脳内情報表現への情報理論的アプローチ

阪口 豊，樺島 祥介 †
電気通信大学大学院 情報システム学研究科
東京工業大学大学院 総合理工学研究科 †

1 はじめに

本稿では，脳内での情報表現の学習原理とその実現に関する研究の中で主に情報理論的なアプローチについて述べる．脳内での情報処理の多くは，入力された情報をそれと異なる表現に変換し次の層に出力するという形をとっていると思われる．それでは，その変換はどのような原理に基づいておこなわれているのであろうか．ここでは，特徴抽出細胞の受容野に関する形状，トポロジーや発火頻度に関するスパース性といった感覚野に特徴的な性質に着目し，生理学的知見や情報の変換原理，またそれを脳内で実現するための制約との関連性といった問題について議論する．

2 古典的な特徴抽出細胞自己形成のモデル

1970 年代には，視覚野に見られる特徴抽出細胞が生後の感覚体験を通じて形成されることを示唆する実験報告が相次いだ（例えば，Blakemore and Cooper[7]）．このような報告を受けて，この時期には，この現象を説明するための神経回路モデルがいくつか提案されている．

特徴抽出細胞の自己形成を説明する学習モデルを始めて提案したのは，Malsburg[11] である．彼は，Hebb 学習と荷重ベクトルの正規化（荷重の総和一定）を組み合わせた神経回路モデルを提案し，それに種々の線分を模した刺激を繰り返し与えると，線分の方位を表現する細胞が自動的に形成されることを数値実験により示した．なお，Hebb 学習とは，ある細胞 A が興奮しているとき，A からの刺激を受け取った別の細胞 B が興奮すれば，両者のあいだの結合荷重が強化されるという学習則であり，心理学者 Hebb が 1949 年にその原理を提案したものである．

その後，Malsburg のモデルを変形した様々なモデルが提案されたが，ここでは，その中で数学的にシンプルな形で定式化されている Amari and Takeuchi [1] を紹介する．このモデルでは，細胞は線形閾値素子として定式化され，その出力は次式で与えられる．

$$y = \mathbf{1}[\sum_j w_j x_j - w_0 x_0] \tag{1}$$

ここで，x_j は細胞への興奮性入力，w_j は荷重，x_0 は抑制性入力，w_0 は抑制性入力に対する荷重である（抑制性入力が通常の閾値に対応する働きをしている）．

一方，学習則は次式で定義される．

$$\tau \frac{\partial w_j}{\partial t} = -w_i + c_1\ y\ x_j \tag{2}$$

$$\tau \frac{\partial w_0}{\partial t} = -w_0 + c_0\ y\ x_0 \tag{3}$$

ここで，τ は荷重減衰の速さを定める時定数，c_0, c_1 は正の定数である．この式の第 1 項は荷重が一定速度で減衰することを表す項であり，第 2 項が Hebb 学習を表す項（入力 x_j と出力 y が同時に正になったときに増加）である．このモデルでは，減衰項と抑制性入力の学習を導入することにより，学習によって興奮性入力に対する荷重が発散してしまうのを防いでいる．

この学習方程式の安定平衡解は，その細胞を興奮させるような入力信号の平均値である．すなわち，細胞は学習により特定の信号のみに反応するようになる．各細胞の受容野の大きさ（反応する信号領域の広がり）は，c_0, c_1, x_0 の

値に依存して決まるので，これらの値が異なる様々な細胞が存在すると仮定すれば，いろいろな大きな受容野をもった細胞が形成されることになる．また，この学習方程式には安定平衡解は多数存在し，その値は荷重の初期値と入力信号の出現の仕方に依存して決まる（多安定性）．したがって，信号源が定常であり，また，荷重の初期値が十分に広く分布していれば，信号空間の様々な領域に対してそれぞれ特異的に反応する細胞が形成されることになる．

さて，Amari らのモデルは単一の細胞の動作を規定するものであり，神経細胞集団がシステムとしてどのようにして情報を表現するかという点については（初期値や係数の分布を除いて）触れられていない．この点に関して，先に述べた Malsburg のモデルでは，複数の特徴抽出細胞の間に，それらの距離に応じて，近いもの同士のあいだには興奮性，遠いもの同士のあいだには抑制性の相互結合（いわゆる側方抑制結合）が導入されており，細胞間の相互作用のメカニズムが組み込まれていた．

このような相互抑制結合は，学習を進める上で以下の二つの効果をもつ．一つは，抑制性の結合により，同時に興奮する細胞の数が制限されるという効果である．この効果により，同一の信号を表現する細胞数が制限されるため，細胞集団全体として様々な信号を表現できるようになり，同時に，以下で議論する sparse coding が実現される．もう一つの効果は，互いに近くに位置する細胞は同時に活動するために，学習の結果，それらはよく似た信号に対して反応するようになるという効果である．実際，Malsburg の数値実験の結果を見ると，互いに近くに位置する特徴抽出細胞はよく似た刺激に対して反応するようになるという性質を示している．このような性質を一般にトポグラフィ (topography) とよぶ．トポグラフィは，視覚野に限らず他の感覚野にも普遍的に見られる構造である．

Malsburg の実験結果から，Hebb 学習と側方抑制結合の組合せがトポグラフィ自己形成の原理であるという議論が生まれた．Malsburg は議論をこの点に絞った論文を残しており [18]，また，Amari [2] は，神経場のモデルに基づいてその性質を数理的に解析している．この理論は，topographic map 上の特徴抽出細胞の分布は，入力信号の出現頻度に比例することを示しており，このような学習モデルが，頻繁に出現する刺激を細かく表現する機構を作り出す能力を有していることを示唆している．さらに，Takeuchi and Amari [16] では，ある条件の下では，同じ学習方程式からカラム構造が安定な解として得られることを示している．

以上の研究で得られたトポグラフィ形成の本質的部分を取り出し，工学的に実現しやすい形で定式化したものが，Kohonen [8] による SOM(Self-Organizaing Map) である．彼は，側方抑制結合による神経力学は，最大入力を受け取った細胞が自分の周辺の細胞を引き連れて活動するという結果に生み出すことに着目した．すなわち，彼は，複雑な非線形ダイナミクスの部分を省略し，結果である「最大値検出」の部分だけを取り出してアルゴリズムを組み立てた．

SOM では，入力ベクトルを $\boldsymbol{x}$，各細胞の荷重ベクトルを $\boldsymbol{w}_i$ としたとき，$||\boldsymbol{x}-\boldsymbol{w}_i||^2$ が最小値をとる（つまり，入力信号と最も近い荷重ベクトルをもった）細胞が活動する．また，学習則は，活動した細胞とその周辺の細胞に関して，荷重ベクトルを入力ベクトルに近づける，という形で与えられる．

$$\Delta \boldsymbol{w}_i = c(\boldsymbol{w}i - \boldsymbol{x}) \tag{4}$$

この学習則により，トポグラフィ学習の本質的部分は実現される．

以上，特徴抽出細胞の自己組織化に関する古典的なモデルについて概観してきた．以上の議論から，これらのモデルの本質は，Hebb 学習と相互抑制結合にあることがわかるであろう．これらの性質は，現代的な定式化の中ではどのようにして現われてくるのであろうか？次章以下の議論の中で，この問いに対する答えを見いだしてほしい．

3　情報理論に基づくモデル化

シナプス効率の変化に基づいて現実の脳に類似した脳の情報表現を導く Malsburg, Amari, Kohonen などの古典的モデルは単純で分かりやすいモデルではあるが，その機能的な意味との関連性が見えにくい．そこで最近ではこれらのボトムアップ的なモデル化に対し，情報表現に関して望ましい機能の形成を学習の原理に据えてトップダウン的に学習則を導くモデル化が増えてきた．このようなモデル化はともすれば現実の脳との遊離が心配されるが，脳のように複雑なシステムを理解するためには有効かつ必要不可欠なアプローチである．

3.1 冗長度圧縮原理

自然界から感覚器官に入力される情報のほとんどはある種の規則性を持っている．そのため，表現が冗長 (redundant) である．視覚情報を例とすれば，晴れた空は一面空色をしており，木の葉はほぼ全面緑色，ビルの壁も普通は一面同一色で塗りつぶされ，人の肌も同一人種ではほぼ同じである．これは時間軸に対するスナップショットをとれば網膜上の広範な領域でほぼ同一な規則的刺激が入力されることを意味する．聴覚情報も然り．我々の音声ではヒトの聴覚系が聞き分けられる音の中のほんの一部しか用いられておらず，音楽の旋律，虫の鳴き声も可能な音の配列パターンのうち限られた規則性を持つものしかない．

ただし，外界からの情報が規則性を含んでいるか否かという問題はそれが表現されている符号化方法に相対的なものであることに注意しよう．たとえとして，以下のようなサイコロ賭博を考えよう．長半賭博では出たサイコロの目に関してそれが偶数か奇数かのみの情報が得られれば良い．そこで，A は数字の目が 1(奇数)，2(偶数) のみでかつ各々の目が等確率で出現する専用のサイコロを作った．賭博を目的とし長半の結果のみに興味のある B はこのサイコロを繰り返し振って得られる長半の系列は十分ランダムに見えるだろう．一方，それが細工されたサイコロであるとは知らず，普通のサイコロの目として認識する C には 1 あるいは 2 のみが繰り返される系列は非常に規則性が高く感じるであろう．つまり，受け手が考慮している可能性に対して現実に生じる可能性が相対的に少ないときに情報源に規則性があると思え，情報の表現方法 (C が長半専用のサイコロを普通のサイコロのように見ていること) に冗長度があると感じ，その裏返しとしてサイコロが細工されているという知識を得るのである．

情報理論的に言えばこれは次のように表現出来る [5]．ある感覚器官 (たとえば網膜) が処理可能な情報量の限界を C bits/sec とする．これに対し，外界から入力される情報量を H bits/sec とすれば一般に H は C よりも小なく

$$C - H \tag{5}$$

だけ冗長であり，それが大きい程対象に関して沢山の知識が得られると言える．

ところで，大量の情報を処理出来る感覚器官を持つためには大量の神経細胞が必要となりそれに比例したエネルギーを要する．仮に細胞の発火・非発火状態を 1 ビットに割り当てるとすれば，上で述べていることは単位時間当たり H 個くらいの細胞の発火・非発火で情報が表現出来るのにも関わらず感覚器官ではそれよりも多い C 個の細胞を使って情報表現を行なっていることを意味する．この表現がそのまま以降の情報処理に伝えられるのであればエネルギー的に無駄が生じ，疑いなく生命維持に不利である．

そこで，脳内では冗長度がなるべく少なくなるように情報が変換されている可能性が尤もらしい．これを脳内情報表現の形成に関する基本原理とする仮説を**冗長度圧縮** (Redundancy Reduction) という．Malsburg, Amari, Kohonen などの古典的モデルも多次元の入力信号に内在する規則性を抽出し，それ以外の成分を除去するネットワークが構成されるという意味で冗長度圧縮にかなったものである．実際、これを原理として採用したところでモデルに殊更新しい側面が付け加わる訳でもない．むしろ，冗長度圧縮を出発点とする現代的理論の利点は評価関数の最大／最小化という形で問題を定式化することで議論が整理され話の見通しが良くなることにある，と考えるべきであろう．もちろん，脳のように研究対象が複雑になればなるほど見通しの良い定式化がブレークスルーを得るための非常に重要な要因になる，ということはここであらためて強調するまでもない．

3.2 確率モデルと情報量

これまでのところ，広く受け入れられている説では脳内では各細胞の活動度の組み合わせによって情報が表現されていると考えられている (Rate coding + 分散表現).

この説に従えば脳内の情報処理とは入力層内での表現

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \tag{6}$$

をそれが投射する特徴抽出層での出力に関する表現

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_M) \tag{7}$$

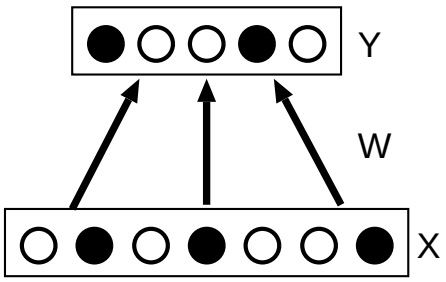

図 1: 脳内の情報変換

に変換することに他ならない (図１). この変換を担うのが２層を結ぶシナプス結合でありその可塑性により多様な変換が可能となる．ここで，各々のベクトルの成分はあるタイムスケールで計った各細胞の平均発火率そのもの，あるいはその平均値からのずれを表していると考えればよい.

ただし，細胞の応答には不確かな要素が多いので入力 $\boldsymbol{x}$ が決まると必ず確定的に出力 $\boldsymbol{y}$ が決まる訳ではない．このような状況下での入出力関係は条件付き確率

$$P(\boldsymbol{y}|\boldsymbol{x}) \tag{8}$$

を用いると都合良く表現することが出来る.

以上のような確率モデルによる記述を用いる最大の利点は情報理論との親和性の高さである．たとえば，情報理論に従えば入力層に与えられる情報量は

$$H(X) = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \ln P(\boldsymbol{x}) \tag{9}$$

によって数量化できる．これと入力層で処理可能な情報量の限界値との差により入力層での冗長度が計算される．ここで，$P(\boldsymbol{x})$ は外界からの刺激により入力層にパターン $\boldsymbol{x}$ が生じる確率を表す．さらに，出力層 (特徴抽出層) での情報量も機械的に

$$H(Y) = -\sum_{\boldsymbol{y}} P(\boldsymbol{y}) \ln P(\boldsymbol{y}) = -\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}) P(\boldsymbol{x}) \ln \left[\sum_{\boldsymbol{x}} P(\boldsymbol{y}|\boldsymbol{x}) P(\boldsymbol{x}) \right] \tag{10}$$

によって求められ，それを用いて入力層と同様，出力層での冗長度を計算することが可能となる．つまり，情報理論を用いることにより情報処理の過程で表現の冗長度が如何に変化するのか，という問題を定量的に扱うことが可能となるのである.

3.3 認識モデルと生成モデル

前節では情報の流れとして入力層から出力層への順方向的なものを例として取り上げたが，脳内には V1 から LGN への結合などのように出力側から入力側へ逆方向の結合が存在する場合もある (図２). 確率モデルの観点からこれらの結合の役割を考えてみよう.

まず，順方向の結合は観測可能なデータ (入力)$\boldsymbol{x}$ を用いて隠れた変数 (出力) $\boldsymbol{y}$ を計算する役割を担っている．これはちょうど提示されたデータを解釈あるいは認識するプロセスを彷彿とさせるためこのような役割を行なう確率モデルは一般に認識モデルと呼ばれる.

それに対し，逆方向の結合は隠れた変数の値を用いて観測可能なデータを生成する役割を持つ．このようなモデルは例えば，外界を模倣したりあるいは予測したりする際には必ず必要であり一般に生成モデルと呼ばれる．認識モデルが前節で導入した条件付き確率 (8) により表されるのに対し，生成モデルは $\boldsymbol{y}$ を与えた際に $\boldsymbol{x}$ が得られる条件付き確率 $P(\boldsymbol{x}|\boldsymbol{y})$ により表現される.

Bayes 公式

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})}{\sum_{\boldsymbol{x}} P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})} \tag{11}$$

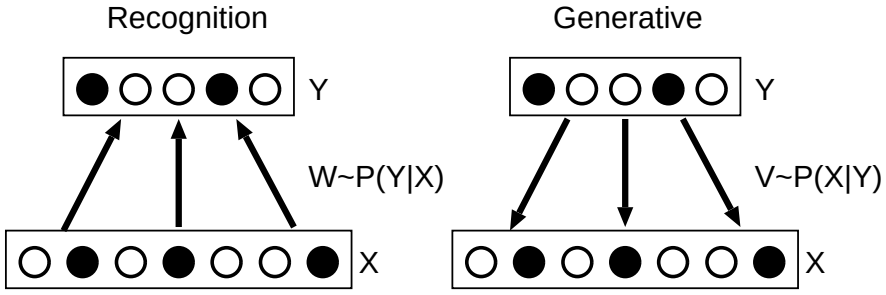

図 2: 認識モデルと生成モデル

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})}{\sum_{\boldsymbol{y}} P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})} \tag{12}$$

を用いれば，生成モデル，認識モデルからそれぞれに対応する “特別な” 認識モデル，生成モデルを求めることが出来る．

もちろん，生成モデル，認識モデルは数式の上では別の役割を担うものであり互いに無関係に存在して構わない．ただし，脳を外界からの情報を用いて学習により適切な応答を作り上げる確率モデルとみなすと，学習則の局所性など合理的な制約条件を満たすためには公式 (11)，(12) を通してこれらは密接な関係を持たなくてはならないことが多い．

3.4 最大／最小化原理

確率モデルによる記述に従うと冗長度圧縮原理を定量的に書き下すことが出来る，ただし，この原理はあくまでも大まかな指針を与えるものでありその具体的実現には様々なヴァリエーションが考えられる．どのヴァリエーションを用いるかは問題設定，ネットワーク構造との相性，計算量などに強く依存する．以下に，いくつかの例を紹介する．

3.4.1 容量最小化

外界からの情報 $\boldsymbol{s}$ は入力 $\boldsymbol{x}$ を通って出力 $\boldsymbol{y}$ へ変換される．ただし，各段階で相加性のノイズが加えられるとする．出力層における表現の冗長度 $\mathcal{R}$ を

$$\mathcal{R} = 1 - I(\boldsymbol{y}, \boldsymbol{s})/C_{out}(\boldsymbol{y}) \tag{13}$$

により定義する．ここで，$I(\boldsymbol{y}, \boldsymbol{s})$ は外界からの情報 $\boldsymbol{s}$ と出力層での表現 $\boldsymbol{y}$ との間の相互情報量

$$I(\boldsymbol{y}, \boldsymbol{s}) = \sum_{\boldsymbol{s},\boldsymbol{y}} P(\boldsymbol{y}, \boldsymbol{s}) \ln \left[\frac{P(\boldsymbol{y}, \boldsymbol{s})}{P(\boldsymbol{s})P(\boldsymbol{y})} \right] \tag{14}$$

を表し，その従属性が高い程大きな値を取る．直観的には $\boldsymbol{s}$ が与えられた際に $\boldsymbol{y}$ に伝わる情報の大きさを表すものと考えて良く，入力から出力へのシナプス結合の関数となる．また，$C_{out}(\boldsymbol{y})$ は出力層において処理可能な情報量の限界値を表す容量である．これも，シナプス結合の関数である．

式 (14) において生物が外界の情報を必要最低限得なければならないことを考えると相互情報量 $I(\boldsymbol{y}, \boldsymbol{s})$ はある程度の値を保たなければならない．そこで，式 (13) で表される冗長性を減らすため $I(\boldsymbol{y}, \boldsymbol{s})$ がある一定の値で拘束された条件下で容量 $C_{out}(\boldsymbol{y})$ を最小にするように結合が決める．このような最小化原理に基づけば，V1 で見られる局在化された受容野がモデルにより形成されることが報告されている [4].

3.4.2 情報量最大化

冗長度圧縮のための最大／最小化原理として広く用いられている評価基準の一つに情報量最大化 (Information Maximization, Infomax) というものもある [9, 6].

Infomax の基本的なアイデアは入力 $\boldsymbol{x}$ と出力 $\boldsymbol{y}$ との間の相互情報量

$$I(Y,X) = H(Y) - H(Y|X) \tag{15}$$

がなるべく大きくなるようにシナプス結合の値を決めるというものである．相互情報量は入力 $\boldsymbol{x}$ が与えられた際に出力 $\boldsymbol{y}$ に伝わる情報量の大きさと考えられるから出力層での容量が一定とすればこの値が大きい程出力層での冗長度は圧縮される．つまり、冗長度圧縮にかなった評価基準である．

ここで，もし出力が入力と可逆な関数 $G(\boldsymbol{x})$(シナプス結合は G の中にパラメータとして含まれている) を用いて

$$\boldsymbol{y} = G(\boldsymbol{x}) + \boldsymbol{\eta} \tag{16}$$

のように与えられる場合，$H(Y|X) = H(N)$ となり式 (15) はシナプスの値によらない定数となることが知られている [12]．ただし，$\boldsymbol{\eta}$ は変換の際に加わる相加性のノイズである．つまり，この場合 Infomax は出力層での情報量最大化と等価になる．

Infomax は後に述べる V1 における受容野形成のモデルの中で Linsker によってはじめて提唱された最大化原理である [9]．最近ではこれを非線形素子に拡張し一般の可逆な $N \to N$ 写像を対象とすることで，この原理は脳のモデルに限らず Blind deconvolution[1]などの信号処理の問題に広く応用出来ることが明らかにされている [6, 3].

3.4.3 独立成分解析

Infomax とは独立な考え方から出発しているが類似した基準として独立成分解析 (Independent Component Analysis, ICA) というものがある．

ICA とは出力 $\boldsymbol{y}$ に関する確率分布に関して成分間の従属性が最も小さくなるように，言い換えれば成分間の独立性が最も大きくなるような入力 $\boldsymbol{x}$ から出力 $\boldsymbol{y}$ への一次変換 M を求める問題である [3].

形式的には次のような最小化問題を考えれば良い．入力 $\boldsymbol{x}$ の分布と一次変換 M によって定まる出力 $\boldsymbol{y}$ の確率分布を

$$P(\boldsymbol{y}) = P(y_1, \ldots, y_N); \quad \boldsymbol{y} = M\boldsymbol{x} \tag{17}$$

とする。これを用いると、その中の 1 つの変数 y_l $(l = 1, 2, \ldots, N)$ のみに着目し他の変数の情報は無視した分布 (周辺分布) が

$$P_l(y_l) = \sum_{y_{j\neq l}} P(\boldsymbol{y}) \tag{18}$$

のように計算できる．この周辺分布から

$$\prod_{l=1}^{N} P_l(y_l) \tag{19}$$

という分布を再構成する．この分布は 1 つの変数に関する統計性に関しては真の分布 (17) と同じ情報を持っているが，変数間の依存関係に関しては完全に独立である．そこで，2 つの分布 (17) と (19) の間に何らかの距離 $\mathcal{D}$ を導入し

$$\mathcal{D}\left(P(\boldsymbol{y}), \prod_{l=1}^{N} P_l(y_l)\right) \tag{20}$$

を最小にするように一次変換を決める．

これがなぜ冗長度圧縮原理に関連するかを理解するためには前述の Infomax との関係を考察すればよい．式 (20) において分布の距離 $\mathcal{D}$ として KL ダイバージェンスを採用すると

$$\mathcal{D}\left(P(\boldsymbol{y}), \prod_{l=1}^{N} P_l(y_l)\right) = \sum_{l=1}^{N} H(y_l) - H(y_1, y_2, \ldots, y_N) \tag{21}$$

[1] N 個の統計的に独立な情報源から生成された信号が混合されて観測される場合に，N 個の独立な観測点から得られる観測情報を用いてそれを元の独立な情報に分解する問題.

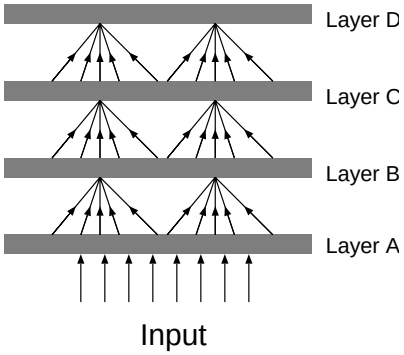

図 3: Linsker モデル

という式が得られる．つまり，この場合 ICA は出力層における相互情報量最小化に他ならない．この式を変形すると Infomax での評価関数である出力層での情報量が

$$H(y_1, y_2, \ldots, y_N) = \sum_{l=1}^{N} H(y_l) - \mathcal{D}\left(P(\boldsymbol{y}), \prod_{l=1}^{N} P_l(y_l)\right) \tag{22}$$

という形で表される．これは Infomax が出力層のおいて I) なるべく個々のユニットが表現する情報量 $H(y_l)$ を大きくする，II) 相互情報量 $\mathcal{D}$ を小さくする，という 2 つの要請から成り立っていることを意味している．

もし仮に，この式において I) よりも II) の要請のほうがより大きなウェートを占める場合には結果的に ICA と Infomax とほぼ同じ評価基準を与えると考えられ，その意味で情報量圧縮にかなった基準になっている [13, 5]．もちろん，状況によっては I) のほうが II) よりも重要になる場合があり，その場合は両者の結果は定性的に異なったものになる。

Blind deconvolution の文脈では入力信号の統計性がガウス分布よりも中心が強調されかつ裾野が長い super-gaussian である際には ICA と Infomax は定性的に同じ結果を与え，その逆の sub-gaussian 場合には定性的に異なる結果を与えると考えられている [6]．ちなみに，音声，音楽などの聴覚情報や自然画像に代表される視覚情報など自然界に存在する情報の多くは super-gaussian 分布に従っており，そのような入力に対しては ICA，Infomax の基準は定性的に同様な結果を与えるものと予想される．

4 情報理論に基づいた V1 における受容野形成のモデル

前節では，確率モデルを用いた記述により，冗長度圧縮原理が一般的な意味での情報量に関する最大／最小化問題として定式化され数学的議論が可能となることを述べた．その中でも少し触れたが，本節ではこのような考え方が実際の研究の中でどのように活かされているのかより詳しく紹介するため，受容野形成に関する代表的な 3 つのモデルを詳説する．

4.1 Linsker のモデル

V1 には中心が明るくその周りが暗いような入力に強く反応する細胞 (center-surround cell) やある方向をもった線分に強く反応する細胞 (orientation-selective cell) など特徴的な形状に反応する細胞が数多く存在する．多くのモデルではこれは自然界に存在する視覚刺激に内在する構造を学習した結果であると考えられていたが，Linsker (1986) は多層の feed-forward 型の情報処理機構に Hebb 則に基づいた簡単な学習を仮定すれば全くランダムな外部刺激からこのような特徴的な反応特性を持つ細胞が形成されることを示した [9].

彼の考えたモデルは図 4.1に示すように A，B，C，D，… と名付けられた層が順方向に結合された feed-forward 型ニューラルネットワークモデルである．各層のニューロンは一つ前の層の限られた近傍 (受容野) に含まれる細胞の

みから入力 V_j を受け

$$O = a + \sum_{j=1}^{K} w_j V_j \tag{23}$$

に従って次の層に出力を伝える．ここで，w_j はシナプス結合，a はパラメータ，$j = 1, \ldots, K$ は着目した細胞の受容野に含まれる細胞を表す．

Linsker は学習則

$$\begin{aligned} \langle \Delta w_i \rangle &= \eta \left[\langle VO \rangle + b \langle V \rangle + c \langle O \rangle + d \right] \\ &= \eta \left[\sum_j C_{ij} w_j + \lambda \left(\mu - \sum_j w_j \right) \right] \end{aligned} \tag{24}$$

に従って結合 w_i は変化すると仮定した．ただし，ここで $\langle \cdots \rangle$ は入力に関する統計的平均を表し $a - d$ が学習則の詳細を決めるパラメータ，C_{ij} は入力信号に関する $K \times K$ の共分散行列，λ ははじめの形を変形して得られる $a - d$ の線形結合をまとめて書いたもの，η は学習率を表す．

さて，学習則 (24) は以下のコスト関数

$$E = -\frac{1}{2} \boldsymbol{w}^T C \boldsymbol{w} + \frac{\lambda}{2} \left(\mu - \sum_j w_j \right)^2 \tag{25}$$

の最小化条件から得られることに注意しよう．この式の第 1 項 $\boldsymbol{w}^T C \boldsymbol{w}$ は出力に関する分散を表し，これを小さくすることが冗長度圧縮に対応している，ただし、それに加えて入力信号のオフセットがゼロでないことから生じるペナルティー項が第 2 項目に存在する．また，このままでは結合パラメータは発散してしまうため同時に

$$w_- \leq w_i \leq w_+ \tag{26}$$

という拘束条件を導入し，この範囲を越えた場合にはパラメータの値を強制的に上下限値 w_+, w_- に戻すことにする．

彼は以上アルゴリズムに従い，まず，A → B の結合を学習した後，それを用いて B → C の結合を学習し，その後 C → D …という手順で計算機実験を行なった．彼の報告によれば図 3 に示すようにこのアルゴリズムに従えば入力層 A にはユニット間で無相関な刺激のみが入力されたにも関わらず center-surround cell が C 層に bilobed orientational-selective sell が G 層に観察された．

ここに述べたオリジナルの Linsker モデルはアルゴリズムの非線形性が強く解析は難しい．それでも，本質的な点を残し，より簡単化したモデルに関しては理論的解析が可能で特徴抽出細胞形成のメカニズムも明らかにされている [10, 17]．彼のモデルは簡単なメカニズムで特徴的な受容野の形成を説明した点で概念モデルとして画期的であった．しかしながら，結合形成のアルゴリズムに刈り込みルールや競合学習機構を “手でいれる” など生理学的妥当性という観点では課題が多い．

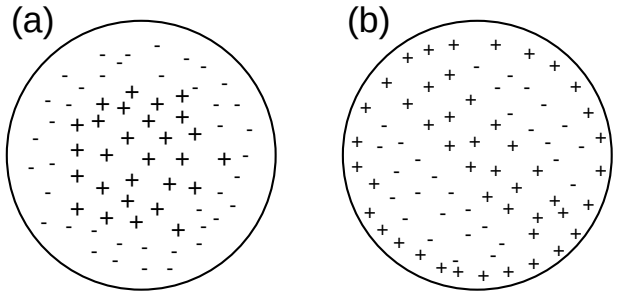

図 4: Linsker モデルで観察された特徴的な正負結合の配置．(A) C 層で観察されたもので周囲の入力が負で中心の入力が正のとき発火する center-surround cell を (B) は G 層で観察されたもので左斜めに傾いた線分に強く反応する orientation-selective cell.

4.2 Olshausen and Field のモデル

広く知られている V1 の特徴抽出細胞の特徴として i) 局所性 (localized)，ii) 方位選択性 (oriented)，iii) スケール選択性 (bandpass) という性質がある．また，V1 における情報は細胞の活動度が低くなるようにいわゆるスパースコーディング (sparse coding) で表現されているという事実も観察されている．これらの観察事実を情報理論的なモデル化によって相互に関連付けようとしたのが Olshausen and Field (1996) による研究である [14].

彼らのモデルでは LGN に対応した視覚情報に関する入力層と V1 に対応した出力層 (特徴抽出層) を仮定する．このモデルの特徴は、Linsker 等の他の多くのモデルとは異なり順方向結合 (認識モデル) ではなく，逆方法結合 (生成モデル) の学習を軸に議論を展開しているところである．具体的には LGN における像情報は V1 での活動度 a_i を用いて

$$I(x,y) = \sum_i a_i \phi_i(x,y) \tag{27}$$

のように表現されると仮定する．ここで，$I(x,y)$ は LGN での座標 (x,y) における細胞の活性度，$\phi_i(x,y)$ は V1 での細胞 i が発火した際に LGN の座標座標 (x,y) に誘発される活動度で V1 から LGN への結合を表現していると思えば良い．ちなみに解剖学的観測事実によれば V1 から LGN への逆向きの結合は LGN から V1 への順方向結合に比べて 5 ∼ 10 倍多いと報告されている．

冗長度圧縮の原理に従い，彼らは V1 では情報は低いエントロピーで符号化されていると考えた．更に，その表現形態がスパースコーディング (sparse coding) である，というのが彼らの主張である．ところが，一方でそれは外界から入力される画像情報をうまく再構成，言い替えると予測出来るものでなくてはならない．そこで，彼らは以下のコスト関数

$$E = -[\text{preserved information}] - \lambda[\text{sparseness of } a_i] \tag{28}$$

に関する最小化問題を考えた．ここで，第一項は自然画像と生成モデルにより再構成された画像との自乗誤差

$$[\text{preserved information}] = -\sum_{x,y} \left[I(x,y) - \sum_i a_i \phi_i(x,y) \right]^2 \tag{29}$$

を表し，第二項は V1 での符号化がスパースコーディングになるように導入された正則化項

$$[\text{sparseness of } a_i] = -\sum_i S\left(\frac{a_i}{\sigma}\right) \tag{30}$$

であり $S(\frac{a_i}{\sigma})$ は $a_i = 0$ で最小となりそれから σ 程度ずれると値が急増する単峰性の関数である．また，λ は学習則へのその影響の度合を調整するパラメータである．

この最小化問題は

$$P(I|a,\phi) \sim \exp[-(\text{preserved information})] \tag{31}$$

$$P(a) \sim \exp[-(\text{sparseness of } a_i)] \tag{32}$$

を導入すれば元画像 I から生成モデルを用いて V1 での活動度 a_i を推定する統計的推定問題と考えることもできる．

a_i についての最小化はコスト (28) に関する最急降下 (微分値を用いてコスト関数が小さくなる方向に状態を変化させるアルゴリズム) を表すダイナミクス

$$\dot{a}_i = b_i - \sum_j C_{ij} a_j - \frac{\lambda}{\sigma} S'\left(\frac{a_i}{\sigma}\right) \tag{33}$$

により行なうことができる．ただし，$b_i = \sum_{x,y} \phi_i(x,y) I(x,y)$，$C_{ij} = \sum_{x,y} \phi_i(x,y)\phi_j(x,y)$ である．これは画像が入力された際の短いタイムスケール ($\sim 100ms$) での細胞の時間応答を表すものと考えられる．

一方，彼らは結合を表す $\phi_i(x,y)$ の時間変化もこの最小化問題により記述されると考える[2]．具体的には $\phi_i(x,y)$ の変化に関する時間スケールは細胞応答の時間スケールと比較して十分長く多数の画像の提示により引き起こされると仮定し

$$\Delta\phi_i(x_m,y_n) = \eta \left\langle a_i \left[I(x_m,y_n) - \hat{I}(x_m,y_n) \right] \right\rangle \tag{34}$$

[2] より正確に言えばエネルギー (28) から得られる自由エネルギーに関する最小化.

を $\phi_i(x,y)$ に関する学習則とするのである．ここで，$\hat{I}$ は学習画像 I を提示された際に短いタイムスケールでのダイナミクス (33) により再構成された画像 $\hat{I}=\sum_i \hat{a}_i\phi_i(x_m,y_n)$，$\eta$ は学習率を表し，$\langle\cdots\rangle$ は多数の学習画像に関する平均を意味する．式 (34) はシナプス結合に関する Hebb 学習則に他ならないことに注意しよう．

Olshausen and Field は自然画像を入力として以上のアルゴリズムに従いモデルの学習を行なった．その結果得られた受容野について冒頭に述べた i) 局所性 (localized)，ii) 方位選択性 (oriented)，iii) スケール選択性 (bandpass) という性質を再現できたというのが彼らの主な結論である．

論文の中では触れられていないが彼らの報告は以下の点に関して興味深い疑問点を残している．彼らは生成モデルに関するコスト最小化原理 (冗長性圧縮) のみに基づいて，短いタイムスケールでの細胞応答のダイナミクス，長いタイムスケールでのシナプス結合の学習則を導出している．ところが，式 (33) を見ればわかるようにこのダイナミクス，学習則は順方向の結合 (認識モデル) を同時に導入しなくては局所的な計算で行なうことは不可能である．また，受容野というのは普通，特徴抽出層のある細胞に投射している入力層の細胞の範囲を示すものであり通常は生成モデルではなく認識モデルに付随する概念である．実のところ，ニューラルネットワーク的な表現をすれば，彼らは生成モデルを表す逆方向の結合に関して必ず同じ値の順方向結合が存在していることを「暗に」前提としているのである．しかしながら，生理学的には 2 層間の結合が対称である必然性はない．

以上の点に関して，Olshausenan and Field のモデルの現実的妥当性を検証する一つの方策は局所計算性を満足させるため生成モデルとともに認識モデルも導入し Wake-Sleep アルゴリズム (池田先生の講義参照) などで双方向の結合に関する学習を同時に行なわせてみることである．事実，線形で簡単なネットワークに関しては認識モデルの結合は生成モデルの結合を逆向きにしたものになることは確認されている．ただし，これはあくまでも非常に単純化した場合の結果であり少なくともシミュレーションレベルでの検証は必要であろう．

4.3 Rao and Ballard のモデル

近年，視覚系の神経細胞の活動が，従来の意味での受容野 (classical receptive field) の周囲に提示された刺激の影響を受けること (contextual modulation という．詳しくは佐藤先生の講義を参照) が多くの研究者によって報告されている．このような神経活動は，テクスチャ分離，隠蔽関係や主観的輪郭といった知覚体制化に使われているという説があるが，この現象が脳の様々な部分で起きているのであれば，現象ベースの解釈ではその意味を統一的に理解することができない．そこで，計算論的なモデルに基づいてその意味を理解しようとしたのが，ここで紹介するモデル [15] である．

彼らの基本的スタンスは，階層的構造をもつモデルを用いて「上位層に表現された内部モデルにより画像を表現する」というモデルベース（つまり生成モデル）の考え方である．そして，予測と画像のあいだに違いのある部分があれば，そこには内部モデルがもっていない新たな情報が含まれていると考えて，それを下位層の活動として表現する．逆に，画像が予測したとおりであれば，それで画像が完全に表現されていることになるので，下位層は活動しない．一方，下位層で検出された誤差情報は上位層に送られ，上位層で表現されている内部モデルはその誤差情報に基づいて修正される．上位層は，さらにその上位層とのあいだで同様の処理をするようになっており，これにより階層的なネットワークの中を双方向に情報が行き来するようになる．

以上の操作を繰り返すことにより，モデルの内部には入力画像に対する階層的な表現が次第に形成されていく．彼らは，このような双方向型階層的モデルによって入力を表現する方法を predictive coding と呼んでいる．双方向型階層的モデルを考える理由は，視覚系では LGN, V1, V2 などの間に双方向性の結合が普遍的に見られるからであり，彼らは，上位層から下位層への連絡は予測を，下位層から上位層への連絡は予測誤差を伝えていると考えたのである．なお，この coding においては，下位層の活動が画像の局所的な特徴を直接表現するのに対し，上位層の活動は画像全体にわたる大域的な特徴，さらには画像に隠れた本質的パラメータを表現しているものと考えられる（隠れ状態の推定に関しては，池田先生の講義を参照）．

モデルの動作について 3 層構造のネットワーク（図 5) を例にとって説明する．

いま，画像 $\boldsymbol{I}$ が，基底ベクトル $\boldsymbol{U}_j$ の荷重和に非線形変換 f を加えたものに，ノイズ $\boldsymbol{n}$ が加わったものであると仮

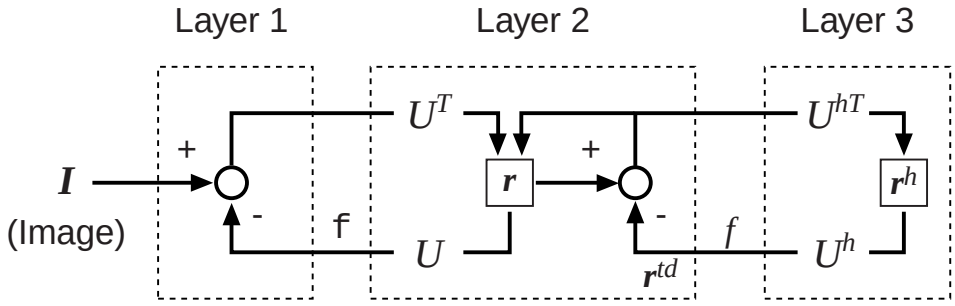

図 5: Rao and Ballard のモデル

定する．

$$\boldsymbol{I} = f(\sum_j r_j \boldsymbol{U}_j) + \boldsymbol{n} = f(U\boldsymbol{r}) + \boldsymbol{n} \tag{35}$$

画像を受け取る層を第 1 層と考えると，ここでの基底ベクトルは第 2 層の各素子が表現する特徴量に，荷重は各素子の活動度に対応する．したがって，上の式の第 1 項は，第 2 層における内部表現（各素子の活動）に基づく予測の項，第 2 項は予測では記述できない誤差の項に対応することになる．なお，U は基底ベクトルを並べてできる行列，$\boldsymbol{r}$ は r_j を並べてできる列ベクトルである．

同様にして，第 2 層の活動度 $\boldsymbol{r}$ は，第 3 層から予測される成分 $\boldsymbol{r}^{td}$ と予測では説明できない成分 $\boldsymbol{n}^{td}$ の和に分解される．上の式と同様に，上位層からの予測分 $\boldsymbol{r}^{td}$ は，上位層の基底ベクトルの荷重和 $f(U^h\boldsymbol{r}^h)$ で与えられる（$\boldsymbol{r}^h$ は第 3 層の活動度である）．

いま，予測と実データとの差を評価するコスト関数として，

$$E_1 = \frac{1}{\sigma^2}||\boldsymbol{I} - f(U\boldsymbol{r})||^2 + \frac{1}{\sigma_{td}^2}||\boldsymbol{r} - \boldsymbol{r}^{td}||^2 \tag{36}$$

を考える．ここで，第 1 項は第 1 層における誤差，第 2 項は第 2 層における誤差である．全体のコスト関数としては，これに活動度の性質を表す事前確率を反映させた項を加えたもの

$$E = E_1 + g(\boldsymbol{r}) + h(U) \tag{37}$$

を考える．ここで，$g(\boldsymbol{r}), h(U)$ はそれぞれ $\boldsymbol{r}, U$ の事前分布の対数尤度の符号を反転したものである．$\boldsymbol{r}$ の分布としてガウス分布を仮定すれば，$g(\boldsymbol{r}) = \alpha\sum_i r_i^2$ となる．また，内部の活動度がスパースになるような状況を考えるのであれば，$g(\boldsymbol{r}) = \alpha\sum_i \log(1 + r_i^2)$ とすればよい．E の最大化は内部モデル空間でのエントロピー最大化と等価である．

以上の準備の下で，第 2 層素子の動作則を導出する．Olshausen and Field と同様に，上のコスト関数の最急降下法により活動度 $\boldsymbol{r}$ の更新則を得ることができる．

$$\frac{d\boldsymbol{r}}{dt} = -\frac{k_1}{2}\frac{\partial E}{\partial \boldsymbol{r}} = \frac{k_1}{\sigma^2}U^T\frac{\partial f}{\partial \boldsymbol{x}}^T(\boldsymbol{I} - f(U\boldsymbol{r})) + \frac{k_1}{\sigma_{td}^2}(\boldsymbol{r}^{td} - \boldsymbol{r}) + \frac{k_1}{2}g'(\boldsymbol{r}) \tag{38}$$

この式の第 1 項は，第 1 層で検出された誤差ベクトルに各素子の荷重ベクトル $\boldsymbol{U}_j^T$ をかけたもの（内積）である．一方，第 2 項は第 3 層からの予測 $\boldsymbol{r}^{td}$ と第 2 層の実際の活動度 $\boldsymbol{r}$ との差を表す項である．第 3 項は活動度の事前分布を反映した項である．この計算は局所的なアルゴリズムによって実現できる．なお，特殊の場合として $f(\boldsymbol{x}) = \boldsymbol{x}$ であるときには，非線形性がなくなるため，同じ動作を単一層構造のネットワークで実現できる．

学習則も同様にして，$\boldsymbol{U}_j$ に関する最急降下法により得られる（各自導されたし）．第 3 層の動作則，学習則も同様である．結果は，一種の Hebb 学習になる．

以上のモデルを基づき，Rao and Ballard は end-stop neuron の形成と contextual modulation 特性の発現を数値実験により示している．end-stop neuron とは，線分，エッジ，grating（縞模様）などに対して反応するものの，それがある程度の長さをもっていると活動が弱まり，それがごく短いや途中で寸断されている場合（つまり，端点やかど）に対して強く反応する細胞である．このような細胞は，境界部分を検出するため働きを担っていると考えられている．

end-stop cell のモデルでは，画像データにガウシアンフィルタ（２次元ガウス分布の形をした空間フィルタ，元画像をぼかす効果がある）をかけたものを第 1 層への入力として与えている．第 2 層は，32 素子からなるユニット三つによって構成され，各ユニットの受容野は互いに一定の画素数分だけずれている．第 3 層は 128 素子のユニット一つからなる．したがって，各素子の受容野位置や，階層とともに受容野が大きくなる構造はあらかじめ埋め込まれていることになる．なお，出力関数は $f(\boldsymbol{x}) = \boldsymbol{x}$ としている．

このモデルに自然画像を与えて学習を行なった後，第 2，3 層素子の荷重ベクトル $\boldsymbol{U}_j$ を調べると，第 2 層には Gabor フィルタ的な構造が，第 3 層にはそれを組み合わせたような構造が形成される．学習後のモデルに一定の長さの線分を提示したところ，第 2 層のある素子は，自分の受容野内だけで閉じている短い線分に対して強く反応したのに対し，受容野を越えた長さをもつ線分に対してはあまり反応しなかった（つまり，end-stop の効果が現われたことになる）．また，線分の長さを横軸，素子の活動度を縦軸にとったグラフを描くと，その素子の活動度は，V1 の 2, 3 層に見られる細胞の活動変化とよく似た反応を示すことがわかった．

このようなモデルの振舞いは，自然画像にはある程度の長さをもった線分が多く含まれており，学習により第 3 層素子がそのような長い線分の特徴を表現するようになったために生じたと考えられる．すなわち，短い線分は第 3 層素子が表現できる内部モデルだけでは記述できず，その内部モデルからの誤差分が第 2 層の活動として生じたのである．実際，第 3 層からの feedback 入力を遮断すると第 2 層素子に見られる end-stop の性質は失われるが，これは，V2 の活動を止めたときの V1 第 6 層細胞の活動と符合する．

一方，contextual modulation のモデルでは，画像データを直接与え，sparse coding に対応する確率分布を仮定して学習を行なっている．第 2 層は，互いに一定の領域だけ重なった受容野をもつ 3 x 3=9 個のユニットからなっているほかは，end-stop cell のモデルとほとんど同じである．学習の結果，第 2 層には，Olshausen and Field で見られるような局所的 Gabor 型の受容野が形成された．Gabor 型とは，受容野の特性がガウス分布の形に三角関数を掛け合わせた形，すなわち，

$$G(x, y) = \exp(-\frac{x^2 + y^2}{2\sigma^2}) \cos(\omega(x \cos\theta + y \sin\theta)) \tag{39}$$

または，

$$G(x, y) = \exp(-\frac{x^2 + y^2}{2\sigma^2}) \sin(\omega(x \cos\theta + y \sin\theta)) \tag{40}$$

で表されるものである．ここで，σ は受容野の広がりを定める定数，θ と ω はそれぞれ方位と空間周波数を表す定数である．

学習後に，1) 一定領域で閉じた grating（一定の方位をもった縞模様），2) 領域全体を覆う grating，3) 中心と周辺で直交する方位をもつ grating，4) 周辺部だけの grating をそれぞれ提示すると，第 2 層素子の活動は，2) の場合に大きく抑制されること，3) の場合には 1) よりも大きくなること，4) では 2) と同様に抑制されることがわかった．これらの結果は，contextual modulation の典型的な実験データと一致する．

以上の結果は，画像がもつ一般的な構造が第 3 層で表現され，それから逸脱した部分が誤差成分として第 2 層で表現されるために生じたものである．テクスチャのポップアウト現象（周囲と異なるテクスチャをもった部分がぱっと浮き上がって見える現象）も同様にして説明できることは容易に想像できよう（周囲からの予測で説明できない部分を表現する部分が第 2 層の活動として抽出できるからである）．なお，彼らは，ガウシアンフィルタの場合も sparse coding を仮定した場合もほぼ同じ結果が得られたことから，このような情報表現機構の形成に sparseness は特に必要ではないと述べている．

この論文では，predictive coding が脳内処理の様々な場面に現われる共通原理であると主張しているが，この考え方自体はとりたてて新しいものではない．対象をモデルベースな方法で表現する枠組みは種々の分野（例えば，画像解析，情報圧縮など）で広く用いられており，そこでは，対象をできるだけうまく説明できるパラメータを推定した上で，誤差部分だけを別に表現するという考え方が一般的である．また，contextual modulation の意味を「周辺からの予測に沿わない部分を取り出す」ことに求める考え方も，この現象が発見された当初より唱えられてきたものである．したがって，彼らの研究の意義は，モデルベースの考え方を初期視覚の自己形成の場に持ち込み，単に「このような解釈が可能だ」というだけでなく，生理学データとつきあわせられるような具体的な結果を示してみせた点にあるといってよいだろう．

5 おわりに

脳内情報表現の実現に関してこれまで情報理論を用いてどのようなアプローチがとられて来たのか，ということに焦点を絞りほぼ時間順序的にモデルの変遷を概説した．この変遷を見ると，冗長度圧縮という統一的な視点に立つことでモデルの見通しが良くなり，現実の脳の構造に則したより複雑なモデル化を行うことが可能になって来たことが分かる．もちろん，これは単にものの見方の変化によるだけではなくこの間計算機の性能が飛躍的に向上したため，ある程度複雑なモデルに関する実験が可能になったという技術革新の影響が大きい．また，冗長度圧縮や確率モデルによる記述といった仮説や表現自体に関してもそれがはたして現実の脳を記述するのにどれほど妥当なものか疑わしいという異論も多い．しかしながら，やみくもに複雑なモデル化を行いそれから意味のある結果を導くことはほとんど絶望的に難しいことも事実である．重要なことは，とりあえずあるものの見方をしてみた上で，その方針でどこまで現実の非自明な現象を説明できるのかとことん調べ尽くすことであろう．そういった意味で，ここで紹介したアプローチも Olshausen and Field，Rao and Ballard らの研究でようやく実験による検証が可能な段階に達してきたというところではないだろうか．今後の展開に注目したい．

参考文献

[1] Amari S and Takeuchi A(1978). *Biological Cybernetics* **29**, 127–136.

[2] Amari S (1980). *Bullutin of Mathematical Biology* **42**, 339–364.

[3] Amari S and Cardoso JF (1997). *IEEE Tran. on Signal Processing* **45**, 2692.

[4] Atick JJ and Redlich AN (1990). *Neural Computation* **2**, 308.

[5] Barlow HB (1989). *Neural Computation* **1**, 295.

[6] Bell AJ and Sejnowski TJ (1995). *Neural Computation* **7**, 1129.

[7] Blakemore C and Cooper GF (1970). *Nature* **228** 477–478.

[8] Kohonen T (1982). *Biological Cybernetics* **43** 59–69; **44**, 135–140.

[9] Linsker R (1986). *Proc. of National Academy of Science, USA* **83**, 7508; (1988) *Computer* March 1988, 105.

[10] MacKay DJC and Miller KD (1990). *Neural Computation* **2**, 173.

[11] von der Malsburg C (1990). *Kibernetik* **14**, 85–100.

[12] Nadal JP and Parga N (1995). *Network* **5**, 565.

[13] Obradovic D and Deco G (1998). *Neural Computation* **10**, 2085.

[14] Olshausen BA and Field DJ (1996). *Nature 381*, 607.;(1997). *Vision Research* **37**, 3311.

[15] Rao RPN and Ballard DH (1999). *Nature Neuroscience 2*, 79–87; (1997) *Neural Computation 9*, 721–763.

[16] Takeuchi A and Amari S (1979). *Biological Cybernetics* **35**, 63–72.

[17] Yuille AL, Kammen DM and Cohen DS (1989). *Biogical Cybernetics* **61**, 326.

[18] Willshaw DJ and von der Malsburg C (1976). *Proceedings of the Royal Society of London B* **194**, 431.