巧いモデル解析とは —情報簡略化の原理について—

鈴木 義一郎*

HOW TO ANALYZE DATA THROUGH MODELING
—ON THE PRINCIPLE OF DATA SIMPLIFICATION—

Giitiro SUZUKI*

Skilful and non-fictional model-building requires simultaneously two contradictory conditions. By the first, the model should be made to fit the actual phenomena, and by the second, it should be handled with ease. The essential part of data-analysis usually amounts to an input-output relation, where the input is the so-called data and the output is decision. In other words, data-analysis is the process of simplifying data for the convenience of decision. Some principles of simplification are presented. Especially in the case of model-fitting, the necessity for giving some subjective judgements is emphasized.

As an example of the common principle of simplifying data, concepts of the simplest sufficient information and of the most efficient information are proposed. To increase fitness of model the necessity for modification of preassigned model is also emphasized.

From the actual point of view, more global process of simplification must be considered. In this process, the input consists of the various types of information and the output is just the final conclusion. There may be no general principle of describing such global input-output relation. The true and reasonable data-analysis, however, must be grasped as a part of this global input-output relationship.

1. モデルの適合性と最適化

世の中すべての事柄にいくつかの原因がありそして帰結がある。ナニワブシの文句ではないが“ナニガ・ナニシテ・ナントヤラー”という次第である。原因の方を入力, 結果の方を出力と言い直してみると, 入力・出力の関係を調べればよいということになる。ところが, 実際の現象の多くは多入力, 多出力ときており, さらに出力のあるものは全体としての最終的な出力にフィード・バックして影響を与えるといった具合で, 複雑極まりない。したがって, ある現象をあまりキメ細かく記述しようと意気込んでみても, しょせん手に負えたものではない。単純化すればそれだけ扱いやすくなるのは理の当然とはいいいえ, どこまで単純化するかさらにどのように単純化

するかが問題になる。結局, よく効いていそうな要因の方から重点的に取り上げていって, あまり効いていないものはどんどん無視すればいいということになる。この辺のプロセスをやや気どった表現で“現象を捨象する”などと言いあらわす。さらに捨象して残ったスケルトンの方を“モデル”等と表現する。結局現象をうまくモデル化できればいいということになる。

ここで現象のモデル化がうまくいったという表現についてもう一度明確にしておく,

- (i) その現象の基本的な側面をある程度の確に表現しているか(適合性の問題)
- (ii) そのモデルを取り扱っていくうえで必要とするデータや計算法等が利用できて最適解を見出し得るか(最適化の問題)

このような二律背反的な2条件を同時的に満足するよ

* 統計数理研究所 (The Institute of Statistical Mathematics)

う配慮することが、絵空事ではないというまいモデル解析のための必須条件となる。

2. 理論家はスタンダード好み

大局的な入出力関係をとらえる前に、主としてデータが関与している部分について考えてみる。データ解析の方法を扱っている入門書でも、データを加工し記述していく方法に触れているとはいっても、そのほとんどが平均とか標準偏差そして相関係数と回帰直線といったものの、導出法を扱っているに過ぎない。それも身長とか体重のようなさしておもしろくもないデータを用いて。ところが実際面で問題にされるデータとなると、もっと複雑多岐な様相を呈しているため、入門書で得た知識だけではどうにも手の施しようがないといった事態に立ち至る。つまり、データを簡略化していくさいの原理的なものに関する説明がなされていないから、入門書のパターンをはみ出したような事例に関しては、処理能力が身につかないことになる。

入門書の導入法に関するこのような罪を裏返してみると、書き手、ひいては理論家があまりにもスタンダードな場面だけを限定し、典型的なモデルに関してだけの解法をねらい過ぎていからである。条件が崩れたときの救済措置とか、“但し書き”についての考察に、無頓着のままでいるからである。

例えば、サンプリング・エラーだけで問題を片づけようとする場合なら、構造の安定した集団からデータが何回でも取り出せるという、再現性の仮定が必要となる。そこで確率というモノサシを使うからには、測度論的な確率論に立脚した、数理統計学のワキからはみ出すことは許されない。ところが、実際の現象の方が再現性に乏しいという場合——そのような場合の方が圧倒的に多いと思うのだが——どう対処するかが問題になる。どんな本をひもといてみてもヒントになりそうなことは見あたらない。いっそあきらめてしまえと放棄できるのなら別状ないが、そうもいかない場合にどうするか。現時点としては、仮定がすべて満たされているものと想定して一応の解を得たなら、その解の信憑性について若干疑問の余地あり、といった但し書き付きで報告しておくぐらいが関の山だろう。

ところで、理論家はなぜスタンダードな場面だけを好むのだろうか。その理論の立脚面の認識のし方が甘いこともあるが、何よりも、仮定を崩すと既製の解析手法が使えなくなるからである。そのようなずれた場合にも使えそうな、代替の解析方法がありそうかという、解け

ようなずらし方をするのならともかく、悲観的ケースの方がほとんどであろう。それは担当者の非力さに起因するというよりは、問題そのものの難解さによるものである。結局絶望的なのかというと、必ずしもさにあらず、解析の手法だけに依存するという考え方を捨てればいい。シミュレーション等を用いて経験的に答えを得ようとするのも一方法だし、すべてをスッパリした数量的表現に求めようとせず、感覚的に実証できたと判断される段階で満足しようという行き方だって考えられる。とかく理論家は、華麗さを好みスッパリさせることを好むといった性癖があるから、解決手段として解析的な方法の方を重視しがちになる。問題の先の先までの答えについては、よう面倒が見きれないことが判明したのなら、解析手法偏重主義の行き方は、どう見ても悪癖として葬り去らなければならない。一度染みついた癖というのは、当人がよほど意識して努力しない限り、なかなか払拭しきれないものである。

3. 主体的表現の重要性

お互いにわかりやすく話すことが大切である。物を話したり書いたりする人種は、相手にわからせようとする努力を怠りがちである。だから、人の話を聞いたり読んだりする場合、“わかる”ということにこだわりすぎるのは賢明な策とは思えない。あまりチャランポランでも困るけれども、適度に聞き流すコツも心得ておきたいものである。わかりやすく物を話したり書いたりできないのは、問題意識の薄弱さに起因することが多い。日本の研究者族は、とかく外国の人間があるイメージのもとに築きあげた理論で、かなり権威づけられたものの微調整だけで済まそうという魂胆がある。このような舶来崇拜主義は、何も研究者だけに限ったことではない。画期的なアイデアを登用した新製品の開発には金がかかりすぎるから、ちょっとした真似事に微調整を加えて、アメリカのような他の先進国でやったことを、残された低開発国を目ざして市場を占有するといった、大企業の姿勢にしても然りである。総体的なエネルギー不足がそうさせるのだといった解釈も成り立つが、日本人の器量の小ささは風土的な影響によるのかもしれない。

日本語がとくにそうであるように、日本の文化は“察し”の文化であるという、うがった表現もなされている。大口でわめいて他人の主張をへこまして、自分の考えを押し通そうとする態度は余り見られない。“含み”のようなものを残しておいて相手に察してもらうこ

とで意志の伝達を図ろうとする態度の方が一般的である。このような、日本で生まれ日本の風土で育った人間には、西洋流の物の考え方がなかなか吸収されないのだろう。もっとも日本人をみな西洋化しようとたくらんで変に圧力をかけてみても、しょせん本物にはなりきれないだろうから、日本人本来の得意な面を伸ばす方向に持っていく方が得策かもしれない。観光旅行なんかでもグループで行うのが日本人の特性なりとあざけられているように、日本人は団体でもって自己主張したが、個で主張しようとする、途端につまはじきの憂き目を見るのがわが国のお国柄であるから、主体的な表現はとかく影をひそめてしまいがちになる。だから、能力を持った若手が急には頭角をあらわし得ないし、非常に独創的な仕事をするのにも不向きにできている。しかし、日本人のグループとして活動しようという性向は、脆弱な個のエネルギーを集集させて、巨大な集団としてのエネルギーに発展せしめるうえで有用になってくる。こつこつと小さな利幅を着実に積み上げた成果がGNP¹⁾世界第2位である。日本人全体としての国力が急激には絶対沈下し得ないだろうという見通しは、他の国にある種の脅威を与えるのにじゅうぶんである。

話がやや脇道にそれたが、とにかく主張しなかったのは、自分自身を他と比較して評価してもらう代わりに、直接的な評価を期待した自己主張のし方が、しかも日本の風土にもうまく溶け合えるようなものが、もっと考案されて然るべきであろうということである。つまり、日本人同志でも通用する主体的表現法が種々開発される必要あり、という次第である。グループの方が受け容れられやすいというのなら、グループとしての主体性を持った表現でもかまわない。このような場合に使えるもの、例えば川喜田氏のKJ法²⁾等というのが考えられる。

さて情報簡略化の原理を述べようとする意図の論説で、なぜ主体的表現の重要性を説いたりしたのかについて触れておこう。最後の節でも述べるが、大局的な意味での情報簡略化のプロセスの中では、客観的に判断しきれない部分とそうでない部分とを、同時的に取り上げていかなければならない。とくに、主観的な判断を加味する

さいには、できる限り論理的にスムーズに導入することが肝要となる。つまり、情報というものの簡略化を大局的な意味で正しく行い得るには、必要な範囲内で、主体的なものをタイミングよく混ぜていくことのできる能力が、必要となるのである。「客観的」というと「絵空事」、「主観的」というと「一人よがり」といった具合に連想するようでは実際問題として困るのである。“客観性”という縦糸に“主観性”という横糸を適宜織り混ぜながら丁寧に仕上げられた織物が正しいデータ解析本来の姿なのであるから。

4. 多元情報のとらえ方

世の中には一元的なものより多元的情報の方が圧倒的に多い。そのような情報が計量値としてとらえられる場合には、いわゆる多次元空間内の点として表現できる。しかしカテゴリーにしか分類できないような情報となるとそうはいかない。ところが計算機を媒体に情報を処理していこうという段になると、多次元空間化というか、情報を数量化しないとうどうにもならないといった考えが根強い。計算機の仕組みから考えてみて、情報をカテゴリーに分類して収納しておいて、必要な時に必要な情報だけを検索できるようにして利用するのが、本来的な使い方ははずである³⁾。多次元空間内に直接表現できる情報ならともかくも、そうでないパタンの情報までやたら数量化して、多次元空間内での線型演算を施して簡略化をこころみる。いろいろなプログラムがパッケージ化され、そんなに頭を使わなくても計算機を手軽に使えるようになった昨今の風潮が、こういった傾向をさらに助長している。もっとも、計算機に線型演算以外のことはできないようにしてしまったとはいっても、それ以外に何か調子のいいことができそうかを考えてみても、多くを期待できる代物ではない。まだまだというより将来においても、データ解析の主役は人間の方が務めなければならないだろう。

多次元空間化が必ずしも適切でないデータは、どう取り扱ったらいいのだろうか。まず順序づけ不可能なカテゴリカルデータを無理やり多次元化しようすると、超高次元空間内でしか表現し得ない。それを低次元化して情報を吸収しようとなると情報量の損失に測りしれないものがある。また、機械的に多次元化を行うとデータ本来の持味がうまく吸収されなくなるといった欠点も考え

1) 国民総生産という、一見いかめしそうな表現が取られているが、必ずしも生産活動に直結した部分だけのトータルではなく、国内のモビリティのある種の測度にすぎない。

2) 例えば、川喜田二郎著「発想法」(中公新書)を参照。

3) 例えば、国鉄の「みどりの窓口」等で取られているような計算機の利用法がそうである。

られる。いわゆる“数量化理論”⁴⁾なるものを用いてデータ解析しようとする立場がこの類にあたる。色彩の感覚などではどうか。これはカテゴリーとしてとらえるより、色環図のような丸い形の連続体と考える方が適切かもしれないが、これを線型空間内に閉じ込めることは恐らく不可能であろう。

多次元空間に閉じ込めることがあまり妥当とも思えないような多元情報の場合には、いろいろな項目ごとにある種の“層別”をこころみて考えていくのが適切ではなからうか。つまり、“二元表”のようなものをいくつか並列的に観察することによって多元情報を吸収しようという考え方である。この層別の考え方は、すでに多次元空間内に表示されたデータを解析していくうえでも、きわめて有用である。形式的な低次元化の手法よりはるかに、データの持味に即した解析ができようというものである。このような意味でクラスタリングの手法が強力な有効性を発揮するであろうし、計算機との“相性”の点からも申し分がない。もっともこれまでに開発されているクラスタリングの手法は、多次元空間内のデータのクラス分けという観点にだけ固執している傾向が見受けられるので、今後はもっと広い視点からの手法の開発が望まれる。

最後に多次元空間にこだわって損をする例をもう一つ挙げておこう。情報の吸収に視覚的なものを援用させることは、だれしも考えることである。それだからこそ、高次元のデータをより低次元化してグラフ表示をこころみる。ここで錯覚が起こる。つまり、グラフ表示は平面内で行われなければならないからデータを二次元空間にまで落とす必要がある、といった先入観である。視覚的にとらえ得る平面と二次元空間とをうっかり同一視してしまっているという誤謬である。一目でわかる手書きの郵便番号を、計算機がなかなか識別できないといった事情から考えてみても、平面内に記された形のパターン、とくに色のついたもの等を認識することが、いかに多次元的なものと異質であるかが理解できる⁵⁾。人間の計算機に勝る卓越性がこのようなパターン認識の能力であるといっても過言ではあるまい。最近になって、チャノフの“顔面表示法”⁶⁾とか脇本の“星座図表”⁷⁾とい

った手法のように、二次元より高次の多次元データを、平面内に表示しようというこころみが開発されだしたことは、データ解析手法の発展史のうえで画期的なエポックとも考えられよう。

5. データ簡略化の原理

これまで議論してきたことをもう一度総括的にながめてみて、“簡略化の原理”なるものが一体存在するものなのか、存在するとすればその探索法をどうするか、等について述べてみよう。

まず、データよりは広義の情報を入力とし、最終的な判断を出力とする、大局的な入出力関係が想定される。その中でのかなり重要な一部分として、データもしくはデータ状のものを入力に、統計的加工を経た図形とか数値を出力とする、データ簡略化のプロセスを記述すべき入出力関係が存在する。簡略化のしすぎは情報の損失を招くし、簡略化不足は判断の決め手となり得なくなるから、どの段階でもって簡略化のプロセスをやめるかがポイントになる。そのような問いに対する答えは、利用目的その他の条件に従って、いろいろなボタンごとに分けて考えるべき部分と、すべての場合に共通した“データ簡略化の原理”によって説明できる部分から構成されることになると思う。後者の部分が多ければ利用者にとっては便利の上もないし、そのような期待に答えるべく努力することが理論家の使命でもある。ところで、“簡略化の原理”なるものの存在を期待できるのだろうか。無いものねだりをして詮ないこと、少なくとも存在の可能性のありそうなポイントを見定めてから、発掘作業に取りかかることが賢明である。

そこで、それぞれの典型的なモデルごとに開発された手法を、そのまま使っても大丈夫な場合と、何らかの意味で手直しをして用いなければならない場合とに分離し、後者の場合には、どう手直ししたらいいかについて整理してみることにから出発すべきである。さらにそのような手直しを、事前情報にもとづいて設定できる場合とそうでない場合とに分けて考えておくことも必要である。前者の場合は解析の段階では考えてみる価値があるが、データ状の事前情報は必ずしも利用できないと考え

4) 例えば、林知己夫著「数量化の方法」(東洋経済新報)を参照。

5) 色相に関する情報は一見カテゴリーカルなもののようにも思えるが、色環図のように連続して変化するものとしてとらえるべきである。ただ、これをユークリッド空間化することはまず無理である。

6) Herman Chernoff “The use of Faces to represent points in k -dimensional space graphically” *J. American Statistical Association* Vol. 68 (1973).

7) 脇本和昌・田栗正章 “2次元図式パターンを用いる判別分析” 応用統計学 Vol. 3 (1974).

の方が実際のだろう。

まずモデルごとに開発された手法の共通的テーマについて考えてみよう。代表的なものは単純十分情報という概念であろう。ある情報 X の利用目的が明確に規定されている場合には、その目的に沿うよう攻撃目標のパラメタ θ を設定できる。そこで、情報 X のそのようなパラメタ θ に関する情報量 $I(X|\theta)$ を減少せしめないという条件でより単純な形式の情報 Y に還元することをこころみる。すなわち、等式、

$$I(X|\theta) = I(Y|\theta)$$

を成立せしめるような $Y=Y(X)$ を (θ に関する) 十分情報と呼ぶことにすれば、もとの情報 X からもっとも単純な十分情報 Y^* を導出するパタンに関して、一般的な特徴づけを与えることが目標である。統計の方の理論では“最小十分統計量”⁸⁾ という概念がこれにあたる。

この十分情報という概念についても若干改良の余地がありそうである。というのは、情報量を全然失わないで簡略化しようというのは、少々虫がよすぎる。つまり情報量を保存することにこだわりすぎるあまり Y^* がそれほど調子のいい簡略化になっていないケースも、多々出てくるのではという疑問が残る。そこで、情報量に関してある程度の損失を許容しながら、簡略化の進行具合もある程度満足するようにもっていかうという考えが出てくる。当然情報量の損失の程度 $\delta(Y) = I(X|\theta) - I(Y|\theta)$ と簡略化の完遂度 $\rho(Y)$ とのかね合いが問題になる。例えば、

$$e(Y) = [\rho(Y) - \delta(Y)] / \rho(Y^*)$$

を最大にするような Y^{**} を有効情報とでも呼称し、これが使いものにならないかと考えてみるのも一方法である。 $e(Y^{**}) > 1$ なら Y^{**} の方を使った方がいいと結論できそうだが、問題は簡略化の完遂度の測度 $\rho(Y)$ の与え方がスムーズにできそうにかかってくる。

次にモデルを手直しして使う場合について考えてみよう。先に、データ状のものは事前情報として利用できないという状況下で考えないと実際のでないことも指摘し

た。モデルからのずれを許容して扱うさい、まず問題になることは、どのモデルをどのような手直しをして用いるかという問いに、答えを用意することが必要となる。つまり、モデルの選択の問題と手直しの問題とを同時に取り上げていかなければならないということである。しかも、手元にあるデータのみを用いることによって、

今、採択し得る狙上のモデルが $\{f_1, f_2, \dots, f_k\}$ のように与えられているとする。さらに各モデルにある種の手直しを加えて適合性を増強させたモデルを $\{\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_k\}$ とする。手元のデータ X をモデル f の下で解析するさいに適合度を測ることのできるモノサン $\gamma(X, f)$ が用意されていたとすれば、

$$\gamma(X, \tilde{f}_1), \gamma(X, \tilde{f}_2), \dots, \gamma(X, \tilde{f}_k)$$

を最大にするようなモデル $\tilde{f}_{i^*}$ を使えばいいということになる。このような行き方で問題となりそうなことを挙げてみると、次のようなことが考えられる。

- (i) モデルの手直しの規則が実際に規定され得るのか。
- (ii) 適合度を測るモノサン γ が自然な形で与えられるか。
- (iii) 同一データ X を使って手直しを規定したり修正されたモデルの選択を行ったりしている点で、独立性のくずれに起因する難点が深刻な問題とならないか。

このような問題を提起した直接的動機は次のような事情による。世に“適合度検定”なるものが頻繁に使われているけれども、この呼称が不適切極まりない。というのは“適合度検定”によって明らかにされることが、その分布を仮定して推論することを否定する根拠がなかったという事実だけで、その分布を積極的に支持していいといった適合度の測度を与えてくれるものではない。むしろ不適合度検定とでもいう方が適切である。したがって、いくつかの分布について適合度検定をやってみて、棄却できなかったものが複数個残ったとしたら、その中でどれを選べばいいかの答えは得られないでとまどってしまう。それなら適合度検定を一度しかやらなければいいではないかと考えるのもおかしい。何故そのような分布を仮定したのかについての根拠が、見あたらないからである。強いて挙げるなら、仮定した分布を用いると、種々解析を進めていくうえで都合であるから、ということになる。やたら否定主義を取ってみても無意味であるから、とにかく利用できるものを使っていこうという姿勢にも一理ある。ただ問題を解決するということだけにこだわりすぎるあまり、情報吸収法に関する適否

8) 例えば、正規分布の平均 μ と分散 σ^2 を推定する目的で与えられたサンプル $X = (x_1, x_2, \dots, x_n)$

において、いわゆる最小十分統計量、 $Y^* = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right)$ が単純十分情報を与えている。通例

平均や標準偏差のスタイルにデータの簡略化が行われるのは、モデルとして暗に正規分布のようなものを仮定しているためとも考えられる。

を確かめなかったという点で、冷静さを欠いたデータ解析の態度と非難されても反論できないだろう。

6. 大局的な簡略化の意義

最後に情報を入力、判断を出力と考えた、より大局的な入出力関係をとらえるさいの簡略化の問題について付記しておこう。計算機にほうりこんで計算する部分はもちろん、データを簡略化していくプロセスも、このような広義の簡略化の過程の一部分にすぎないという事実を認識することが肝要である。とくに、データ以前の情報の吸収法とか導入法が問題になる。さらに、簡略化されたデータを最終判断とどう結びつけるかもである。

このような広義の全過程を、一挙に簡略化するさいの原理的なものは、存在しないだろうと思われる。というのは、現実に取りられる判断の多くがデータ状の情報以外のものにもかなり依存していると考えられるからである。つまり、ある判断を下すときには、当人の先験的な情報を何とはなしに使ってしまっていることが多い。例えば、ある目的のためにデータ解析を行ったというときに、そのような解析パターンを規定するさいに、その人の先験情報が使われているということである。さらに、解析結果がどうもおかしいとして、すぐには最終判断と結びつけることをためらったりする態度なんかにも、当人の先験的に得た情報に関与している、といった具合である。各人の持っている先験情報の中には、きわめて主観的なものもあれば、客観的なものに近いものもある。広義簡略化のさいに、前者のタイプのものが多く使われるのは困りものであるが、後者のタイプの方を貴しとすることにも問題はあろう⁹⁾。このようなわけで、広義簡略化の全過程を、客観的に記述しようというところが無理であることがわかる。

かくて、データ解析を行うさいに、より大局的な簡略化のパターンについて考えてみることは大切であるが、そこ

に共通した原理的なものを求めようとしても無駄であることがわかった。ただ、大局的な簡略化の問題を考えていくうえで、実際に使う側と理論を構成していく側とに、それぞれ共通した注意書きを述べることはできる。

大局的な簡略化を行うさいに、各個人の先験情報を確かに投入して進行しているのだといった事実を、認識しているかどうかが問題である。つまり、無意識で先験情報なるものが投入されている場合には、そのような情報の性格にどの程度の客観性を期待していいのか、したがってそれを用いて行った最終判断が、どのくらい客観的であるとみなせるのか、といった問いには答えられないことになる。これが、先験情報を持ち込むさいそのことを認識しているとなると、可能な限り客観的な先験情報を使うことができないか、さらに先験情報投入のため客観的なものがどのくらい犠牲にされたのか、といった問いにある程度の答えは用意しておくことが可能となる。結局、得られた結論の主観的なウエイトのようなものも、大ざっぱながら見積もることができる。

一方、理論の立場から接近していこうという人たちにも、反省する余地がいろいろある。もし形式的なもの、客観的なものにこだわりすぎていると、大局的な簡略化の過程に対して何ら有益なことを示唆し得ないばかりか、より狭義のデータ解析のパターンに関しても、ピントのぼやけた命題を羅列するだけに終わってしまうのではなかろうか。理論の本筋は客観化が第一とばかり、あいまいな主観的なものを許容した理論の存在などあるはずがない、といった幻想を抱いてはいないだろうか。自分の得意とする守備範囲が、理論の構成と展開であると自らのワク組みを固執するあまり、より一般的というより、常識的なものの考え方をするのを怠ってはいないだろうか。さらに本当の生データから最終判断に至るまでの道筋を、概観できる経験がないとの理由で、純理論の部分にのみ関心を払いすぎてはいないだろうか。

以上、情報簡略化のプロセスについての“原理”らしきものに関する、私なりの考えを述べてみた。このような考えに対して、いろいろな難点やら異論やらが当然でてくるものと予想されるので、読者諸兄からの御叱正を待つ次第である。

9) 例えば結婚相手を決めるのにファースト・インスプレッションがかなり効いていて、それがかなり個人差のあることは定説である。このように、客観的な先験情報で利用可能なものがかなりあったとしても、それらがほとんど効いていないことがよくある。