモデルパラメータの集合であり，$\theta_i \equiv \{\mu_i, S_i, W_i, B_i\}$ である．この確率分布から，入力 x が与えられた時の出力 y の期待値 $E_\theta[y|x] \equiv \int dy P(y|x, \theta) y$ を求めると，(1) 式で定義された NGnet の出力と一致する．すなわち，確率分布 (2) 式は NGnet の確率モデルを定義している．

(2) 式より，観測変数の組 (x, y) に対する同時分布は以下の混合モデルとして定式化できる．

$$P(x, y|\theta) = \sum_{\{z\}} P(x, y, z|\theta) \qquad (3a)$$

$$P(x, y, z|\theta) = \exp\left[\sum_{i=1}^{M} z_i \log P(x, y, i|\theta) \right] \qquad (3b)$$

ここで，$z \equiv (z_1, \cdots, z_M)'$ は M 項変数であり，観測変数 (x, y) がユニット i から生成されたことを $z_i = 1$，$z_j = 0\ (j \neq i)$ によって表す．定義より $\sum_{i=1}^{M} z_i = 1$ である．(3) 式による定式化では，隠れ変数は i から z に変換されている．

3. 学習アルゴリズム

3.1 ベイズ推定法とモデル構造評価

T 個のデータセット $(X, Y) \equiv \{(x(t), y(t))|t = 1, \cdots, T\}$ が観測された時，ベイズ推定の目的は未知変量に関する事後分布 $P_{post}(Z, \theta|X, Y)$ を求めることである．ここで，$Z \equiv \{z(t)|t = 1, \cdots, T\}$ は学習データの系列に対応する隠れ変数の系列である．ベイズの定理から，未知変量に関する事後分布は以下で与えられる．

$$P_{post}(Z, \theta|X, Y) = \frac{P(X, Y, Z|\theta) P_0(\theta)}{P(X, Y)}$$

ここで，$P(X, Y, Z|\theta) \equiv \prod_{t=1}^{T} P(x(t), y(t), z(t)|\theta)$ は完全データセット (X, Y, Z) に対するモデルパラメータ θ の尤度である．$P_0(\theta)$ はモデルパラメータ θ の事前分布である．正規化項 $P(X, Y) \equiv \sum_{\{Z\}} \int d\theta P(X, Y, Z|\theta) P_0(\theta)$ はモデル周辺化尤度と呼ばれている．

ベイズ推定において計算されるモデル周辺化尤度は，最適なモデル構造，すなわち，NGnet のユニット数 M を決定するための定量的な評価基準を与える．モデル構造 M への

依存性を明確にするためにモデル周辺化尤度を $P(X,Y|M)$ と記述すると，これは観測データセットに対するモデル構造 M の尤度を意味している．したがって，$P(X,Y|M)$ が最大となるモデル構造 M を選択することはモデル構造に関する最尤推定に対応する．

しかしながら，NGnet のような非線形混合モデルに対して，積分計算を含むモデル周辺化尤度を解析的に求めることは困難であり，近似手法が必要である．以下では，自然共役分布を用いた階層的な事前分布を導入し，VB 法に基づいて，パラメータ事後分布とモデル周辺化尤度を近似するためのアルゴリズムを導出する．

3.2　階層事前分布

本研究では，モデルパラメータの事前分布として以下で定義される自然共役分布を与える．

$$\begin{aligned} P_0(\theta|\xi) &= P_0(g)\prod_{i=1}^{M} P_0(\mu_i|S_i)P_0(S_i|\sigma_i) \\ &\quad \times P_0(W_i|B_i,\Upsilon_i)P_0(B_i|R_i) && \text{(4a)} \\ P_0(g) &= \mathcal{D}_M(g|\gamma_0) && \text{(4b)} \\ P_0(\mu_i|S_i) &= \mathcal{N}_N(\mu_i|m_{0i},\gamma_{0i}S_i) && \text{(4c)} \\ P_0(S_i|\sigma_i) &= \mathcal{W}_N(S_i|\gamma_{s0},\gamma_{s0}\sigma_i I_N) && \text{(4d)} \\ P_0(W_i|B_i,\Upsilon_i) &= \prod_{j=1}^{D}\mathcal{N}_{N+1}(w_{ij}|0,\beta_{ij}\Upsilon_i) && \text{(4e)} \\ P_0(B_i|R_i) &= \prod_{j=1}^{D}\mathcal{G}(\beta_{ij}|\gamma_{\beta 0}/2,\gamma_{\beta 0}\rho_{ij}/2) && \text{(4f)} \end{aligned}$$

ここで，$\mathcal{D}_M(\cdot|\cdot)$，$\mathcal{W}_N(\cdot|\cdot,\cdot)$，および，$\mathcal{G}(\cdot|\cdot,\cdot)$ は，それぞれ，M 次元 Dirichlet 分布，N 次元 Wishart 分布，および，ガンマ分布を表記している (付録 A を参照)．I_N は $N\times N$ 次単位行列を表記している．$\gamma_0 \equiv (\gamma_{01},\cdots,\gamma_{0M})'$ であり，γ_{s0} と $\gamma_{\beta 0}$ はスカラーである．m_{0i} は N 次元ベクトルであり，事前分布における μ_i の期待値に対応する．$\Upsilon_i \equiv diag(\upsilon_{i1},\cdots,\upsilon_{i(N+1)})$ は，事前分布における W_i の逆分散を制御するパイパーパラメータである．σ_i と $R_i \equiv diag(\rho_{i1},\cdots,\rho_{iD})$ は，それぞれ，事前分布における S_i と B_i の期待値の逆数を表現するハイパーパラメータである．Ueda らの研究[11] では，以上のハイパーパラメータのうち Υ_i のみを未知変量として扱い，階層的事前分布を与えているが，本研究では，$\xi \equiv \{\sigma_i,\Upsilon_i,R_i|i=1,\cdots,M\}$ を未知変量として扱い，以下の階層事前分布を与える．

$$\begin{aligned} P_0(\xi) &= \prod_{i=1}^{M} P_0(\sigma_i)P_0(\Upsilon_i)P_0(R_i) && \text{(5a)} \\ P_0(\sigma_i) &= \mathcal{G}(\sigma_i|\gamma_{\sigma 0}/2,\gamma_{\sigma 0}\tau_{\sigma 0}^{-1}/2) && \text{(5b)} \\ P_0(\Upsilon_i) &= \prod_{j=1}^{N+1}\mathcal{G}(\upsilon_{ij}|\gamma_{\upsilon 0}/2,\gamma_{\upsilon 0}\tau_{\upsilon 0}^{-1}/2) && \text{(5c)} \\ P_0(R_i) &= \prod_{j=1}^{D}\mathcal{G}(\rho_{ij}|\gamma_{\rho 0}/2,\gamma_{\rho 0}\tau_{\rho 0}^{-1}/2) && \text{(5d)} \end{aligned}$$

以上の事前分布において，添字 "$_0$" を持つ全てのハイパーパラメータは定数である．この定式化では，ハイパーパラメータ ξ も未知変量であるので，ベイズ推定は事後分布 $P_{post}(Z,\theta,\xi|X,Y)$ を求めるものとして拡張される．このような階層的事前分布を導入する利点については，3.4 節で述べる．

3.3　変分法的ベイズ推定法

VB 法[7)~9)] では，ある試験事後分布 $Q(Z,\theta,\xi)$ を用いて真の事後分布 $P_{post}(Z,\theta,\xi|X,Y)$ の近似を行う．この近似は，以下で定義される自由エネルギーの最大化問題を解くことによって実現される．

$$\begin{aligned} F[Q] &\equiv \left\langle \log\frac{P(X,Y,Z|\theta)P_0(\theta|\xi)P_0(\xi)}{Q(Z,\theta,\xi)}\right\rangle_Q \\ &= \log P(X,Y) - \mathrm{KL}(Q\,\|P_{post}) && \text{(6)} \end{aligned}$$

ここで，$\langle\cdot\rangle_Q$ は試験事後分布 Q に関する期待値であり，以下で定義される．

$$\langle f(Z,\theta,\xi)\rangle_Q \equiv \sum_{\{Z\}}\int d\theta d\xi Q(Z,\theta,\xi)f(Z,\theta,\xi)$$

$\mathrm{KL}(Q\,\|P_{post})$ は，試験事後分布 Q と真の事後分布 P_{post} の KL divergence であり，以下で定義される．

$$\mathrm{KL}(Q\,\|P_{post}) \equiv \left\langle \log\frac{Q(Z,\theta)}{P_{post}(Z,\theta|X,Y)}\right\rangle_Q$$

KL divergence は，$Q(Z,\theta)$ と $P_{post}(Z,\theta|X,Y)$ とが確率分布として一致する時に，最小値 0 になる．また，(6) 式右辺第 1 項の $\log P(X,Y)$ が Q に依存しないことに注意すれば，自由エネルギー $F[Q]$ を Q に関して最大化すると，試験事後分布 Q は真の事後分布 P_{post} に等しくなり，最大化された後の自由エネルギーは最適モデル構造選択基準である（対数）モデル周辺尤度と等しくなることが分かる．

本研究では，隠れ変数 Z，モデルパラメータ θ，およびハイパーパラメータ ξ が互いに独立となるような試験事後分布 Q を用意し，それによって事後分布 $P_{post}(Z,\theta,\xi|X,Y)$ の近似を行う．すなわち，$Q(Z,\theta,\xi) = Q_Z(Z)Q_\theta(\theta)Q_\xi(\xi)$ と因子分解できるものとする．この時，(6) 式の自由エネルギーは以下のように変形できる．

$$\begin{aligned} F[Q_Z,Q_\theta,Q_\xi] &= L - (H^\theta + H^\xi) && \text{(7a)} \\ L &\equiv \left\langle\left\langle \log\frac{P(X,Y,Z|\theta)}{Q_Z(Z)}\right\rangle_\theta\right\rangle_Z && \text{(7b)} \\ H^\theta &\equiv \left\langle\left\langle \log\frac{Q_\theta(\theta)}{P_0(\theta|\xi)}\right\rangle_\xi\right\rangle_\theta && \text{(7c)} \\ H^\xi &\equiv \left\langle \log\frac{Q_\xi(\xi)}{P_0(\xi)}\right\rangle_\xi && \text{(7d)} \end{aligned}$$

ここで，

$$\langle f(Z)\rangle_Z \equiv \sum_{\{Z\}} Q_Z(Z)f(Z),$$

$$\langle f(\theta)\rangle_\theta \equiv \int d\theta Q_\theta(\theta)f(\theta),\quad \langle f(\xi)\rangle_\xi \equiv \int d\xi Q_\xi(\xi)f(\xi)$$

の表記法を用いた．

(7) 式において，L は期待対数尤度に対応し，観測データに対するモデルの適合度を表す項である．一方，H^θ と H^ξ は，それぞれ，モデルパラメータとハイパーパラメータに関するエントロピーに対応し，これらはモデルの複雑さに対するペナルティ項となる[8),9)]．したがって，VB 法はモデルの複雑さに対する正則化を導入した期待対数尤度の最大化手法とみなすことができる．モデルの複雑さに対する正則化項の効果は，事前分布における定数ハイパーパラメータの影響を大きく受ける．しかし，データの分布に関する事前知識がない場合，これらのハイパーパラメータを直接設定し，正則化項の効果を制御することは困難な場合がある．そこで本研究では，より積極的にモデルの複雑さに対する正則化を制御するために，確信度 κ ($\kappa > 0$) を導入する．これにより，(7) 式の自由エネルギーは以下で修正される．

$$F^\kappa[Q_z, Q_\theta, Q_\xi] \equiv \kappa L - (H^\theta + H^\xi) \tag{8}$$

(8) 式はデータを κ 回重複して観測したことに対応し，VB 法における観測データに対する信頼度が通常に比べて κ 倍になっていることを意味している．

自由エネルギー (8) 式の最大化は，以下の 3 ステップを反復手続きによって行う（実装法は付録 B を参照）．

1. VB-E ステップ: $\quad Q_z \leftarrow \underset{Q_z}{\operatorname{argmax}} F^\kappa[Q_z, Q_\theta, Q_\xi]$

2. VB-M ステップ: $\quad Q_\theta \leftarrow \underset{Q_\theta}{\operatorname{argmax}} F^\kappa[Q_z, Q_\theta, Q_\xi]$

3. VB-H ステップ: $\quad Q_\xi \leftarrow \underset{Q_\xi}{\operatorname{argmax}} F^\kappa[Q_z, Q_\theta, Q_\xi]$

上記の反復手続きを ML 法における EM アルゴリズム[5),13)]との類似性から VB-EMH アルゴリズム(注 2)と呼ぶ．VB-EMH アルゴリズムの各ステップで自由エネルギーは単調増加するので，自由エネルギーは局所最大値に大域的に収束する[9)]．

学習後，NGnet の出力 $\hat{y}$ は予測分布に基づいて以下で与えることができる．

$$\begin{aligned}\hat{y} &= \int dyd\theta P_{post}(\theta|X,Y)P(y|x,\theta)y \\ &\approx \int dyd\theta Q_\theta(\theta)P(y|x,\theta)y = \langle E_\theta[y|x]\rangle_\theta \end{aligned} \tag{9}$$

ML 法や最大事後確率推定（Maximum A Posteriori estimation, MAP）法では，最適なモデルパラメータを θ^* として点推定し，NGnet の出力は $E_{\theta^*}[y|x]$ となる．一方で，VB 法では (9) 式のように各モデルパラメータ θ に対する $E_\theta[y|x]$ が事後分布で重み付き平均された形が NGnet の出力となる．これはアンサンブル学習の一種であり，ML 法や MAP 法に比べて過学習を防ぐ効果が期待でき，かつ，近似精度が高くなる．

3.4 階層事前分布の利点

本研究では，事前分布ハイパーパラメータ ξ を未知変量として扱い，階層的な事前分布を与えたが，これには VB-EMH アルゴリズムの実装の観点から以下のような利点がある．

（1） 各ユニットの線形回帰行列 W_i は $D \times (N+1)$ 個の自由パラメータから構成される．一方で，ユニットによっては出力回帰に無関係な入力次元が存在する場合がある．この場合に少ない観測データから W_i を推定すると過剰な自由度のために過学習を招きやすい．これを防ぐために，Mackey らによって提案された ARD 法[12)]に基づく階層事前分布を (4e) 式と (5c) 式によって導入している．この階層事前分布は，事前分布に関する W_i の期待値を 0 とし，その逆分散を Υ_i に比例させて与えるものである．ベイズ推定を用いて Υ_i を推定すると，出力回帰に無関係な入力 x_j に対応する逆分散 v_{ij} は非常に大きくなる．これにより，v_{ij} に対応する回帰パラメータは実質的に定数とみなせるので，W_i の自由パラメータ数が自動的に決定できることになる．

（2） (4d) 式における γ_{s0} は，入力逆共分散行列 S_i に関する事前知識がデータ何個分に相当するかを決める定数ハイパーパラメータである．一方，入力逆共分散行列を推定するためには少なくとも N 個のデータが必要になることに対応して，Wishart 分布は $\gamma_{s0} \leq N-1$ の時に発散する．そのために，γ_{s0} は $(N-1)$ より大きくなければならない．特に入力次元 N が大きい場合，この制限はパラメータ事後分布の推定に非常に強いバイアスを加えることを意味する．この困難を避けるために，我々は S_i の事前分布としてその期待値が $\sigma_i^{-1} I_N$ となる Wishart 分布を与えている．この事前分布は S_i に関する事後分布を推定する際に $\langle S_i \rangle_\theta$ が縮退するのを防ぐ正則化の役割を果たし，数値計算上安定した結果を得ることができる（付録 B における (B. 1) 式と (B. 2) 式を参照）．さらに，σ_i を未知変量として扱いベイズ推定を行うことによって，正則化によるバイアスが観測データに適応して自動的に決定される．これによって推定における事前分布の依存性を極力抑えることが期待できる．

（3） 上記と同様の理由により，出力誤差逆分散 B_i の推定に関わるハイパーパラメータ ρ_{ij} に対しても階層事前分布を導入し，ベイズ推定によって事前分布によるバイアスを観測データに基づき自動的に決定する．

上記の第 (1) 項については Ueda らの研究[11)]でも導入されているが，第 (2) 項と第 (3) 項については本研究で初めて導入されたものである．

4. 階層的モデル選択法

VB-EMH アルゴリズムは局所的な最適解に収束するので，NGnet のような非線形モデルでは試験事後分布の初期値によって収束先が異なる．この局所最適解の良さは，さらにモデル構造にも依存する．そのため，悪い局所最適解を避けながら良いモデル構造を選択するための手法は実用上重要である．本節では，データセットの大域的な分布を考慮に入れ

（注 2）通常，VB-EM アルゴリズムと呼ばれるが，ハイパーパラメータに関する推定ステップを含んでいるために，ここでは VB-EMH アルゴリズムと呼んでいる．

た階層的モデル選択法を提案する．この手法は，NGnet のみならず一般の混合モデルに対しても適用可能である．

4.1　実装法

本手法は 2 つのフェーズから構成される．第 1 フェーズ（トップダウンフェーズ）では，学習データセットを階層的に分割し，2 分木を構成する．根ノードでは $M = 2$ 個のユニットを持つモデルを用意し，VB-EMH アルゴリズムを用いて学習データセット $\mathcal{S} \equiv (X, Y)$ に対する学習を行う．VB-E ステップにおいてユニットインデクス i に関する事後分布を得ることができるので，その MAP 基準に従って $\mathcal{S}$ を 2 つのサブセット $\mathcal{S}_1$ と $\mathcal{S}_2$ に分割する．すなわち，$\mathcal{S}_i$ は，$i = \mathrm{argmax}_j\, Q_z(z_j(t) = 1)$ を満たすデータ $(x(t), y(t)) \in \mathcal{S}$ の集合である．分割されたデータサブセット $\mathcal{S}_1$，$\mathcal{S}_2$ のぞれぞれに対して上記の手続きを再帰的に行うことにより，データについての 2 分木が構成できる．なお，この再帰手続きは，与えられたデータサブセットについて $M = 1$ 個のユニットを持つモデルが $M = 2$ 個のユニットを持つモデルよりも局所自由エネルギー（そのデータサブセットに対してのみの自由エネルギー）が大きくなったところで停止する．図 1 は 2 次元混合正規分布の推定に階層的モデル選択法を適用した時のトップダウンフェーズの様子を示したものである．ここで，“△” 印と “x” 印は各階層で分割されたデータサブセットを表しており，楕円は推定に用いられた各ユニットの正規分布を表している．

第 2 フェーズ（ボトムアップフェーズ）では，トップダウンフェーズで構成された 2 分木の各ノードを下から上へと深さ優先で訪問しながら，ユニットの統合を試みる．葉ノードでは，トップダウンフェーズで用いられた $M = 1$ 個のユニットを持つモデルを保持しておく．上位ノードでは，子ノードが持つ全てのユニットを統合したモデルを用意する．ユニット統合によって生成されたモデルは，このノードに割り当てられたデータサブセットを使って VB-EMH アルゴリズムによる再学習を行う．次にこのデータサブセットに対する最適モデル構造を選択するために混合比 $\langle g_i \rangle_\theta$ が小さいユニットについて順次削除操作を行いながら，局所自由エネルギー最大化基準により局所最適モデル構造を決定する．この手続きは下位ノードから上位ノードに向かって順番に行われる．以上の手続きを下位ノードから根ノードに到達するまで深さ優先で行うことによって，全データセットについての最適なモデル構造を持つモデルが構成される．

以上で説明した階層的モデル選択法は図 2 で示される再帰的手続きを用いて実装することができる．ここで，ステップ (1)-(7) がトップダウンフェーズに対応し，ステップ (8)-(10) がボトムアップフェーズに対応する．

4.2　階層的モデル選択法の特徴

混合モデルでは，一般にモデルが複雑になる（混合数 M が大きくなる）にしたがって局所最適解の数が増加する．また，悪い局所最適解では，あるデータ領域を複雑過ぎるモデルを用いて説明している一方で，別のあるデータ領域では簡

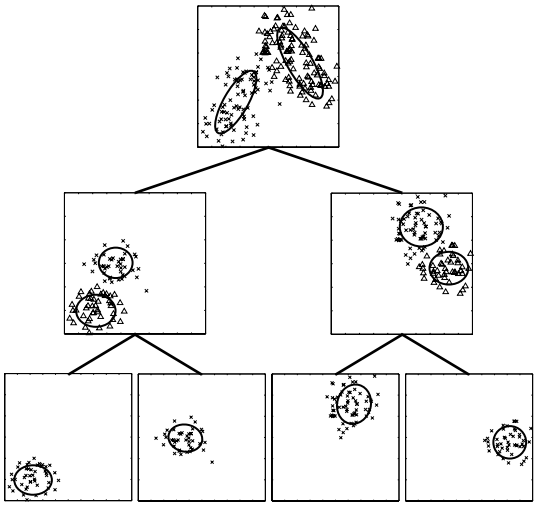


Fig. 1　Binary data tree produced in top-down phase

Function *HMS*(Dataset $\mathcal{S}$)
Output: Posterior Q^*_θ
（1） $M = 1$ のモデルを用意し，VB-EMH アルゴリズムによって $\mathcal{S}$ に対する学習を行う．これによって得られる自由エネルギーを F_1^λ とし，θ に関する試験事後分布を $\bar{Q}_\theta$ とする．
（2） $M = 2$ のモデルを用意し，VB-EMH アルゴリズムによって $\mathcal{S}$ に対する学習を行う．これによって得られる自由エネルギーを F_2^λ とし，隠れ変数 $z(t)$ に関する試験事後分布を Q_z とする．
（3） もし，$F_1^\lambda > F_2^\lambda$ ならば，この関数を終了し，$Q^*_\theta = \bar{Q}_\theta$ を出力として返す．そうでなければ以下の処理を行う．
（4） $\mathcal{S}_1, \mathcal{S}_2$ を空集合として用意する．
（5） $t = 1, \cdots, T$ に対して以下の処理を行う．
　（a） もし $Q_z(z_1(t) = 1) \geq 0.5$ ならば，$\mathcal{S}_1$ にデータ $(x(t), y(t))$ を加える．そうでなければ，$\mathcal{S}_2$ にデータ $(x(t), y(t))$ を加える．
（6） 関数 *HMS*($\mathcal{S}_1$) を適用し，それによって得られる試験事後分布を $Q_{\theta,1}$ とする．
（7） 関数 *HMS*($\mathcal{S}_2$) を適用し，それによって得られる試験事後分布を $Q_{\theta,2}$ とする．
（8） θ_i に関しては $Q_{\theta,1}$ と $Q_{\theta,2}$ の直積として与えられ，g に関しては $Q_\theta(g) = P_0(g)$ で与えられる試験事後分布 Q_θ を用意し，データセット $\mathcal{S}$ に対する学習を行う．これによって得られる自由エネルギーを F^λ とし，削除候補となるユニットインデクスを $i^- = \mathrm{argmin}_i \langle g_i \rangle_\theta$ とする．
（9） Q_θ からユニット i^- を削除した試験事後分布 Q^-_θ を用意し，データセット $\mathcal{S}$ に対する学習を行う．これによって得られる自由エネルギーを $\hat{F}^\lambda$ とする．
(10) もし，$F^\lambda \geq \hat{F}^\lambda$ ならばこの関数を終了し，$Q^*_\theta = Q_\theta$ を出力として返す．そうでなければ，$Q_\theta = Q^-_\theta$ とし，削除候補となるユニットインデクスを $i^- = \mathrm{argmin}_i \langle g_i \rangle_\theta$ とする．その後，ステップ (9) を行う．

Fig. 2　Procedure for hierarchical model selection

単過ぎるモデルを用いて説明しようすることがしばしば起こる．トップダウンフェーズはこうした状況を避ける役割を果たす．トップダウンフェーズでは，高々$M = 2$のモデルを用いて推定するために，局所最適解は少ない．このため，初期値に対する依存性が減少し，安定した解が得られることが期待できる．また，推定後の事後分布を用いてデータ分割を行うためにデータとモデルの双方に適合した形でデータ領域分割を行うことができる．さらに，各データサブセットに対しても自由エネルギーが計算できるために領域分割を進めていくことが適切かどうかを定量的に評価することができる．

ボトムアップフェーズは，分割されたデータ領域についての最適化を行う役割を果たす．トップダウンフェーズにおける前処理によって，あるノードに割り当てられたデータ領域はその 2 つの子ノードに割り当てられたデータ領域の和集合として表現できる．このため，あらかじめ子ノードが担当するデータ領域に対して最適な分布が分かっていると仮定すると，親ノードでは 2 つの分布の混合分布が準最適な分布の 1 つであると考えられる．この考察に基づいて，本手法は各ノードが担当するデータサブセットに対する最適分布を葉ノードから逐次的に推定し，親ノードの学習を行う際には，2 つの子ノードの推定結果を混合させた分布を初期値として用いている．これにより，良い初期値から学習を行うことができるので，悪い局所解に陥ることを防ぐとともに，VB-EMH アルゴリズムの反復回数を低く抑えることができる．一方，子ノードでは親ノードの部分領域に関する最適化しか行わないために，2 つの子ノードの領域境界付近の推定が悪くなる場合がある．この状況は領域境界近傍で本来は局所的に 1 つのクラスタを形成しているデータがトップダウンフェーズで 2 分割されてしまった時によく起こる．このように過剰に分割されたデータ領域を担当するユニットは他のユニットに比べて混合比が小さくなると考えられるので，混合比の小さなユニットの削除を試みる．これによって，局所的な領域分割による副作用が解消されやすいと考えられる．

Ueda らのユニットの分割・併合 (Split and Merge, SM) 法[11] も悪い局所最適解を避けながらモデル構造探索を行う手法であるが，それと比較して本手法の長短は以下のようにまとめられる．

- SM 法におけるユニットの分割は本手法のトップダウンフェーズで行われる処理に対応する．SM 法ではデータに対する局所的な適合度にしたがって分割を試みるユニットが選択される．一方で，本手法はデータ空間全体から逐次・階層的な分割を行っているので，大域的な分布を考慮した分割方法となっている．
- SM 法において併合を試みるユニットの選択基準は事後分布 $Q_z(z_i(t))$ のユニットインデスク i に関する相関によって与えられている．この計算量は $O({}_MC_2 \times T)$ となり，かつ，${}_MC_2$ 通りの組合せに対して順位付けが必要となる．そのため M や T が大きくなるにしたがって，この処理に必要な計算量が大きく増加する．一方，ユニット併合操作に対応する本手法のモデル探索法は，ボトムアップフェーズのユニット削除操作である．削除を試みるユニットの選択基準は混合比 $\langle g_i \rangle_\theta$ で与えられており，M 通りの順位付けのみを行えばよい．したがって，ユニット数 M が大きくなってもこの処理に必要な計算量はそれほど増加しない．
- 最適モデル構造が M^* とすると，本手法のデータサブセットは最低でも $2M^* - 1$ 通り生成され，それぞれに対して VB-EMH アルゴリズムを適用する必要がある．これが本手法の最大の欠点である．しかしながら，トップダウンフェーズでは高々2 個のユニットを持つモデルの学習を行うだけであり，下位ノードに進むにつれて学習データも少なくなるので，それほど計算量は多くならない．ボトムアップフェーズでは，多くのユニットを持つモデルに対する学習も必要となるが，各ユニットは下位層において局所的な最適化が行われているため，収束までの EHM アルゴリズムの反復回数はそれほど多くならない．この点は，5. 1 節の実験結果によって示す．

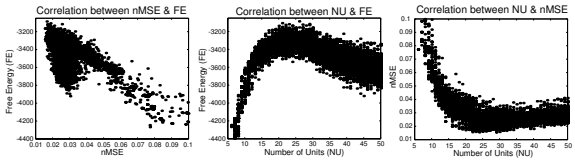


Fig. 3 Correlation diagrams among generalization error ($nMSE$), free energy (FE) and the number of units (NU).

5. 実験

5. 1 関数近似問題

最初の実験では，以下の関数[14] を用いて本手法の関数近似能力とモデル選択能力を評価した．

$$y = \max\left\{e^{-10x_1^2}, e^{-50x_2^2}, 1.25e^{-5\left(x_1^2+x_2^2\right)}\right\}$$

定義域 $-1 \leq x_1, x_2 \leq 1$ から入力変数 $x \equiv (x_1, x_2)$ を一様にサンプルして得られた 500 個の入出力変数の組 (x, y) を学習データとする．ただし，各出力変数には平均 0，標準偏差 0.1 の大きなガウス雑音を付加する．

まず，混合数 M が 5 から 50 まで 46 種類のモデル構造を用意し，各モデル構造に対して VB-EMH アルゴリズムの初期条件を変えて 100 回学習を行った．ここで，確信度 $\kappa = 7$ を用いている．図 3 は，モデル構造と学習後の汎化誤差，および，自由エネルギーの相関図を示したものである．汎化誤差は，入力変数をその定義域から 41×41 個の点として格子上に等間隔でとり，そこでの出力変数に関する真値と近似値との平均自乗誤差を真値の分散で正規化した値（正規化自乗誤差, $nMSE$）によって評価されている．この図から，汎化誤差の小ささと自由エネルギーの大きさの間に強い相関があることが分かる．

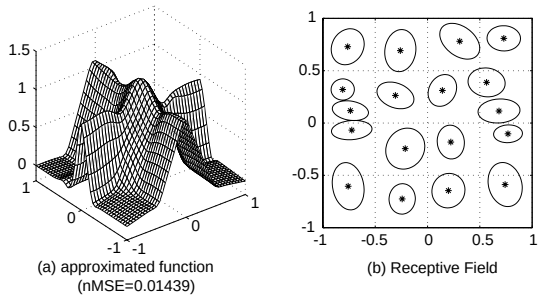


Fig. 4　(a) Function shape approximated by the NGnet; (b) receptive fields of units in the trained NGnet.

次に階層的モデル選択法を用いてモデル選択を行った．その結果，混合数 $M = 18$ のモデル構造が選択され，汎化誤差は $nMSE = 0.0144$ となった．収束するまでに要する VB-EMH アルゴリズムの反復回数は 4807 回であった．図 4 は学習後の NGnet が出力する関数形と各ユニットの受容野を示したものである．ユニットが定義域内に均等に配置され，その結果，精度の高い関数近似が実現されている様子が分かる．

同じ実験条件において，我々が以前に提案したオンライン EM 法によって得られた結果は $nMSE = 0.0189$ であった [6]．VB 法と階層的モデル選択法を用いることにより近似精度は約 24% 改善された．また，モデル選択能力を比較するために SM 法によるモデル選択アルゴリズム [11] を実装した．ここで，初期モデル構造は階層的モデル選択法と同じ $M = 2$ とした．その結果，最善の場合でも汎化誤差は $nMSE = 0.0168$ であり，この最善の結果を得るために要する VB-EMH アルゴリズムの反復回数は 4317 回であった．階層的モデル選択法は SM 法に比べて若干多くの計算回数を必要とするが，精度の高い関数近似を実現できているといえる．

次に，関数近似問題のベンチマーク集である DELVE データセット [15] の中から “pumadyn8” という課題を用いて，高次元データに対する性能を評価した．この課題は，ロボットアームの 8 次元の状態変数から残りの 1 次元の状態変数を推定するものである．データセットはデータ数とノイズの大きさに応じていくつかの条件に分かれているが，この中から “nm256”，“nh256”，“nm1024”，“nh1024” を実験に用いた．ここで，“nm” は中程度のノイズを，“nh” は大きなノイズを含んでいることを意味している．“256” と “1024” は学習データセットのデータ数である．各データセットに対して本手法（VB-NGnet）を適用した時の $nMSE$ を表 1 の上段に示す．また，オンライン EM アルゴリズム（OEM-NGnet）[6] や，NGnet と類似した混合回帰モデルである Mixture of Experts network（MEnet）[16),17)]，および，階層的 MEnet（HMEnet）[17),18)] による結果も表 1 に示す．nm1024 を除く全ての条件において，本手法は他の手法より優れており，特に，データ数が少なくノイズが大きい場合に，本手法が良い性能を示していることが分かる．以

Table 1　$nMSE$ of each method in “pumadyn8” tasks

	nm256	nh256	nm1024	nh1024
VB-NGnet	0.1256	0.4901	0.0517	0.3503
OEM-NGnet	0.2104	0.5152	0.0518	0.3896
MEnet	0.1270	0.5370	0.0461	0.3603
HMEnet	0.1893	0.5375	0.0490	0.3639

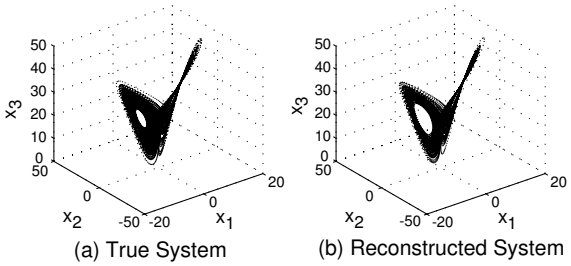


Fig. 5　(a) True Lorenz attractor; (b) an attractor reconstructed by the trained NGnet

上の比較から，本手法は既存の手法に比べて関数近似能力に優れた効率良いモデル選択を実現しているといえる．

5. 2　非線形力学システムの同定問題

未知の非線形力学システムの同定問題に本手法を適用した．本実験では，以下の微分方程式で定義される Lorenz システムを取り上げる．

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} a(x_2 - x_1) \\ -x_2 + (b - x_3)x_1 \\ -cx_3 + x_1 x_2 \end{pmatrix}$$

ここで，$x \equiv (x_1, x_2, x_3)$ はシステムの状態変数である．$\dot{x}$ は状態変数 x の時間微分を表している．システムパラメータは $a = 10.0$, $b = 28.0$, $c = 8/3$ というしばしば用いられる値に設定する．図 5(a) は Lorenz システムのアトラクタを示したものである．

アトラクタ上で時間間隔 $\Delta\tau = 0.01$ でサンプリングされた $T = 5000$ 個の時系列 $X \equiv \{(x(k\Delta\tau)|k = 1, \cdots, T)\}$ を観測時系列とする．観測時系列 X に対する離散化ベクトル場は以下で定義される．

$$V(x(k\Delta t)) = \frac{x((k+1)\Delta t) - x(k\Delta t)}{\Delta t}$$

NGnet は x から $V(x)$ への写像を近似するように学習を行う．学習後には，任意の初期状態 $\hat{x}(0)$ から離散化された時系列を以下の式を用いて自動的に生成することができる．

$$\hat{x}(t + \Delta t) = \hat{x}(t) + \hat{V}(\hat{x}(t))\Delta t$$

ここで，$\hat{V}(x)$ は NGnet によって近似された離散化ベクトル場である．

図 5(b) は NGnet によって再構成されたアトラクタを示している．Lorenz システムが持つ奇妙なアトラクタ構造を良く再現できていることが分かる．ここで，アトラクタ上における離散化ベクトル場の汎化誤差は $nMSE = 0.007\%$ であった．図 6 は 100 個の異なる初期状態から NGnet によって予測された時系列と真の時系列との誤差の平均を示している．平均して約 $t = 2$ まで正確な予測を行っていることが

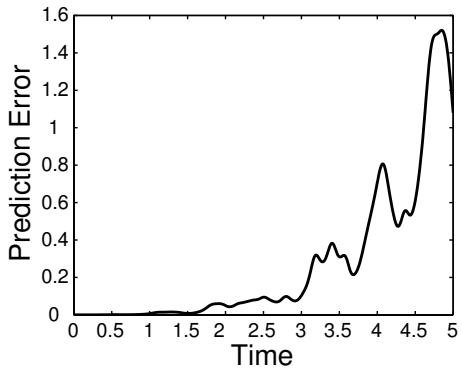


Fig. 6 Average prediction error of the trained NGnet

分かる．我々は以前オンライン EM アルゴリズムを用いて同じ実験を行っている [3]．この時，ほぼ同等の結果を得るためには $T = 100,000$ 個の観測系列が必要であった．20 分の 1 のデータ量から同程度の結果を得ていることから，提案手法が高い汎化性能を持っていることが分かる．

6. おわりに

本論文では，正規化ガウス関数ネットワークに対する新しいモデル選択法を提案した．本手法では，モデルパラメータについての階層的事前分布が導入され，変分法的ベイズ推定法に基づいて学習が行われた．この時に計算される自由エネルギーは最適モデル構造を決定するための評価基準として用いられた．また，効率良いモデル構造選択を実現するための階層的モデル選択手法が導入された．本手法は関数近似問題と非線形力学システムの同定問題に適用された．その結果として，本手法が近似精度が良く，汎化能力に優れていることが示された．本手法はバッチ学習のためのモデル選択手法であるが，オンライン学習に適用できる手法へ拡張することが今後の課題である．

謝辞

本研究は、財団法人テレコム先端技術研究支援センターおよび通信・放送機構の研究委託により実施された．また人工知能研究振興財団の研究支援を受けた．

参 考 文 献

1) L. Xu, M. I. Jordan and G. E. Hinton: An alternative model for mixtures of experts, *Advances in Neural Information Processing Systems 7* (eds. G. Tesauro, D. S. Touretzky and T. K. Leen), 633/640, MIT Press (1995)
2) 石井 信, 佐藤 雅昭: 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, **6**-1, 30/40 (1999)
3) S. Ishii and M. Sato: Reconstruction of chaotic dynamics by on-line EM algorithm, *Neural Networks*, **14**-9, 1239/1256 (2001)
4) 吉本 潤一郎，石井 信，佐藤 雅昭: 連続力学システムの自動制御のためのオンライン EM 強化学習法, システム制御情報学会論文誌, **16**-5 (2003)
5) 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*, **39**-1, 1/38 (1997)
6) M. Sato and S. Ishii: On-line EM algorithm for the normalized Gaussian network, *Neural Computation*, **12**-2, 407/432 (2000)
7) S. Waterhouse, D. Mackay and T. Robinson: Bayesian methods for mixture of experts, *Advances in Neural Information Processing Systems 8* (eds. D. S. Touretzky and M. C. Mozer and M. E. Hasselmo), MIT Press, 351/357 (1996)
8) H. Attias: A variational Bayesian framework for graphical models, *Advances in Neural Information Processing Systems 12* (eds. S. A. Solla, T. K. Leen and K. R. Müller), 206–212, MIT Press, 206/212 (2000)
9) M. Sato: Online model selection based on the variational Bayes, *Neural Computation*, **13**-7, 1649/1681 (2001)
10) S. Amari: Natural gradient works efficiently in learning, *Neural Computation*, **10**-2, 251/276 (1998)
11) N. Ueda and Z. Ghahramani: Bayesian model search for mixture models based on optimizing variational bounds, *Neural Networks*, **15**, 1223/1241 (2002)
12) D. J. C. Mackay: Bayesian non-linear modelling for the prediction competition, *ASHRAE Transactions*, **100**, 1053/1062 (1994)
13) R. M. Neal and G. E. Hinton: A view of the EM algorithm that justifies incremental, sparse, and other variants, *Learning in Graphical Models* (ed. M. I. Jordan), Kluwer Academic Publishers, 355/368 (1998)
14) S. Schaal and C. G. Atkeson: Constructive incremental learning from only local information, *Neural Computation*, **10**, 2047/2084 (1998)
15) C. E. Rasmussen, R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra and R. Tibshirani: *The DELVE Manual*, http://www.cs.toronto.edu/~delve (1996)
16) R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton: Adaptive mixtures of local experts, *Neural Computation*, **3**, 79/87 (1991)
17) S. R. Waterhouse: *Classification and regression using mixtures of experts*, Ph. D. Thesis, Department of Engineering, University of Cambridge (1997)
18) M. I. Jordan and R. A. Jacobs: Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, **6**, 181/214 (1994)

《付 録》

A. 確率分布

A. 1 正規分布

p 次元確率ベクトル x が p 次元中心ベクトル μ，$p \times p$ 次逆共分散行列 S をパラメータとする正規分布に従う時，その確率分布 $\mathcal{N}_p(x\,|\mu, S)$ は以下で与えられる．

$$\mathcal{N}_p(x\,|\mu, S) \equiv (2\pi)^{-p/2}|S|^{1/2} \times \exp\left[-\frac{1}{2}(x-\mu)'S(x-\mu)\right] \quad \text{(A. 1)}$$

(A. 1) 式を x の関数として用いる場合には，これをガウス関数と呼ぶ．

A. 2 Dirichlet 分布

確率条件 $\sum_{i=1}^{p} x_i = 1, x_i \geq 0$ を満足する p 次元確率ベク

トル $x \equiv (x_1, \cdots, x_p)'$ が p 次元ベクトル $\gamma \equiv (\gamma_1, \cdots, \gamma_p)'$ をパラメータとする Dirichlet 分布に従う時，その確率分布 $\mathcal{D}_p(x|\gamma)$ は以下で与えられる．

$$\mathcal{D}_p(x|\gamma) = \frac{\Gamma(\gamma_1 + \cdots + \gamma_p + p)}{\Gamma(\gamma_1 + 1) \cdots \Gamma(\gamma_p + 1)} x_1^{\gamma_1} \cdots x_p^{\gamma_p}$$

ここで，$\Gamma(x) \equiv \int_0^\infty e^{-t} t^{x-1} dt$ はガンマ関数である．

A. 3　ガンマ分布

確率変数 x $(x \geq 0)$ が a と b をパラメータとするガンマ分布に従う時，その確率分布 $\mathcal{G}(x|a, b)$ は以下で与えられる．

$$\mathcal{G}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx]$$

A. 4　Wishart 分布

$p \times p$ 次正定対称行列 X がスカラー値 λ と $p \times p$ 次正定対称行列 Δ をパラメータとする Wishart 分布に従う時，その確率分布 $\mathcal{W}_p(X|\lambda, \Delta)$ は以下で与えられる．

$$\mathcal{W}_p(X|\lambda, \Delta) = c \cdot \exp\left[-\frac{1}{2} \mathrm{Tr}[\Delta X]\right]$$

ここで，

$$c \equiv \frac{|\Delta|^{\lambda/2} |X|^{(\lambda-p-1)/2}}{2^{p\lambda/2} \pi^{p(p-1)/4} \prod_{n=1}^{p} \Gamma((\lambda + 1 - n)/2)}$$

である．

B.　VB-EMH アルゴリズム

3. 3 節で導出した VB-EMH アルゴリズムは以下のように実装できる．

（1）　事前分布定数パラメータを設定する．

（2）　ステップ (3a) と (3b) の計算に必要な事後分布 Q_θ, Q_ξ に関する期待値 $\langle\cdot\rangle_\theta, \langle\cdot\rangle_\xi$ を適切に初期化する．

（3）　自由エネルギー F^λ が収束するまで以下のステップ (a)-(c) を繰り返す．

（a）　VB-E ステップ: Q_Z を以下で更新する．

$$\begin{aligned}
Q_Z(Z) &\equiv \prod_{t=1}^{T} Q_z(z(t) = 1) \\
Q_z(z_i(t) = 1) &\leftarrow \frac{\exp[U_i(x(t), y(t))]}{\sum_{j=1}^{M} \exp[U_j(x(t), y(t))]} \\
U_i(x, y) &\equiv \langle \log g_i \rangle_\theta - \frac{1}{2} x' \langle S_i \rangle_\theta x + x' \langle S_i \mu_i \rangle_\theta \\
&\quad - \frac{1}{2} \langle \mu_i' S_i \mu_i \rangle_\theta - \frac{1}{2} y' \langle B_i \rangle_\theta y \\
&\quad + y' \langle B_i W_i \rangle_\theta \tilde{x} - \frac{1}{2} \tilde{x}' \langle W_i' B_i W_i \rangle_\theta \tilde{x} \\
&\quad + \frac{1}{2} \langle \log |S_i| \rangle_\theta + \frac{1}{2} \langle \log |B_i| \rangle_\theta
\end{aligned}$$

Q_Z に関する十分統計量の期待値は以下で求められる．

$$\langle z_i f(x, u) \rangle = \frac{1}{T} \sum_{t=1}^{T} Q_z(z_i(t) = 1) f(x(t), y(t))$$

（b）　VB-M ステップ: $Q_\theta(\theta)$ を以下で更新する．

$$\begin{aligned}
Q_\theta(\theta) &\equiv Q_\theta(g) \prod_{i=1}^{M} Q_\theta(\mu_i, S_i) Q_\theta(W_i, B_i) \\
Q_\theta(g) &\equiv \mathcal{D}_M(g|\gamma) \\
Q_\theta(\mu_i, S_i) &\equiv \mathcal{N}_N(\mu_i | m_i, \gamma_i S_i) \\
&\quad \times \mathcal{W}_N(S_i | \gamma_{si}, \gamma_{si} \Delta_i) \\
Q_\theta(W_i, B_i) &\equiv \prod_{j=1}^{D} \mathcal{N}_{N+1}(w_{ij} | v_{ij}, \beta_{ij} \Xi_i) \\
&\quad \times \mathcal{G}(\beta_{ij} | \gamma_{\beta i}/2, \gamma_{\beta i} \lambda_{ij}/2) \\
\gamma &\equiv (\gamma_1, \cdots, \gamma_M)' \\
\gamma_i &\leftarrow \kappa T \langle z_i \rangle + \gamma_{0i} \\
\gamma_{si} &\leftarrow \kappa T \langle z_i \rangle + \gamma_{s0} \\
\gamma_{\beta i} &\leftarrow \kappa T \langle z_i \rangle + \gamma_{\beta 0} \\
m_i &\leftarrow \frac{\kappa T \langle z_i x \rangle + \gamma_{0i} m_{0i}}{\gamma_i} \\
\Delta_i &\leftarrow \frac{1}{\gamma_{si}} \Big(\kappa T \langle z_i x x' \rangle + \gamma_{0i} m_{0i} m_{0i}' \\
&\quad + \gamma_{s0} \langle \sigma_i \rangle_\xi I_N - \gamma_i m_i m_i' \Big) \qquad \text{(B. 1)} \\
\Xi_i &\leftarrow \kappa T \langle z_i \tilde{x} \tilde{x}' \rangle + \langle \Upsilon_i \rangle_\xi \\
V_i &\leftarrow \kappa T \langle z_i y \tilde{x}' \rangle \Xi_i^{-1} \\
\Lambda_i &\leftarrow \frac{diag\left(\kappa T \langle z_i y y' \rangle + \gamma_{\beta 0} \langle R_i \rangle_\xi - V_i \Xi_i V_i' \right)}{\gamma_{\beta i}}
\end{aligned}$$

ここで，$V_i \equiv (v_{i1}, \cdots v_{iD})$, $\Lambda_i \equiv diag(\lambda_{i1}, \cdots, \lambda_{iD})$ である．$Q_\theta(\theta)$ に関する期待値 $\langle\cdot\rangle_\theta$ は以下で与えられる．

$$\begin{aligned}
\langle \log g_i \rangle_\theta &= \psi(\gamma_i + 1) - \psi\left(\sum_{j=1}^{M} \gamma_j + M \right) \\
\langle S_i \rangle_\theta &= \Delta_i^{-1} \\
\langle S_i \mu_i \rangle_\theta &= \Delta_i^{-1} m_i \qquad \text{(B. 2)} \\
\langle \mu_i' S_i \mu_i \rangle_\theta &= m_i' \Delta_i^{-1} m_i + N/\gamma_i \\
\langle \log |S_i| \rangle_\theta &= -\log |\Delta_i| - N \log(\gamma_{si}/2) \\
&\quad + \sum_{n=1}^{N} \psi\left(\frac{\gamma_{si} + 1 - n}{2} \right) \\
\langle B_i \rangle_\theta &= \Lambda_i^{-1} \\
\langle B_i W_i \rangle_\theta &= \Lambda_i^{-1} V_i \\
\langle W_i' B_i W_i \rangle_\theta &= V_i' \Lambda_i^{-1} V_i + D \Xi_i^{-1} \\
\langle \log |B_i| \rangle_\theta &= -\log |\Lambda_i| - D \log(\gamma_{\beta i}/2) \\
&\quad + D \psi(\gamma_{\beta i}/2)
\end{aligned}$$

ここで，$\psi(x) \equiv \frac{d}{dx} \Gamma(x)$ はダイガンマ関数である．

（c） VB-H ステップ: $Q_\xi(\xi)$ を以下で更新する．

$$Q_\xi(\xi) \equiv \prod_{i=1}^{M} Q_\xi(\sigma_i) Q_\xi(\Upsilon_i) Q_\xi(R_i)$$

$$Q_\xi(\sigma_i) \equiv \mathcal{G}\left(\sigma_i | \gamma_{\sigma i}/2, \gamma_{\sigma i}\tau_{\sigma i}^{-1}/2\right)$$

$$Q_\xi(\Upsilon_i) \equiv \prod_{j=1}^{N+1} \mathcal{G}\left(\upsilon_{ij} | \gamma_{\upsilon i}/2, \gamma_{\upsilon i}\tau_{\upsilon ij}^{-1}/2\right)$$

$$Q_\xi(R_i) \equiv \prod_{j=1}^{D} \mathcal{G}\left(\rho_{ij} | \gamma_{\rho i}/2, \gamma_{\rho i}\tau_{\rho ij}^{-1}/2\right)$$

$$\gamma_{\sigma i} \leftarrow N\gamma_{s0} + \gamma_{\sigma 0}, \quad \gamma_{\upsilon i} \leftarrow D + \gamma_{\upsilon 0},$$

$$\gamma_{\rho i} \leftarrow \gamma_{\beta 0} + \gamma_{\rho 0}$$

$$\tau_{\sigma i}^{-1} \leftarrow \left(\gamma_{s0}\,\mathrm{Tr}\left[\langle S_i \rangle_\theta\right] + \gamma_{\sigma 0}\tau_{\sigma 0}^{-1}\right)/\gamma_{\sigma i}$$

$$\tau_{\upsilon ij}^{-1} \leftarrow \left(\left(\left\langle W_i' B_i W_i \right\rangle_\theta\right)_{j,j} + \gamma_{\upsilon 0}\tau_{\upsilon 0}^{-1}\right)/\gamma_{\upsilon i}$$

$$\tau_{\rho ij}^{-1} \leftarrow \left(\gamma_{\beta 0}\left(\langle B_i \rangle_\theta\right)_{j,j} + \gamma_{\rho 0}\tau_{\rho 0}^{-1}\right)/\gamma_{\rho i}$$

ここで，$(\cdot)_{j,j}$ は行列の j 行 j 列目の要素を表す．$Q_\xi(\xi)$ に関する期待値は以下で与えられる．

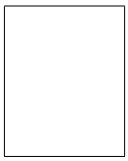
$$\langle \Upsilon_i \rangle_\theta = diag\left(\tau_{\upsilon i1}, \cdots, \tau_{\upsilon i(N+1)}\right)$$

$$\langle R_i \rangle_\theta = diag\left(\tau_{\rho i1}, \cdots, \tau_{\rho iD}\right)$$

$$\langle \sigma_i \rangle_\theta = \tau_{\sigma i}$$

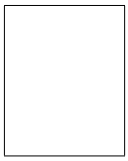
［著 者 紹 介］

吉 本 潤一郎（正会員）



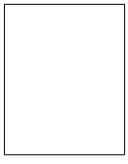
2002 年 9 月奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．同年 10 月より科学技術振興事業団 CREST 銅谷プロジェクトの研究員となり，現在に至る．博士（工学）．ニューラルネットワーク，統計的学習理論，強化学習の研究に従事．

石 井 信



1988 年 3 月東京大学大学院工学系研究科修士課程修了．（株）リコー中央研究所研究員，ATR 人間情報通信研究所研究員，奈良先端科学技術大学院大学情報科学研究科助教授を経て，現在，同教授．工学博士．非線形力学系，ニューラルネットワーク，強化学習，統計的学習理論，バイオ情報学の研究に従事．

佐 藤 雅 昭



1980 年 3 月大阪大学大学院理学研究科物理学専攻博士課程修了．ニューヨーク大学助手，フロリダ大学助手，ATR 視聴覚機構研究所主任研究員，ATR 人間情報通信研究所主任研究員を経て，現在，ATR 人間情報科学研究所主任研究員．理学博士．非線形力学系，カオス，ニューラルネットワーク，強化学習，ロボット制御，統計的学習理論の研究に従事．