統計数理 (2001)
第 49 巻 第 1 号 1–7

地図を描く・風景を眺める

統計数理研究所　伊庭幸人（オーガナイザー）

1. はじめに：見ることの誘惑

本特集では,「地図を描く・風景を眺める」と題して，主成分分析や多次元尺度法の周辺について，さまざまな視点からの報告を集めた．統計の研究者やヘビーユーザーにとっては，こうした手法は，応用的に使い尽くされた，目新しくないものとして映るかもしれない．しかし，収録された諸論文をみればわかるように，それは必ずしも真ではない．これらの方法は，思いがけない分野でデータ解析の手法として再発見されたり，あらためて注目を集めたりしている．高次元空間に散らばる対象の隠された秩序を目で捉えたいという欲望は根が深いのである．また，こうした方法の数理が，3 次元空間での再構成問題と共通しているという点も重要である．あとで見るように，物体の再構成問題は，たんぱく質分子の姿を捉えることから，ロボットの視覚まで，さまざまな分野で出現する．

近年のひとつの流れとして，パターン認識を中心とする応用サイドからの多変量解析への再評価がある．これは，ニューラルネットの流行から,「人工ニューラルネットとは多変量統計モデルである」という認識を経て，いわば逆方向から統計的手法の重要性が再認識されたものだと理解できる．この中で，カーネル主成分分析をはじめとする諸手法の発見あるいは再発見があった．本特集では，こうした方法論的発展を直接問題にする代わりに，各応用分野で方法論的な要求がどのように生じてくるかを眺める，という立場を取った．結果として，独立成分分析のように全く触れることのできなかったテーマもあるが，理論や方法論からの展望は，別の機会に譲りたい．

本特集は 7 篇の論文で構成される．以下，それらについて，テーマ別に紹介して行くが，その前に，多変量解析に通じていない読者のための用語の解説と整理を挿入する．不要な向きは飛ばして先に進まれたい．

2. 用語の解説と整理

主成分分析 (PCA) とは，共分散行列（または相関行列）を対角化することで，データの分散の多くを説明できる「軸」の組，いいかえれば線形部分空間を求める操作を意味する．データが多次元空間で楕円体の形に分布していれば，固有値の大きい順番に並べた固有ベクトルが，長い順に並べた楕円体の軸に対応する．一般に，固有値が大きいほうから数個の固有ベクトルの張る空間にデータを射影することで，その様子を「見る」ことができ，次元の縮約が可能である．共分散行列を用いた主成分分析を統計モデルの立場からみると，変量の種類（心理学ならテストの種類）ごとに違う重み（因子負荷）を掛けた少数個のベクトル（因子）の寄与とガウス雑音の和でデータを表現する問題に翻訳できる．このようなモデルを一般に**因子分析**モデルという．普通の主成分分析に対応付けられるのは，雑音が独立同分布の場合であるが，この仮定

を除いて一般化することも可能である（本特集の**川崎**氏の論文を参照）.

一方，**多次元尺度法（MDS，多次元尺度構成法）**とは，対象間の（非）類似性がデータとして与えられたとき，それらをできるだけ保存するように，低い次元の空間に射影して，「地図」を作る方法のことを指す．素朴なイメージとしては，世界地図を都市間の移動の所要時間に合わせてゆがめて描いた「地図」を想像されたい．多次元尺度法は，こうした地図表現を追求し，もともと 3 次元空間にない対象を低次元空間で表現することで，データの構造を「見る」こと，データの解釈を見い出すことを目的に考えられた．主成分分析は多次元尺度法の一種である**古典的多次元尺度法（主座標分析）**と対応付けられる（注 1）． 一方，距離の定義をいろいろ変えたり，特定の変換のもとで不変な性質のみを考慮した「地図」を考えることによって，多様な多次元尺度法が考えられる．特に，対象間の距離の大小順序のみを考えた多次元尺度法を**非計量多次元尺度法 (NMDS)** と呼ぶ（**田口**氏らの論文参照）.

主成分分析を出発点とすれば，上で述べた手法と 7 節で紹介する手法は

・確率モデル化とその方向での一般化 → 因子分析
・距離や「地図」の概念の一般化 → 多次元尺度法
・局所化と非線形構造の抽出 → カーネル主成分分析
・複数のクラスタを考える → 有限混合分布モデルによる分析
・射影や成分分解の特徴付けの変更 → 射影追跡，独立成分分析

とまとめられるかもしれない（最後の二つについては 7 節参照）．但し，各方法にはそれぞれ異なった背景や哲学があり，このまとめ方ではそれらが無視されていることを注意しておく．

3. 頭の中身と心の内側

さまざまな多変量解析の手法が，そのルーツを心理学にもっている．特に，多次元尺度法 (MDS) が開発され，応用されてきたのが心理学と社会調査法の周辺であったことはよく知られている．

本特集では，伝統的な心理学とは少し違った方向で，人間の認知に関係するような題材をいくつかとりあげた．まず，脳研究との関連である．これは**岡田**氏の論文で論じられている．視覚認知の心理学的研究では，被験者にいろいろなパターンを見せて，それに対する被験者の反応を見る．脳の電気生理学的研究でも，パターンを見せるのは同じであるが，調べるのは個々のニューロンの反応である．これは，社会調査とのアナロジーでいうなら，人間でなくニューロンが相手の「世論調査」と考えられる．**岡田**論文にあるように，比較的単純な反応をするニューロンが集まっていると考えられている領域を調べる場合（初期視覚の研究）には，仮説を立てて演繹的に議論を進めるのが有効であった．より複雑な情報処理が行われる領域，「記憶」や「思考」と「見ること」が交錯するような領域の研究（高次視覚の研究）では，「アンケート」の設問自体が複雑であり，「回答」の解析は困難になる．顔や文字の認知はこの代表的な例といえる．ここで，多変量解析の手段を発見法的に利用し，頭の中の記憶や思考の世界を「見る」ことができないか，というのが，**岡田**論文のテーマである．

顔や文字などの認識は，情報処理技術としても重要である．特に，顔の認識は個人識別の手段として注目され，パターン認識やコンピュータビジョンの実験場となっている．この分野を主成分分析やそれに関連した手法の応用という観点から論じたのが**坂野**氏の論文である．**坂野**論文では，最近注目されている手法であるカーネル主成分分析（核非線形主成分分析）についても解説がなされている．情報処理技術の研究は心理学とは別の価値観に依拠しているが，「人間が簡単にやっていることがコンピュータにできない」という驚きと「それでは人間はどうやっ

ているのか」という問いを通じて，両者は接している．この意味では，**坂野**論文のテーマも，人間のまわりの「意味の世界」の探求と考えられる (注 2).

本特集の最初の構想では，より正統的な心理学ないし社会調査の分野での多次元尺度法に関する総合報告を含める計画があった．具体的には，統計数理研究所の中心的研究課題のひとつであった「数量化理論」の歴史をふりかえる解説，あるいは，心理学での「心を見る」試みを回顧する論文，などいくつかの案があったが，様々な事情や掲載論文の本数の兼ね合いなどで実現しなかった．

4. シミュレーション結果の解析

計算機の進歩は，大規模なシミュレーションによる複雑なモデルの解析を可能にした．特に，物理科学においては，対象のモデル化が比較的容易であることもあって，スーパーコンピュータ等を用いた計算がさかんに行われている．しかし，こうした計算の結果得られるものは，たとえば，多数の原子の座標の時系列であり，これをそのままビデオにして眺めてみても，十分理解することができないという事態が起こってくる．そこで求められるのが，「シミュレーションデータの多変量解析」とでもいうべきジャンルの研究である．これは，ある意味で，物理学が心理学や社会学が直面してきたような複雑さを扱うところまで進歩した証かもしれない．

本特集の**北尾**氏の論文と**肥後**氏らの論文の後半は，こうした分野における主成分分析とその周辺の手法の応用を扱ったものである．具体的には，たんぱく質・ペプチドの folding の問題が扱われている．巨大な鎖状分子であるたんぱく質は，複雑に折りたたまれた形で機能を果たしているが，必ずしもひとつの形状のみを固定的にとっているわけでなく，複数の形の間をジャンプしつつ，あるいは連続的に，揺らいでいる．適当なモデルをとって，このさまをシミュレートした場合に，その結果をどのように「見る」か，「地図」にするかというのが，ここで問われている問題である．

シミュレーションで得たたんぱく質の形状のサンプルが正しく熱平衡状態を代表しているなら，それらから得られた「主成分」は，該当する温度でのギブス分布の共分散行列の固有モードに一致する．もし，その温度で重要なたんぱく質の形状が一種類しかなく，ギブス分布がそのまわりのガウス分布で近似できるのなら，この固有モードは問題を記述するために適した「集団座標」である．実際のたんぱく質は沢山の局所的なエネルギー極小を持つ系であるため，ギブス分布は多峰性となり，ガウス近似はしばしば破綻する．このような様相の「景観」(landscape) を主成分分析で得られた部分空間への射影で「見る」というのがひとつの立場である．別の立場としては，より踏み込んで，景観を「典型的な形状」を各成分の中心とするようなガウス分布の混合でモデル化するという立場もある．**北尾**氏の論文でいう JAM (Jumping Among Minima) モデルはその文脈で理解することができる．なお，JAM モデルでは各成分への分類規準は物理的考察によってあらかじめ与えられているが，これを含めてデータから推定することも考えられる (注 3).

こうした方向の研究は，もちろん，たんぱく質と主成分分析という組み合わせに限らない．**北尾**氏，**肥後**氏らの論じているモデルは物理法則に近いレベルから組み立てられた「リアル」なモデルであるが，より抽象化されたモデルや人工生命などの概念的なモデルでも同様のことが問題になる．また，本特集の**田口**氏らの論文も，論じている題材は違うが，最近の理論物理研究者の関心の方向を示すものとして，通じる面があると思う．基礎物理学研究所で 2 回にわたって行われた研究会 (1999, 2000) では物性物理・化学物理の分野を中心にこうした問題をひろく論じた．この方向にさらに興味のある方は，1999 年の研究会の報告 (「物性研究」，74–2 (2000 年 5 月号))，及び，そこに所収の著者による概観 (伊庭 (2000)) を見ていただきたい．2000

年度の研究会の報告も，近いうちに「物性研究」に載る予定である．

5. 物体の再構成 —— たんぱく質からロボットの視覚まで

以上の例では，高次元の「意味空間」「形状空間」のようなものを低次元に表現することを考えた．数理的に類似の問題として，もともと 3 次元空間中にある物体を距離情報や射影情報から再構築するという問題がある．これは，素朴な意味で「地図を描く」というのに近いが，以下で見るように，意外な応用がある．

肥後氏らの論文の前半で取り上げられているのは，実験データから，たんぱく質の 3 次元形状（高次構造）を復元するという問題である．現在主流の実験方法である NMR（核磁気共鳴）法で得られるのは，「原子間の距離」のデータなので，そこからたんぱく質を構成している原子の相対位置を推定する問題は多次元尺度法に類似したものになる．通常の多次元尺度法との違いは，(1) 特定の種類の原子対がある限界より近い距離にある場合のみ距離のデータが得られるので，欠測がきわめて多いデータとなる (2) 対象がたんぱく質という鎖状の高分子であること，および，その原子の並び方（一次構造）が，事前にわかっている，の 2 つである．これから，形状推定問題は，事前知識が拘束条件として与えられる多次元尺度法とでもいうべき，興味深い問題に帰着されることがわかる．**肥後**論文では，これを解く方法として，幾何学ベースのアルゴリズム（**肥後**論文でいう多次元尺度法）と制約条件を組み込んだシミュレーションによる最適化法の二つが紹介され，後者が詳しく論じられている．

藤木氏の論文では，複数の視点からの画像による 3 次元物体の再構成という問題が論じられている．ここでの問題設定の特徴は，各視点の位置も未定であることである．具体的には，部屋の中をぐるぐる歩き回るロボットがいて，いくつかの位置での見え方をもとに，物体（部屋全体でも良い）の形状と自分の移動の軌跡の両方を同時に推定するというケースが考えられる．一見すると難しそうであるが，物体にいくつかのしるしが付けられている（または，なんらかのアルゴリズムによって異なる画像間で複数の特徴点の対応が与えられる）という条件下では，特徴点の座標のベクトルをデータとして因子に分解することで，比較的簡単に目的を遂げることができる．**藤木**論文では，画像の射影法の解説からはじめて，このタイプの方法（因子分解法）の数理について詳細な解説が行われている．

6. 時系列

主成分分析・因子分析や関連手法を実際に使う場合，時系列データが問題になることは多いと思われる．時系列データの場合，遅れ（ラグ）を伴って相関が伝播することが普通であり，各時点でのサンプルが独立ではない点を考慮した手法が必要となる．本特集では**川崎**氏の論文が時系列データに対する主成分分析と因子分析の手法を扱っている．**川崎**論文の前半では，定常時系列を周波数成分に分解することで独立データの場合に帰着させるという古典的手法が，後半ではより柔軟な実時間での因子分析モデルが扱われている．後半の議論は，統計モデルとしての定式化（記述法としての主成分分析でなく因子分析モデルを考えること）の意義を示す例としても興味深い．

本特集の他の論文でも時間的要素があちこちに顔を出していることにも注意されたい．たとえば，**岡田**論文の最後の部分は，視覚認知の過程が刺激からの時間に応じて進行していく様子を，電気生理データから多次元尺度法を通じて理解する試みの報告である．**藤木**論文の問題でも，各観測の間の移動距離が短いというような条件があれば，時間的要素を含んだ問題が生ずると思われる．**田口**論文の例のひとつでも，時系列データが考察の対象になっている．

7. 方法論からの展望

本特集の大部分は具体的な応用例を中心として構成されているが，以下では，その中から出てくる方法論的課題について簡単に論じる．

古典的な主成分分析やそれに対応する多次元尺度法などにおいて最も不満なのは，線形空間への射影では局所的な構造を十分とらえられないという点であろう．この点を主成分分析の枠内でうまく扱おうとしたのが**坂野**論文で紹介されている**カーネル主成分分析**である．多くのケースで，この方法は，計量的な多次元尺度法の一種として理解することもできる（**カーネル多次元尺度法**）．局所的な性質を捉えるという意味では，主成分分析の背後にある多次元ガウス分布をガウス分布の混合（正規混合分布）に置き換えるのが，もうひとつの戦術である．応用サイドからこれを導入したのが，**北尾**論文における JAM モデルである．なお，クラスタ分析と主成分分析の融合としては，他の方向，たとえば，「与えられたデータの射影をクラスタ分析したときに最もクラスタの分離が良くなるような射影を求める」といった方法も考えられる (Nakamura and Baba (2000))．

線形性を越えるための別のアプローチとしては，すでに言及した非計量多次元尺度法 (NMDS) がある．非計量多次元尺度法は心理学や社会調査法などでは古典的な手法となっているが，他の分野，特に物理科学での応用は少ないと思われる．本特集の**田口**氏らの論文では，非計量多次元尺度法の手法と適用例が論じられている．田口氏らの問題提起自体も興味深いが，重要なのは，非計量多次元尺度法のような手法に注目している理論物理や非線形科学の研究者がいるという事実であろう．

本特集に登場しなかった重要な話題として，**独立成分分析 (ICA)** がある．独立成分分析は成分の分布の非ガウス性などを利用して，データを独立成分の和（と雑音）に分解するものである．従来からある手法として，分散の多くを説明する部分空間にデータを射影するかわりに，スケールに依存しない規準で「もっとも面白い構造が見られる」部分空間への射影を求める方法があり，**射影追跡法**と呼ばれている．独立成分分析の問題意識はこれと共通する点がある．本稿で論じられた各例についてこれらの方法の適用可能性を考察するのも興味深い．

8. 終わりに： 見ることの限界へ

近代は「一望のもとに」見ることへの欲望の時代であったといわれる．その意味では，主成分分析や多次元尺度法による可視化は，近代の欲望を忠実に実現しようとするものである．実際，ヨーロッパ文明による世界支配は，「地図を描く」ことと並行して行われたのであった．もっとも，主成分分析や多次元尺度法の応用の多く──3 次元空間での再構成問題は除く──において，見られる対象が抽象的な空間の中に位置づけられる点に着目すれば，そこにプレ近代/ポスト近代的な要素を見出すこともできる．近代の成立において重要であった「見る」は，「意味の空間」のようなものを排除して，物事を物理空間の中に位置づけるということだったとも考えられるからである（フーコー (1969)）．「意味の空間」がどこまで意味を持つのかという問いは，主成分分析や多次元尺度法，数量化が始まって以来の問題であるが，新しい世紀の研究はこれにどう答えるのか，興味はつきない．

技術的な面から考えると，素朴な主成分分析や多次元尺度法では重要であった「見る」という要素は，カーネルを用いたり，独立成分を導入したりすることで薄れる方向にあるのかもしれない．有限混合分布モデルでは，各成分分布（クラスタ）に分けるという意味で，モデルにシンボル的要素が導入されており，既に「見る」という範囲を超えて，記号的に「考える」ことへの第 1 歩を踏み出しているようにも思われる．「見る」ことと「考える」ことの間で，今後

の統計学はどう揺れ動いていくのだろうか（伊庭（1996, 1998））.

9. 研究会と特集の連携

この特集に関連して，特集の執筆者と関連分野の研究者を招待して行う研究会を企画している（2001 年度後半に統計数理研究所で開催予定）. 研究会の結果を出版物にまとめることは多くても，出版の企画にタイアップしてあとから研究会を行うのは，珍しいかもしれない．統計数理の特集号としてははじめての試みだと思うが，面白い会にしたいと考えている．

注.

（注 1）データ $\{x_i\}, i = 1, \ldots, N$ に対して，ユークリッド距離 $\delta_{ij} = ||x_i - x_j||$ を計算し，古典的多次元尺度法を用いて d 次元空間に埋め込んで得られる布置は，同じデータを主成分分析して大きいほうから d 個の固有値に対応する固有ベクトルの張る空間にデータを射影して得られる布置と一致する．

ここでいう古典的多次元尺度法とは，距離 δ_{ij} の 2 乗に「2 重中心化」の操作を行ったもの $\tilde{\delta}_{ij}^{(2)} = \delta_{ij}^2 - 1/N \sum_i \delta_{ij}^2 - 1/N \sum_j \delta_{ij}^2 + 1/N^2 \sum_{ij} \delta_{ij}^2$ と布置された対象の間のユークリッド距離 d_{ij} の 2 乗を同様に 2 重中心化したもの $\tilde{d}_{ij}^{(2)}$ について，$\sum_{ij} (\tilde{d}_{ij}^{(2)} - \tilde{\delta}_{ij}^{(2)})^2$ を最小にする布置を求める手法のことである．ユークリッド距離を用いた場合でも，他のあてはめ方，たとえば，$\sum_{ij} (d_{ij} - \delta_{ij})^2$ を最小化した場合には，主成分分析との厳密な対応は成り立たない．

（注 2）岡田論文での「認知の過程を見る手段としての多変量解析」と坂野論文の「より良い能率で認知的な処理を行う手段としての多変量解析」の間のずれと関係に注意すべきである．「研究手段」と「研究対象」の微妙な関係は心理学や脳研究の特性であると同時に，パターン認識や統計科学一般の特性でもある．これは一種の自己言及性といえるかもしれない．

（注 3）別のシステム（スピングラスモデル）についてのこの種の試みについては Iba and Hukushima (2000)，伊庭・福島 (2000) 及び Marinari *et al.* (2001), Hed *et al.* (2001) 参照.

謝　辞

各論文の著者の方々には，お忙しいところ，面倒なお願いに快く応じて頂き，深く感謝しております．著者および査読者の皆様には記して感謝の意を表します．また，統計数理研究所の石黒真木夫氏と前田忠彦氏にはさまざまな場面で貴重な助言を頂き，本特集の成立に大きく貢献して頂いたことを特記します.

参 考 文 献

フーコー，M. (1969).『臨床医学の誕生』（神谷美恵子 訳），みすず書房，東京．（原著： Foucault, M.（1963），Presses Universitaires de France）

Hed, G., Domany, E., Palassini, M. and Young, A. P. (2001). Correlated spin domains generate hierarchical structure in state space in short range spin glasses, e–print, http://arxiv.org/abs/cond–mat/0104264

伊庭幸人 (1996). 基礎的問題から見た情報統合，人工知能学会誌, **11–2**（1996 年 3 月号），193–200.
伊庭幸人 (1998). 反身体の思想（「無時間の思想」の付録），研究会「認知・行動の基底としての力学と論理」報告，物性研究，**71–2**（2000 年 11 月号），198–208.
伊庭幸人 (2000). 「解析法」セッションのイントロダクション，物性研究，**74–2**（2000 年 5 月号），115–121.
伊庭幸人，福島孝治 (2000). 有限混合分布モデルによる有限温度の準安定状態の表現，物性研究，**74–2**（2000 年 5 月号），134–142.
Iba, Y. and Hukushima, K. (2000). Identification of metastable states in finite temperature simulations: A self–organization approach, *Proceedings of ICCP5, Progress of Theoretical Physics Supplement*, No. 138, 462–463.
Marinari, E., Martin, O. C. and Zuliani, F. (2001). Equilibrium valleys in spin glasses at low temperature, e–print, http://arxiv.org/abs/cond–mat/0103534
Nakamura, T. and Baba, Y. (2000). An algorithm for clustering based on homogeneity, *The Tenth Japan Korea Joint Conference of Statistics Proceedings*, 63–68.