大量複雑データ解析の先端研究における数学への期待

鷲尾　隆
大阪大学 産業科学研究所

§１　大量複雑データ解析の背景

本稿では，筆者を含む大量複雑データに関する解析手法の研究者が，その研究発展において数学の研究成果に大きな期待を寄せていることを述べさせていただきます．

私たちの日常には，パソコン，携帯電話などの個人情報機器が溢れています．産業界に視野を広げるならば，業務用パソコン，各種サーバーから工作ロボット，大型計算機，スーパーコンピュータ，大規模な計算機集合体であるクラウドなどに至るまで，さらに多種多様で大仕掛けの情報機器が使われています．しかも，それらの多くがキーボードやディスプレイなどの人間との情報入出力装置に加えて，ＧＰＳのように自らの位置を測定したり，消費電力，気温，湿度の測定や，音声や画像の取得を行う色々なセンサを持っています．また，多数の遺伝子発現やたんぱく質の読み取り装置，気象や地震の観測システム，最新の大型望遠鏡など，科学技術分野においても新たに多様な測定装置が開発されています．そして，これらから取得されたデータを記憶保持するための大容量メモリーやハードディスクが備えられていることが殆どです．そればかりではなく，それらの機器は家庭内や大学内，社内のＬＡＮで結ばれ，さらにはインターネットにより広く世界中と結ばれています．こうして我々が住む社会では，広範に存在する様々な情報機器の膨大な取得情報やその記憶情報に，ネットワークを通じてアクセスすることができます．

これらの情報には，個別の単純なデータばかりでなく，音声などの時系列データ，画像などの２次元以上の構造を持ったデータが含まれます．さらに，多数の機器をまとめて扱えば，機器の位置関係やネットワーク上の接続関係，気温の分布など，相互の何等かの関係までを含めた大規模な構造データが取得されます．このように社会に溢れる情報は，単に量的に膨大なだけでなく，内容の多様性や構造の複雑性など質の面においても顕著なデータを多く含んでいます．ここでは，これらを大量複雑データと呼ぶことにします．

大量複雑データには，もちろん各機器の位置や周囲の気温，写真といった個別標本の貴重な情報が含まれています．しかしそればかりではなく，広範囲な地域の気温分布や機器を携帯する人々の混雑具合など，複数センサ情報で捉えられる事象の情報も含まれています．さらには，各標本や事象をそれらの特徴に関するデータから幾つかの種類に分類したり，多数の標本や事象の時間的・空間的・因果的関係などを推定するに足る情報が含まれている場合も多くあります．大量複雑データがより手軽に収集可能な社会になりつつある中で，このようなより高度な事象把握やその関係分析を行う社会的，学術的ニーズが増大

しています．例えば，インターネットの各ホームページに書かれた文章中の特徴的な単語の組み合わせによって，数百億ものページ各々の内容が特定の話題にどれだけ近いかを数値で自動評価したり，それらを幾つかの話題ジャンルに自動分類したいというニーズがあります．これには個々の標本の特徴から，標本の評価値を推定したり標本を分類する解析が必要です．また生化学の分野では，各々ある１時点の１細胞の遺伝子発現状態を表す多数の DNA マイクロアレイ標本から，細胞内において遺伝子の発現によって生成されるたんぱく質が次に引き起こす遺伝子発現の半順序を表す遺伝子発現ネットワークを推定するニーズがあります．これには個別の標本データを総合して事象間の因果関係を捉える解析が必要です．

計算機科学分野の多くの研究者が，このような大量複雑データの解析原理・手法・技術の先端研究に取り組んでいます．人手を主体として大量または複雑なデータの高度な解析を行うことは，人間の能力や労力，時間の制約上，到底困難であるため，計算機によって自動的ないし半自動的に実行可能な解析手法やその効率的アルゴリズムを探求することが主な研究内容です．このような研究を行う上では，まず解こうとする問題の数学的性質を明らかにすることが重要です．さらに，その内容に応じて解析目的のために必要な規範を導入し，解法や効率的なアルゴリズムを問題と規範の数学的性質を踏まえて作り上げていきます．本稿では，これら先端研究の内容が大きく数学の研究成果に負っていることを紹介し，今後の研究発展のためにさらに数学に大きな期待を寄せていることを述べます．

§２　大量複雑データ解析手法の概要

大量複雑データの解析は，それが用いる原理や機能に拠って統計的手法や機械学習，データマイニングなどと呼ばれます．これはそれぞれの手法が発展してきた数理統計や人工知能，データベースなどの研究分野によって呼び名が違っているためですが，何れもが前節で説明した大量複雑データを解析対象としており，最近では明確な区別は無くなって来ています．大量複雑データ解析には，細かく分ければ数千種類を超える実に多様な原理や機能，目的を持つ手法が含まれ，それらを解説する幾つものハンドブックが発行されています[1]. 最も代表的な解析のカテゴリーとしては, 回帰分析や分類学習が挙げられます[2]. 言うまでもありませんが回帰分析は，各標本 i が標本空間 X 上の点$\mathbf{x}_i$で与えられ，かつそれに実数の目的値y_iが付与されたデータ $D=\{(\mathbf{x}_i, \mathrm{y}_i) \mid i=1,\ldots,n\}$を基に，$\mathbf{x}_i$から$\mathrm{y}_i$の予測値$\hat{\mathrm{y}}_i$を出力する回帰式と呼ばれる実関数$\hat{\mathrm{y}}_i = F_X(\mathbf{x}_i)$を推定する解析です. 例えば先ほどの文章集合の一部の文章について, 各々の内容を代表する幾つかの単語の組み合わせ$\mathbf{x}_i$と我々が評価したある話題への各々の近さの実数値表現y_iが与えられれば, 新たに与えられた文書のその話題

への近さを我々にほぼ代わって予測する回帰式を回帰分析によって得ることができます．分類学習もこれと似ていて，各標本 i を表す$\mathbf{x}_i$にクラスと呼ばれるカテゴリカルなラベル$\mathbf{y}_i$が付与されたデータ D=$\{(\mathbf{x}_i, \mathrm{y}_i) \mid \mathrm{i}=1,\ldots,\mathrm{n}\}$を基に，$\mathbf{x}_i$から$\mathbf{y}_i$の予測値$\hat{\mathbf{y}}_i$を出力する分類器と呼ばれるカテゴリカル関数$\hat{\mathbf{y}}_i = F_X(\mathbf{x}_i)$を推定する解析です．例えば同様に一部の文章について，各々の内容を代表する幾つかの単語の組み合わせ$\mathbf{x}_i$とそれが既定の数種類の話題のどれに属するかを我々が評価したラベル$\mathbf{y}_i$を与えれば，新たに与えられた文書の話題を我々にほぼ代わって予測する分類器を分類学習によって得ることができます．

回帰分析や分類学習の他にもクラスタリングなど代表的なデータ解析手法が存在し，何れの研究も計算機が使えるようになった２０世紀後半から発展ました．しかしはじめに述べた背景により，大量複雑データを対象とした解析手法の研究が盛んになったのは，１９９０年代後半からです．特に近年では数学的原理を積極的に導入し，大量データを高効率，高精度に解析する手法の研究が行われています．幅広いデータ解析に適用可能なカーネル法はその最たるものと言えます[2]．例えば，回帰式$\hat{\mathbf{y}}_i = F_X(\mathbf{x}_i)$を線形なものに限定すれば，様々な高速アルゴリズムを利用して大量データの解析で可能です．しかしながら，線形回帰式では標本を説明する変数の複雑な関係の高精度学習が困難であることが多く，適用に限界があります．一方，非線形回帰式を大量データに適用しようとしても，精度の良い妥当な回帰式を得るための高速アルゴリズムが一般には存在しないというジレンマがありました．これに対して，カーネル法では標本空間 X からもっと高次元な空間 H への写像 ϕ を導入し，H 上で新たに回帰式$\hat{\mathbf{y}}_i = F_H(\phi(\mathbf{x}_i))$を学習することを考えます．この時，$F_H$が線形になるように$\phi$を選び，X 上では非線形な関係を H 上の線形関係に置き換えて，高速なアルゴリズムによって回帰式の学習を行います．特に H をヒルベルト空間に選べば，多くの解析においてすべての標本ペア$\mathbf{x}_i, \mathbf{x}_j$間の H 上の内積で定義されるカーネル関数$\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = <\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)>$のみを扱えば良く，高次元空間 H 上の陽な計算を省略できることが知られています．さらに再生性という性質を持つカーネル関数を使えば，X 上の任意の連続な非線形関数F_Xを H 上の線形関数F_Hに変換可能であることも知られています．このように数学における関数解析の研究成果を導入することにより，大量複雑データ解析のジレンマを解消した様々な解析が可能となりました．

これ以外にも筆者の研究分野において，数学の研究成果が活躍している例は多数あります．例えば，我々の研究室で取り組んでいる回帰分析や分類学習の変数選択問題は劣モジュラ関数最大化問題として定式化でき，それを解くために最新の最適化理論の知見が使われています[3,4]．また，同じく統計的なデータ生成過程推定（因果推論）問題を標本分布の非ガウス性を利用して解くために，近年の独立成分分析や上記カーネル法の研究成果が使われています[5,6]．次節では，これらについて述べます．

§３　大量複雑データ解析における数学の活躍例

・劣モジュラ関数最大化による回帰分析や分類学習の最適変数選択[3,4]

回帰分析や分類学習において，各標本 i の表現$\mathbf{x}_i$を構成する上で利用可能な d 個の説明変数からなる空間 V が与えられた時，回帰式または分類器$\hat{y}_i = F_X(\mathbf{x}_i)$の学習に用いたデータ以外のテストデータ $D_T=\{(\mathbf{x}_i, y_i) \mid i=1,\ldots,m\}$に関する対数尤度$\sum_{i=1}^{m} log\, p(y_i|\hat{y}_i)$を最大化する k 次元(k≪d)以下の標本空間（即ち説明変数集合）X_{opt} ($X_{opt}\subseteq V, |X_{opt}|\leqq k$)を選ぶ問題，即ち

$$X_{opt} = \underset{X\subseteq V,|X|\leq k}{\operatorname{argmax}} \sum_{i=1}^{m} \log p(y_i \mid \hat{y}_i)$$

を最適変数選択問題といいます．学習データとテストデータが異なるので，たとえ k=d であっても一般に必ずしもX_{opt}=V とはなりません．これはいわゆる NP-困難問題であることが知られており，d や k が大きければしらみ潰しにすべての変数組み合わせ X(⊆V)を調べることは計算上困難です．そこで，従来は最良優先探索（貪欲探索）や正則化などの手法によって，X_{opt}の近似として満足すべき X を V 中で探すことが行われてきました．

しかし，大量複雑データにおいて$\mathbf{x}_i$から y_iが決まる過程を理解するために，y_iの予測に寄与する説明変数集合X_{opt}を正確に求めることが重要な場合は多くあります．この場合には，上記の最適化問題を厳密に解かねばなりません．ある与えられた標本集合の下で$\hat{y}_i$は X の選び方によって決まるので，最適変数選択の観点から尤度$p(y_i|\hat{y}_i)$を集合 X を引数とする集合関数$p(X|y_i)$に置き換え，上記問題を

$$X_{opt} = \underset{X\subseteq V,|X|\leq k}{\operatorname{argmax}} G(X) \quad \text{s.t. } G(X) = \sum_{i=1}^{m} \log p(X \mid y_i)$$

と書き換えることができます．この時，対数尤度 G(X)は以下の劣モジュラ性と呼ばれる性質を持つことが知られています．

$$G(X_1 \cup \{x\}) - G(X_1) \geq G(X_2 \cup \{x\}) - G(X_2)$$
$$\text{ただし}, X_1, X_2, \{x\} \subset V \text{かつ} X_1 \subseteq X_2.$$

これは図１に示すように，ある集合 X_1 にある要素を加えた場合の関数値の増分を，その X_1 を含むより大きな集合 X_2 に同じ要素を加えた場合の関数の増分が超えないことを意味します．この性質は尤度に限ったものではなく，相互情報量や情報量基準など他の多くの統計的指標に見られます．また，経済における限界効用逓減効果のように，他の様々な分野でも扱われる関数の性質です．

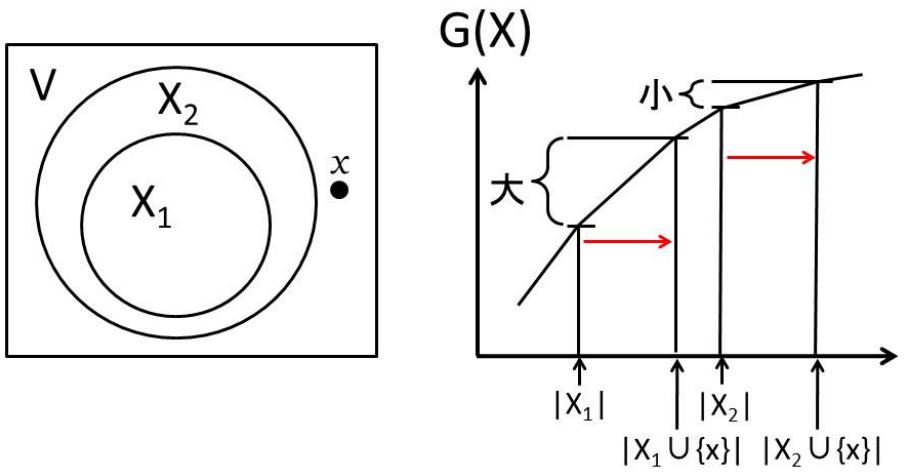

図１　劣モジュラ関数の性質

劣モジュラ関数は集合という離散構造上の関数ですが，このように連続関数で言うところの凸性に類似した性質を有します．離散数学研究において，この性質やそれを利用した関数最適化に関する様々な研究が行われてきました．しかしながら詳細は割愛しますが，多項式オーダーで解ける劣モジュラ関数最小化問題の研究に比較すると，ここで述べている NP-困難な最大化問題の実用的な厳密解法研究はあまり行われていませんでした．そこで我々の研究室では，問題によっては V が数十変数以上を含む場合であっても厳密解を高速に求めることができるアルゴリズムを研究して来ました[3,4]．このように，これまでの離散数学の研究成果を生かすことによって，大量複雑データにおける目的値の決定過程の理解に寄与する貴重な手段が得られます．劣モジュラ関数ないしはそれに関連する関数の最適化問題は，これに限らずデータ解析の多くの局面において見られ，今後，ますます深く関連してくるものと考えられます．

・非ガウス性を利用した統計的因果推論[5,6]

統計的因果推論や人工知能のベイジアンネットワーク構造学習の分野では，各標本 i が標本空間$X = \{x_1, \ldots, x_p\}$上の点$\mathbf{x}_i$で表される統計的データ $D=\{\mathbf{x}_i \mid i=1,\ldots,n\}$が与えられた時，各変数$x_j \in X$ の値が他の何れの変数の値から統計的に生成されたのか，即ち統計的データの生成過程を推定する問題が研究されてきました．特にある置換写像$k:\{1,\ldots,p\} \to \{1,\ldots,p\}$が存在して，以下のような線形かつ非巡回な過程から生成されるデータに関する研究が数多く行われて来ました．

$$
\begin{aligned}
x_{k(1)} &= e_{k(1)} \\
x_{k(2)} &= b_{k(2)k(1)} x_{k(1)} + e_{k(2)} \\
&\vdots \\
x_{k(p)} &= b_{k(p)k(1)} x_{k(1)} + \cdots + b_{k(p)k(p-1)} x_{k(p-1)} + e_{k(p)}
\end{aligned}
$$

ここで，e_jはx_jに関する非観測なガウス性外乱であり，他の外乱とは独立です．また，各bは結合係数です．しかしながら，この比較的単純なデータ生成過程でさえも，しばしばデータ D からの逆推定が困難であることが知られています．例えば，最も単純な２変数標本空間$X = \{x_1, x_2\}$の場合には，そのデータ D から以下(1),(2)の何れが正しいのかを識別することができません．

$$\left.\begin{aligned} x_1 &= e_1 \\ x_2 &= b_{21}x_1 + e_2 \end{aligned}\right\} (1) \qquad \left.\begin{aligned} x_2 &= e_2 \\ x_1 &= b_{12}x_2 + e_1 \end{aligned}\right\} (2)$$

このような識別性の限界は長年の課題とされて来ました．

これに対して我々の研究室では，非観測外乱e_jが非ガウス性である場合には，線形非巡回なデータ生成過程が一意に識別可能であることを示し，その推定アルゴリズムを研究して来ました．例えば，真のデータ生成過程が上記(1)であるデータ D について，x_2をx_1に回帰した場合には回帰残差$r_2^{(1)} = e_2$でありx_1とは無相関かつ独立です．これに対して，x_1をx_2に回帰した場合には，回帰残差

$$r_1^{(2)} = \left\{1 - \frac{b_{21}\,\mathrm{cov}(x_1, x_2)}{\mathrm{var}(x_2)}\right\} x_1 - \frac{\mathrm{cov}(x_1, x_2)}{\mathrm{var}(x_2)} e_2$$

はx_2と無相関ではありますが独立ではありません．従って，外乱e_jが非ガウス性である場合には回帰変数と残差の独立性を調べれば，上記(1),(2)を識別することができます．この事実は，統計的因果推論やベイジアンネットワーク構造学習の研究において従来見過ごされて来たことですが，数学において盛んに研究されてきた独立成分分析に深く関係しています．

また，上記のように２変数 x,y 間の独立性を評価するために，しばしば以下の非線形相関係数が使われます．

$$C_{f,g}(x, y) = \frac{E_{x,y}(f(x)g(y))}{E_x(f(x)^2)E_y(g(y)^2)}$$

任意の非線形関数 f,g について $C_{f,g}(x,y)$が十分ゼロに近ければ，x と y は独立であると判断できます．しかしながら，あらゆる f,g の組み合わせについて $C_{f,g}(x,y)$を計算することは現実的ではありません．そこで，与えられた x,y のデータについて $C_{f,g}(x,y)$が最大になる f と g を前述したカーネル法を用いて推定し，その $C_{f,g}(x,y)$が十分ゼロに近いことを以って独立性を判断する手法がよく使われています．

§ 4　数学によるイノベーションと連携への期待

以上，大量複雑データに関する解析手法の先端研究分野では，数学的原理を積極的に導入した研究が行われていることを述べました．多くの場合，これは実用的計算時間かつ実用的精度を有する大量複雑データ解析手法を得るために行われており，今日のデータ解析手法の研究においてなくてはならないものです．言うまでもなくデータ解析手法の研究は，伝統的に確率・統計の研究分野と深いつながりを持ってきました．しかし，今日ではそれに限らず，関数解析や離散数学，位相幾何学など，幅広い数学分野の諸原理が使われるようになって来ています．

ご存じのように歴史を振り返れば，数学は幅広い科学・技術分野に対してそれらの基礎を提供して来ました．データ解析に限らず，より大規模化，複雑化する現代の社会ニーズに答えるためには，多くの分野でますます数学の高度な基礎研究成果が必要とされて行くでしょう．データ解析手法を研究する立場から数学者の方々へ何より期待する点は，一層幅広く深い基礎研究の推進です．我々が扱うデータの種類や規模が拡大するに連れて，データ解析の研究分野でこれまで扱われたことのない新たな数学の基礎研究の成果が導入されて来ています．数学の基礎研究成果に広がりと厚みがあるほど，我々データ解析の研究者は，より柔軟に深くそれらの成果を利用して研究を発展させて行くことができるのです．

ただし異分野の研究者間には，必ずと言っていいほど研究成果のトランスファーを妨げる知識や問題意識のギャップが存在します．そのギャップを埋めるためには，我々データ解析手法研究者の一層の数学に関する勉強が必要であるのはもちろんですが，数学者の方々にも研究活動において一部応用に向けた認識を持っていただけると助かります．また，このように双方の理解を深めるためには，数学者とデータ解析手法の研究者の間の出会いの場が必要となります．ここに述べた多くの研究内容も，何らかの意味でそのような出会いをきっかけにして生まれてきたものです．日本では大量複雑データ解析手法に関する研究が，日本統計学会をはじめとした統計関係の学会はもちろんのこと，電子情報通信学会や人工知能学会，情報処理学会など多くの数理系，工学系の学会にまたがって展開されています．このような研究コミュニティーには，計算機科学や統計，物理に関する研究者が多く含まれていますが，まだ数学者はあまり多くありません．今後，数学者との一層の連携・協力の場が生まれることを期待して筆を置きます．

参考文献

[1] S. Ranka (ed.), Handbook of Data Mining, Computer & Information Science Series, Taylor & Francis Group Ltd (2007).

[2] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and

Statistics) , Springer-Verlag (2006).

[3] Y. Kawahara, K. Nagano, K. Tsuda and J.A. Bilmes, Submodularity cuts and applications, Advances in Neural Information Processing Systems 22, pp.916-924, Cambridge, MIT Presss (2010).

[4] Y. Kawahara and T. Washio, Prismatic Algorithm for Discrete D.C. Programming Problem, Advances in Neural Information Processing Systems 24 (2011).

[5] S. Shimizu, P. O. Hoyer, A. Hyvarinen and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7, pp.2003-2030 (2006).

[6] S. Shimizu, T. Inazum, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P.O. Hoyer and K. Bollen, DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model, Journal of Machine Learning Research, 12, pp.1225-1248 (2011).