Special Edition

多変量解析と多次元データ解析

—データの科学の中で見る—

林　知己夫 | Hayashi Chikio

統計数理研究所名誉教授
兼㈳輿論科学協会会長

■1942年東京大学理学部数学科卒。74～86年統計数理研究所所長。86～91年放送大学教授。専攻は数量化、標本調査法、データの科学、社会調査法、日本人の国民性調査、国民性の国際比較方法論。

三つの異なった言葉が出てくるのでこの説明から始めよう。多変量解析は統計学で言うmultivariate analysisのことである。多次元データ解析—データ分析と言いたいところであるがあまりにも広すぎるのでデータ解析とした—は、多変量解析の技法を用いることも多いが、異なった考えに立つもので、数量化（定質的なものの数量化）分類、多次元尺度解析法（multidimensional analysis, MDS）、対応分析（correspondence analysis, CAやdual scaling, DUS）その他の関連方法を含むもので、数理統計学でいう多変量解析と異なる発展を示しているものである。データの科学はデータによる現象理解という立場を徹底的に押し進めて、データ分析を進める方法論的概念である。ここではデータの科学の立場に立って話を進めることにする。

1．データの科学とは何か

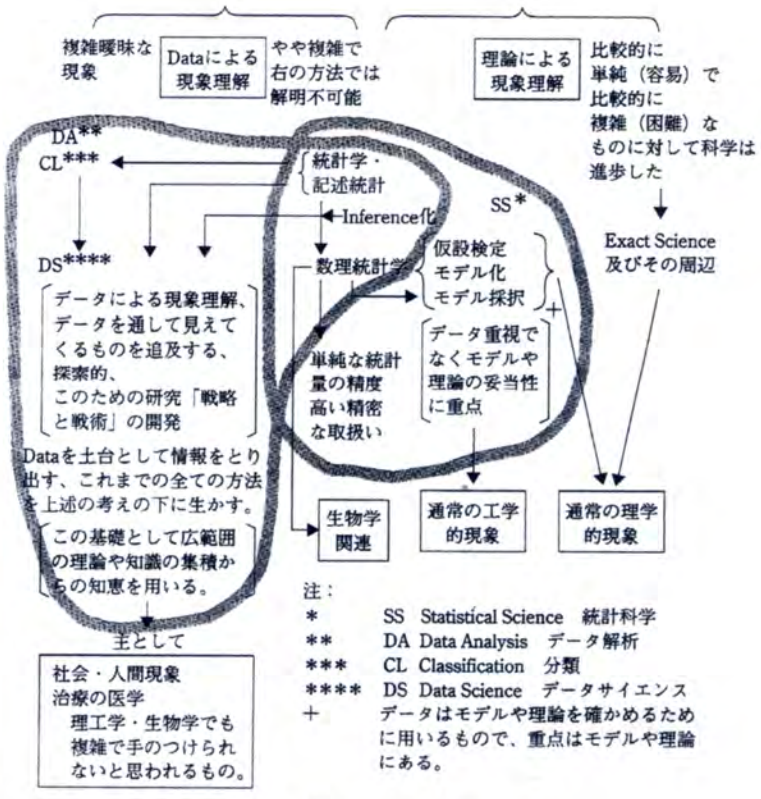
まず、データの科学とは何かから説明してみよう。複雑・曖昧な現象を科学的に取り扱うにあたっては従来の統計学の方法では正鵠を得た姿が見えてこない。これを含むさらに大きな研究戦略を必要とする。これは、「理論」によってではなく「データ」によって現象を理解しようとする立場である。またことの性質上データは必然的に多次元的なものにならざるを得ない。

これを考えるに到った経緯から話そう。

ここ十数年来、調査の科学ということを考え、社会調査による現象の理解を考えてきた。従来の見方と異なった見方で調査を作りあげていくという行き方で、これまでの方法で見えてこなかったものが見えてきたということである。しかし、これにも限界が見えてきた。そこで、さらに高い・広い立場で考えなくてはいけないということで、今度は再び統計学

の原点に戻って考え直し、新しいものの見方をしようと試みたのである。我々の考える根源はデータそのものであり、データを中心に据えてものを見ていこうとするものである。これがデータの科学の根本理念である。

データの科学というのは、データによって現象を理解するというものである。私は「理解する」という言葉の方が好きであるが、それはどうも主観的すぎるということであれば、調査を「解明する」という表現を使ってもよい。この方が一般的であるが、気持ちとしては、データにより現象を理解すると言った方が情熱が伝わる。さらにこれは、「データ」を通して何が見えてくるかを追及することも含むものである。理論によって現象を理解しようとするのではなく、「データ」(これを得るためには、過去の知識をポテンシャルとしてこれをフレキシブルに用いるのである)によって理解しようとする立場である。



科学的方法論の中の統計的方法

「データの科学は調査の科学を包含する概念である」。こういうふうにとらえるわけである。つまり、調査の科学よりもう一つ上に上がって、データによってものを理解するという形になるわけである。

これではまだ、データの科学の位置付けがはっきりしないので、科学方法論と広義の統計的方法との関連性の中に位置付けて考察してみよう。図を参照されたい。

科学は前述のように比較的に単純で複雑な現象を対象にして大いに発展し、科学的方法論が確立していった。これが物理学に代表されるexact scienceと言われるものであり、この方法がexact scienceに近い領域に多く活用され成果をあげてきた。この考え方がいわゆる“科学”という観念を植えつけた。しかし、複雑な現象はこれでは取り扱えない。ここに新しい方法として統計学、統計的方法が発展してきた。exact science的な考え方では取り扱えないものを取り扱う科学的方法である。これが記述統計学となり、確率を導入してinferenceを中核とする数理統計学になる。数理統計学が確立されると方法の数学化、精密化に進むのが一方向であり、もう一方にはいわゆる伝統的科学方法を志向し、理論に対する仮説ー検証、モデル化、統計的モデル化を

指向し、モデル選択の理論が生じ、伝統的科学方法論のにおいが強くなる。統計学の原点を離れた逆戻りの傾向を示してきた。exact science及びその周辺やモデル化は“理論による現象理解”という科学観（像）になる。ここでは、理論やモデルが大事で、データはそれを確かめるためのいわばソフトである。これが昂じると、データ棄却の誘惑にかられ統計的方法の陥穽におちいる。複雑曖昧な現象を取り扱うとなると統計学の原点に帰り、データによる現象理解に徹せざるを得なくなる。データをどうとり、どう分析するかが中心となる。データそのものがハードとなる。分析に重みがかかるとデータ解析・分類という方法論になり、これだけでは、不十分となるとデータの科学となる。この内容については、以下に述べるが統計学の原点に戻りデータによる現象理解、データにより何が見えてくるかという発想が大事となる。データを盲信せず、その由来を尋ね徹底した操作的立場からデータを探索的に見ていくのである。このために蓄積された科学的知識のみならず、あらゆる知識、知恵をポテンシャルとして活用することになるのである。

2．多変量解析

測定したものが実数で与えられている場合が取り扱われる。記述的に言えばiなるものがR個の測定で$X_{1i}, X_{2i}, \cdots X_{Ri}$で与えられている場合を考える（一般に添字$i$を落として表現する）。$i$は$R$次元ユークリッド空間内の点として与えられている場合である。Rが２の場合が最も単純な場合である。しかし、Rが大きい場合がその本領である。こうしてR個の数量的測定データの間の何等かの関係を求めようとするところに多変量解析が誕生した。

この考えそのものが科学の世界において革新的なことであった。従来の科学は１.で説明したように、いわゆる理論を土台に少ない変数の間の関係を想定し、少数のパラメータをデータから求め、因果関係を明らかにするところにその特色があった。Rの数が大きく、複雑な関連現象になるとこうした考え方では問題が処理できないのである。ここに多変量解析の魅力があった。

$R=2$の場合は、X_1とX_2との間に直線的な関係を仮定、X_1とX_2との関連の程度をあらわす相関係数が発生した。大事なことは直線的な関係を仮定した上での関連の程度を測ろうとするという点である。

つまり$X_1=aX_2+b$，a, bはパラメータを考え、(X_{1i}, X_{2i})，$i=1, 2\cdots n$のデータを用い最小二乗法によってa, bを決め、X_2からX_1を予想することを考えるのである。これを回帰直線という。X_2を用いないときはX_1の予測の最適解は算術平均値であり、$\overline{X_1}=\frac{1}{n}\sum_i^n X_{1i}$であり、その誤差は、$X_1$の分散$\sigma_1^2=\frac{1}{n}\sum_i^n (X_{1i}-\overline{X_1})^2$によって与えられ、これが精度である。$X_2$の知識を用い$aX_{2i}+b$によって$X_1$を予測するとき、平均的にその精度を表現すると、つまり誤差の分散$\frac{1}{n}\sum_i^n (X_{1i}-aX_{2i}-b)^2$を計算すると$\sigma_1^2(1-\rho^2)$；$\rho$は$X_1$と$X_2$との相関係数：となる。$X_2$から上述の直線関係を用い、$X_1$を推定したときの誤差分散がそれを用いなかった場合の誤差σ_1^2にくらべ、どの位少なくなるかを示すのが相関係数であるわけである。

Rが３になっても同じである。$X_1=aX_2+bX_3+C$と考えるのである。このときは、相関

係数の代わりに重相関係数というものを考えることになる。意味は全く同じで、X_1を予測するのにX_2, X_3の知識を用いて回帰直線を用いて予測したとき、どの程度誤差分散が小さくなるかの程度をあらわすものになる。ここで一つの新しい概念として偏相関係数というものが現れてくる。X_2を一定としたときX_1とX_3との相関係数、またはX_3を一定としたときX_1とX_2との相関係数と言われるものである。前者は、X_2からX_1を直線関係で推定したときの残差とX_2からX_3を直線関係で推定したときの残差との相関係数である。またX_1をX_2から推定したときの相関係数とX_1をX_2, X_3から推定したときの重相関係数を較べ、どの程度相関係数が上昇したかに関係する量（X_1とX_3との偏相関係数となる）として計算することもできる。ここまでは新しい概念として生々としている。$R>3$以上では新しい概念は生じない。これらが重相関（回帰）分析論となるが、変数選択やmulticolinearityの話が展開されているが、私は重相関（回帰）論の鬼っ子であると思っている。データを取り扱うデータの科学の分野から見ると脇道の議論である。

また、X_iが数量でなく、分類、右か左か、白か黒か、当選か落選か、AかBかという形で与えられている場合が判別分析と言われる。分類の判断成功率（的中率）がこの場合、相関係数に代わるものとなる。計算の上では相関比が用いられる。X_1という分類が二つに限らずS個あっても全く同じであるが、このときは多次元的に思考しなければならない。$a_0+\sum_j^R a_jX_j$という形を用い、X_1という分類を予測するのである。具体的には$a_0+\sum_j^R a_jX_{ji}$を用いてX_{1i}つまり右か左か、白か黒か等を予測する。このとき、判断成功率が最も高くなることが眼目となる、$a_0+\sum_j^R a_jX_j$の分布を仮定しない限りこれは難しいので相関比最大という基準が用いられる。これが線型判別関数論である。分類を多次元化するという拡張も無理なくできる。しかし、一般によく用いられる場合はX_1という分類の数が2か3、また3以上ではその間に順序のついている場合である。

これらの重相関（回帰）論・線型判別関数論が記述統計学における議論の中心であった。共に線型の関係を中心に据えたものであるがcurve linearであっても同様に扱うことができる。

$$X_1=a_0+\sum_j^R a_jX_j$$

であっても$X_2^2\rightarrow X_2$, $X_2X_3\rightarrow X_3$, $X_2X_3X_4\rightarrow X_4$, $X_2^2X_3^2X_4\rightarrow X_5$…とおけば全く同様に取り扱うことができるが、意味はますます解り難くなる。多変量解析の理論は、実用的には、たったこれだけの底の浅いものであった。

推測（Inference）つまり確率を導入して数理統計学となると$X_1, X_2, \cdots X_R$を確率変数と考えるのである（判別のときX_1は分類の形で確率変数でない）。このときX_jの確率分布が考えられ、一般に$X_1, X_2, \cdots X_R$はR次元ガウス分布が、Inferenceを主体とする統計学では（一次元のときと同様に）主体となっているのである。一次元のときのガウス分布でも現実的に見出すのは難しい。R次元ガウス分布となると現実的に私は経験したことがない。ガウス分布は現実に即物的に存在するものではなく、中央極限定理として存在するものと思っている。もしくは全くの仮定であって、現実的に確かめようもないものである。このようにし

てガウス分布に基づく複雑な計算に基づく諸理論が展開されているのが数理統計学における多変量解析論なのである。データに関係あるところでは、二つの（あるいはいくつかの）母集団の差の検定分散、共分散行列の同一性、標本重相関係数の標本分布、重相関係の検定、標本偏相関係数の標本分布、回帰係数の標本分布、判別分析における特性根の分布、その有意性、係数の有意性、正準相関に関する諸理論、分散分析に関する理論等多くの問題が主として、多変量ガウス分布の下に計算されている。また諸統計量の極限分布なども計算されるし、因子分析（後述）の因子負荷の有意性、因子数の決定なども上述の仮定の下に計算されている。非ガウス分布の場合も取り扱われているが、制約が多すぎて現実に応用できるものを聞かない。正に複雑で論文作成の宝庫であり、通常のレフェリーだと雑誌掲載を拒否できず採択となってしまう。また、マハラノビスの距離というものがもてはやされたこともあった。これは多次元ガウス分布をする二つの分布があり、同一の分散・共分散行列を持ち平均値のみが異なる場合、分布間の距離を定義したものであり、この限りにおいて分布の距離の意味を明確にしたものであるが、あまりにも条件が窮屈であり、一般の二つの分布の距離を適切に表現するものではないのでデータ解析の方法として発展しなかった。数理統計学内の一つのあだ花に過ぎなかった。Inferenceの数理統計学の手にかかるとどうしてこうもつまらなくなってしまうのか。記述的には革新的で面白いものであったが、データ解析の立場に立つと、底が浅いと言わざるを得ない。底の浅いものをInference化しても画期的なものが出るはずはない。

もう大分前のことであるが、高速計算機の普及から因子分析や主成分分析がデータの分析に用いられるようになり、多変量解析論に因子分析や主成分分析が入り込んできた。底の浅いものにやや厚みが出てきたように見えた。因子分析は二因子法に始まり多因子法、ラデックス法にまでつながるもので、計算の問題は末梢的で、モデル構成の考え方そのものが核心なので、ここが勝負所である。多因子法はデータの科学から見れば疑問の多いモデルであるが、多変量解析がこれに喰いついてきた。モデル構成を問題にするのでなく、計算を中心にしたもので、ガウス分布が入り込むし、統計量の計算、因子数の検定、有意性にこだわってしまい面白味がなくなってしまった。統計学がどうして本来の筋を外れて、現象解析に対して目が見えなくなってしまったか不思議でならない。これに関連して主成分分析法が入りこんできたが、ここでも特性根の有意性、主成分の数の推定ということに力が入ってしまう。

統計量の計算を中心に置いた数量統計学のなれの果ての姿を見るような気がしている。

現在、多変量解析に関する大方の期待は、形式的には重相関係数、（重）回帰分析、判別分析、因子分析法、主成分分析法（この他は以下3に述べるものである）に関する記述的統計学の部分である。つまり、データ解析におけるコンピュータソフトとしての方法である。これをどう生かしていくかは、データ解析の考え方にゆだねられているのである。多変量解析への期待は、膨大な発展をとげている多変ガウス分布に基づく数理統計学でもな

ければ、統計量の極限分布やそれに基づく仮説検定論でもないことだけは確かである。

3．多次元データ解析（数量化・分類・MDS・CA(DUS)その他関連する方法）

MDS以外のものは多変量解析の技法を必要な限り取り入れているのであるが、発想は甚だ異なるものである。数量化やCAはデータ分析の技法でなくその考え方や方法論が重要なのである。数量化の核心は数のないもの（質的なもの、カテゴリカルデータ）をどう測定で探り出し、これに数量を与えてデータ分析し、それ以外の方法では妥当性を以て取り出し得なかった知見を探り出すかにかかっている。「数はものそのものに内在するものではない、我々が科学的に目的を達するために与える道具である」というところに核心があるわけである。本稿で述べることは、詳しくは林(1993)：数量化―理論と方法―（朝倉書店)、具体的には林・鈴木(1997)：社会調査と数量化（増補版）（岩波書店）にあるのでそれに譲るが、ここでは形式的に見て多変量解析との対応を示しつつ話を進めてみる。しかし、対応づけたとしても、数量化の思想抜きにしては有効なデータ解析はできないのである。これを生かすにはデータの科学の文脈において、数量化の思想をもとにデータを取り扱わねばならないのである。もう一つ数量化において大事なことは、外的基準のある場合と外的基準のない場合を分けて考えることである。この二つはデータ分析の立場に立つとき大きな違いとなるのである。もう一つ注意すべきことは、多変量解析では最初に座標が存在し、この空間の中で議論を展開することになるが、数量化では量がないから座標が存在しないのである。分析した後で座標ができ上がってくるのである。メトリックが分析のあとで決まるということになる。考え方が逆になるので理解しがたい点にもなっている。座標の意味が後で決まるというところに注意されたい。その座標の意味は、数量化の方法に依拠して出てくるものなのである。

また、数量化は形式的に見れば、相関表の形（一つまたは多数）が与えられているとき、表頭、表側のカテゴリーに数を与えつつ知見を探るという見方もできるが、これは形の上だけの整理であって、生きた―前向きの―データ分析の考え方ではない。数量化では、いかに方法が組み上げられるかという考え方そのものに生産的な方法の源があるわけである。

数量化Ｉ類ということで巷間よく通っているものがある。外的基準があり、それが数量である場合であり、要因が質的な分類（アイテム・カテゴリー）で与えられている場合である。要因のアイテム・カテゴリーに数量を与えて外的基準を最もよく推定しようとする場合である。一見重相関分析と似ているが、根本的な違いは線型の考え方にある。多変量解析では直線的関係を仮定して分析を進めるのであるが、数量化の場合は、「線型にする」という考え方でアイテム・カテゴリーに数量を与えるという考え方である。「線型にする」と「線型である」とは大変な違いで、量のないものを線型にしたらどんな数量が出てくるかを知ることが大事な問題なのである。よく、調査項目の回答で非常に賛成に５、賛成に４、どちらでもないに３、反対に２、非常に反対

に１を与え、調査項目間の相関係数を計算し、因子分析をしていることが国の内外を通じよく行われているが、これは数量化の考えに全く相反している考え方である。数量化はアプリオリに数を与えるものではない。この任意性を排除し、且つ、線型性を量的なもの同士に仮定しないところにその特色があることに注意されたい。数量化とこうした通常の考え方と相反するものであるが、よく報告書では、あるときは数量化、あるときは上述の考えに基づく因子分析を使っているものを見かけるが、データ解析の根本を理解していない態度と言わざるを得ない。

外的基準が一次元数値でなくベクトルであっても取り扱うことができるし、要因に数量的なものがあり明らかにX_1と線型の関係が見出せる場合（curve linearであっても同様である)、量質の要因をあわせて分析することも可能である。また要因が量であってもこれをカテゴリー化しこれで質的要因に変え、Ｉ類をそのまま用いることもよい。この場合、計算されて出てきた数量ともとの数との関係を調べてみるともとの量がいかなる意味を持っていたかをよく理解することができる。

外的基準があり、これが分類で与えられている場合で判別分析に対応する。数量化Ⅱ類と言われるものである。ここでは外的基準の分類の数が２である場合よりむしろ、３以上であってそれに対応する数量化された数量が多次元（一般に分類の数より一つ少ない）の場合に興味ある知見が取り出される。要因のアイテム・カテゴリーは多くてもよいが標本数との釣合いが大事な情報となる。これを決める一般法則は残念ながら見出せない。外的基準のある場合この他面白い方法もあるが、これについては前書を参照されたい。

外的基準のない場合の一つに数量化Ⅲ類がある。今日これほど用いられている方法はない。きわめてフレキシブルな方法で、因子分析するのに必要になる仮定は一つもいらない。より的確な知見を安心して得ることができる。人と回答（アイテム・カテゴリー）との並べ替えによる知見探索に尽きる―人と回答との同時分類を志向する―。これを解析的に行うのが数量化Ⅲ類である。この方法が社会調査データに活用されるようになってきて以来、多次元データ分析の不可欠の方法となってきた。この社会調査への適用の最初は1971年のハワイ日系人調査であった（方法自体は、コンピュータが未発達時代の1955年に完成されていた)。日本人と日系人の回答比率があまり異ならないが、比率の多寡の出方が何かおかしいことに気がついた。回答の結び付きが異なるのではないかということでⅢ類を用いてみたところ、実際結び付きが異なることが解った。つまり日本人と日系人の考えの筋道の違いが描き出され、社会調査、特に比較調査（さらに国際比較調査）の基本的方法と考えられるに到った。これ以来、Ⅲ類が単純集計による解釈以前にすべきこととして社会調査に不可欠なものとなったのである。

一方70年代になるとフランスの文献に数量化Ⅲ類のような図がよく目につき出した。調べてみるとJ. P. Benzécri（ベンゼクリ）教授が、correspondence analysis（CA）と称し、数量化Ⅲ類と同じ方法を開発していた。73年に図書も出版されている。現在の数理統計学（多変量解析も当然含む）に反旗を翻し、独自

の方法をその思想と共に発展させていたのである。しかし、数量化全般に亘るものではなくⅢ類の範囲にとどまっているのは、数量化の源泉となる思想と異なっているためと思われる。カナダのニシサト教授のdual scalingも同じ系列のものである。数量化Ⅲ類も、数量化の世界から言えば、源泉から流れ出た一つの流れに過ぎないのである。

数量化Ⅳ類は多変量解析と全く異なる発想から出ている。二つのものの間の関係が数量（あるいは数値）で与えられている場合—但しいわゆる数学的意味のメトリカルでなくてよい。むしろ大小をあらわすファジィなものと見てよい—この情報を基に要素（$i, j=1, 2 \cdots n$）の関連性を見出そうとする方法である。つまり要素のユークリッド空間（2次元あるいは3次元）内の位置、それを総括する内部構造を求めようとする考え方である。これはソシオメトリーのデータを用い集団構造を見ようとしたもので、数量化のごく初期に完成されていた方法である。ファジィな関係を基に大体の構造を知ろうとするところにその特色がある。方法もそれに応じて甘い測度が用いられているのである。これも要素の並べ替えに端を発するものである。要素の間の関係がランクオーダー（対称なら$n(n-1)/2$個ある）として与えられている場合がMDSとなるのである。Ⅳ類はMDSの原型となっていたのである。

ここで注目すべきことは、外国では判別分析、CA、MDSはまま専門分野を異にして別物のように考えられているが数量化の考えに従えば全く同じ種類のものである。統一的考え方から流れたものに過ぎず、自然に同様なものと考えられるのである。数量化Ⅳ類からよく実際に用いられるMDA-OR（MDA-UOはⅡ類から）やAPMという方法が派生しているのである。これらの言葉を説明する余裕はないので前書を参照されたい。

外的基準のない場合はこの他多くの方法が—一見異なるように見えながら、数量化の考えに従うとき一つの流れに乗ってしまう—発展してきている。

分類というのも新しい方法である。分類内はなるべく似ているように、分類間はなるべく異なるように、分類するのであるが、何が似ており、何が異なるかから考えを進め、そのために必要な測度を作ることから始まる。$X_{1i}, X_{2i}, \cdots X_{Ri}$という多次元的データあるいは要素間の多次元的関係から始まるのであるが、これをどう目的に沿って分類の形に（外的基準のない場合である）まとめあげるかを講究するものであるが、数理統計学の多変量解析とは、全く発想を異にするものである。

本説で述べた諸方法それとの関連諸方法は、ここで述べたことに止まらず大いに発展してきているのである。

多次元的データを取り扱う方法は、いわゆる多変量解析をはるかに超えて、数理統計学の埒外で、発展してきているのである。これらは、これらなりの弱点を持っているので、現象解析に対しこれらを適切有効に活用するために、データの科学の下に統合し、新しい発展方法を見出すべきときに来ていると考えている。

ESTRELA

2000年10月号
2000／No.79

CONTENTS　目　次