潜在変数モデルに基づく確率ニューラルネットワーク

上田 修功
NTT コミュニケーション科学研究所
京都府相楽郡精華町光台
e-mail: ueda@cslab.kecl.ntt.jp

1 まえがき

統計的推定とは、データの背後にある確率分布の推定と言える。ニューラルネットワークによる分布推定のモデルとしてボルツマンマシン [1] が著名であるが、近年、表現能力と学習効率の両面でボルツマンマシンの性能を上回るヘルムホルツマシン [2][3] が提案されている。これらは、隠れユニットの出力値が、通常の多層ニューラルネットワークの様に確定的に定まるのではなく、確率的に定まるところから、"確率ニューラルネットワーク" と呼ばれている。

ヘルムホルツマシンは、ボルツマンマシン同様、入力データの空間上で分布推定を行なわず、隠れユニットの出力の分布として間接的に入力データの確率分布を推定する。いわば、隠れ層を用いて外界を "まねる" 学習機械の実現を目的とする。この点で、混合正規分布モデルに代表される通常の統計的確率分布推定手法とは一見異なるアプローチをとる。

しかしながら、隠れユニットを潜在変数と見なすと、実は、興味深い事に、ヘルムホルツマシンは、因子分析等の従来の多変量統計解析で考案された潜在変数モデル [4] の一形態（正確には非線型潜在変数モデル）と見なせることが分かる。つまり、ヘルムホルツマシンは潜在変数モデルに基づく確率ニューラルネットワークと位置づけることができる。

本稿では、ヘルムホルツマシンにおける分布推定の基本原理および学習アルゴリズムを概説すると共に、潜在変数モデルというより上位の視点でヘルムホルツマシンを捉え、従来の多変量統計解析手法との関係について概観する。

2 ヘルムホルツマシン

2.1 ヘルムホルツ自由エネルギー

ヘルムホルツマシンはその名の通り統計力学からのアナロジーを用いてヘルムホルツ自由エネルギーを定義し、その最小化として学習則が導出されるが、その本質は非観測データ (隠れユニットの状態) を含む不完全データからの最尤推定問題として定式化されている。以下これについて概説する。

全隠れユニットの状態を α、ネットワークパラメータを Θ、入力データを d とすると、対数尤度関数は

$$\mathcal{L}(\Theta; d) = \log P(d; \Theta) \tag{1}$$

で定義される。一方、Bayes 則：

$$P(d; \Theta) = \frac{P(d, \alpha; \Theta)}{P(\alpha|d; \Theta)}$$

の両辺の対数をとり、新たな未知分布 $\tilde{P}(\alpha|d)$ に関する条件付き期待値をとると次式を得る。

$$\begin{aligned}\log P(d; \Theta) &= \mathrm{E}_{\tilde{P}(\alpha|d)}\{\log P(d, \alpha; \Theta) - \log \tilde{P}(\alpha|d)\} \\ &\quad + \mathrm{KL}(\tilde{P}(\alpha|d) \,\|\, P(\alpha|d; \Theta))\end{aligned} \tag{2}$$

$\log P(d; \Theta)$ は α に依存しないので、分布 $\tilde{P}(\alpha|d)$ で期待値をとっても変化しないことに注意。KL() は KL 情報量 (Kullback divergence) で次式で定義される。

$$\mathrm{KL}(\tilde{P} \,\|\, P) = \mathrm{E}_{\tilde{P}(\alpha|d)}\left\{\log \frac{\tilde{P}(\alpha|d)}{P(\alpha|d; \Theta)}\right\} \tag{3}$$

故に、式 (1),(2) より、対数尤度関数は

$$\begin{aligned}\mathcal{L}(\Theta; d) &= \mathrm{E}_{\tilde{P}(\alpha|d)}\{\log P(d, \alpha; \Theta) - \log \tilde{P}(\alpha|d)\} \\ &\quad + \mathrm{KL}(\tilde{P}(\alpha|d) \,\|\, P(\alpha|d; \Theta))\end{aligned} \tag{4}$$

と書き表せる。

ここで、$\mathcal{U} = -\mathrm{E}_{\tilde{P}(\alpha|d)}\{\log P(d,\alpha;\Theta)\}(\geq 0)$ を内部エネルギー、$\mathcal{H} = \mathrm{E}_{\tilde{P}(\alpha|d)}\{-\log \tilde{P}(\alpha|d)\}$ をエントロピーと見なすと、統計力学のアナロジーから次式に示す温度 $T=1$ におけるヘルムホルツの自由エネルギーが定義できる。

$$\mathcal{F} = \mathcal{U} - \mathcal{H} \tag{5}$$

式 (5) を用いると式 (4) は次の様に書き換えられる。

$$\mathcal{L}(\Theta;d) = -\mathcal{F}(d,\Theta,\tilde{P}) + \mathrm{KL}(\tilde{P}_{\alpha|d} \parallel P_{\alpha|d}) \tag{6}$$

上記の KL 情報量は真の事後確率を別の事後確率 $\mathcal{P}$ で置き換えた際の penalty 項と見なすこともできる。

KL $\geq$ 0 に注意すると、$-\mathcal{F}$ は対数尤度関数の上限となっていることが分かる。従って、最尤推定、即ち、$\mathcal{L}$ の最大化は $\mathcal{F}$ の最小化として実現できる。換言すれば、ヘルムホルツ自由エネルギーの最小化により得られたパラメータ Θ は対数尤度関数の最大化として得られる最尤推定値:

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \mathcal{L}(\Theta;d)$$

に一致する。

更に、$\mathcal{F}(d,\Theta,\tilde{P})$ を coordinate descent 法で最小化すると、興味深いことに、最尤推定の逐次解法として著名な EM アルゴリズム [5] が導出できる。coordinate descent 法とは、

$$\begin{aligned}\mathcal{F}(d,\Theta^{(t)},\tilde{P}) &= -\mathrm{E}_{\tilde{P}(\alpha|d)^{(t)}}\{\log P(d,\alpha;\Theta)\} \\ &\quad +\mathrm{E}_{\tilde{P}(\alpha|d)^{(t)}}\{-\log \tilde{P}(\alpha|d)\}\end{aligned}$$

とし、$\mathcal{F}(d,\Theta^{(t)},\tilde{P})$ の $\tilde{P}$ に関する最小化 ($\Theta^{(t)}$ は固定) により、$\tilde{P}^{(t+1)}$ を推定し、次いで、$\mathcal{F}(d,\Theta,\tilde{P}^{(t+1)})$ の Θ に関する最小化 ($\tilde{P}^{(t+1)}$ は固定) により $\Theta^{(t+1)}$ を推定する、という処理を収束するまで繰り返す最適化手法である。この時、前者が EM アルゴリズムの E ステップに相当し、後者が M ステップに相当する。何故なら、前者の最小化、即ち、$\partial\mathcal{F}/\partial\tilde{P}=0$ より、

$$\begin{aligned}\tilde{P}(\alpha|d)^{(t+1)} &= P(d,\alpha;\Theta^{(t)}) / \sum_{\alpha} P(d,\alpha;\Theta^{(t)}) \\ &\equiv P(\alpha|d;\Theta^{(t)})\end{aligned}$$

が得られ、これは E ステップにおける期待値計算の際に用いる事後確率そのものである。次に、$\tilde{P}^{(t+1)}$ を固定した下での $\mathcal{F}$ の Θ に関する最小化は、$\mathrm{E}_{P(\alpha|d;\Theta^{(t)})}\{\log P(d,\alpha;\Theta))\}$ の最大化となり、これは EM アルゴリズムの M ステップと一致する。詳細は、文献 [6] を参照されたい。

2.2 ネットワークモデル

ヘルムホルツマシンは、図 1 に示す様に入力層 ($l=1$) と隠れ層 ($l>1$) から成り、各層は 0,1 の 2 値を出力するユニット群 (s) から構成される。隠れ層は多層構造で、各層のユニット間はボトムアップの重み (ϕ) とトップダウンの重み (θ) により結合され、各々認識重み、生成重みと呼ばれる。従って、ネットワークパラメータは

$$\Theta = \{\phi,\theta\}$$

である。(各ユニットは認識、生成に対応して 2 種類のバイアスを持つが以下では簡単のためバイアスも含めて重みと呼ぶことにする。)

ヘルムホルツマシンにおける隠れユニットは、ボトムアップの認識モデルに対応する認識確率とトップダウンの生成モデルに対応する生成確率の 2 種類の確率を持つ。この認識モデルと生成モデルとの分離がボルツマンマシンとの最大の相違点であり、この分離によりボルツマンマシンに比べ極めて効率の良い学習が実現されている。

ヘルムホルツ自由エネルギーを計算する際、隠れユニットの状態 α の全ての組み合わせを考慮しなければならない。隠れユニットが N 個の場合、2^N 通りの組み合わせがあり現実的な計算は絶望的となる。そこでヘルムホルツマシンでは、各ユニットの状態は互いに独立であると仮定しこの問題に対処している。

[認識確率の計算]

第 l 層の第 j ユニットに着目すると、このユニットが状態 1 となる認識確率 q_j^l は第 $l-1$ 層の各ユニットの状態 $s_i^{l-1} \in \{0,1\}$、およびそれらと着目ユニット

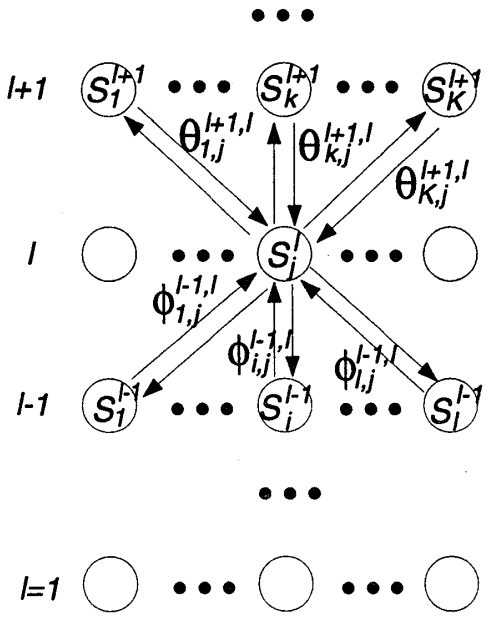

図 1: ヘルムホルツマシンのネットワーク構造

間の認識重み $\phi_{i,j}^{l-1,l}$, $i=1,2,\ldots$ の関数として

$$q_j^l(\phi, s^{l-1}) = \sigma(\sum_i s_i^{l-1}\phi_{i,j}^{l-1,l}) \qquad (7)$$

として定義される。但し、$s=(s_1,s_2,\ldots)$ とする。また、$\sigma()$ は Sigmoid 関数とする。各ユニットの独立性の仮定より、ネットワーク全体の認識確率、即ち、$\tilde{P}_{\alpha|d}(\phi,d)$ は

$$\tilde{P}_{\alpha|d} = \prod_{l>1}\prod_j (q_j^l)^{s_j^l} \times (1-q_j^l)^{1-s_j^l} \qquad (8)$$

として計算できる。但し、入力層 ($l=1$) でのユニットの認識確率は入力データの値 (s_j^1) に応じて $q_j^1 = s_j^1$ とする。

[生成確率の計算]

次に、第 l 層の第 j ユニットが状態 1 を生成する確率は、第 $l+1$ 層の各ユニットの状態 $s_k^{l+1} \in \{0,1\}$、およびそれらと着目ユニット間の生成重み $\theta_{k,j}^{l+1,l}$, $k=1,2,\ldots$ の関数として

$$p_j^l(\theta, s^{l+1}) = \sigma(\sum_k s_k^{l+1}\theta_{k,j}^{l+1,l}) \qquad (9)$$

として定義される。従って、独立性の仮定から、隠れ層のユニットの生成確率 $P(\alpha;\theta)$ は次式で計算できる。

$$P(\alpha;\theta) = \prod_{l>1}\prod_j (p_j^l)^{s_j^l} \times (1-p_j^l)^{1-s_j^l} \qquad (10)$$

また、入力層のユニットの生成確率 $P(d|\alpha;\theta)$ は

$$P(d|\alpha;\theta) = \prod_j (p_j^1)^{s_j^1} \times (1-p_j^1)^{1-s_j^1} \qquad (11)$$

となる。式 (8),(10),(11) を式 (5) に代入し整理することにより、ヘルムホルツマシンの自由エネルギーは次式のようなる。

$$\mathcal{F}(d,\theta,\phi) = \mathrm{E}_{\tilde{P}(\alpha|d)}\{\sum_{l\geq 1}\sum_j s_j^l \log\frac{q_j^l(\phi,s^{l-1})}{p_j^l(\theta,s^{l+1})} + (1-s_j^l)\log\frac{1-q_j^l(\phi,s_{l-1})}{1-p_j^l(\theta,s^{l+1})}\} \qquad (12)$$

2.3 学習アルゴリズム

学習アルゴリズムは、基本的には式 (12) の θ,ϕ に関する最小化として導出されるが、以下に述べるように 2 つのアプローチによる効率化のための近似解法が提案されている。

(1) 平均場近似法

式 (12) では $\tilde{P}(\alpha|d)$ の期待値の対象となる確率変数 s が分母と分子に存在し、このままでは計算が困難となる為、期待値を各々独立に実行して近似計算する。簡単の為、$\tilde{P}(\alpha|d)$ による期待値を $<>$ で表すと、式 (12) の期待値計算は次式の様に近似される。

$$\mathcal{F}(d,\theta,\phi) \simeq \sum_{l\geq 1}\sum_j (< s_j^l > \log\frac{< q_j^l >}{< p_j^l >} + (1-< s_j^l >)\log\frac{1-< q_j^l >}{1-< p_j^l >}) \qquad (13)$$

s の期待値は認識確率を用いると、$< s_i^{l-1} >= q_i^{l-1}$, $< s_k^{l+1} >= q^{l+1}$ と計算できるので、式 (13) は

$$\mathcal{F}(d,\theta,\phi) \simeq \sum_{l\geq 1}\sum_j q_j^l(\phi,q^{l-1})\log\frac{q_j^l(\phi,q^{l-1})}{p_j^l(\theta,q^{l+1})} + (1-q_j^l(\phi,q^{l-1}))\log\frac{1-q_j^l(\phi,q^{l-1})}{1-p_j^l(\theta,q^{l+1})} \qquad (14)$$

となる。但し、$\boldsymbol{q}=(q_1,q_2,\ldots,)$とする。

式(14)の右辺は一種のKL情報量と見なせるので、

$$\mathcal{F}(d,\theta,\phi)\simeq\sum_{l\geq1}\sum_{j}\mathrm{KL}(q_j^l(\phi,\boldsymbol{q}^{l-1})\parallel p_j^l(\theta,\boldsymbol{q}^{l+1}))$$

と書ける。結局、ネットワークパラメータθ,ϕは

$$\sum_{l\geq1}\sum_{j}\mathrm{KL}(q_j^l(\phi,\boldsymbol{q}^{l-1})\parallel p_j^l(\theta,\boldsymbol{q}^{l+1}))\rightarrow\min_{\theta,\phi}$$

の最小化として求めらる。

(2) Wake-Sleep アルゴリズム

Wake-Sleep アルゴリズムでは、θとϕを交互に学習する。即ち、下記の Wake-phase と Sleep-phase とを収束するまで繰り返す。

学習は確率的近似法[7]に基づく。確率的近似法とは、目的関数が期待値として$J(\Theta)=\mathrm{E}_{p(\boldsymbol{x})}\{f(\boldsymbol{x};\Theta)\}$と表現されている時、パラメータを

$$\Theta^{(t+1)}=\Theta^{(t)}-\eta(t)\left[\frac{\partial f}{\partial\Theta}\right]_{\Theta=\Theta^{(t)}}$$

として逐次推定する方法である。通常の勾配法と異なり、Jは単調に減少しないが平均的に減少する。即ち、

$$\sum_{t=1}^{\infty}\eta(t)=\infty,\ \sum_{t=1}^{\infty}\eta(t)^2<\infty,\ \lim_{t\rightarrow\infty}\eta(t)=0$$

を満たせば、$\mathrm{E}\{\Delta J\}<0$が理論保証される。

[Wake-phase]

Wake-phase では認識モデルの重みϕを固定し、ボトムアップに、現在の認識モデルの重みϕに基づく認識確率$q_j^l(\phi,\boldsymbol{s}^{l-1})$に従って$s_j^l$の値を逐次、確率的に定める。そして生成モデルの重みθをトップダウンに学習する。

ここで、式(12)の右辺の対数の部分を注意深く見ると、部分的に$\mathrm{KL}(q_j^l\parallel p_j^l)$と見なせる。従って、$\mathcal{F}$を最小化することは、$\mathrm{KL}(q_j^l\parallel p_j^l)$の最小化、即ち、認識モデルを真の分布と見なし、生成モデルを認識モデルに“近づく”ように学習させることと解釈できる。そこで、式(12)を目的関数として、ϕを固定し確率的近似法を適用することにより次式の生成モデルの重みの学習則を得る。

$$\theta_{k,j}^{l+1,l}(t+1)=\theta_{k,j}^{l+1,l}(t)-\eta(t)s_k^{l+1}(s_j^l-p_j^l)\quad(15)$$

式(15)を見ると、$s_k^{l+1}=1$の時、$(s_j^l-p_j^l)$に比例して$\theta_{k,j}^{l+1,l}$が更新されるという直観的に自然でかつ単純な学習則となっていることが分かる。計算も局所的に実行できるため極めて効率的である。

[Sleep-phase]

Sleep-phase では生成モデルのパラメータθを固定し、トップダウンに、現在の生成モデルのパラメータθに基づいて、生成確率$p_j^l(\theta,\boldsymbol{s}^{l-1})$に従って$s_j^l$の値を逐次確率的に定める。そして認識モデルのパラメータϕをボトムアップに学習する。

故に、ここでは$\mathrm{KL}(q_j^l\parallel p_j^l)$の最小化ではなく、$\mathrm{KL}(p_j^l\parallel q_j^l)$の最小化を実現しなければならない。そこで、式(11)でpとqを置換した：

$$\mathcal{F}(d,\theta,\phi)'=\mathrm{E}_{\tilde{P}(\alpha|d)}\{\sum_{l\geq1}\sum_{j}s_j^l\log\frac{p_j^l(\theta,\boldsymbol{s}^{l-1})}{q_j^l(\phi,\boldsymbol{s}^{l+1})}+(1-s_j^l)\log\frac{1-p_j^l(\theta,\boldsymbol{s}_{l-1})}{1-q_j^l(\phi,\boldsymbol{s}^{l+1})}\}\quad(16)$$

を目的関数とし、Wake-phase と同様に、θを固定して確率的近似法を適用することにより、以下の学習則を得る。

$$\phi_{j,k}^{l,l+1}(t+1)=\phi_{j,k}^{l,l+1}(t)-\eta(t)s_j^l(s_k^{l+1}-q_k^{l+1})\quad(17)$$

式(17)は式(15)と同様に、$s_j^l=1$の時、$(s_k^{l+1}-q_j^{l+1})$に比例して$\phi_{j,k}^{l,l+1}$が更新される。

Wake-Sleep アルゴリズムは、上記から分かるように若干直観的な導出を含み、かつ、θとϕの目的関数が異なるという問題があるが、得られた学習則は極めて簡明で、かつ、実用面での有効性も確認されている[10]。

以上、ヘルムホルツマシンについて概説したが、次節では、多変量解析の分野で非観測変数を含む分布推定法として知られる潜在変数モデルについて概説し、ヘルムホルツマシンと従来の多変量解析手法との関係について考察する。

3 潜在変数モデル

3.1 一般形

潜在変数モデルとは、観測データ $\boldsymbol{x}$ と観測されない潜在変数 $\boldsymbol{z}$ との間に

$$\boldsymbol{x} = f(\boldsymbol{z}; W) + \boldsymbol{\epsilon} \tag{18}$$

なる関係を仮定する。ここで、$\boldsymbol{x}, \boldsymbol{z}$, および $\boldsymbol{\epsilon}$ は確率変数 (確率ベクトル) で $\boldsymbol{z}$ の次元は $\boldsymbol{x}$ のそれより小さいかまたは等しいとする。$\boldsymbol{\epsilon}$ はノイズモデルで、W はモデルパラメータである。即ち、観測データ $\boldsymbol{x}$ は $\boldsymbol{z}$ と $\boldsymbol{\epsilon}$ の分布に従って確率的に生成されることになる。

3.2 線形ガウスモデル

潜在変数モデルの具体例として、多変量解析の分野で著名な因子分析モデル [4] がある。観測データを $\boldsymbol{x} \in \mathcal{R}^p$ とすると、因子分析とは、観測データの p 個の各変数を、それらに共通する、$q(< p)$ 個の変数からなる共通因子 ($\boldsymbol{z}$ の各要素) と、観測データの各変数に固有な独自因子 ($\boldsymbol{\epsilon}$ の各要素) とに分解し、潜在変数である共通因子を通してデータの簡潔な記述を行う統計手法である。因子分析モデルでは、線形関数：

$$f(\boldsymbol{z}; \Theta) = W\boldsymbol{z} + \boldsymbol{\mu}$$

を用い、$\boldsymbol{z}, \boldsymbol{\epsilon}$ は通常、以下の多変量正規分布を仮定する。

$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\psi_1, \ldots, \psi_p))$$

更に、$\boldsymbol{z}$ と $\boldsymbol{\epsilon}$ は統計的に独立とすると、$\boldsymbol{x}$ の共分散行列 Σ_x は、簡単な計算により

$$\Sigma_x = WW^T + \mathrm{diag}(\psi_1, \ldots, \psi_p)$$

となる。従って、$\boldsymbol{x}$ の分布は以下で得られる。

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma_x)$$

因子分析はデータの簡約化 (次元圧縮) という点で、主成分分析 (PCA) と共通するが、後者は主成分の抽出に留まり、分布推定を実現するモデルとなっていないという点で前者と本質的に異なる。尚、最近、分布推定のための PCA モデルも提案されている [8]。また、紙面の都合上詳細は省略するが、近年ニューラルネットの分野で精力的に研究されている独立成分分析 (ICA) も潜在変数モデルに属する。

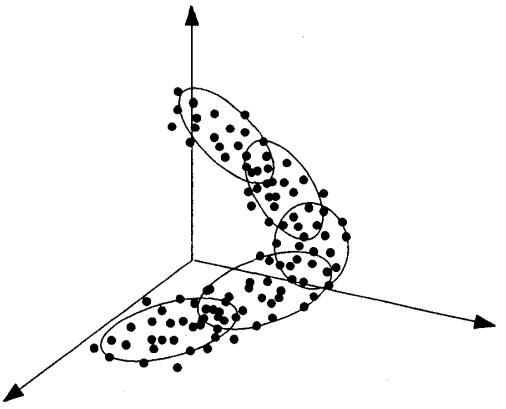

図 2: 混合線形ガウスモデル。黒点は 1 つのデータ $\boldsymbol{x}$ に、楕円体は 1 つの潜在変数モデルに各々対応する。p 次元空間中で複数の潜在変数モデルによりより低次元の多様体が抽出される。

3.3 非線型モデルへの拡張

前説の議論から明らかなように、潜在変数モデルとは、隠れ変数 (非観測変数) の分布を通して観測データが生成されるとする確率モデルである。換言すれば、次元圧縮された $\boldsymbol{z}$ の空間で $\boldsymbol{x}$ の分布 $p(\boldsymbol{x})$ を推定していると言える。最近、より精度良く $p(\boldsymbol{x})$ を推定すべく、線形ガウスモデルのような大局的な次元圧縮ではなく、有限個の線形ガウスモデルを混合した、混合線形ガウスモデルが提案されている [9]。

混合線形ガウスモデルは、直観的には、図 2 に示すように、$\boldsymbol{x}$ の空間中から低次元の多様体を抽出していると解釈できる。図 2 を見ると、いわゆる通常のガウス混合分布モデルと混同するかも知れないが、図 2 の各楕円の次元は $\boldsymbol{x}$ の次元より小さくなっているという点で異なる。即ち、クラスタリングと次元圧縮を同時に実現する斬新な確率モデルとなっている。学習アル

ゴリズムはEMアルゴリズムに基づいて定式化されている[9]が、現在の性能は十分ではない。つまり、多くの局所最適解に収束し、十分な推定結果が得られていないことを付記しておく。

ヘルムホルツマシンの場合、x, zが離散値をとるという制約はあるが、式(18)でxを観測データ、zを隠れユニットの状態、fをネットワークのユニットの出力、Wを結合重みに各々対応させると、ヘルムホルツマシンを、潜在変数モデルの一形態と見なすことができることは明らかであろう。この場合、大局的に非線形なモデルとなっている。即ち、ヘルムホルツマシンは、非線形潜在変数モデルのネットワーク実現と言える。

4 まとめと今後の課題

本稿では、分布推定のための確率ニューラルネットワークとして近年提案されたヘルムホルツマシンについて概説し、更に、潜在変数モデルというより上位の視点でヘルムホルツマシンを概観し、従来の多変量統計解析手法との関係について概説した。

ヘルムホルツマシンの学習則は統計物理のアナロジーを用いて自由エネルギー最小化の枠組みで定式化されているが、その理論的妥当性は不完全データからの最尤推定という統計の基礎理論に基づいている。

ヘルムホルツマシンは従来の線形潜在変数モデルを非線形に拡張したものとして興味深いモデルであるが、線形潜在変数モデルの混合モデル(混合線形ガウスモデル)も強力な対抗馬と言える。大局的な非線形モデルと局所線形の混合モデルとの差はアプローチの差であり、優劣を一般的に論じることは困難であり、今後実用面で評価すべきであろう。いずれにしても、重要な事は、これらが従来の統計モデルの焼き直しではなく、非線形に拡張した新たなモデルとして位置づけることができることである。

しかしながら、潜在変数モデルに基づく確率ニューラルネットワークの研究はまだ始まったばかりで、次元の決定法、学習法等の実用面での本質的な問題を残しており、今後の更なる発展が期待されている。尚、本稿で概説したヘルムホルツマシンの学習は明らかに教師無し学習であるが、教師有り学習等の様々な変種も提案されている。これについては文献[3]を参照されたい。

参考文献

[1] Hinton, G.E. and Sejnowski, T.J., "Learning and relearning in Boltzman machines," *Parallel Distributed Processing: Explorations in Microstructure of Cognition* (D.E. Rumelhart and J.L. McClelland, eds.), Cambridge, MA: MIT Press, 1986.

[2] Dayan, P., Hinton, G.E., Neal, R.M, and Zemel, R.S., "The Helmholtz machine," *Neural Computation*, vol.7, pp.889–904, 1995. MA, pp. 681–687, 1995.

[3] Dayan, P. and Hinton, G.E., "Varieties of Helmholtz machine," *Neural Networks*, vol.9, no.8, pp.1385–1403, 1996.

[4] Anderson, T.W., "An introduction to multivariate statistical analysis," John Wiley& Sons, New York, chap.14, 1984.

[5] 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. 39, pp. 1–38, 1977.

[6] Ueda, N. and Nakano, R., "Deterministic annealing EM algorithm," *Neural Networks*, in press, 1998.

[7] Robbins, H. and Monro, S., "Stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp.400–407, 1951.

[8] Bishop, C. M., Svensen, M. and Williams C. K. I., "EM optimization of latent-variable density models," *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge MA, pp. 465–471, 1996.

[9] Ghahramani, Z. and Hinton, G. E., "The EM algorithm for mixtures of factor analyzers," Tech. Report CRG-TR-96-1, Univ. of Toronto, 1997.

[10] Frey, B.J., Hinton, G.E., and Dayan, P., "Does the wake sleep algorithm produce good density estimators?" *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge MA, pp. 661–667, 1996.