解説/Review

統計数理の誕生とその広がり

樋口 知之*

The Birth of Statistical Mathematics and its Expanding World

Tomoyuki HIGUCHI*

Abstract– This document gives a brief explanation for what is "Statistical Mathematics," and describes how its notion has been expanding to a wide variety of research fields. The recent interests of statistical mathematics are focused on generating more flexible mathematical models for describing complex phenomena in terms of data generation mechanism and/or its function. Such efforts demand collaborative works with the machine learning community in a big data era which we are now faced with.

Keywords– statistical modeling, information criterion, prediction performance, Bayesian modeling, MCMC, particle filter, personalization, induction and deduction, simulation, data assimilation, uncertainty quantification, machine learning, sparse modeling, kernel method

1. はじめに：数理モデルと現象

『統計数理』とはいったい何であろうか？統計数理研究所が設立されたのは戦中の 1944 年であるが，それ以前に統計数理の言葉の定義はあったのであろうか？本稿ではまずそのあたりから考察してみたい．統計数理は数理科学とオーバーラップするところは相当大きいが，その関係を包含で示すことはできず，応用分野・隣接領域はむしろ数理科学よりも広いように思える．数学の言葉で記述された研究対象やその機能を数理モデルと呼んだとき，数理モデルの評価は実際の現象の観測・計測データのある特徴量との乖離で行なわれる．数理モデルが表現・予測する量が実際の観測・計測の値にぴったりと合うことは通常ありえず，その乖離の程度を表現する新たな数理モデル（目的関数）が実際の研究においては欠かせない [1]．ある意味，数理モデルとリアルとの間の緩衝材が研究には欠かせないのである．この緩衝材は，数理モデルという一種の仮想空間をリアル空間に埋め込む存在のため，細胞膜的な機能を果たしていると例えられよう．統計数理がつかさどるのは，この数理モデルと細胞膜の両方を同時に数学でモデル化する作業である．細胞膜は細胞内外を単に隔てている静的な構造体ではなく，細胞の活動にとって重要な機能を担っている．生物同様，細胞膜を通して数理モデルはリアルと情報交換可能なため，統計数理が果たす役割は本質的である．

2 章では，統計数理研究所の成り立ちを振り返りつつ，大先輩の諸発言に統計数理の定義を見いだしてみたい．そのことで統計数理の特性を明らかにし，統計数理が数理モデルの構築において果たした役割を明らかにする．特にモデル評価や統計的モデリングの意味について考えてみたい．3 章では，モデル評価の基盤的な考え方となる予測性能や，機能のモデル化について考察する．予測性能の向上に欠かせない柔軟な表現能力を備えたモデルの枠組み，特にベイジアンモデリングについても触れる．4 章では，事前情報として，シミュレーションのような支配方程式に立脚した情報を活用し，データ解析作業に直接シミュレーションを溶け込ませる枠組みについて触れる．その作業は大きな計算コストを要求するため，現代的な統計数理の研究においてはスーパーコンピュータが欠かせない．5 章では，この 1，2 年，大きな注目を浴びているビッグデータと統計数理との関わり，およびビッグデータが促す統計数理の研究動向について概説する．6 章はまとめとして，人材育成の重要性を説いてみたい．

2. 統計数理と統計的モデリング

2.1 統計数理研究所の設立

数理統計（Mathematical Statistics）と統計数理（Statistical Mathematics）の意味するところの違いは英語で表記すれば明らかであるが，日本語表記だと横幹連合の会員の方々でも両者を混同されているケースを時折見聞きする．統計数理の言葉の起源をどこまで遡れるのか正

*統計数理研究所 東京都立川市緑町 10-3

*The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa-shi, Tokyo

Received: 16 February 2014, 26 February 2014

確にはわからないが，たぶん統計数理研究所の設立以前は学術界においても一般的な用語ではなかったと推察する．統計数理研究所は戦争終結間近の 1944 年に設立されたので，今年（2014 年）はちょうど設立 70 周年になる．20 世紀になって，人口統計などの社会統計を主としていた統計学は工業生産などへの応用面が開かれ，その情勢のもとに統計学に関する研究所を設立しようという機運が生じた [2].「統計数学を中心とする統計科学に関する研究所の設立」が 1943 年に学術研究会議の建議として提出され，翌年の 1944 年 6 月 5 日に文部省の直轄研究所として統計数理研究所が発足した．

戦中時に設立された研究所は必ずしも戦力増強を直接の目的としたものではなかったが，科学戦といわれた情勢下に特に必要性の高い特定の専門分野の研究を促進することを目的としていた [3]．実際，官制の第 1 条によれば，その目的は「確率に関する数理およびその応用の研究を掌り，並びに研究の連絡，統一および促進を図る」（現代文表記に修正）とされていた．研究所の名称は，初代所長の掛谷宗一が提案したもので,「この研究所の扱う学問の内容が，従来の単なる統計でもなければ，いわゆる純粋数学的な数理統計でもなく，両者を含むもの」として，新しい統計数理という名称が採用されたとある [2]．戦前のため今となってはその正否は文献情報にたよるしかないが，いずれにせよ，今の統計数理に連綿とつながる考えである．あまり知られていないが，正所員 6 名の他に 7 名ほどの兼任所員がおり，伊藤清も兼任職員として 6 年半ほど統計数理研究所に在籍した．戦中ながら設立早々にも統計数理研究所講究録の出版が開始され，1945 年 1 月 15 日に発行された第 1 巻第 13 号に伊藤清も 2 本論文を発表している [4].

2.2 統計数理の特性

現実との接点を非常に意識した設立時の研究における志向性は，研究所において脈々と受け継げられていく．数量化理論で著名な 7 代所長の林知己夫は，主として社会調査の共同研究を通じて，従来の記述統計学でもない，推測統計学でもない,『統計数理』の確立に尽力した [5]．合理的，実証的な方法論の体系としての統計数理は，理論・応用の区別のない，林によれば「まず何が肝要であるかを見抜き，これをフォーミュレイトし，実験や調査の計画をたて，この下にデータを獲得し，分析し，予測し，行為の指針を得ることを志向する．これらに関する方法的成果はもとより，これを編み出すまでの全過程を包含するものである」として適用分野を時代と共に拡大する [6]．一貫して現実の問題に根ざした統計学の発展に尽くした 8 代所長の赤池弘次は，統計数理の意味を,「統計数理とは統計的な概念，これはいろいろ統計的な見方，考え方をことばで表現したものでありますが，そういうものがだんだん豊かになっていくような，そういう展開を目的とした数学的な理論といったらよいのではないかと思われます.」と口述している [7]．つまり統計数理は，統計的な思考を数学の力を借りて体系的に取り扱う学問である．赤池は統計的思考を具体的に,「今入ってきた新しいものを，必ず自分の持っている経験やその他の情報と対比する．これが統計的な見方の本質であります.」と同書内で解説し，経験やデータなど手にしている情報を統合的に取り扱い，予測そして意志決定に至るプロセスを体系的に科学する学問が統計数理であるとしている．

また,「統計数理には，抽象的な性格と具体的な性格の二面がありまして，具体的なものとの接触を通じて抽象的な考えあるいは方法が発展させられていく，これが統計数理の研究の特質であります.」とも赤池は述べている．この関係に関連して，興味深い記述を最近見つけた．統計数理研究所の設立にも大きく貢献した初代兼任所員である北川敏男は，著書 [8] の 8 章の『統計学と私』の中で,「私は，幼少の頃，地理，歴史というような記述的な学問が大変好きであった．一方，数学のような論理的な思惟の学問も得手な子供であったから，妙なコンビネーションである.」と回顧している．具体性と抽象性，記述的と論理的など，脳機能の分担論で言うところの右脳と左脳の機能ともバランスよく働くことが，どうも統計学の研究に必要な資質でなかろうかと思われる．小学生で言えば，社会に強い興味を持ちつつ算数も得意である児童である．世間でよく言う “理系” 対 “文系” の色分けでは，たぶん，統計学の意味で優れた能力は特定されにくいと思われる．

2.3 統計モデルと情報量規準

統計数理は複雑・不確実・動的な対象を研究対象とし，不完全情報の状況での判断を探る．10 代所長（2002〜2010 年度）の北川源四郎は統計数理を「実世界の現象を解明するために，本質的な情報を抽出し，予測・知識獲得や意思決定を行うための方法を対象とする学問」と定義づけた [9]．この情報処理における推論の根拠を示すものが統計モデルとなる [10]．従って統計モデルは，データに対してユニークに決まるものではなく，いろいろな立場や知識によって多数のものが考え得ることになる．このことから統計数理研究所では，統計モデルのことを “数理のめがね”（材質が数理でできた眼鏡の意味）と呼んできた．

領域科学（統計数理が応用される分野）では，真のモデルがあって，それをデータから推定したいと考えがちであるが，統計数理では真の構造の忠実な再現を目指すのではなく，永続的に統計モデルの改良を行なう．この行為が統計的モデリングなのである．赤池の言葉によれば,「構造が確認されている確率的な機構という特殊な場合を除き，期待の構成の仕方は我々の持つ知識や経験

の使い方に大きく依存する．したがって，唯一無二の真の構造のようなものは存在しない．‥‥‥ したがって，我々はより良いモデルの探求を通じて，常に未知の状態にある究極的な真理あるいは真の構造に迫るのである.」となる [11]．手元にあるモデルの性能が悪ければ潔く見切ってしまいモデルを乗り換える態度は，“モデルの使い捨て” とも言えよう．

“めがね” は個人の視力に調整されており，時にはサングラスを着用する人もいるため，対象は固定されていても “めがね” 毎にものの見え方は変化する．正しくモデルを乗り換えるためには，モデルの性能を適切に比較する道具が必要になってくる．ここで我々は，将来のデータをより良く予測する統計モデルが良いモデルとの立場をとる．北川の言葉を借りれば,「真のモデルに含まれる真のパラメータを推定することが目的ではなく，本当の世界から将来出てくるであろうデータを予測するために適当なモデルをつくることを目的とする視点」となる [9]．この予測能力を一般化した概念は汎化能力と（特に機械学習の分野で）呼ばれる．

予測分布および汎化能力に注目すれば，AIC（Akaike Information Criterion）や小西・北川によって提案されたGIC 等の情報量規準が自然な形で導出される [12]．情報量規準は，統計モデルを比較する上で必須の物差しであり，モデル群の大海原を航海するためのコンパスと言える．赤池は，情報量規準を参照しながらモデルを使い捨てる時代を切り開いた先駆者なのである．ここにいたっては，一般にモデルの真贋を判断する作業と考えられているモデルヴァリデーションも，モデルを進化させる契機の作業でしかなくなる [13].

3. 機能のモデル化と設計科学

3.1 予測が主導する発見と機能表現

統計数理では，予測能力の高いモデルを求め続ける一連の過程から，統計的モデリングの目的である情報抽出・知識発見や意志決定において，良い結果が得られることを期待している [1, 14]．人類が構築してきた様々な科学的理論は，あらゆる場面にも適用可能な普遍的なものというよりは，我々の認識の範囲内の現象を適切に表現するものと考えるべきである．従って，実在の理論と現実のデータのずれにこそ新しい発見の可能性が潜んでおり，特に，特異的振る舞いを示すずれは新たな因果関係やイノベーションの発露と考える．このように統計数理においては，予測と発見は連動している [15].

予測性能を高めるには，対象そのものを初めから実体的に精緻に記述するのでなく，対象に関する情報の入出力関係に代表されるような，機能自体を模倣する数理モデルを構築するほうがシンプルでかつ効果的である．近年の著しい発展があるロボティクスをささえるのも，『機能のモデル化』の概念である．例えば，悲しい顔を見た時に相手を元気づけるようなロボットを作りたいとする．その目的のために従前は，眼の生理的構造，そこからの信号伝達，脳の信号処理の理解，そして運動方程式に基づくロボットの制御など，すべての研究作業が，素過程の積み上げ方式，つまり演繹的プロセスのつなぎ合わせであった．残念ながらこのような研究アプローチでは目的の達成にはなかなかいたらない．ところが最近のロボティクスの主たる作業は，インプットデータとアウトプットデータの関係にのみ注目し，その機能（入出力関係）を近似する数理モデルの構築である．数理モデルに含まれる諸パラメータは統計的学習手法によって決定できる．素過程の積み上げによる演繹的手法より，データにもとづく帰納的手法のほうが目的の達成には効率がいいことは，子供がニュートンの運動方程式を知らなくても，自転車に乗れたり，鉄棒を逆上がりできるようになる事実を想起すれば明らかであろう [16].

この，予測を良くすることを目的とした機能のモデル化にまずは注力する姿勢と，本当のものをなるべく正確に記述する姿勢とは，実はかなり違うということを認識しなくてはならない [9]．この違いは，学術会議で以前に議論された，設計科学と認識科学の違いと同相である．新しい学術の体系として,『設計科学』が学術会議から提言されたのは今から約 10 年前のことである [17]．設計科学は，人間の全体性を現す人工物システムを研究対象とし，目的や価値を正面から取り込んだ新しい科学である．この思想は，工学分野においては自然に受け入れられるが,『あるものの探究』を主な目的として発展してきた『認識科学』，特に知的好奇心を重んずる自然科学の価値観とは対立するものである．しかしながら，自然科学の取り扱う対象が地球環境や生命体といった複雑なシステムになってくると，そのシステムの理解には，むしろ設計科学の方法論が適している．そもそも，人間の個別あるいは集団的活動が大きく影響する地球システムや生命体においては，“あるもの” 自体の定義が困難であり，結果として認識科学のアプローチは筋が悪い．事実，地球科学の分野では，世界的な大規模災害の頻度の増加にも後押しされる形で，知的好奇心を満足させる現象の理解よりも予測性能の向上を第一義的な目的とする研究が増えている．また，その考え方はゆっくりではあるが学術全体に広く浸透しつつあるように見受けられる．統計数理の，予測能力を高めるモデルの探索を通じて知識発見を実現する考え方は，広く受け入れられつつあると言えよう．

3.2 ベイジアンモデリング

モデルを頻繁に乗り換えるために必要なもう一つの鍵は，柔軟にモデルを創造できる数理的な枠組みである．統計数理では，データや対象に対して持っている知

識や経験，期待感を統計モデルに十分に反映させることが推奨されている．同時にモデルの予測能力も高めることが期待されている．従前は，活用するモデルの表現能力がそれらに対応できず，比較的少数のパラメータを持つ数理モデル群の利用でそれらの実現をどうにか図っていた．しかしながら，手元のデータの表現能力を安易に高めようとすると膨大なパラメータの導入を招き，それらを安定して推定することは事実上不可能であった．時としてデータ数よりもパラメータ数が多くなり，その推定にかなりの計算資源を必要とするとともにオーバーフィットの本質的な問題が生じ，データ解析の現場では大きな欲求不満が鬱積していた．

この問題を解決する糸口は以前から知られていた．データを柔軟に表現するために膨大なパラメータを使うのは同じだが，そのパラメータに対して我々が持っている知識や経験を直接反映させるのである．具体的には，パラメータを確率変数として取り扱い，既存の知識や経験はその確率変数の事前分布として定義する．これにより莫大な未知数の値を安定して（オーバーフィットを合理的に回避しつつ）推定することが可能となる．この確率構造は，ベイズ統計学そのものであり，確率変数は直接観測（計測）できないことから潜在変数と呼ばれる．ただし，潜在変数間や，データと潜在変数間の関係が非線形となることも，また事前分布が非ガウスとなることも普通である．その結果，潜在変数の高次元条件付分布は解析的に表現できず，アンサンブル近似を採用せざるを得ないことが難点であった．このため，ベイズ統計学の有用性は以前から理解されていたが，この問題の抜本的な解決は 1980 年代まで待たざるを得なかった．それ以前は，ベイズの定理自体は 18 世紀に早々に発見されたにもかかわらず，長い間，確率の解釈，事前分布の設定，事後分布の計算の困難さのために哲学的議論に終始し，実用化にはほど遠かったのである [18]．実用化の扉の鍵となったのは，一つは計算機の急速な発達，もう一つは計算集約的な画期的アルゴリズムの提案である．

パラメータ数の爆発は大容量メモリーを，またベイズ推論のあらゆるところに出てくる膨大な場合の数の積和操作は高速 CPU を要求する．1980 年代の計算機の発達は，主にスーパーコンピュータを軸に著しいものがあり，ベイズ統計学はこの恩恵を被った．さらに，MCMC（マルコフ連鎖モンテカルロ）法や，粒子フィルタのような高次元のパラメータベクトルを推定する画期的な計算アルゴリズムが提案され，ベイズモデルの有効利用が大きく進んだ [19–21]．

3.3 新 NP 問題と個人化技術

『設計科学』の考え方はゆっくりではあるが学術全体に広く浸透しつつあることを 3.1 において指摘したが，一方社会に目を向けると，価値観の多様化などを受け，“コ”（個人，個性，固有，個別）に特化したサービスが求められている．例えば，オーダーメイド医療，副作用の研究，テーラーメイド教育，マイクロマーケティング（One-to-One marketing）などすべて “個人化” という言葉で概括できる，“コ” に特化したサービスあるいは製品である．21 世紀は，20 世紀の大量生産・大量消費をめざした科学から，個人に焦点をあわせる科学へ確実にシフトしつつある [22–25]．

最近の計測技術やスマートフォンの発達により，ケース毎に膨大な属性変数が（同時）測定可能になり，ケース数 N よりも属性変数の数 P が圧倒的に大きい状況が普通になってきた．例えば，人間ドックの検診を受けると，さまざまな医療検査診断項目があり，その結果通常のデータ解析の環境では $N \ll P$ となる．属性変数の値が近いケースのペアを探せば，個人化サービスの実現は容易そうであるが，実はほぼ不可能である．$N \ll P$ が生む困難さに拍車をかけるのが，データの欠損である．先ほどの例で言うと，人間ドックのタイプにより，簡易検査では多くの検査項目が未測定となる場合である．このような個人化サービスを阻むデータの特性は，計算数学の分野で広く知られている “NP 困難”（Non-deterministic Polynomial time hard）の言葉にひっかけて,『新 NP 問題』と呼ぶことも多い．この困難を減じるためには，ある属性変数（特徴量）で似た値をとるものは，他の属性変数でも似た値をとることが期待できるといったような先見的知識を活用することで，表でデータが抜けているところを埋めていく作業―つまり，インピュテーション―が効果的である．この情報処理にもベイジアンモデリングが重要な役割を果たすことは言うまでも無い [21, 25]．

4. シミュレーションからデータ同化へ

4.1 帰納と演繹

さまざまな科学の領域における複雑な現象の解明手段として，実験，理論，そしてシミュレーションは自然科学の研究手法の三本柱である．近年の計算機の発達に伴い，研究開発を行うあらゆる場においてシミュレーションの占める役割分が増大してきている．通常シミュレーションは，当該分野の支配方程式を計算機に実装するために数理モデルに変換した，いわゆるシミュレーションモデルの開発から始まる．もしシミュレーションモデルが動的な時間発展形式ならば，初期条件・境界条件等を与えれば，解は淡々と計算され更新されていく．つまり演繹的推論，言い換えれば順問題思考である．得られた計算結果から，高度化された可視化技術等を利用して当該分野における科学知を発見していく作業，それがシミュレーションを用いた科学的推論のスタイルである [23]．

一方，統計学においては，現象を支配している規則，

関係式といった経験則を観測や計測データから推測していく．すなわち帰納的推論，他の言葉では逆問題思考である．研究対象の現象が示す複雑さに比してデータ量が少ない場合は，知識や経験をモデルに反映させるのが統計数理である．3.2 で示したように，ベイズ統計の枠組みで，それらを事前分布として組み込む術を我々は知っている．シミュレーションが生み出す時間発展情報も事前情報として取り扱えば，データだけでは得られない新しい知識が得られるのではないかと容易に想像される．従来は，シミュレーションの結果のヴァリデーション用の材料として，データがシミュレーションと対峙し，お互いに干渉しあうことなく独立に存在してきた．両者を混ぜることなど禁手だとまで断言する人もいる．その発言には，モデルの真贋判定を第一義的に要求する『認識科学』の思想が色濃く投影されている．

4.2 データ同化とスーパーコンピュータ

近年の人工衛星観測に代表される地球観測システムにより，かつてシミュレーションによってしか推定できなかった地球規模の観測データが得られるようになってきている．一方で，計算機能力の上限や参考とする観測データの不足により，これまでのシミュレーションは現実を大幅に理想化・単純化した設定で実施されてきたことも事実である，そこで，広範囲で得られ始めた観測データとシミュレーションを統合する情報処理がデータ同化である [21, 26, 27]．シミュレーションは，初期条件と境界条件をいったん与えてしまえば独自に計算が進むものであるが，データ同化ではシミュレーションの各タイムステップの計算結果を，データを参考にして修正し，修正された結果を使って次のタイムステップの計算に臨む．この修正のプロセスを挟むことにより，単純なシミュレーションモデルでもより現実に合った予測が可能になる．

データ同化の問題は，シミュレーションによるシステムの時間更新をシステムモデル，シミュレーションに内在する諸変数のさまざまな観測法による観測を観測モデルと定式化することで，制御および統計科学において長年研究されてきた状態空間モデルの一般化版として定式化できる．すると，データ同化の範疇の諸問題は，状態空間モデルの視点から，システムモデルとしてシミュレーションモデルが用いられているもとでの，状態推定ならびにパラメータ推定の問題として見なすことができる．また，シミュレーションモデルのデータとの組織的な照らし合わせ法を統計学の観点から考察することにより，従来シミュレーション科学において副次的問題とされてきたシミュレーションモデルの評価法に統一的視点を与えられる．ただし気象・海洋分野のデータ同化においては，シミュレーションモデルが含む変数の個数が数百万次元程度に，また観測の次元も数万程度にもなるため，計算の限界に挑戦しなくてはならない．従って，計算アルゴリズムの研究とともに，最先端の計算機環境整備も欠かせない．統計数理研究所のデータ同化研究開発センター [28] の研究チームは，スーパーコンピュータ「京」上のアプリケーション開発研究プロジェクトに当初から参加するとともに，2014 年度から稼働する世界最大の共有メモリーを備えた所内のスーパーコンピュータも利用する [29].

シミュレーションなどの物理（演繹）モデルベースでの物理量状態の時間発展更新と，さまざまな観測装置からの実際の物理量の観測に基づく状態補正（帰納）の二つを適切に組み合わせる作業は，さまざまな研究領域において必須である．つまり，データ同化手法はあらゆる分野への応用が可能である．我々の研究チームのデータ同化の多様な応用事例については YouTube に活動紹介ビデオ（10 分）をアップしてあるので，ぜひご覧いただきたい [30].

4.3 品質管理と計測・観測デザインの高度化

前述したように，データ同化の主たる目的の一つは，同化によって得られた諸変数を初期値として用いることにより，高精度の予測を行うことである．これ以外にもいろいろな使い方が考えられるが，例えば観測点数が固定の場合，どの地点でのどのような未観測量をもし観測できたならば，全体としての予測性能が向上できるか，つまりセンサー配置の最適化など，センシングの高度化法の提案も可能である [31]．つまり，効率良い合理的な計測デザインへの道が開けてくる．日本のものづくりを支えてきた統計的品質においては，従前，帰納法によるアプローチが主であったが，シミュレーションの結果を積極的に利用する試みも活発であり，それらの動きは，UQ (Uncertainty Quantification) と呼ぶ，システムに内在する不確実性を体系的に取り扱う研究分野の形成として束ねられつつある．アメリカの巨大学会である SIAM と ASA が共同で 2013 年に新しい学術雑誌を出版開始したこともその証左であろう [32].

産業界の製品開発においては，スーパーコンピュータを利用した設計部門と，試作品を用いてさまざまな実証実験を行なう部門が分離したまま，相互の情報交換が同時的でない場合もよくある．設計部門の要求にそのまま答える形で多数の試作品を作るには金型等の大きなコストがかかり，また一方，先進的なシミュレーションアプリを動かすにはスーパーコンピュータの長時間利用がかかせない．この両者の折衷案として流体科学分野においては，室内実験と数値シミュレーションを一体化する計測融合シミュレーションの研究が 1990 年代からすすめられてきた [33]．計測融合シミュレーションの基本コンセプトは，制御工学におけるオブザーバの概念を流体の数値シミュレーションに応用するものであり，その文

脈からも明らかなように，データ同化の特殊形として定義できる．従って，データ同化は，計測融合シミュレーションが取り扱ってきた流れに関する問題，具体的には，呼吸や血流など生命現象から自動車や航空機開発のための風洞実験，気象など地球規模の現象等，幅広い分野の流体現象に適用および新展開が可能である．最近のセンシング技術の進化は，生体分子の挙動を分子レベルで観察・測定し，操作することを可能にしている．多くの場合，計測データは多様な計測ノイズを含む動画データであるため，シミュレーション結果を動画から導出した流速ベクトルに同化させることで直接計測できない量を推定する試みも始まっている [34].

5. ビッグデータと機械学習

5.1 産業のサービス化と統計数理

ビッグデータに関する記事は，経済系の新聞では毎日のように，また一般紙においても科学技術の欄だけでなく社会面でも頻繁に目にするようになった．2012 年 3 月末にホワイトハウスから出された「ビッグデータ・イニシアティブ」声明もあり，現在，産学でビッグデータの研究開発プロジェクトが進行中である．ビッグデータの利活用にかかわる方法と技術として次の 3 つが大切とされている：データ解析法，データ可視化，データ工学 [35]．ビッグデータのサイズはペタバイト級も普通なため，データは物理的に分散して格納し，アプリケーションが数千ノード上で動く必要がある．従って，データへのアクセス効率性を考えた上での格納法を含めた，業界標準的なデータ工学技術開発が重要であることは論を待たない．MapReduce や Hadoop という言葉を耳にされたことがあるかも知れないが，それらは巨大なデータの管理と取り扱いに関する計算技術である [36, 37].

データ解析法は，統計数理，機械学習，データマイニング，最適化など，広義の意味での「データ科学」（データを直接・間接的に扱う手法を研究する領域）の手法群を指す．機械学習の原点は前世紀半ばまで遡るが，統計的機械学習とやや範疇を狭めれば，その活発な研究がスタートしたのは 2000 年代に入ってからと言え，Amazon や Google 等の情報サービス産業の興隆と歩調を一にする．ビッグデータとなるとややもすれば，圧倒的に増えたサンプル数によって 3.3 で述べた新 NP 問題が解決されるのではないかと期待されがちである．しかしながら我々は，属性変数ベクトルの高次元がもたらす“次元ののろい” からは決して逃れることはできない．従って，3 及び 4 章で述べた，事前情報を活用したインピュテーション法の研究はもちろん，クリギング等の時空間解析法やデータリンケージ法の研究も今後存在感をさらに増す [38].

サンプル数が増えると，発見の意味で偽陽性を手にする可能性が本質的に高まる点に細心の注意がより求められる．例えば，膨大なデータから共起する事象を列挙し，その出現確率が高いだけでは因果を示したことにはならない．ビッグデータ時代にこそ因果推論の研究が大切になる [37, 39, 40]．高精度センサーの廉価化及び無線インターネット網の高速化という ICT インフラ網の充実と，橋梁やトンネルといった社会インフラの維持管理及び犯罪対策などの社会的要請により，今後，センサー同士がインターネットを介して自動的に会話を行なう，M2M (Machine to Machine) 通信が爆発的に増えることは疑いようがない．従って，膨大なセンサーデータをクラウドに丸ごと転送せず，ある程度その場で処理するオンライン計算（ストリーミング計算と呼ばれる）が今後主役になる．さまざまな産業の現場にセンサーを配置し，ストリーミング計算技術で加工された情報をクラウドへ転送することにより，リアルタイムのセンサス情報が提供できる．農業，漁業等の一次産業への応用は，これまで ICT の本格的導入が遅れていた分，積極的に取り組む価値がある．このようにストリーミング計算技術は，我々が生活している社会システム全体を知能化するために必須の技術である [24].

5.2 スパースモデリングと最適化

前述したように，ビッグデータでも，膨大な数の属性変数が作る超高次元空間内に高々有限個のサンプルが埋め込まれている状況である．このような “スカスカ”（疎な）空間にも，多くの場合構造が存在する．例えば，マーケティングでも明らかなように，人間の思考や行動は多様性はありつつも，特定の状況だと同様のパターンを示すことがよくある．これがデータ空間に構造を生じる．自然現象においても，同様のことは普通に生じる．ある地点の降水量に寄与するのは主として空間的に近傍の物理量であるのは明らかであるが，どこまでとどれかは状況にも大きく依存するため，簡単には把握できない．つまり，超高次元空間内の構造を目で見ることは不可能であるため，膨大な属性変数をいくつかの群にまとめつつ，構造を掴む上では不要そうな変数を大幅に削減する方法が必須である．つまり次元削減を行ないつつ，予測と現象理解の上で有効な超平面を探す作業が本質的になる．この作業自体は，そもそも統計学における多変量解析のお家芸とも言えるが，情報量規準等を用いた変数選択のアプローチ（離散最適化）では組み合わせ爆発を誘引し，ビッグデータに対しては現実的でない．ビッグデータからの効果的かつ効率的な変数選択法の研究課題により，2.3 で述べたモデル選択の問題が新しい形で再登場したのである．

この課題の解決には，概念的には古い，汎化能力を高める一つの方策である制約（正則）化の枠組みが改めて有効であった．正則化は，3.2 で述べたベイジアンモデ

リングの一つの表現形である．通常は，パラメータに対して $L2$（二乗誤差の最小化）制約を加えることで正則化を実現していたが，一方，$L1$ 正則化（絶対値誤差の最小化）の採用により目的関数の解析的な性質を病的にするかわりに，得られるパラメータの最適解に癖が出ることを意図するアプローチも以前からあった．ここでの癖というのは，値が小さいパラメータ値は “大胆にも” ゼロに自動的にセットされてしまう性質を指す．これにより，属性変数に関するパラメータの値がゼロとなれば，それは事実上その属性変数は不要であることを意味し，変数選択が最適化により実現できるのである．このアプローチは，病的な最適化関数に対して高速に最適解を求めるアルゴリズムが提案されたことを端緒として，1990 年半ば以降この 20 年間，理論的にもまた応用の広がりの意味でも精力的に研究されてきたテーマと断言できる [41]．スパースモデリング研究の隆盛を後押しした別要因として，最適化法の計算高速化と汎用的ソフトウェアの実用化がある．次に述べるサポートベクターマシンなどの非線形識別器の実現においても，最適化手法の高度化は大きな役割を果たした [42]．

信号処理の分野で話題の圧縮センシング（CS: Compressed Sensing）は，スパース性（零成分が多いという性質）を持つ高次元の信号を少ない観測から復元する枠組みであるが，その定義からして，$L1$ 正則化や変数選択の研究と数理的素材がほとんど共通になることは明らかであろう．

5.3 識別モデルとカーネル法

統計数理では，3.1 で述べたように，真の構造の忠実な再現を目指すのではなく，予測性能の向上を第一義的な “目的” として永続的に統計モデルの改良を行なう．未来のデータの予測ということは，統計モデルとしてはデータの発生機構を模擬するモデル，つまり（データ）生成モデルを取り扱うことを通常意味する．データ生成モデルを利用して，そのデータを二群に識別したい場合は，データを判別するモデルやルールがさらに必要になる．ここで，もし判別そのものが目的ならば，データ生成機構の推理を通した間接的な目的の実現でなく，“目的” に即した統計モデルを直接取り扱うことも有用である．この統計モデルのことを識別モデルと呼ぶ．古くはロジスティック回帰，近年だとサポートベクターマシンがそれに相当する．サポートベクターマシンの登場は，それまでの生成モデルを経由した分類アルゴリズムの研究に一石を投じるとともに，線形データ解析からの本格的な離陸を研究面においては加速した [43]．

サポートベクターマシンは，通常，属性変数をさらなる高次元の別空間（無限次元でもよい）に非線形写像（特徴写像と呼ばれる）したとしても，必要な計算の複雑度を劇的に押さえつつ分類器の構成が実現できるという，カーネルトリックに立脚する [44, 45]．パターン認識手法の多くはデータどうしの内積計算を含むが，一方，カーネルトリックはそれをデータ相互の類似度関数（カーネル関数）に置き換えても，その後の計算を保証する数理的な枠組みである．つまりデータ空間内における非線形識別面の構成が，複雑なデータの識別を実現する肝なのである．このような非線形識別器の成功を受けて，データ解析手法の多くが，データどうしの内積計算をカーネル関数で置き換える等の作業により，カーネル○○ 法として非線形化され適用範囲を拡大中である．

6. おわりに：人材育成

統計学，機械学習，データマイニング，最適化，OR，情報処理など，広義の「データ科学」を卒業・修了した学生数は減少の一途である [35]．統計学に目を向けると，日本の統計学に関する人材育成策は，統計学の学際性に鑑み，大学・大学院に統計学科あるいは統計学専攻を設けず，その各応用分野での具体的課題に取り組ませる中で統計に関する専門的人材を養成する，分野点在型方式である．統計数理研究所が統計科学専攻を担う基盤機関として 1988 年に総合研究大学院大学へ参画して初めて，日本に統計学を専門とする専攻ができた [46]．これ以外，未だに国内の大学には統計学を専門とする学科・専攻は存在しない．主要な大学には必ず統計学科があるという欧米諸国やアジア先進諸国の状況からすると，日本の現状は奇異と言わざるを得ない．

今後，多様な価値観が入り交じり，かつ地球規模のグローバルな競争が一層はげしさを増す中では，データ科学のような分野横断的な（横串型）学問を専門とし，複数の応用分野を副専門とする教育（T 型あるいは Π 型育成）組織が効果的である．この理念を実現すべく，統計数理研究所では 2012 年 1 月に統計思考院を設置し，ポスドク等の若手人材育成に向けたさまざまなプログラムを実施している [47]．あわせて現在，文部科学省から委託された『数学協働プログラム』[48] および『データサイエンティスト育成ネットワーク形成』[49] の二つの事業でもって，日本の置かれた状況の改善に少しでも貢献すべく奮闘中である．

参考文献

[1] 北川源四郎，樋口知之: 予測とモデル，数理科学，Vol.36, No.9，pp. 11-18，1998.

[2] 日本の数学 100 年史，第 5 章　昭和前期，pp. 42-43，1984.

[3] 文部省: 「学制百年史」，第 1 編，第 5 章，第 2 節「大学・研究機関等の設置と拡充」，1972.

[4] 伊藤清: 統計数理研究所講究録，第 1 巻，第 13 号，pp. 361-383，1945.

[5] 森本栄一: 数量化理論の形成過程に関する研究，紀要『技術文化論叢』，第 1 号，pp. 51-54，1998.

[6] 林知己夫: 統数研の創世記，統計数理研究所 50 年のあゆみ，p. 13，1994.
[7] 赤池弘次編: 科学の中の統計学　現代科学と統計数理の接点，講談社ブルーバックス，B-692，1987.
[8] 北川敏男: 統計科学の三十年―わが師わが友―，共立出版，1969.
[9] 北川源四郎: 統計的モデリングによる情報抽出，ゆらぎの科学と技術，pp. 23-44，東北大学出版会，2004.
[10] 赤池弘次: 真理への近さを測る，ゆらぎの科学と技術，pp. 13-22，東北大学出版会，2004.
[11] 赤池弘次: 時系列解析の心構え，時系列解析の実際 (II)，pp. 197-203，朝倉書店，1995.
[12] 小西貞則，北川源四郎: 情報量規準，朝倉出版，2004.
[13] 北川源四郎: モデル進化の契機としてのヴァリデーション，モデルヴァリデーション，pp. 181-205，共立出版，2005.
[14] 北川源四郎，樋口知之: 知識発見と自己組織型の統計モデル，bit 別冊「発見科学とデータマイニング」，3 月，pp. 159-168，2000.
[15] 樋口知之編: 「特集 予測と発見」，統計数理，Vol.54，No.2，2006.
[16] 樋口知之: 機能と帰納: 情報化時代にめざす科学的推論の形，情報・システム研究機構　新領域融合研究センター平成 17 年度実績報告書，2006.
[17] 第 18 期日本学術会議・新しい学術体系委員会，新しい学術体系 – 社会のための学術と文理の融合 –，平成 15 年 6 月 24 日，2003，
http://www.scj.go.jp/ja/info/kohyo/18youshi/1829.html
[18] 北川源四郎: 情報量規準 AIC からベイズモデリングへ―赤池弘次氏がたどった道，統計数理は隠された未来をあらわにする―ベイジアンモデリングによる実世界イノベーション，東京電機大学出版局，pp. 119-130，2007.
[19] 伊庭幸人他: 計算統計 2 マルコフ連鎖モンテカルロ法とその周辺（統計科学のフロンティア 12），岩波書店，2005.
[20] 山本 拓（監修）: 数理・計算の統計科学（21 世紀の統計科学），東京大学出版会，2008.
[21] 樋口知之: 予測にいかす統計モデリングの基本―ベイズ統計入門から応用まで，講談社，2011.
[22] 北川源四郎: 個の時代の統計学，月刊誌「統計」，9 月号，pp. 18-19，2008.
[23] 樋口知之: データ同化によるシミュレーション: 計算と大規模データ解析の融合，サービス工学の技術，東京電機大学出版局，2012.
[24] 樋口知之: ビッグデータと個人化技術，月刊誌「統計」，9 月号，pp. 2-9，2012.
[25] 佐藤忠彦，樋口知之: ビッグデータ時代のマーケティング，講談社，2013.
[26] 中村和幸，上野玄太，樋口知之: データ同化: その概念と計算アルゴリズム，統計数理，Vol.53，No.2，pp. 211-229，2005.
[27] 樋口知之編著，上野玄太，中野慎也，中村和幸，吉田亮: シリーズ<予測と発見の科学>6 データ同化入門 – 次世代のシミュレーション技術 –，朝倉書店，2011.
[28] データ同化研究開発センター，
http://daweb.ism.ac.jp/contents/
[29] 統計数理研究所のスーパーコンピュータ，
http://www.ism.ac.jp/ura/ism-nickname/ssda_list.html
[30] データ同化研究開発センターの活動紹介ビデオ，
http://www.youtube.com/watch?v=UE7899O9Ouo
[31] 樋口知之，中村和幸: データ同化によるオンラインセンシングの高度化，計測自動制御学会誌，Vol.51，No.9，2012.
[32] Journal on Uncertainty Quantification のサイト，
http://www.siam.org/journals/juq.php
[33] T. Hayase and S. Hayashi: “State estimator of flow as an integrated computational method with feedback of online experimental measurement,” J. Fluids Eng. Trans. ASME, Vol.119, pp. 814-822, 1997.
[34] H. Koyama, T. Umeda, K. Nakamura, T. Higuchi, and A. Kimura: “A high-resolution shape fitting and simulation demonstrated equatorial cell surface softening during cytokinesis and its promotive role in cytokinesis,” PLoS ONE, Vol.7, Issue 2, e31607, 2012.
[35] ビッグデータに関するマッキンゼー・グローバル研究所のレポート，
http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation
[36] 樋口知之: データ解析の真髄とは，ハーバード・ビジネス・レビュー，2 月号，pp. 98-108，2013.
[37] 樋口知之: データ・サイエンティストがビッグデータで私たちの未来を創る，情報管理，Vol.56，No.1，pp. 2-11，2013.
[38] N. Cressie and C. K. Wikle: “Statistics for Spatio-Temporal Data (Wiley Series in Probability and Statistics),” John Wiley and Sons, 2011.
[39] 宮川雅巳: 統計的因果推論―回帰分析の新しい枠組み（シリーズ・予測と発見の科学），朝倉書店，2004.
[40] J. Pearl: “Causality,” Cambridge University Press, 2013.
[41] 科研費新学術領域研究「スパースモデリング」，
http://sparse-modeling.jp/
[42] T. Hastie, R. Tibshirani, and J. Friedman: “The Elements of Statistical Learning: Data Mining, Inference, and Prediction,” Second Edition (Springer Series in Statistics), Springer, 2009.
[43] C. M. Bishop: “Pattern Recognition and Machine Learning (Information Science and Statistics),” Springer, 2006.
[44] 赤穂昭太郎: カーネル多変量解析―非線形データ解析の新しい展開（シリーズ確率と情報の科学），岩波書店，2008.
[45] 福水健次: カーネル法入門―正定値カーネルによるデータ解析（シリーズ 多変量データの統計科学），朝倉書店，2010.
[46] 総合研究大学院大学統計科学専攻，
http://www.ism.ac.jp/senkou/
[47] 統計数理研究所統計思考院，
http://www.ism.ac.jp/shikoin/index.html
[48] 文部科学省: 「数学協働プログラム」，
http://coop-math.ism.ac.jp/
[49] 丸山宏，樋口知之，竹村彰通: データサイエンティスト育成ネットワークの形成」事業の概要，第 5 回横幹連合コンファレンス講演要旨，2013.

樋口 知之

東京大学大学院理学系研究科博士課程了，理学博士．2011 年より情報・システム研究機構理事および統計数理研究所長．総合研究大学院大学統計科学専攻教授．専門はベイジアンモデリング，特にデータ同化の研究に従事．日本学術会議・情報学分野の連携会員．