The Institute of Statistical Mathematics, 2006 Open House

統計数理の世界

―研究紹介とエッセイ―

統計科学
―未来予測と知識発見の文法―

The Institute of Statistical Mathematics

大学共同利用機関法人
情報・システム研究機構
統計数理研究所

「 研究紹介とエッセイ集 」目次

ご挨拶

モデリング研究系

データ科学研究系

数理・推論研究系

予測発見戦略研究センター

リスク解析戦略研究センター

プロジェクト研究員

ご挨拶

日ごろから統計数理研究所の活動に関して、ご理解ご支援を賜り誠に有難うございます。当研究所は統計数理に関する我が国唯一の大学共同利用機関として、統計数理の研究拠点、共同利用そして日本で唯一の統計科学専攻の基盤機関としての機能を果たしてまいりました。

１９４４年の設立以来、当研究所は現実の問題に根ざした研究をモットーに、様々な分野における学術研究や実社会における困難な問題に挑戦し、その問題解決の過程から時代に即した新しい知的情報処理の方法を開発することを目指してまいりました。その結果、情報量規準 AIC、ベイズモデリング、数量化理論などの独創的で実用的な方法を生み出すとともに、社会調査、地球科学、生命科学、工学などの様々な領域において、独自の特徴ある成果を挙げてきました。このような成果の一つの表れとして、我々の先輩である赤池元所長が本年度の京都賞の受賞者に選ばれ、その基礎科学における偉大な貢献が世界的に再確認されると同時に研究所の基本的姿勢の正しさが示されたことは、所員一同にとっても喜びに耐えません。

当研究所では、自らの研究推進と並行して総合研究大学院大学の基盤機関のひとつとして、次世代を担う人材養成のために博士後期課程の大学院教育にも取り組んでまいりました。既にこれまでに６０人以上の博士取得者を輩出し、多くは国内外の大学、研究所、企業の第一線で活躍しています。さらに本年度からは博士前期からの 5 年一貫制のコースを加え、より充実した体制にいたしました。統計数理は、常に最先端の科学研究に取り組みながら、科学研究の最も基礎的部分を開拓していこうとする稀有で魅力的な学問分野です。統計数理研究所はこの両面に挑戦する覚悟を持った意欲的な若者の参入を歓迎いたします。

今回のオープンハウスは大学院を目指す大学生、社会人の方々に統計数理研究の魅力を体感していただくとともに、地元近隣の皆様にも統計数理研究所を身近な存在としてご理解いただきたいという思いで開催いたしました。ぜひ研究室にお立ち寄りいただき、日本人の国民性調査、生物の系統樹の推定、自然現象や経済の予測、話者認識、高速無線データ通信、じゃんけんソフト、モンテカルロ手法や乱数による計算などの興味ある具体的成果をご覧いただければ幸いです。

統計数理研究所は、情報化社会・リスク社会に即した新しい科学的研究の基盤構築を目指して挑戦を始めようとしているところです。日ごろよりのご支援を改めて感謝申し上げるとともに、今後ますますのご支援ご理解をよろしくお願い申し上げます。

２００６年７月

統計数理研究所長
北川　源四郎

非線形時系列解析とその応用

尾崎　統

A)今年度もCausal Modelingの視点から上記研究課題で以下のような研究活動をした。

A-1) 方法論

今年度は特に以下の2点に関して大きな進展があった:

1) 赤池ノイズ寄与率解析の状態空間化:

ノイズ間の同時相関が強い多変量時系列のノイズ寄与率解析に有効な方法を見出した。状態空間モデルの枠組みの中に観測されない隠れた共通変数を導入し全状態変数の駆動ノイズの分散行列を対角化することで困難を解決。

2) Compartment Model のGARCH化:

金融データ解析で必須のGARCHモデリングの考えをEEG-Compartment状態空間モデルに拡張し、効果的パラメタリゼーションとモデル同定手法を導入。

A-2) 応用

1)意識(昏睡と覚醒)のHeteroschedastic時系列モデリング:

Roy John教授(ニューヨーク大学医学部)提唱の「意識の理論」をダイナミックな時系列モデリングの立場から裏づける試みの共同研究に取り組んだ。

2)ミクロとマクロの時系列融合モデル研究:

東北大学未来科学技術共同研究センター、川島研Jorge Riera博士らと協力して、脳神経活動のミクロ(EEG)とマクロ(fMRI)を融合する神経血流動力学モデル(確率微分方程式)とその同定法を導入.

3)JST研究プロジェクト「脳科学と教育」:

日本科学技術振興機構 (JST)プロジェクト研究「脳科学と教育」(研究総括小泉英明博士)に参加、生理学研究所定藤研、京都大学心理学科板倉研に協力して幼児の視線時系列データの非定常非線形モデリングに取り組む。

4)N-TIMSAC非線形システム同定法の普及活動

企業に於けるN-TIMSACの導入を容易にするため、ベンチャー企業「21世紀技術」(石井代表)と共同で比較的小規模のN-TIMSACシミュレーターの製品開発研究(新居浜高専、豊田幸裕教授、上智大学、V.Ozaki教授らと共同)。

5)年金資産の最適制御問題

年金資産を最適運用するためのHeteroschedasticityのもとでの最適制御問題を研究。Time-deformed process, Micro-market Structure Model、Jump Diffusion Modelなどの有効性を実データで検証。

6)その他:

i) 工業プロセスの排煙中の脱NOx制御、脱CO_2制御、ボイラー温度制御等に関連して、太平洋セメント技術本部、住友共同電力、日本ベーレーに技術指導。

ii) 生理学研究所金桶博士とMEGデータによる感覚認知機能解析で共同研究。

iii) 理研 Nikolaev 博士と EEG データの Evoked Potential 推定問題で共同研究。

iv) 理研中原博士と神経パルス時系列データの Causality 解析で共同研究。

v) 金融時系列のボラティリティ推定問題に関連してみずほ信託銀行資産運用本部に技術指導。

vi) 統数研、公開講座「金融データの非線形時系列モデリング入門」連続講義。

vii) 一橋大学大学院、国際企業戦略研究科で非線形時系列解析連続講義。

viii) 今年度から電子メディアと個人的ネットワークを通じて尾崎研究室を外、世界に開放、時系列データ解析問題を抱える**あらゆる科学分野**の研究者と、**新しい時系列解析の方法**を模索する研究者がデータを前に共に考え議論する**「場」**を提供する試みを開始。

B ） 今年度の国際誌掲載論文 （国際会議プロシーディング論文は省）:

[1] Wong, K. F., Galka, A., Yamashita, O. and Ozaki, T.(2006) "Modelling non-stationary variance in EEG time series by state space GARCH model", *Computers in Biology and Medicine*, in press.

[2] Jimenez, J.C. and Ozaki, T. (2005) "An approximate innovation method for the estimation of diffusion processes from discrete data". *J. Time Series Analysis*, Vol 27, 1, 77-97.

[3] Riera,J., Aubert,E., Iwata,K., Kawashima,R., Wan,X. and Ozaki,T.(2005) "Fusing EEG and fMRI based on a bottom-up model: inferring activation and effective connectivity in neural masses", *Phil. Trans. of Royal Society, Biological Sciences, Vol. 360*, No.1457, 1025-1041.

[4] Jimenez, J.C., Biscay, R. and Ozaki, T.(2005) "Inference methods for discretely observed continuous-time stochastic volatility models: A commented overview", *JAFEE Journal*, to appear.

[5] Peng, H., Tamura, Y., Gui, W., and Ozaki, T.,(2005) "Modeling and asset allocation for financial markets based on a stochastic volatility microstructure model". *Int. J. of Systems Science*, 36, No.6, 315-327.

[6] Yamashita, O., Sadato, N., Okada, T. and Ozaki, T., (2005) "Evaluating frequency-wise directed connectivity of BOLD signals applying relative power contribution with the linear multivariate time series models", *Neuroimage, Vol.25*, 478-490.

球面上に一様ランダムに点を配置する

モデリング研究系　　種村　正美

わたしたちが住んでいる地球はほぼ球形をしている。ところが、ふだん、わたしたちは球面上に居ることを意識せずに暮らしている。しかし、地球的規模で生態学的調査をするとか、地球上に気象観測施設などを配置するという問題になると球面をまともに取り扱う必要が出てくる。

そのような問題でしばしば必要とされるのが、球面の上に一様にランダムに点をばらまく方法である。一様とは、どの場所でも平均して同程度の個数の点があることをいう。地球上の位置は緯度と経度で表されることが多いが、緯度と経度を互いに独立にランダムにサンプルすれば目的の配置が得られるのだろうか。この小論では、一定個数の点を球面上のある有界領域の中に一様ランダムに配置する方法を考えてみる。

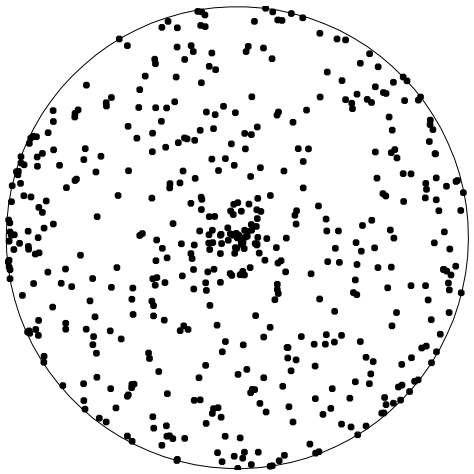
中心が座標原点にあり、半径が R の球を考える。そのとき、球面上の各点は方位角 θ と偏角 ϕ で表される。方位角は北極から測った球面角で $0 \leq \theta \leq 180^\circ$ の範囲をとり、偏角は赤道に沿った球面角（$0 \leq \phi \leq 360^\circ$）である。

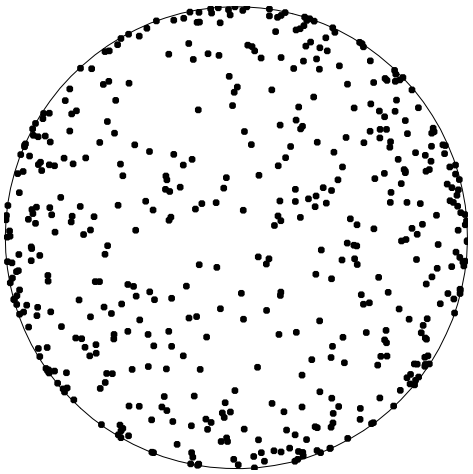
さて、いまこの球面上に、北極を中心とする球面半径 r の円（球帽という）を考える。そして、この球帽の内部に一様ランダムに点を配置したい。素朴に思いつくのは、上に述べたように方位角 θ（緯度に対応）を $0 \leq \theta \leq r$ の範囲で、偏角 ϕ（経度に対応）を $0 \leq \phi \leq 360^\circ$ の範囲で互いに独立にとることである。このやり方で、$r = 90^\circ$ として 500 個の点をランダムにサンプルした例が左の図である。これは球面の北極側からの正射影である。一見して、北極近くに点が集中している。実はこの素朴なやり方は一様という条件を満たしていないのである。

一様性をもつランダム点の生成法は次の通りである（詳細は例えば種村 [1998] 参照）。ξ, η をそれぞれ 0 と 1 の間の一様乱数とする（統計数理研究所の乱数ポータルサイトから容易に入手可能）。このとき、上記の球帽内で一様ランダムに分布する点のサンプル値 (θ_0, ϕ_0) は

$$\theta_0 = \cos^{-1}\{1 - (1 - \cos r)\xi\}, \qquad \phi_0 = 2\pi\eta$$

で与えられる。右の図はこの方法で 500 個の点を北半球でサンプルした例であり、一様性が満たされていることが見てとれる。





文献

種村　正美 (1998): 球面上の最適配置の問題．**統計数理**, **45**, 359–381.

種村　正美 (2004): 球面の上に点を均等に配置する．『**形の科学百科事典**』, 朝倉書店, pp.634-635.

フィールド生態データの科学

島谷健一郎(モデリング研究系)

野外に生えている植物に何らかの疑問を感じて科学的に解明しようと思った時、その原点のひとつにモニタリング調査があります。まず、調査する場所を決め、その中の全個体にラベルを付けて区別し、位置を測って記録します。それからそれぞれの太さ（直径）や高さ、葉や花の数を数えたり、土壌や光などの環境条件を測ったり、最近ではサンプルを取って実験室へ持ち帰り、遺伝子を分析したりします。その後も調査地を訪れて、同じ項目を測量し、死んだ個体と新しく生まれた個体を確認します。これを定期的に何年も繰り返していくと、植物の持つ様々な形質が時間と共に変化して行くサマが数値で追跡され、それをグラフ化したりすることで、植物たちの特徴や変化を立証したり、逆に予想だにしなかった現象を発見できたりします。

さらに 1 歩進めて、並んでいる数値を「データ」とみなして統計的に分析したり数理モデルを作ったりすると、フィールド生態データの科学となります。人が 1 個 1 個手で測って作ったデータですから、科学するほどの価値もないように思えますが、実際はその反対で、人が直接測っているだけに数が少なく（でも 1 人で 5000 本 10000 本測るのはごく普通です)、しっかりした統計処理を施さないと、本当に成長量に違いがあるのか、本当に開花個体の数が増えているのか、本当に光環境が生死に影響を与えているのか、等々を断定できません。中でもデータの中に個体位置という空間情報があり、これはそこでの環境やその辺りの密度など、植物同士の競合と直結します。こういったデータには、空間パターン解析とか点過程モデルとかを駆使する必要があります。

例えば対馬の照葉樹林で、5cm 以上の成木 10 数種について直径と死亡率の関係を調べると、減少型(小さいほど死にやすい)、U 字型（小さい時と大きい時に死にやすい)、一定型に分類されました。さらに光環境や成長まで加えると、どの 2 種も同じパターンは示しません。こんな多様性が、ここでの 40 近い樹種の共存を醸し出しているようですが、こんなパターンもデータの数値を見ているだけでははっきりせず、平滑化とその情報量規準による平滑度選択を経てはじめて明瞭になりました。

何本かの大木を残して伐採・収穫をし、その母樹から落下する種子で自然に森を再生させる方法がありますが、その結果はマチマチです。例えば岩手県北上市の 35 年前のブナ林伐採跡には森林が復元しかけていますが、ブナは保残母樹の下に限られています。でも、真下でなく樹冠周辺に多い林分も見られます。青森県八甲田の 25 年生ブナ林には保残母樹と関係なくブナが林立していますが、木曽ヒノキ林の 1 部を伐採した跡はヒノキでなくアスナロが更新しています。かつアスナロは保残木の間でなく、むしろ保残木の近くに多い。こんな 2 次林の姿も、最近接解析とか非定常ポアソン過程と呼ばれる手法で明瞭に示せ、さらに遺伝子データと合わせる事で、伐採後に種子が保残母樹から散布された様子から、その後、しだいに死亡して行った経過を推定できたりします。

じゃんけんソフト募集

石黒真木夫
統計数理研究所
ishiguro@ism.ac.jp

統計数理研究所は，統計科学への理解を深めていただくひとつの手段として「じゃんけんソフト」を開発し, 小中学生を対象とするイベントなどでご披露してきました. このじゃんけんソフトで遊ぶ中で, 統計科学的データ解析の働き, 有効性を感じていただくことができます.

このじゃんけんソフトは, もともと人間とじゃんけんするように作られたものですが, 別のじゃんけんソフトとの対戦が可能です. 以下のサンプルを参考に, 統計数理研究所のじゃんけんソフトと対戦するソフトを作ってみませんか. 高校パソコンクラブのメンバーなら能力に不足はないはずですし, 中学生, 小学生も挑戦できると思います.

(詳しくは

http://www.ism.ac.jp/~ishiguro/Profiss/pfs.@990827.@204933.dir/contest.g1frm.html

をご覧下さい.)

[C によるサンプルルーチン]

```
jankensub(input,output)
int *input,*output;
{
/*
      input = 人間の「前回」の手
           1 = グー   2 = チョキ  3 = パー
       この情報をサブルーチン内部に蓄積して利用するのは，もちろん,
       自由です.

       main routine では，一回のゲーム開始の時点で
        input = -1 を渡します. 必要に応じて内部で利用しているワ
       ークエリアなどを初期化するように組んでおいて下さい.

       なんらかの方法で output に整数値を設定して返して下さい.

       main routine では,
           output  < 2     グー
                   = 2     チョキ
                   > 2     パー
           と認識します. たとえば,
*/
      if(input == 1) output = 3;
      if(input == 2) output = 1;
      if(input == 3) output = 2;
/*
           とすると，相手が前回と同じ手を繰り返すと
           想定して勝ちにいく戦略になります.
*/
}
```

バベルの塔と確率の科学　モデリング研究系　伊庭幸人

２１世紀のキーワードのひとつは「確率」である．前世紀の最後の１５年ほどで，パターン認識，人工知能，進化生物学，誤り訂正符号，コンピュータグラフィクス，金融など，さまざまな分野で，確率の考え方にもとづく理解の仕方や計算手法が大きく発展し，また，あらためて見直されてきた．なかでも，マーケティング，生態学，自然言語処理などでは，確率的な枠組みに基づいて対象の個性や環境の非定常性をとらえる手法が展開され「平均値の発想」を超えるものとして注目されている．

「確率」というと「○○の確率は××パーセントである」の「××」の部分のみが注目されがちであるが，ほんとうに重要なのは「○○」の部分に何を持ってくるかである．その意味で，確率モデルとは単にものごとの不確実性を示すだけでなく，目の前の世界を切り分けて，それを記述する「ことば」に対応するものである．筆者の所属する「モデリング研究系」の「モデリング」とは概略そのことを意味している．

「確率」の含意するものはそれだけではない．統計物理では，確率モデルを用いて，多数の要素からなるシステムが条件によってまったく違う様相を示すことが研究されている.「水と氷」「交通渋滞と正常な状態」などがその例である．また，乱数発生やカオス，乱流のように「確率的な振舞いの起源」を問う研究も「確率の科学」の一側面である．

統計物理で開発された計算手法が人工知能，マーケティング，進化生物学，ビット誤り訂正，コンピュータグラフィクスなどで利用されていることでわかるように，さまざまな分野が確率という糸で結ばれて関係していることがわかってきている．今日，諸科学は対象によって分割され，お互いに言葉が通じない「バベルの塔」的な状況が出現しているが,「確率の科学」は，そこに横糸を通すものとして，さらに重要になるだろう．

統計数理研究所は，わが国には珍しく「モノ」以外のテーマを中心に据えた独立の研究所である．そこでは上で述べた意味の「確率の科学」をひとつの柱として，実世界との結びつきを重視しつつ，さまざまな数理的な方法論の探求が行われている．共同利用施設として，複数の専門分野を理解できる好奇心に溢れた人々が集う場所でもある．

現代数学は高度に精緻なものとなったが，ほかの科学や実社会とのギャップはそのぶん拡大している．ここでいう「確率の科学」に関することを，数学の「確率論」の専門家にたずねても，多くの場合，期待するような答は得難いだろう．その間を埋めるべき，既成の「応用数学」は，ともすれば実世界との関連を見失い，しばしばミニチュアの純粋数学となってしまっている．その中で，統計数理研究所は，小さな施設であるが，独自の個性を持ち，今後のわが国の数理科学の研究の核となりうると期待している．

モデル化と高効率データ処理に基づく無線データシステムの研究

モデリング研究系　　助教授 瀧澤由美
客員教授 深澤敦司

本研究は、雑音と帯域制限を有する無線チャネルにおける高速データ通信を目的としている。研究の具体的内容は、(1)方式に関する基礎的研究、(2)高速低消費電力データ処理のための回路および LSI の研究、および(3)検証のためのプロトタイプハードウエアの開発よりなる。

広帯域 CDMA 方式は北米向け 2GHz パーソナル通信サービスへの適用を目的として、筆者らによって 1992 年に提案され、1996 年には北米標準 IS-665 として受理された。この方式は部分的変更を経て NTT DoCoMo によって国際標準化された。

筆者らは次世代の公衆移動無線通信（携帯無線）に着目し、非分割広帯域変調による帯域有効利用率の向上とデータの高速化をめざして研究してきた。筆者らは次世代の携帯電話として、現在サービス可能なデータ速度（無線帯域幅 5MHz で 384kbit/s）の 10 倍以上で、かつ高品質なセルラシステムの開発を目標としている（図１）。

ディジタル装置では無線基地局からの時刻信号を基準として全ての演算が行われる。このためには受信した電波からターゲット基地局の時刻信号(クロック)を抽出する必要がある。これを同期という。広帯域 CDMA によるディジタル無線では最高速度の演算はこの部分である。一般にデータ速度が 10 倍に高速化されると消費電力も 10 倍となる。今回マッチトフィルタを用いたプロトタイプハードウエアを試作し、スペクトル拡散変調、コヒーレント復調など、主要機能とデータ速度 512kbit/s までの動作の確認を行った（図２）。現在はさらに無線の広帯域化とデータの高速化の実現をめざす。

統計数理研究所では､従来のディジタル演算ではなくアナログデータ処理による同期システムの実現の研究を行った。またこれに基づき、東京大学新領域創成科学研究科では、生体の応答特性を模擬する C MOS LSI によって演算量の軽減と極低消費電力化を実現した。試算によれば同期に要する消費電力はディジタル技術の約 1/100、受信系として 1/10 以下を達成した。

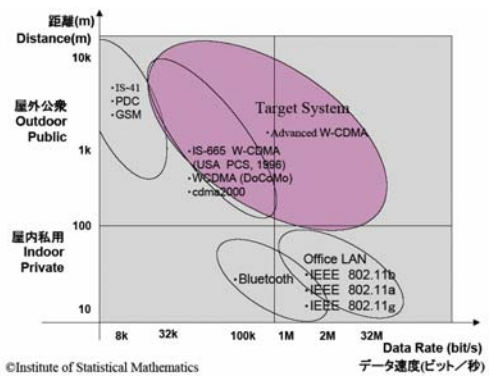


図１　本研究成果の適用領域

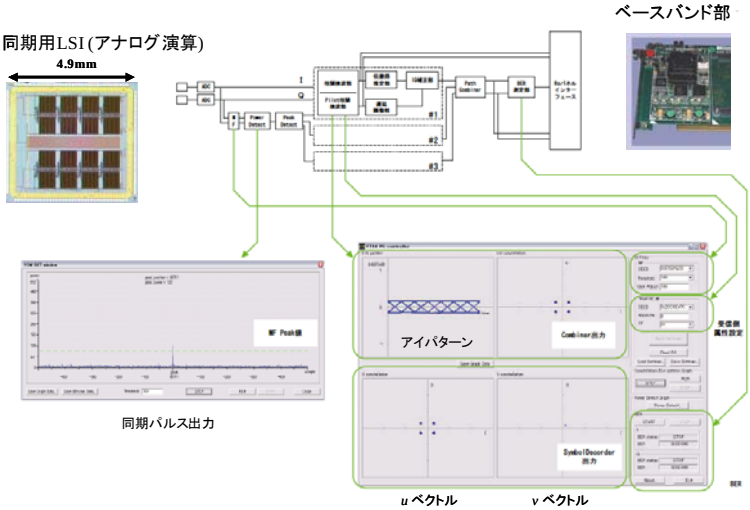


図２　検証用プロトタイプハードウエア（右上）
アナログ演算 LSI（左上）と自動評価システム（下）

「機能と帰納プロジェクト」サブプロジェクト：

マルチモーダルデータからの不変情報の発見とその方法論の研究

代表研究者：松井知子　（統計数理研究所）

共同研究者：田邉國士　（早稲田大学）

佐藤真一，古山宣洋，井上雅史　（国立情報学研究所）

花田里欧子　（京都教育大学）

入野俊夫　（和歌山大学／統計数理研究所）

福水健次，Marco Cuturi　（統計数理研究所）

21 世紀の知識社会では，インターネットや大容量の電子媒体を通して，多様なマルチモーダルデータが一層利用できるようになることは確実です．その中で，それらのデータをいろいろな目的でうまく処理する技術が強く求められています．本プロジェクトでは，各目的に合わせて，マルチモーダルデータから重要な情報（ここでは“不変情報”と呼ぶ）を自動的に発見するための新しい帰納的方法論について検討しています．マルチモーダルデータを扱う具体的な課題をいくつか取り上げ，それぞれに不変情報の発見を試み，それらのアプローチを横断的に解析します．

現在のところ，映像検索，空間音源定位，自然対話解析について，それぞれに不変情報の抽出を試みています．例えば，空間音源定位に関してはまず，さまざまな方向・距離から到達する音を内耳モデルを用いてコーディングし，PLRM（Penalized Logistic Regression Machine）などの帰納的学習機械を用いて，空間音源定位のための「統計モデル」を自動学習します．PLRM はカーネルマシンとして，データのみに基づいて広い範囲のモデルを帰納的に獲得できる能力があります．カーネルマシンは，無限の数のカーネル回帰変数を内包していて，それらにより無限の数のモデルを表現することができます．次いで，そのモデルのポスト分析を行い，そのモデルに表されている情報を探ります．その得られた情報を空間音源定位のための不変情報と考えます．最終的には，この不変情報に基づいて神経回路を同定することを目指しています．PLRM により，パルス音もしくは純音について，聴覚フィルタの出力の一次結合情報に基づいて音源方向が検出できることを確かめました．今後は，学習データとして雑音データを複数作成して用い，PLRM による音源方向検出の汎化性について調べていく予定です．また，PLRM を改良して，大規模なデータの処理や有効なデータの選択的利用が行えるようにしたいと考えています．

本プロジェクトには学習理論，最適化，音声・画像処理，聴覚計算論，認知心理など，多岐にわたる専門家が参加し，帰納的アプローチを柱として，新しい科学方法論の新しいパラダイムの創造を目指しています．従来の枠に捕らわれないで，新しい分野を開拓したいと考える皆さん，修士学生・博士学生・研究者として，本プロジェクトに参加しませんか？　具体的に解きたい問題を持っている方，重要な問題自体を創出したいと思っている方，大歓迎です．広範で確かな知識を持つプロジェクトメンバーが，あなたの研究をサポートし，有望な研究者に育てます．帰納的アプローチで挑戦して行きましょう！

〔モデリング研究系　助教授　松井知子〕

研究紹介: データからの学習と推論

モデリング研究系　助教授　福水健次

最近行っている研究内容の概説をしてみたい。大きく言うと、確率的要素を含む高次元で複雑なデータから、さまざまな情報を抽出する方法や、データが陰に表現している確率的な推論ルールを獲得するための方法を研究している。以下、具体的なテーマのいくつかを解説する。

[1] 低い次元によるデータの効率的表現

高次元のデータを扱う際には、そのまま処理にかけるのではなく、事前に低次元空間にマッピングするとよいことが多い。たとえば、データを視覚化して分析を行うには、多くとも３次元空間で表現する必要がある。また、データの次元が数千、数万にのぼるような場合には、低次元にマッピングすることにより不必要なノイズを除去することが重要となることもある。このような低次元表現の問題の中で、特に以下の２つの問題に対して研究を行っている。

(A) 入力 X から出力 Y への確率的関係（回帰や識別問題）を効率的に表現するための、X の低次元特徴ベクトルを求める手法を研究している。これらを有効遺伝子発見などの問題に適用して効果を確認している。

(B) 二つの変数 X と Y の依存関係を捉える低次元表現を求めるための、カーネル正準相関分析という手法の性質をあきらかにし、その手法の理論的正当性を確立した。

[2] 異種の構造を持つデータの確率的統合

近年のデータ解析では、ベクトル的な数値データだけでなく、グラフ、ツリーなどといった構造を持つデータを扱う必要が多くなっている。例えば、遺伝情報に関わるデータでは、遺伝子ネットワーク、種の系統関係を表すツリーなど、構造化データを処理する必要が頻繁に現れる。また、種類の異なる構造を持ったデータを統合して推論に用いる必要も多い。このような問題に対して、カーネル法と呼ばれる方法論を統計的な観点から見直すことにより、異種の構造化データを同列に扱うための統計的方法を開発し、ネットワーク推定の問題などに応用を試みている。

[3] 関数空間を用いた統計的手法

上の２つの研究の理論的基盤をなすのは、無限次元の関数空間のデータを扱うための方法である。このような基盤を整備するため、統計数理的・情報幾何的な理論研究も推進している。

さらに詳しく知りたい方はホームページ www.ism.ac.jp/~fukumizu/ を参照いただきたい。

– 研究紹介 –

グリッド環境に適した遺伝的アルゴリズムによる最適化

染谷　博司* （統計数理研究所モデリング研究系, some@ism.ac.jp）

A Genetic Algorithm Optimization on Computational Grid

Hiroshi Someya (Department of Statistical Modeling, The Institute of Statistical Mathematics)

近年, 最適化手法としての進化型計算, 特に遺伝的アルゴリズム (GA) の有効性が報告されています. しかし, 生命情報科学など解評価に多大な計算量が要求される分野にも GA は応用され始めており, その計算資源は十分ではありません. GA を有効に活用するには, 次世代のより大きな計算資源を用いた GA を実現可能にする必要があります. 一方, 近年, 高速ネットワークを利用しインターネット上の膨大な計算資源を共有利用した超並列計算を可能とするグリッド技術が注目されています. 本研究では, グリッド環境上にて実現可能な進化型計算を示しその性質および有効性について調査しました.

グリッドでの計算では, (1) セキュリティ, (2) 不均質性, (3) 不確実性, (4) 非同期性, (5) 超並列性, を考慮する必要があります. (1)~(3) は, ミドルウェア等での工夫が可能ですが, (4)(5) に関してはアプリケーションレベルでの工夫が必要であり, グリッド環境において実装される GA はこれらについて考慮されていなければなりません. また, これらの他に, 通信量についても考慮する必要があります. グリッドでは通信の遅延時間が大きいため, 通信量が小さく通信頻度が少ないことが望ましいと考えられます. 非並列な高性能な GA と同程度の性能を維持したままで計算時間を短縮できることも望ましい条件として挙げられます.

著者らは, 適応的な探索をする進化型計算の一手法として, *Genetic algorithm with Search area Adaptation* (GSA) を提案しています. GSA は, 代表的ないくつかのベンチマーク問題において, その有効性が確認されています.

グリッド環境における GSA の実現可能性に関する知見を得るために, 簡略化した GSA を実装しその計算時間短縮効果に関する性質を調査しました (図 1). その結果, サーバ数が多い場合でも並列数に対してほぼ比例した計算時間の短縮が可能であることを示唆する実験結果が得られました (図 2).

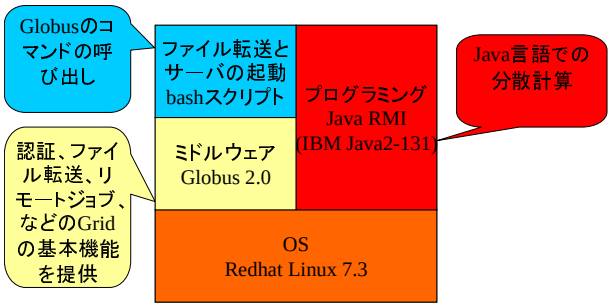


図 1　実装構造

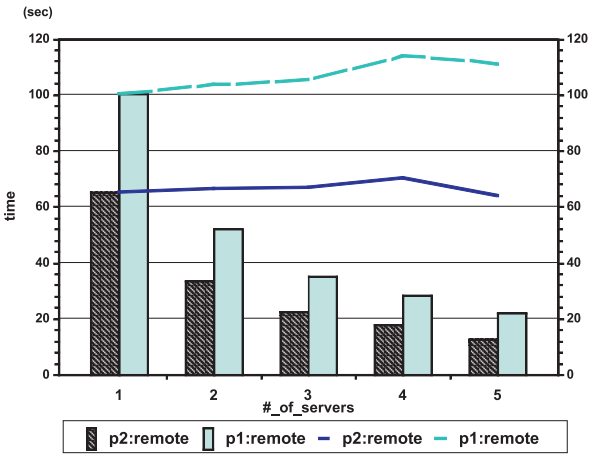


図 2　実験結果

本研究についての詳細は, 以下の文献をご参照ください [(1)~(3)].

文　献

（1） 染谷博司：グリッド環境に適した遺伝的アルゴリズムによる最適化, 統計数理, Vol. 52, No. 2, pp. 381–391 (2004).
（2） 染谷博司：進化型計算による適応的探索およびグリッド環境への応用, 最適化：モデリングとアルゴリズム 17, 統計数理研究所 (2004).
（3） 染谷博司：グリッド環境に適した遺伝的アルゴリズムに関する考察とその実現, 電子・情報・システム部門大会 2003 講演論文集 (CD-ROM), 電気学会 (2003).

あなたにとって一番大切と思うものはなんですか？

統計数理研究所　データ科学研究系
坂元　慶行

これは、統計数理研究所が、1953(昭和28)年以来５年おきに、50年にわたって行っている「日本人の国民性調査」の質問のひとつで、自分にとって一番大切なものを自由に挙げてもらう質問です。この「国民性調査」は、日常的な場面における普通の日本人の態度や心情等について統計調査を行い、日本人のものの見方や考え方の特徴を統計的に明らかにするために続けられてきました。

さて、この質問に対する結果です。「家族」という答は、1968(昭和43)年調査まではわずか13%に過ぎませんでしたが、1970年代からどんどん増え始め、最新の2003(平成15)年調査では45%に達しました。これは、金・財産、愛情・精神など、他の全ての答が減るか停滞する中で見られた極めて特徴的な現象で、答は「家族」に、いわば一極集中を続けています。この質問から見る限り、価値観の多様化ではなく、価値観の一様化、単一化ということになりそうです。私はこれまで、「国民性調査」から得られた50年間の日本人の意識の動きの基調のひとつは、私生活を優先する価値観が強まってきたことであると指摘してきましたが、最近は、この段階を超えて、内向き志向とでも社会離れとでも言うのでしょうか、社会性そのものが後退する傾向が強まっているように思われます。上の質問もこのような傾向を示す典型的な例のひとつです。

ところで、「国民性調査」では、（くじびきのような方法で）ランダムに選ばれた人の家を訪問して当人の答を聞き取って来るのですが、何%位の人が調査に答えてくれる(回収率という)と思いますか？　実は、50年前の1953(昭和28)年調査での回収率は83%だったのですが、その後どんどん下がり、最新の2003年調査では56%でした。今や、やっと２人に１人が答えてくれる状況です。このように回収率が低下した最大の原因は調査を拒否する人が激増したことにあります。拒否は調査対象者の意志表示のひとつという点で、他の不在とか転居といった、いわばやむを得ない理由とは性格が異なります。また、回収率は、若い世代ほど低くなる傾向がありますので、回収率の低下傾向は今後も続くものと予想されます。

調査に協力してくれる人が減っただけではありません。調査に応じてくれた人に関しても、最新の調査では、「わからない」という答や、「場合による」・「どちらでもない」といった「中間的な答」の選択率が最も多く、したがって、その分、それ以外の明確な意見を表現した回答の選択率が小さくなり、はっきりした意見が調査結果に表れにくくなっています。これらの現象は、調査、ひいては社会に関わることへの消極的な態度を示唆しているのではないかと思われます。

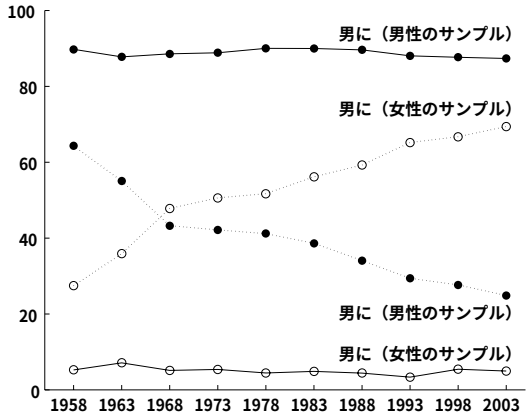
新聞などの報道によりますと、国の一番の基礎である人口を調べるための国勢調査でさえ、昨2005年調査の未回収率は東京都全体で11%に上ったそうです。未回収率が高くなれば、得られた答は本当の姿とは食い違ってきます。このため、現在、「国民性調査」だけでなく、いろいろな社会調査が存続できるか否かのぎりぎりの岐路にあると言っても過言ではありません。しかし、調査なくして合理的な行動の決定は不可能です。クラスの人数が分からなければ給食を何人分用意したらいいかも分かりません。どんな意見が多いかを知らないでいろいろなことを誰かが勝手に決めたら困ったことになるでしょう。国の場合も同じです。多くの国民が家族が一番大切だと思っています。私は、「家族が大切だから、社会に背を向けて自分の殻にこもる」のではなく、「家族が大切だから、家族がいつまでも平和に暮らせるよう、社会に向かってきちんと意見を述べる」ようにしてもらいたいと思います。

継続的な調査データから社会の変化を捉える コウホート分析の方法

中　村　　隆　（データ科学研究系 調査解析グループ）

日本人の国民性調査データ

統計数理研究所では，昭和 28 年 (1953 年) から 5 年ごとに「日本人の国民性調査」を実施している (平成 15 年秋に第 11 次調査が実施された)．この調査の中に「男女の生まれかわり」という質問項目がある．男女別のこの意識の変化は下の図のようになっている．男性はほとんど変わっていないのに対して，女性は大きく変わったことが読み取れる．



コウホート分析

社会の変動を意見や意識の変化を通して捉えようとするとき，主にどのような要因によっているのかを見極めることが大切である．時代や世代によらず人が歳をとることによって変わる加齢の要因，時代につれ人々全体の意見がある方向に変わっていく時勢の要因，生まれ育った時代背景が違うことによる世代の要因などが考えられる．これらの要因の働き方によって，将来の社会の様相は大きく異なってくる．

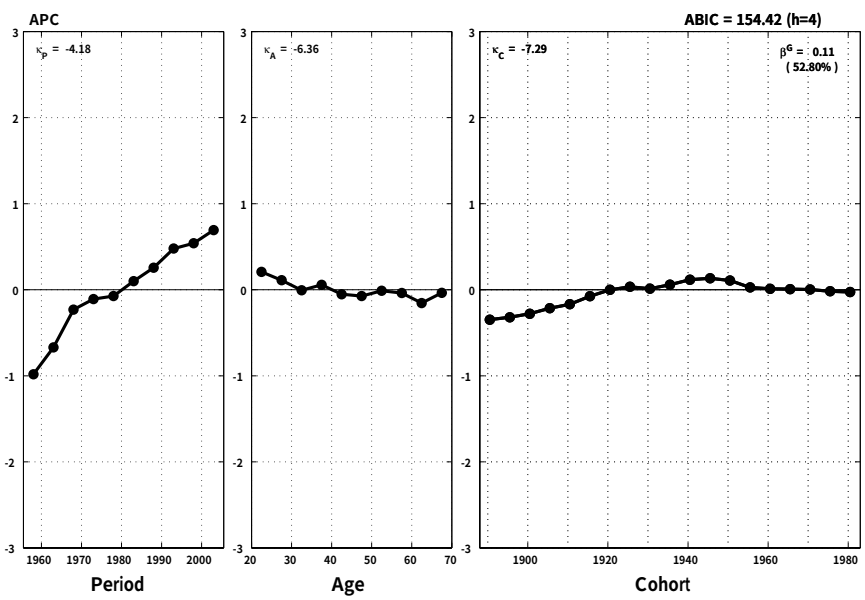
社会の変動を捉え将来の動向を探るために，継続的に調査されたデータから上に述べた各要因の影響の大きさ，すなわち，年齢・時代・世代 (=コウホート=同時出生集団) 効果を分離する統計的方法がコウホート分析と呼ばれる方法である．

困難点

年齢・時代・世代の効果を分けて考えるというコウホート分析の考え方は魅力的であっても，いざ実際に分析を行おうとすると，3 つの効果が分離できないという困難が待ちかまえている．たとえば，若い時の意見を，異なる世代について比較しようとするときに，時代の違いを持ち込まないようにするのはとても難しい，ということである．これがコウホート分析における識別問題と呼ばれる困難点である．

克服法—ベイズ型コウホートモデル

困難点を克服するために，年齢，時代および世代効果が緩やかに変化するという付加条件を取り込み，赤池のベイズ型情報量規準 ABIC を最小にするようなモデルを選ぶベイズ型コウホートモデルを開発した．このモデルにより，先に示した女性のサンプルの「女に」をコウホート分析してみると，下の図のようになり，女性の意識の変化は，第一に時代効果 (Period) によるものであり，年齢効果 (Age)，世代効果 (Cohort) もそれぞれ特徴を見せていることがわかる．



文化多様体解析（CULMAN）　－　意識の国際比較　―

データ科学研究系　吉野諒三

1.「日本人の国民性」調査

統計数理研究所では1953年以来、「日本人の国民性」に関する調査を続けている。この調査の先駆として、1948年に関連分野の研究者による「日本人の読み書き能力調査」がある。この背景には、GHQの一部が民主化政策を考える際に、教育と日本語の関係を問題視し、ローマ字化すべきと考えた経緯があったが、調査の結果、日本人の能力が十分高い事実が確認され、国語のローマ字化が阻止されたと言われている。実際には世界の情勢や占領下の検閲と無関係ではなかったであろうが、いずれにせよ、この調査は、統計的「標本抽出理論」の実践的重要性を確認させた。

他方で、これは戦後民主主義を発展させる科学的「世論調査」の基盤を整える契機ともなった。マスメディア各社はGHQの指示により、統計数理研究所の指導の下で、科学的な世論調査を確立していったのである。戦時中にできた機関が次々と廃止されていく中で、統計数理研究所（開所1944年）は、戦後民主主義の科学的基盤を支える使命を担い、新たに出発したのであった。この流れで「日本人の国民性」調査が開始され、今日では、内閣府の「社会意識に関する世論調査」、ＮＨＫの「生活時間調査」と共に日本の三大標本調査として有名になり、さらに、米国の「一般社会調査」や「世界価値観調査」など、世界各国の大規模な調査を開始させる刺激となった。

2.「意識の国際比較調査」

この研究は、1971年頃から、国民性をより深く考察する目的で日系人を初め、他の国の人々との比較調査へと拡張されてきた。言語や民族など、重要な共通点がある国々を比較し、似ている点、異なる点を判明させ、その程度を測ることによって、初めて統計的「比較」の意味がある。この比較の連鎖を徐々に拡張し、やがてはグローバルな比較を可能とする方針の下で、「連鎖的調査分析」の方法論を確立した。国際比較では、翻訳の問題、各国固有の調査方法の違いの問題など、そもそも国際比較など可能なのかが大問題となる。我々はこの「国際比較可能性」を追求し、計量的文明論を確立するため、「データの科学」（吉野、2001）を試行錯誤している。

これまでの我々の主要な調査には、「意識の国際比較」（日米欧の７カ国）,ハワイやブラジルや米国西海岸の日系人調査、「東アジア価値観国際比較」などが含まれる。（本研究所の研究リポートや、ホームページ参照。）

なお、余談ではあるが、総務庁（現内閣府）の「青少年の意識の国際比較」は、1972年以来の時系列国際比較調査として今日まで継続している貴重な事業であり、これは当時、総務庁青少年対策本部に在職されていた千石保氏（現青少年問題研究所・所長）、遠山敦子氏（元文部科学大臣）が、本研究所の西平重喜所員（現名誉所員）と共に、開始されたのであった。

3.「信頼の世紀」に　―― 計量的文明論の確立に向けて ――

新世紀を迎え、伝統的な産業社会から高度情報化社会へと移りつつある世界において、これまでの人間関係や人々の信頼感のあり方にも急激な変化が見られる。我々は国際比較研究の対象として、近い将来、世界の一極になると想定される東アジアに着目するようになった。この研究を推進する枠組みが、「文化の多様体解析（cultural manifold analysis,**CULMAN**）」（吉野、2005）であり、その確立のために試行錯誤している最中である。

参考文献

林知己夫.（2000）.　これからの国民性研究―人間研究の立場と地域研究・国際比較研究から計量的文明論の構築へ―．統計数理, 48(1), 33-66.

林知己夫, 鈴木達三, 吉野諒三他.（1998）.　国民性７か国比較.　出光書店.

Inkeles, A.（1996）. *National Character*. Transaction Publishers.「国民性論」吉野諒三訳(2003).（付章,吉野原著「日本における国民性研究の系譜」）.　出光書店.

吉野諒三.（2001）.「心を測る」ー個と集団の意識の科学ー. データの科学シリーズ 朝倉書店

Yoshino, R.（2002）.　A time to trust. Behaviormetrika. Vol.29 No.2,pp.231-260.

吉野諒三.（2005）. 東アジア価値観国際比較調査ー文化多様体解析（CULMAN）に基づく計量文明論の構築へ向けてー.　Vol.32, No.1, pp.133-146.

インターネット電話で遊んでみよう！

データ科学研究系　伊原　一

最近は、自宅やオフィスで ISDN やブロードバンドなどの高速インターネット環境が当たり前の時代になってきていますが、ほんの１０年前まではモデムを電話回線に接続してのんびりとパソコン通信をしていたものです。当時は、画像を１枚ダウンロードするのにコーヒーが一杯飲めるくらい時間がかかっていたものですが、このところわずか１０年でインターネットは目覚しい進化を遂げて、ついにテレビや映画も見ることができるようになってきています。

そんな中で、近頃はインターネットを電話回線で接続するのではなく、逆に電話をインターネットに接続して利用できるサービスが提供されてきています。これまでも IP 電話と呼ばれる機能は既に提供されており、自宅の電話を ISDN や ADSL といったデジタル回線にすることでインターネットサービスの一環として低料金の電話が利用できるようになっていましたが、最近では高速インターネットの接続環境さえあれば、どの PC 端末からも自由に電話をかけられるようになってきています。

そこで、研究室のインターネットを利用してインターネット電話をかけてみる実験をしてみました。必要な機材は、インターネットに接続しているパソコンと、USB フォンと呼ばれている電話機の２つです。USB フォンは、最近は家電量販店などでも２～３千円程度で市販されており、また、インターネット通販でも購入することができます。今回、テスト用に購入した USB フォンは見た目は携帯電話にそっくりですが、電源を USB から取るため電池が不要となっている分、携帯電話よりも重量が軽くなっています。本体からはケーブルが延びており、ケーブルの先端部分に USB、イヤホン、マイクの３つの端子がついていて、それぞれパソコンに接続するようになっています。さっそく接続して付属の CD-ROM に入っているソフトをインストールしてみました。

ソフトのインストール自体は CD-ROM を入れると自動的に始まりますが、CD-ROM のセットアップファイルから直接実行してインストールすることもできます。必要なソフトは２種類で、ひとつは USB のドライバー機能を提供するためのもので、USB ドライバーをインストールすることで電話機から電話番号を直接入力することができるようになります。もうひとつはスカイプ(Skype)と呼ばれるインターネット電話のソフトで、これはインターネットのホームページから最新版を無料でダウンロードすることができます。

＜スカイプ(Skype)のホームページアドレス＞
http://www.skype.com

インターネットで Skype の最新版をダウンロードして実行してみると、希望する ID やパスワードに加えて、e-mail アドレスを聞かれるのでそれぞれ画面で入力します。Skype フォンにログインできればインストールは完了です。あとは電話をかけたい人の ID を探して接続ボタンを押すだけとなります。さっそくもう一台のインターネット電話につないでみました。

＜インターネット電話の通話＞

A「トゥルルル、トゥルルル...」

B「クリック！」

A「もしもし、聞こえますか？」

（..... もしもし、聞こえますか？）

B「はいはい、よく聞こえます！」

（..... はいはい、よく聞こえます！）

インターネット電話の通話はこのような感じで、普通の電話と全く変わりません。ただし、実際の声と電話から出る声に若干の時間差があるため、すぐ近くからかけていても国際電話のような会話になるところがおもしろいところかもしれません。また、インターネットに接続していれば追加の電話代がかかることもないのでインターネット環境があれば実質的に無料で電話をかけることができるというわけです。これなら研究の打ち合わせに何時間かかっても電話代の心配をする必要がなく、しかもインターネットに接続していれば世界中どこからかけても電話代がかからないので、特に海外にいる研究者との打ち合わせにはうってつけです。今後は国際的な共同研究などには不可欠のツールになりそうです。

さらに、このシステムの最大の特徴は一般の電話にも電話をかけることができるという点でしょう。一般電話への通話にはクレジットカードが必要で、試しに 10 ユーロ分（約 1,480 円）の度数を購入してオフィスの電話にかけてみたところ、普通の電話と変わらない音質で通話することができました。電話番号は国際電話と同じ方法で国番号＋８１からダイヤルしますが、操作は USB フォンからできるので普通の電話とかけ方は全く変わりません。

＜一般電話との通話＞

A「トゥルルル、トゥルルル...」

電話「カチャ！」

A「もしもし、明日の会議の件ですが」

（..... もしもし、明日の会議の件ですが）

電話「はいはい、資料はできてます」

（..... はいはい、資料はできてます）

インターネット電話から一般電話への通話もこのような感じです。通話料金は日本中どこにかけても１分約３円となっており、北は北海道から南は沖縄まで市内通話並みの料金で電話をかけることができるということになります。また、海外にいても日本への通話料金は全く同じなので、国際電話を市内通話並みの料金でかけることができます。逆に海外へ電話をかける場合の通話料金は国によって異なりますが、インターネットが普及している国へはいずれも市内通話並みの料金となっており、相手がインターネットに接続していなくても一般電話に、しかも市内通話程度の料金で国際電話をかけることができるというのは、つい最近まで１分２００円以上もかけてアメリカに国際電話をかけていたことを考えると画期的なことだといえそうです。国際的な研究者の方々には、ぜひインターネット電話の利用をお勧めします。

階層構造を持つデータと生態学的推論

データ科学研究系　前田 忠彦

1　階層構造を持つデータ

社会は複雑な階層構造を持つシステムであり，社会現象を研究対象とする場合に，そのような複雑な階層構造に対応した複雑な構造を持つデータが取得される。統計数理研究所データ科学研究系の研究者は，さまざまな社会現象に対するデータの取得法と解析法を研究対象としている。私もそのような研究者の一人である。

社会に生きる人間は，ある地域—たとえば市区町村の町丁字—の中に暮らしている。あるいは特定の集団—たとえば会社の部・課や学校のクラス—に所属している。本来は階層構造はより深く，また一人の個人がさまざまな側面で複数の帰属先を持っているが，ここでは単純のために帰属先の集団が単一で階層構造は二段階であると考えておこう。 集団 $j(=1,2,\ldots J)$ の中に個人 $i(=1,2,\ldots,n_j)$ がネストした階層構造を持ち，変数 X に対する測定値は x_{ji} のように表現される。教育を主題とする調査は，クラスの中の生徒を対象者として，あるいは学校に帰属する教員を対象者として，実施される。世論調査では，層化多段無作為抽出法で対象者が標本として抽出されるが，これは全国の地域から特定の町丁字などがまず無作為抽出され，さらにその地域内の個人が無作為抽出されて得られる。これらは二段階の階層構造を持つデータの例である。

個人の行動は帰属する集団の影響を受けるので，個人の行動を分析する際には，そのような階層構造を明示的に反映させたデータ解析法が必要な場合も多い。そのような解析法は総称的に多水準分析 multilevel analysis と呼ばれることがある。

文脈は全く異なるが，同一個人の（身長のような量的変数に関する）時間変化を追った繰り返し測定データも，個人の中に測定時点がネストした一種の階層構造を持つと見なすことができる。

2　生態学的推論

データの階層構造を無視した分析が常に誤りということではない。ただ，明確に注意を喚起しなければならない問題に，生態学的推論 ecological inference の危険性がある。生態学的推論とは地域や集団を測定単位としたデータから得られた相関から個体間の関係を推測することを指し，例えば，市区町村別に大学進学率と離婚発生率の間の相関を求めて仮に負の相関があった場合に,「学歴の高い人は離婚しにくい」といった推論を行う，などである。このような推論が一般には成立せず，むしろ個人レベルでは逆の関係を持つケースさえあり得ることは古くから知られ，考察対象となってきた。研究所の先輩の長谷川政美さんが翻訳された Langbein & Lichtman (1978) などに解説されている。

このような生態学的誤謬を避けるための多水準分析の問題は，十分に議論しつくされておらず社会科学研究者の間でも関心は持たれても充分に活用されていないような印象を持っている。最近少し興味を持ったので研究してみようと思っている。

文献

Langbein, L.I. & Lichitman, A.J. (1978). *Ec　　ca I e e ce*, Sage publications, 長谷川政美 (訳) 「生態学的推論」 朝倉書店, 1980 年.

混合モード調査法の可能性を探る

データ科学研究系
土屋　隆裕

統計数理研究所が 1953 (昭和 28) 年から 5 年ごとに実施している「日本人の国民性調査」は、2003 年秋に第 11 次調査が実施され、先日その結果が公表された。第 1 次調査の時には 83%という今では考えられない高水準にあった回収率も、特に 1980 年代後半から急激に下落し、今回の第 11 次調査ではついに 56%にまで落ち込んだ。用意したサンプルの実に半分近くが調査不能であり、回収サンプルだけの単純集計結果をもってはたして「日本人の国民性」と言ってよいのか、という疑念は当然生じる。

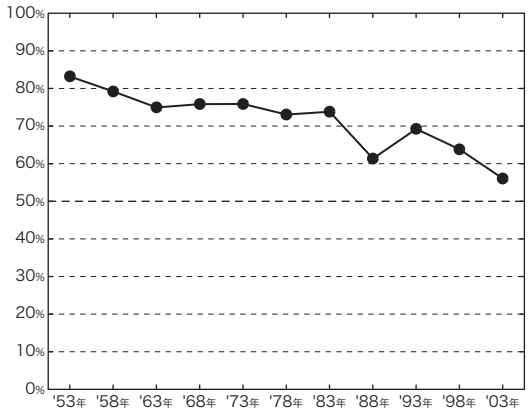


図 1: 「日本人の国民性調査」の回収率

回収率の低下に悩むのは国民性調査に限ったことではない。そのため、調査不能補正の是非やその方法は長く研究されてきた。特に最近では、調査不能を補完する方法の一つとして、混合モード調査法の可能性が考えられている。つまり、個別面接聴取法や郵送調査法、電話調査法やさらにはインターネット調査法などといった調査モードのうち、一つのモードだけではなく、複数のモードによる調査を並行して実施するのである。いずれのモードも回収率は高くなく、それぞれの結果は断片的な偏った情報しか与えないかもしれない。しかしそれらをうまく組み合わせることで、より「偏り」が少ない推定値が得られるのではないか、というのである。

見ず知らずの他人が調査員として訪ねてきても調査への協力は断るが、調査票がホームページ上にあれば回答してもよい、という人や、紙に印刷された質問文を自ら読むのは面倒だが、口頭で質問されれば答える、という人など、調査への協力・非協力は調査モード次第という人は少なくない。今後回収率のさらなる低下こそあり得ても、その回復はまずのぞめないことを考えると、混合モード調査法は、調査不能を補完するための魅力的な方法の一つのように思われる。

そこで数年前から、先に挙げたような複数のモードによる調査を実施し、モード間の比較可能性について検討してきた。例えば以下は、同一の質問項目を個別面接聴取法と Random Digit Dialing による電話調査法とで実施した結果を比較したものである。

もし自分の子供が、「外国人と結婚したい」と言ったとしたら、あなたは、賛成しますか、それとも、反対しますか？

	面接	電話
賛成する	41%	53%
反対する	27%	22%
場合による	29%	20%

明らかに電話調査法では、外国人との結婚に「賛成する」という回答が多く出ている。しかしだからといって直ちに、電話調査法ではリベラルな人をより多く回収できる、と結論づけることはできない。両モードの間では、性別や年齢といった回答者の人口統計学的属性の分布が異なるばかりでなく、調査員が面前にいるのか電話口から声だけが聞こえるのか、回答者が匿名か否か、調査員の性別・年齢層・経験・態度・雰囲気、さらには回答者の動機づけなど、モードに依存する多くの点が異なるからである。

言い換えれば、異なるモードで調査を実施したとき、はたして同一の回答者が同一の回答をするのだろうか、ということである。実際に同一人に対して複数のモードで調べてみればよいではないか、と思われるかもしれないが、問題はそう簡単には解決しない。同一モードで調査しても、個々人の回答が非常に不安定で変動することは昔からよく知られているし、複数回調査することによる回答への影響も無視できない。

実際のところ、回答者は調査モードの影響を少なからず受けているようである。とすれば、異なるモードの間では、調査項目は同じでも調べているものは違うということになる。そのような異質なモードを組み合わせた結果数値というのは、はたして何を表すことになるのであろうか。混合モード調査法を具体化するにあたっては、どのように結果を組み合わせればよいのか、という統計的な技術だけではなく、社会調査の結果は何を表しているのか、社会調査とは何なのか、何のために社会調査を行うのか、という研究者の調査観も問われているのである。

Open House のための研究紹介（2006 年 5 月）

データ科学研究系　助手　松本　渉

他の多くの研究者の方がそうであるように，私の場合も，普段携わっている研究は，何種類かにまたがります。しかし，少し乱暴ですが，それらを一言で言ってしまえば，「組織の調査・分析に関する方法論の開発とその実践」に，私の研究の特色があると言えるでしょう。ここで言う「組織」は，生物の話ではありません。経営学や社会学，場合によっては経済学などがその研究の対象としてきた，企業などの組織のことで，人々の集まりによって形成され，機能する社会のシステムとして理解されうるものです。

この「組織」に関する研究は，長らく企業（厳密には営利企業，会社）の研究が中心でした。しかし，ここ十数年ほど前から，企業以外の組織にも目を向ける動きが出てきました。いわゆる NPO（非営利組織）はその典型ですが，阪神大震災以降，社会的にも認知度を高め，政策的・社会的な観点からも重要性を増してきました。

私の研究生活は，この NPO についてのデータを取得し，経営組織論的な視点から分析することから出発しました。こういったテーマでは，実際に現場に訪れる研究もしばしば重要となります。しかし同時に，何らかの法則性を科学的な知見として見出したり，現場で直感的に理解されている事実を，説得力を伴って確かめたりするためには，やはり何らかの数量化を行い，統計的なデータを利用して，計量的な分析をすることも重要となってくるのです。そして，計量的な分析に耐えうるデータを得るためには，その基礎となる統計的な調査がしっかりしている必要があります。そのため，現在では，データを解析して組織に関する研究を進めると同時に，調査の方法についても研究を進めているところです。

なお，組織の研究のための統計的な調査の方法としては，以下の三つがあると考えています。

1）組織を単位として，多数の組織の調査・分析を進めるアプローチ

2）個人を単位とする調査結果を利用して，組織の調査・分析を進めるアプローチ

3）組織内部における個人の調査から，単一の組織の調査・分析を進めるアプローチ

これまでは，1）の方法をよく利用していましたが，日本では，2）の方法が，比較的学問的に確立していることもあり，最近では，2）の方法，いわゆる社会調査データを利用して分析を行うと同時に，その方法についての研究に力を注いでいます。将来的には，2）で用いられた方法論研究の成果を，何らかの形で 1）に還元させたいと考えています。3）は，個人的には，ほとんど経験がありませんが，機会があれば関わりたいと考えています。

なお，方法論の研究とは別に，分析する具体的な研究対象の範囲も少しずつ広げてきています。その結果，現在では，ソーシャル・キャピタル（社会関係資本）という枠組みで，組織に関する研究を行っています。このソーシャル・キャピタルとは，近年，社会科学で学際的に取り扱われてきているテーマで，具体的には，信頼や人的ネットワーク，規範等を意味するものです。政治学，社会学，心理学，人類学などで扱われてきたこれらの諸概念を一つの枠組みで捉え，経済的な研究との関係を視野に把握しようとすることに特徴があります。NPO もこのソーシャル・キャピタルの一部として考えられるものです。ソーシャル・キャピタルは，開発学や国際協力の場面などで注目されていることもあり，ここ数年は，（NPO だけでなく，）様々な組織に対しての信頼感などの国際比較調査研究に取り組んできています。

NPO にしてもソーシャル・キャピタルにしても，研究としては，発展途上の新規開拓分野であり，頑強な計量的な分析の進め方が十分に確立しているとは言えないでしょう。冒険的な部分は多々あるかと思います。そもそも日常的に変化の激しい組織という研究対象自体に，取り扱いの難しさがあるとも言えます。しかし，それだけに「組織の調査・分析に関する方法論の開発とその実践」は，研究としての面白さはあると思いますし，研究としての社会的意義も大きいのではないかと思っています。

江戸時代のデータ解析－二宮尊徳－

馬場康

二宮尊徳、通称、金次郎が生まれたのは、天明 7 年 7 月（1787 年 9 月）である。この年の 1 月には、将軍が家斉にかわり、6 月には松平定信が老中に就任している。当時、既に武士階級の経済的な破綻が生じていた。そこに、飢餓などの突発的な出来事が状況を益々悪くしていた。

相馬中村藩は、陸奥国相馬地方を領有する外様の中藩であった。元禄から享保にかけては、最高で 17 万俵を超える年貢米の収入があり、人口も約 9 万人を擁し、安定していたが、天明の飢饉後は年貢も人口も大幅に減少している。天保 4 年（1833 年）の飢饉の後は藩財政は破綻していた。

こういう状況下で、相馬藩は二宮尊徳に藩財政の建て直しを依頼する。尊徳は、180 年間の租税データを持参した家老たちの熱意に感じて相馬藩の財政再建を引き受けることになる。尊徳は、この租税データの時系列を分析して再建の基礎を作る。尊徳の時系列の認識は以下のようなものである。年貢米収入を 60 年ごとの 3 期に分けると、平均が 14 万俵、以下 11 万俵、6 万俵と明らかに下がっている。また、90 年ごとの 2 期に分けると、前半が 13 万俵、後半が 7 万俵とやはり減少傾向が見られる。さらに尊徳は、この減少傾向を示すために直近の 10 年間の年平均収納高 57,205 俵を算出している。

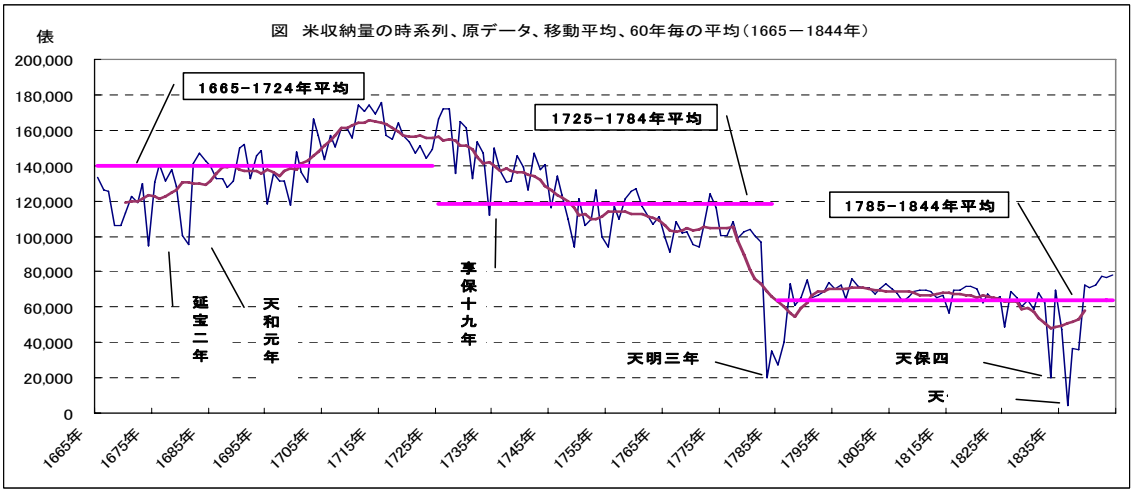
90 年平均と 10 年平均の平均

（76,347＋57,208）÷2＝66,777　俵

をここ 10 年間の年平均租税の目安とし、それに基づいて財政再建の計画を立てたのである。

尊徳の時代は、折れ線グラフなどは用いていないが、現代風に年貢米の時系列をプロットすると図のようになる。折れ線が年貢米の推移を示し、水平の 3 本の直線が 60 年ごとの平均を表す。また、比較的滑らかな実線は 11 年間のデータによる移動平均である。尊徳の分け方が、非常によく時系列の特徴を捉えていることが分かる。

そろばんでしか計算ができない時代に、すでに統計を用いた解析が行われていた。江戸時代のデータ解析の話である。



家系数の変遷

データ科学研究系
上田 澄江

次のような家系のモデルを考える．
1 家系は 2 世代からなり，家系の構成は次の 3 通りとする．

- 両親と子供 2 人(息子と娘)をもつ 4 人家系，
- 両親と子供 1 人(息子または娘)をもつ 3 人家系，
- 子供の無い親だけの 2 人家系

である．時刻 t におけるそれぞれの構成数を $n_4(t), n_3(t), n_2(t)$，$n(t)=n_4(t)+n_3(t)+n_2(t)$ とおく．時刻が 1 増すごとに新しい 4 人家系が 1 つ誕生する．すなわち，時刻 $t+1$ では，兄妹(姉弟)でない男女の子供を 1 人ずつランダムに選び，その 2 人を両親とし子供 2 人(息子と娘)から構成される新しい 4 人家系が誕生すると想定する．出生率は 2，死亡率は 0 である．このとき $n_4(t)=k$ となる確率 $p(n_4(t)=k)$ は n に関する漸化式で表現できる．一方，4 人家系の数に注目すると次時刻における家系数は，1 減，変化なし，1 増のいずれかになり，それぞれの確率は次の式で与えられる．

① $p(n_4(t+1)=n_4(t)-1) = {}_{n_4(t)}C_2 \,/\, {}_{(n_4(t)+n_3(t))}C_2$

② $p(n_4(t+1)=n_4(t)) = ({}_{n_4(t)}C_1 \cdot {}_{n_3(t)}C_1) \,/\, {}_{(n_4(t)+n_3(t))}C_2$

③ $p(n_4(t+1)=n_4(t)+1) = {}_{n_3(t)}C_2 \,/\, {}_{(n_4(t)+n_3(t))}C_2$

今，時刻 0 において 4 人家系のみからなるモデルを想定する．$n(0)=n_4(0)$ である．上記 3 通りの確率のうち複数が 0.25 以上であるならば，乱数でいずれかを選択する．0.25 を超える確率が 1 つだけならばそのものを選択する．このとき $n(0)=100$ として家系数の変化を時刻 100 まで図示したのが図 1 である．4 人家系と 3 人家系の数は急速に接近する．
図 2 は $n(0)=n_4(0)=100$ の 4 人家系から始めて，乱数によって選ばれた男女 1 組を両親とし新しい 4 人家系が誕生するとしたときの $t=1000$ までのシミュレーションによる家系数の変遷を示す．図 3 は次世代の子供の数を 0 から 6 人としその比率を与え，1 人家系～8 人家系の構成を想定した場合のシミュレーションによる家系数の変遷である．$n(0)=2000$，$t=20000$ とし，子供無し，子供 1 人，…，子供 6 人の比率を 1：2：3：3：3：2：1 で与え，親と子供の死亡率を 2：1 とした(出生率と死亡率は同率)．初期値は 1 人家系～8 人家系数を 0，200，300，300，300，300，300，300 とし，曲線は上から順に 1 人家系，2 人家系，…を示す．出生率を変化させれば家系数の変遷は変化するが，t が大きくなれば概ね同比率を保持する．

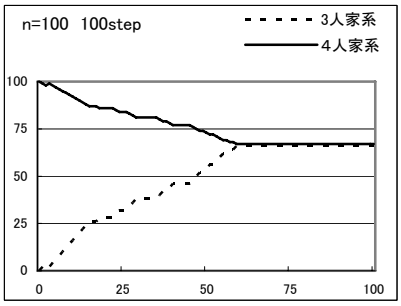


図 1

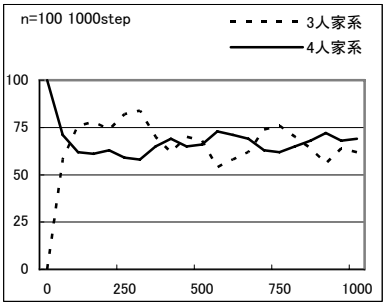


図 2

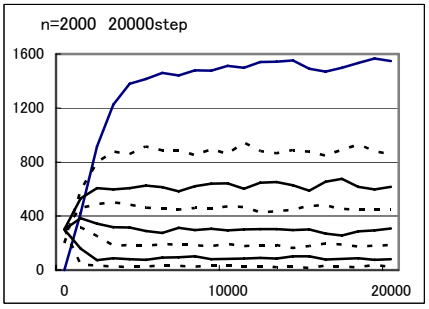


図 3

一般化線形モデルの共役解析

データ科学研究系　助手　大西俊郎

経済データの解析や生物統計学などの応用分野で広く使われている一般化線形モデルを，Bayes 統計の一分野である共役解析を用いて解析することついて理論的な研究を行っています．

一般化線形モデルとは正規線形モデルを次のように拡張したものです．

(i) 誤差分布…正規分布から指数型分布族へ

(ii) 線形予測子と平均パラメータを結ぶ関数関係…恒等関数から一般の単調関数へ

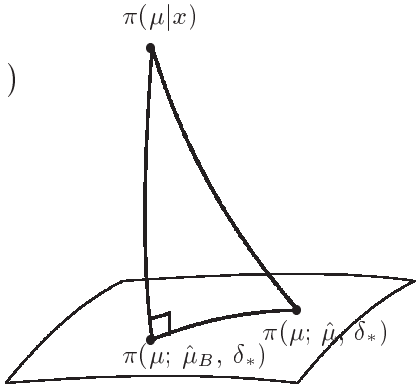
指数型分布族は，正規分布だけでなく，ポアソン分布・二項分布・ガンマ分布などの重要な分布を含む分布族です．2 方向への拡張によって，一般化線形モデルは非常に大きな汎用性を持ちます．

上記の汎用性とともに一般化線形モデルが広範囲に用いられる理論的根拠は，指数型分布族がもつ最小情報量性にあります．平均パラメータを μ とし，分散を平均の関数とみて $v(\mu)$ と書くことにします．指数型分布族は，平均が μ であり，分散が $v(\mu)$ である分布の中でフィッシャー情報量を最小化する分布であることが知られています．この性質は，分布の仮定が間違っていた場合にも推論が大きく間違わないという頑健性につながっているのです．

共役解析とは，共役事前分布を仮定して行う Bayes 解析です．共役事前分布は事後分布と事前分布が同じ分布族に属する事前分布であり，Bayes 推定量の計算が簡単になることが知られています．共役解析のメリットは解析的アプローチが可能な点にあります．また，必ずしも共役でない事前分布を仮定した場合の推定問題を考察することにより，共役解析が非共役解析の射影として理解できることが分かります．

下図において曲面は共役事前分布 $\pi(\mu;\, m,\, \delta)$ の空間を表します．一般の事前分布 $\pi(\mu)$ に対応する事後分布を $\pi(\mu|x)$ とすると，この事後分布から共役事前分布の曲面に垂線を下ろすことで Bayes 推定量 $\hat{\mu}_B$ が得られます．また，この垂線は共役事前分布の曲面と局所的に直交しているだけでなく大域的にも直交していて，任意の推定量 $\hat{\mu}$ とすると次の等式（ピタゴラス関係）が成り立ちます．ここで，$\mathrm{KL}(\pi_1,\, \pi_2)$ は分布 π_1 から分布 π_2 への Kullback-Leibler 分離度です．

$$\begin{aligned}&\mathrm{KL}\big(\,\pi(\mu|x),\, \pi(\mu;\, \hat{\mu},\, \delta_*)\,\big) - \mathrm{KL}\big(\,\pi(\mu|x),\, \pi(\mu;\, \hat{\mu}_B,\, \delta_*)\,\big)\\ &= \mathrm{KL}\big(\,\pi(\mu;\, \hat{\mu}_B,\, \delta_*),\, \pi(\mu;\, \hat{\mu},\, \delta_*)\,\big).\end{aligned}$$



上図のピタゴラス関係が含意するのは，ある一定の条件を満たす事前分布の中で共役事前分布が最も情報量が小さいということです．

以上のように，一般化線形モデルおよび共役解析はともに最小情報量性というよい性質を持ちます．相性がよい 2 つを組み合わせることで優れた推定手法が導かれることが期待されます．

物理乱数発生について

データ科学研究系　田村義保

乱数（Random　Number）は科学の研究には不可欠のものであり、シミュレーションのみならず、暗号やセキュリティの分野でも利用されている。数式を用いて発生させる「擬似乱数」が使われることが多い。しかしながら、簡単な発生方式を使って、乱数の最初の値（初期値）に、計算機内蔵時計を使うようなことをすると、次に発生される乱数の値を第三者に予測されることが起こりうる。このような方法で、セッション番号などを発生させると、なりすましが可能となり、セキュリティ上、問題が多い。また、これから始まる予定の裁判員制度においても世論調査においても、擬似乱数というある意味では操作可能なランダムさを用いた場合、その無作為性が疑われることが生じうる。大規模なシミュレーションは並列計算機を用いた計算が主流となっているが、複数の CPU で並列に発生させた擬似乱数の性質について、詳しく議論されていないように思う。すなわち、乱数として望ましい性質を有しているかどうかについて顧みられてないように考える。

擬似乱数もメルセンヌ・ツイスタのように、周期性のみならず、一様性も優れた発生方法が提案されているが、真性乱数と呼ぶことができるのは、物理乱数のみであると考える。統計数理研究所においては、1956 年に、国産商用計算機の第一号である FACOM128 に放射線を乱数源とする発生装置を世界に先駆けて接続したのを端緒として、1963 年、1971 年、1989 年、1999 年、2004 年と順次、ダイオードの熱雑音を乱数源とする物理乱数装置を接続した形で計算機整備を行って来た。1989 年までの三代にわたる装置は、信号を計数する方式で速度も遅く、形も大型であった。1999 年のものは日立と共同で特許を有しているが、信号を計量する方式に変更している。2004 年のものは、東芝製の市販品を用いているが、この製品の開発にも統計数理研究所は協力している。2001 年には東芝製の乱数ボードをそれぞれに装着したパソコン 100 台からなるクラスターシステムも構築している。

このように、物理乱数の発生装置開発において、統計数理研究所は、常に世界をリードして来た。さらに、高速、高性能にするために、昨年度から発生源・発生方式の再検討を行っている。また、FDK が開発した乱数チップを多数個使用した高速かつ小型の USB 接続型乱数発生装置も FDK と共同で開発し、8 月頃に市販を開始する予定である。また、乱数ポータルを構築し、乱数のダウンロードを可能にしている。ポータルサイトの記事については、近々、アップロードする予定である。

図 1　試験中の乱数発生ボード

図 2　FDK の乱数発生装置のサブボード

研究の国際交流について

中野純司

私が学生のときは､外国に行ってその研究のまねをするよりは日本で自分独自のものを作りだすほうが重要である､と言われており､私もそう考えていた｡もちろんこれは正しい意見である。ただし今では、これを真に受けすぎて若い時に積極的に国際交流の機会を持たなかったことを少々後悔している。

現在では、ほとんどの研究者（のみならずほとんどの人）が地球は狭くなったと実感していると思う。まず、情報の流通の面ではインターネットの普及により全地球規模で距離感がほぼ 0 になった。インターネット上の情報は瞬時に世界中を飛び交う。また、物理的な面では航空運賃の低価格化がある。私が 20 代で初めてアメリカへ行ったときには夏のボーナスをすべてつぎ込んだが、現在では時期を選べば数万円でアメリカやヨーロッパへ行ける。また、アジアの近隣諸国だと国内旅行より安い運賃も珍しくない。これらの影響によって、現在では国際研究交流の機会は非常に増えており、若い研究者も気軽に国際学会に出席している。それでも、真の交流はまだまだ不十分でより積極的な交流が必要であると思う。

私が実質的な研究交流を外国で行ったのは、40 才くらいになってしまってからであった。その後、徐々にその楽しさと有用性を実感していった。その中でも、特に多くのことをドイツのベルリンフンボルト大学の Wolfgang Haerdle 教授から学んだと思う。ある国際会議で私の発表したセッションの座長が彼であり、そのテーマについて討論したことがきっかけであった。彼のやり方には賛否両論あるが、彼が優れた学者、研究室運営者であることは間違いない。彼の学問に対する積極性、国際交流から多くのことを生み出す方法、セミナーの運営法などは私がそれまで経験したものとはずいぶんと異なっていた。そして、訪問者を研究室の運営会議に自由に参加させ、また研究室の全員と話すように誘導し、会議の記録を契約書と考えること、などを経験させてもらった。それらを実感できたことは非常に有用であった。優秀で積極的な研究者はそのような研究スタイルを自分で編み出すのかもしれない。ただ、私の性格からはそれは難しいと思うので、彼からそれらを学べたことは非常によかったと思っている。そして統計数理研究所に赴任してからは、それらの中でも有用なものを自分の研究スタイルに取り入れようとしてきた。すなわち、国内・海外の研究者と積極的に交流することを心がけてきた。そして、若い研究者が多くの海外の研究者と交流できるようにしたいと考え、そのための機会を提供できるようにした。

その活動のひとつとして 2008 年 12 月に統計数理研、日本計算機統計学会などの主催で IASC2008(http://www.iasc-ars.org/IASC2008/)という国際会議を開催する予定である。これは統計の学会としては歴史の古い International Statistical Institute (国際統計協会)の下部組織である International Association for Statistical Computing (IASC)の第４回世界大会と IASC のアジア部会の第 6 回大会を兼ねたものである。この分野の海外の研究者と日本の研究者の実質的な交流の場にしたいと考えている。

関数主要点の性質について

データ科学研究系　計算機統計グループ　清水　信夫

与えられた確率分布の密度関数を k 個の領域に分割するような各領域のある種の中心点として、クラスター分析における k-means 法と本質的に同様の規準に基づき、主要点(principal points)が提案されている[1]。一方、時間や地理情報などをパラメータとして系統的に観測されるデータを関数データとして解析する手法として、関数データ解析[4][5]が提案されている。

関数データに対するクラスター分析としては、関数クラスタリングが提案されており、様々な実データに応用されている。また、これと関連し、関数データに拡張した主要点[6]の研究も行われている(従来の主要点と区別するため、以下では**関数主要点**と呼ぶ)。Flury[1]は確率分布が存在する仮定のもとで主要点を定義しているが、関数主要点の場合もこれに対応し、ランダム関数[3]が存在することを仮定して定義されている。

従来の主要点およびその導出については、様々な理論的考察([1][2][7][9]など)および実データへの応用([8]など)がなされてきたが、関数主要点についても、いくつかの理論的考察ならびに実データへの適用がなされている[6]。また、関数主要点については、ランダム関数を正規直交基底で展開することにより、それらの係数により張られる空間における主要点の議論に帰着する [6]。しかしながら、特殊な確率分布に従うランダム関数における主要点の性質についての理論的考察や、その結果に基づいた実データの解析への応用については未解明の部分が多く、これらの内容の発展を主たる目的として研究を行っている。

参考文献

[1] Flury, B. (1990). Principal points. *Biometrika*, **77**(1), 33-41.
[2] Flury, B. (1993). Estimation of principal points. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, **42**(1), 139-151.
[3] Ibragimov, I. A. and Rozanov, Y. A. (1978). *Gaussian Random Processes*. New York: Springer-Verlag.
[4] Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47, 379-396.
[5] Ramsay, J. O., and Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.
[6] Tarpey, T. and Kinateder, K. (2003). Clustering functional data. *Journal of Classification*, **20**, 93-114.
[7] 清水信夫, 水田正弘, 佐藤義治. (1998). Principal Points の性質について. 応用統計学, **27**, 1-16.
[8] 村木千恵・大瀧　慈・水田正弘. (1998). 主要点解析法による極東夏期天気図の解析. 応用統計学, **27**, 17-31.
[9] 清水信夫, 水田正弘, 佐藤義治. (1999). Principal Points の対称性に関する定理について. 計算機統計学, **12**, 45-53.

パターンの待ち時間問題

統計数理研究所 数理・推論研究系 平野勝臣

1. 問題

正六面体に 0 が 3 ヶ所，1 が 2 ヶ所，2 が 1 ヶ所書かれているサイコロを投げる．0,1,2 の出る確率をそれぞれ $p(=1/2), q(=1/3), r(=1/6)$, $p+q+r=1$ とする．このサイコロを投げて，パターン 020 がはじめて現われるまでに何回サイコロを投げるか.

2. なぜ厳密分布か

020 が N 回目にはじめて現われたとすると，その平均値 $E(N)$ はマルチンゲールの考えを用いた方法で $E(N)=26$ を求めることができる (Li, S. ((1980), *Annals of Probability*, 8, 1171-1176)．この結果，このパターンが現われるまでの試行数は ”平均値 26 の前後 ”と考えるだろう．しかしながら，以下で述べる方法で調べると，N の分散は 1204 であることがわかり，平均だけの情報では，いつ現われるかの見通すことは難しい.

はじめて現われるまでの試行数が問題であれば分散の情報は有用ではあるが，十分ではない．またよりよい推測のためには N の分布を知ることが重要になる．

3. 条件付き確率生成母関数法

条件付き確率生成母関数の方法で求めてみる．N の確率母関数を $\varphi_N(t)=E(t^N)$ とし，直前の結果が i や ij で，そこから 020 が起こるまでの待ち時間の条件付き確率生成母関数をそれぞれ $\varphi^{(i)}(t)$, $\varphi^{(ij)}(t)$, $i,j=0,1,2$ とかく．020 を待つことから

$$\begin{aligned}
\varphi_N(t) &= pt\varphi^{(0)}(t)+qt\varphi^{(1)}+rt\varphi^{(2)}(t)\\
\varphi^{(0)}(t) &= pt\varphi^{(0)}(t)+qt\varphi^{(1)}+rt\varphi^{(02)}(t)\\
\varphi^{(1)}(t) &= pt\varphi^{(0)}(t)+qt\varphi^{(1)}+rt\varphi^{(2)}(t)\\
\varphi^{(2)}(t) &= pt\varphi^{(0)}(t)+qt\varphi^{(1)}+rt\varphi^{(2)}(t)\\
\varphi^{(02)}(t) &= pt+qt\varphi^{(1)}+rt\varphi^{(2)}(t)
\end{aligned}$$

の方程式系を得る．$\varphi_N(t)$ について解くと

$$\varphi_N(t)=\frac{p^2rt^3}{1-t+prt^2-pr(1-p)t^3}$$

を得る．これを微分すれば，平均や分散などの積率が求まり，また，これを級数に展開すれば N の分布が求まる．N の分布を図に示した.

条件付き確率生成母関数の方法で様々な待ち時間の分布の導出を試みている．工学など，さまざまな応用がある．DNA は 4 文字で書かれている．ある文字列が現われるまでどのくらいの文字を調べればよいか．もちろん，このような例では文字列は独立ではない．条件付き確率生成母関数法はマルコフ系列のような依存系列にも対応できる.

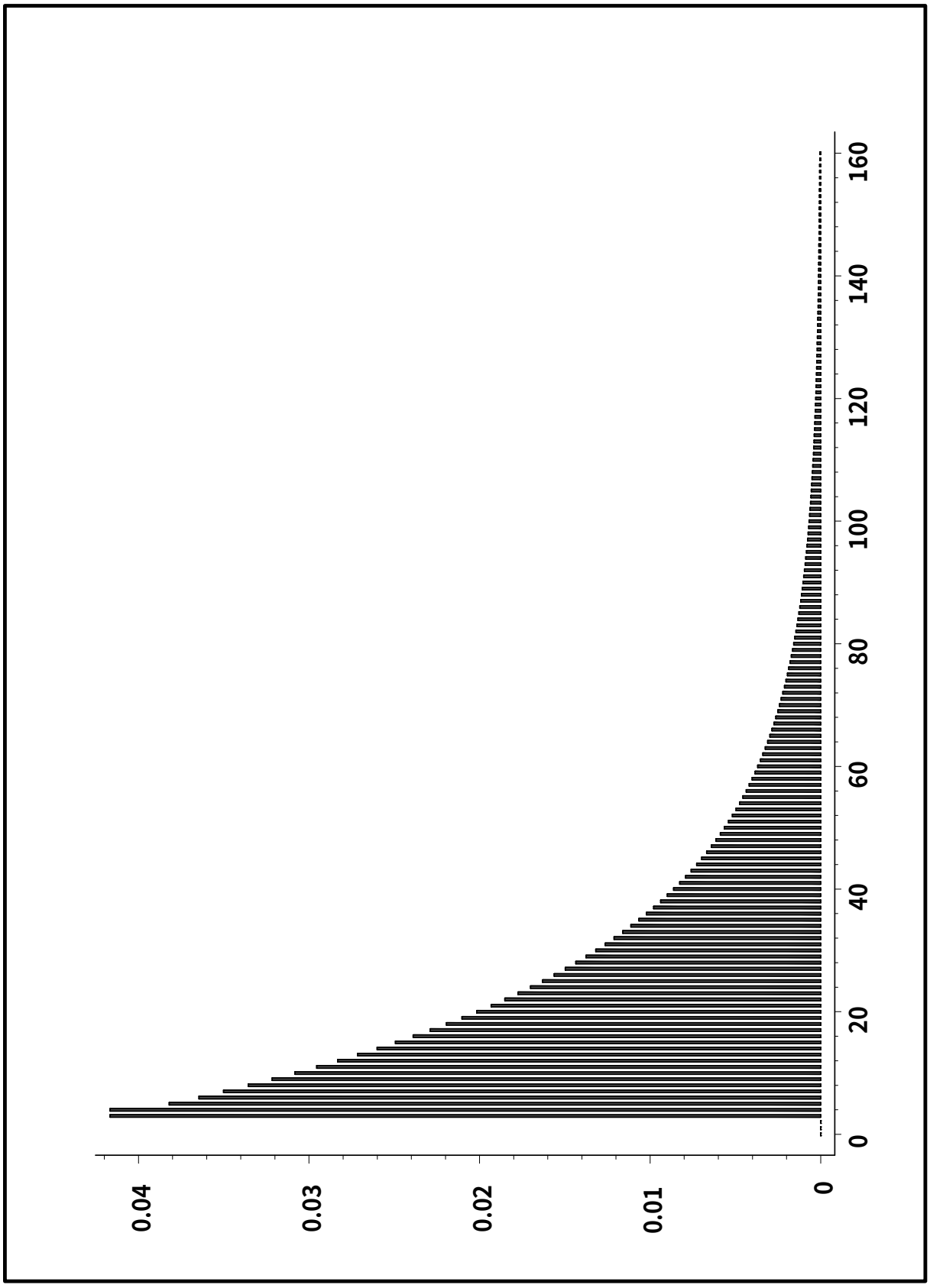


Figure 1: {0,1,2}-値独立系列で $(0,2,0)$ がはじめて出現するまでの待ち時間分布

「モスクワ訪問記」

数理・推論研究系　西山陽一

もう半年以上前のことになりますが、２００５年１２月に一週間ほど、モスクワ大学に出張いたしました。そのときの事を記したいと思います。

旧ソ連の崩壊後、ロシアは治安が悪いとの噂を聞いていたので、不安な気持ちで出かけました。事実、空港は国際空港とは思えぬほど殺伐とした雰囲気で、あまり歓待ムードという感じではありませんでした。ホテルの入り口でも金属探知機による検査があり、大学の建物も、職員ですらＩＤカードを提示しないと入れないという調子でした。でもチェックが厳しい分、かえって安心できるとも思いました。

受け入れのウリヤノフ教授に、学科のオフィスに案内していただいたときは、ある意味で衝撃でした。「我が学科にはオフィスが２部屋しかない。そのうちの１部屋はセミナー室として使っている。もう１部屋がこれだ。」と案内されたその部屋には、お茶を飲むテーブルが１台、ソファーがいくつか、そしてコンピュータが２台置いてあるだけです。あっけにとられて、「あなたの学科は何人ですか？」と聞くと、１０人だと答えます。つまり教授ですら専用のオフィス（机すら！）を持っていないのです。このような環境でも、ロシアは科学研究に頑張っています。私は東京でありがたくも専用のオフィスを持っています。もっと頑張らねばと思いました。彼らには失礼かもしれませんが、彼らのオフィスを見ただけでも遙々モスクワまで来た甲斐があったと思いました。

悪い面ばかりではありません。訪問先では講演をさせていただきましたが、確率論の創始者コルモゴロフ教授も講義をされた荘厳な講堂でやらせて頂けるという夢のような体験をいたしました。昨今のＩＴの発達で、発表形態も 黒板 → ＯＨＰ → ＰＣと変化してきており、私も時代の流れに乗り遅れまいとして最近ではＰＣで講演するようにしていますが、実は私は内心では伝統的な黒板による講義が一番かっこいいと考えるようなメンタリティをもっています。（古い考え方ですみません。）日本の統計学関係の学会では誰も黒板を使いませんが、モスクワ大学では、今も黒板が主流なのだそうです。であるにもかかわらず私は、心の準備ができていなかったので突然変更するわけにいかず、輝かしい歴史を誇る由緒正しい黒板で講演をする機会を逸してしまったのは、少し残念でした。

また、歴史上の偉大な科学者の肖像画が飾ってあるのにも感動しました。私は偏見で、ロシアではロシア人研究者だけが崇拝されているとばかり思いこんでいましたが、ニュートンもガウスも立派な額に入って最高級の敬意を払われていました。モスクワ大学には質実剛健という言葉が似合います。帰り際に、ウリヤノフ教授に「モスクワ大学は私にとって世界中で最もファンタスティックな場所のひとつだった」と正直な感想を述べると、彼はさも嬉しそうに、また来い、と言ってくださいました。２度目の計画はまだありませんが、私は知り合いの後輩にモスクワを推薦しました。

確率モデルの発見

伊藤栄明

確率論は現実の現象と深く関わりあっている。水に浮かべた花粉のこきざみな不規則運動は Brown 運動として知られている。Einstein (1905) は思考実験から Brown 運動の確率モデルを考えた(アインシュタイン選集 1、共立出版)。その研究は実験にもとづいた Perrin(1908) による Avogadro 数の計算にもちいられ、当時問題になっていた原子論への有力な根拠をあたえた。硬貨を投げて表が出れば正の方向に 1 歩、裏が出れば負の方向に 1 歩進むランダムウォークを連続化した確率モデルが Brown 運動であると考えることができる。Brown 運動の確率モデルは Einstein 以前に Bachelie(1900) により株価の変動の理解のためにもちられていたことが知られている(楠岡(2001), 数学通信、第 6 巻、2 号)。Bachelie の著作のタイトルは投機の理論であったそうである。確率論において賭博の問題はふるくから議論され破産の問題は確率論の教科書にのべられている。Bachelie は株価の変動の研究から数理の新しい分野に足をふい踏み入れた。Bachelie の研究も株式市場というものの発案がなければでてこなかったものである。経済活動以外にも人間社会の制度、技術の変化、等に関連して魅力的な確率論の問題が次々にでてくると考えられる。Bachelie や Einstein ほどでない研究者でも運よく面白い課題にであえば新しい数理の分野をひらけるかもしれないと思っている。Brown 運動ほど基本的でなくても確率論には様々な興味深い確率モデルがある。新種の確率モデルの発見を目標とする確率論の研究を進めたいと考えている。

Boltzmann のエントロピー増大の法則をわかりやすく説明する確率モデルとして, Ehrenfest のモデルがある。私は東アジアでひろく行われているじゃんけんと Ehrenfest のモデルを融合したじゃんけんモデルというものを 1969 年に思いついた。このモデルについて考えているうちに、Lotka-Volterra 系、集団遺伝学の Fisher-Wright モデル等に関連していることがわかった。モデルを拡張して行くと, 保存量を変数の数の半数もつ可積分系の確率モデル、というようなものになることに 1976 年に気づいたのだが、非線形可積分系という概念を知らなかったので、不思議な性質だと思った。特殊な確率モデルであるからそのような性質がでてくるのか、あるいはなにか深い理由があるのかがわからなかった。戸田格子（1967）は 1970 年代にすでに有名であり、私も名前は知っていたのだが、どのようなものかを知ったのは 1980 年代のなかごろであった。戸田格子についての M.Henon（1974）及び H.Flaschka（1974）の結果から非線形可積分系という世界があることがわかった。不思議さの理由がわかり、長年の謎がとけたように思った。じゃんけんモデルを常微分方程式系で近似したものは非線形可積分系としても典型的なものの一つであると考えられる。じゃんけんモデルは相互作用のある破産の問題としても興味あるものであり、破産の確率、共存の確率という問題について現在も考えている。モデルを発展させ、種分化の確率モデル、株価の振動の多数決モデル等についても数理的に研究している。多数決モデルからは自然に van der Pol 方程式が得られる。現在、国内、国外の優れた研究者との共同研究、協力において研究をすすめており、このことは大変な幸運であると感謝している。

現在の研究課題と関心のある分野 (2006年6月)

(論文あるいはテクニカルリポートがあるものは枠で囲んである)

数理・推論研究系
土谷　隆
(tsuchiya@sun312.ism.ac.jp)

P=NP?

計算複雑度
数値計算・最適化
アルゴリズム
データ構造

情報幾何

凸最適化アルゴリズムの計算複雑度を問題に関連する幾何学的量(「曲率積分」の一種)で評価

凸最適化のアルゴリズムと計算複雑度（方法論の核）

線形計画問題
半正定値計画問題
2次錐計画問題
凸2次計画問題
logdet 関数和最大化問題

多項式時間アルゴリズムの開発と理論的解析

非線形フィルタリング
スムージングのためのアルゴリズム

粒子フィルタ

ベイズ的推論
学習

情報量規準AIC

グラフィカルモデル
ガウシアングラフィカルモデル
（半正定値計画法の応用）

モデリング

マウスの遺伝子と社会性行動の解析（遺伝研との共同研究）

半正定値計画法による確率密度推定法の開発

2次錐計画法によるリニアモーターカーの磁気シールド最適設計（数千ー数万変数）

ガウシアングラフィカルモデルの時空間モデルへの適用
（数千次元正規分布の推定）

ガウシアングラフィカルモデルによる裁判官と弁護士の関係のデータ解析（極多数のモデル探索）

システム最適化と数理計画法

工学システムのみならず様々な社会システムにおける決定問題の多くは、与えられた条件のもとで目的を達成するようにシステムパラメータを決定する、いわゆる最適化問題に帰着します。システム最適化の数学的基礎を与えるのが数理計画法であり、1940 年代にオペレーションズ・リサーチの必要性を契機に始まった線形計画法の研究、特にダンツィックによる単体法の開発以来、２次計画法、非線形計画法、ネットワーク計画法、整数計画法、微分不可能計画法、２レベル計画法、確率計画法、多目的計画法など枚挙に暇ないほど、理論・計算手法ともに豊富な内容を持つ学問分野として体系化されてきました。特に，線形計画問題の新解法として 1984 年に発表されたいわゆるカーマーカー法を契機として急速に発展した内点法は，線形計画法だけでなく数理計画法の分野全体に大きな影響を与え，凸最適化の重要性を高めました。

一方、決定すべき変数が関数である場合の方がむしろ古い歴史を持っています。等周条件や微分方程式などの種々の制約条件のもとで汎関数を最小または最大にする関数を求める問題を変分問題といいますが、1696 年ベルヌーイによって提起された最速降下線に端を発する変分問題の研究すなわち変分法は、1750 年代にオイラー、ラグランジュらによって始められ、ルジャンドル、ヤコビ、ワイエルシュトラス、クレブシュ、ヒルベルト、ボルザ、ブリス、マックシェーン、ヘスティネスらにより、解析力学などとの関連において発展しました。1950 年代に入ると、ベルマンによる動的計画法やポントリャーギンによる最大原理を皮切りに、バーコヴィッツ、ガムクレリーゼ、ノイシュタット、リオンスらにより最適制御の理論が盛んに研究され、変分問題の内容は著しく豊かになりました。また 1960 年代にはカルマンによる最適レギュレータおよびフィルタの理論により状態空間における現代制御理論が確立されました。

変分法および最適制御の理論は無限計画法として一般化かつ抽象化され、最適解が満たすべき必要条件や十分条件が得られています。しかし、システムが線形で最小化すべき目的関数が２次のいわゆる LQ 問題や特定のフィードバック形式に限定した場合などを除いて、最適解をその必要条件から解析的に求めることは困難であり、一般の非線形の場合には数値計算に頼らざるを得ません。最適制御の数値解法はこれまで微分動的計画法など多数の手法が提案されてきましたが、最も汎用的でかつ効率的といえるのは非線形計画法に基づく解法でしょう。

例えば、連続時間の最適制御問題は、時間関数である制御入力を決定変数とし、常微分方程式や偏微分方程式などを介して、終端状態や過渡状態を評価する目的関数、また同様に終端状態や過渡状態に関する等式あるいは不等式制約条件を持つ無限計画問題であり、関数空間上の非線形計画問題として定式化されます。制約条件がない場合や終端状態に関する等式条件あるいは不等式条件のように有限個の汎関数制約条件のみが存在する場合には、関数空間における準ニュートン法、双対逐次２次計画法がそれぞれ有効です。状態軌道に関する不等式条件のように無限次元の制約条件が存在する場合は数値解を求めるのがより難しくなります。このような問題を解く一つの有力な方法は、制御入力あるいは制約条件を何らかの方法でパラメトライズすることにより離散化し、決定変数あるいは双対変数（ラグランジュ乗数）のいずれかが有限次元となる、いわゆる半無限計画問題に変換して解くことです。

半無限計画法は、歴史的にはチェビシェフ近似などのミニマックス問題を解くための手法として研究が始まりましたが、他にも分権的システムにおける資源配分、競争状況下での意思決定、多目的最適化、信号処理、制御系設計などに広い応用を持っています。半無限計画問題は、決定変数あるいは双対変数のいずれか一方が有限次元であるため、一般的な仮定のもとでは他方の変数が離散的になるという性質があり、これに基づいて等価な有限次元の非線形計画問題に書き直すことができます。半無限計画問題は、外乱などの不確実性を内包するシステムの最適化、すなわちロバスト最適化に関連し、近年盛んに研究されています。

（数理・推論研究系　計算数理グループ　伊藤　聡）

統計科学における制御理論の研究

数理・推論研究系　　宮里義彦

１．制御理論の背景

状態空間法による制御理論の体系化が一段落したのは1970年前後であり、そこから生まれた制御手法の１つが２次評価規範による最適レギュレータである。同時期に統計数理研究所が統計的制御として制御の応用研究（TIMSAC）を行った際も、制御方式はこの最適レギュレータであるが、研究所としてのは主眼はモデリングにあって、制御理論に関してはユーザーの立場にとどまり、基礎理論の観点から制御方式自体についての研究は行われなかった。その後も最適レギュレータ以降の制御理論の多大な発展について、一部の限定された制御手法を除いて、研究所としては関わりを持たずに今日に至っている。しかし制御系の構築においては、制御対象をどのようにモデル化するか、またそのモデルに含まれる誤差をどう評価するかによって、適用される制御手法や制御性能が規定される。従って高性能の制御系を実現するために、制御を意識したモデル化やモデル化誤差を考慮した制御という視点が重要で、モデル化と制御は切り離して考えられない。その意味から単に制御理論のユーザにとどまらず、モデリングと整合性のある制御手法を基礎理論の立場から研究していく必要性がある。

２．制御理論の研究

このような制御理論と統計科学の関係を考慮して、モデルリングと制御を同時に行う適応制御の研究を行っている。適応制御は制御器のオンライン調整のために制御系全体の安定解析が困難で、適用上の様々な制約を受ける。その制約を緩和し適応制御の適用範囲を広げる研究を進めてきた。現在は適応制御に関連して、非線形制御と線形制御の立場から研究を行っている。さらに関連する非線形$H\infty$制御やゲインスケジューリング制御、反復学習制御などの研究を行っている。また応用研究に関しても、共同研究で、スライディングモード制御と周波数整形法（$H\infty$制御）および双線形オブザーバを用いて車両のセミアクティブサスペンションの制御系設計を行い、実機（高速バス）による走行試験から良好な結果を得ている。

３．制御科学と統計科学

統計科学では、有限時間の現象を再現する開ループ的なモデルを構築することに主眼があり、制御はモデルの有効性を検証する項目の１つと見なされることが多い。しかし制御系を良好に動作させるには、すでに制御科学の分野では繰り返し指摘されてきているように、モデルと制御の総合的な考察が必要となる。それには制御科学の深い知見が必要であり、制御科学と統計科学の緊密な連係が必要になる。また制御理論がこれまで取り扱ってきた動的システムという枠組みを越えて、より広いクラスの離散事象システム、生産システム、通信ネットワークシステム等も対象として発展していくためには、さらにシステム科学や情報科学全般も含めた横断的な研究が今後ますます重要になっていくと思われる。

哺乳類の進化

長谷川政美

生物進化の歴史は、現在生きている生物のもつ DNA のなかに刻まれている。いろいろな生物の DNA の間の違いが、進化の歴史を反映しているからである。進化の過程において DNA の塩基が置き換わる現象をモデル化することにより、いろいろな生物の DNA 塩基配列のデータから進化の系統樹を推定することができる。このような研究分野は分子系統学と呼ばれ、DNA の配列データが比較的簡単に得られるようになってきた近年盛んになってきた。

分子系統学が最も活発に行なわれているのが、哺乳類のなかでもメスが胎盤をもった真獣類と呼ばれるグループについてである。真獣類には、われわれヒト以外に、サル、イヌ、ネコ、アザラシ、ウシ、ウマ、コウモリ、クジラ、ゾウ、ジュゴンなど実に多様な動物が含まれる。これらの多様な動物がどのように進化してきたかについて、多くのひとが興味をもつわけである。真獣類の進化について、最近明らかになったことは、図１で示すように、このグループの動物が進化的に大きく３つのグループに分かれることである。１つはボレオ真獣類と呼ばれるものであり、ウシ、ウマ、クマ、ハリネズミ、ヒト、サル、リスなどが含まれる。２つめはアフリカ獣類で、ゾウ、ツチブタ、ハネジネズミ、テンレックなどが含まれ、３つめの南米獣類にはナマケモノ、アリクイ、アルマジロが含まれる。これらの動物のグループが系統樹の上で枝分かれしたのは今からおよそ 1 億年前であり、南半球のゴンドワナ超大陸が分裂を続けており、アフリカと南米が分かれた頃であった。その頃、北半球にはローラシア大陸があった。第１のグループはローラシア大陸で、第２、第３のグループはその名の通りそれぞれアフリカ、南米が孤立した大陸であった時代に、そこで独自の進化を遂げたものと考えられている。このなかには、ボレオ真獣類に属するハリネズミとアフリカ獣類に属するハリテンレックのように、非常によく似た形態が独立に進化した例（収斂進化）も多く見られる（図２）。

このように、分子系統学は哺乳類の進化にとって、大陸移動が大きな役割を果たしてきたことを明らかにした。

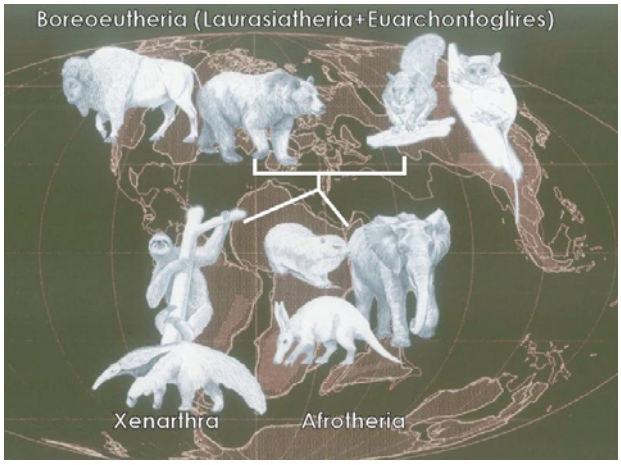


図１. 真獣類全体の系統樹.

図２. ハリネズミ（左）とハリテンレック（右）

ゲノム情報から進化のメカニズムを探る

足立 淳

（モデリング研究系、予測発見戦略研究センター・ゲノム解析グループ）

生命が持つゲノムは長い進化の歴史の産物であり、そこには突然変異に起因する進化の履歴が刻まれている。これまで生物間の相同な遺伝子の比較から系統関係を推定する分子系統学の発展により、生物の進化の道筋は系統樹として推定されてきたが、同時に幾つかの限界点も見えてきた。相同な遺伝子を比較するだけでは非常に短期間に種分化が起きた場合や、種分化の時に長い間にわたって種間交雑が続いた場合などでは、系統関係の推定が困難になってしまう。また系統関係がわかってもそれは進化の道筋が明確になっただけであり、進化のメカニズムについては何も知ることはできない。

近年のゲノム・プロジェクトから生み出される大量のデータは、進化的な視点から解析することによって、新しい意味を与えることができる。遺伝子の機能もまた進化の産物であり、これを理解するためにも進化的な視点は不可欠なのである。ゲノム情報の急速な蓄積により、さまざまな生物間でゲノム全体の比較ができるようになった意義は大きい。遺伝情報全体を扱えるようになったということは、情報が単に量的に増えたということではなく、質的にも全く新しい基盤の上に立った議論が可能になったということを意味する。部分だけを見た議論ではなく、全貌を把握しながら議論が進められるからである。

現在、ゲノム構造を比較することによって種間の遺伝的な相違を調べる研究が盛んである。種間の生物学的な機能の違いを、遺伝情報の違いとして把握できるからである。これをさらに一歩進めて、共通祖先のゲノム構造を再構築することを考えてみる。複数の生物間の共通祖先のゲノム構造が推定できれば、ゲノム構造の変化の歴史を記述することが可能となる。しかし個々の遺伝子とは異なり、ゲノム構造は突然変異の頻度が高くて速く変化するために、その変異を遡ることは簡単なことではない。特にゲノムの50%以上を占める反復配列と高頻度で起こる配列の重複によって、同じ様な配列がゲノム上に散在していることが問題を複雑にしているからである。

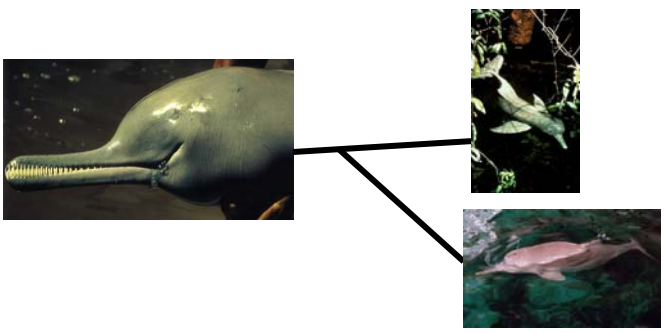
反復配列はゲノム構造の再構築には厄介な存在であるが、短期間に種分化を起こした生物間の系統関係を知る上では貴重な情報源となり得る。そこでゲノム構造の再構築のようなマクロ的な視点では反復配列を無視し、短期間に種分岐を繰り返した系統関係のようなミクロ的な視点では反復配列を重要なマーカーとして扱わなければならない。一方で、配列の重複は、遺伝子をコードしている部分に着目することによって、有用な情報を引き出すのに用いることができる。遺伝子の重複は遺伝子ファミリーの進化の引き金になるので、その歴史を遡ることはとても重要であるが、互いに似た配列が多数存在するために、単純に比較しただけでは重複の歴史的順番を推定することは困難である。なぜならば、ある遺伝子重複の後に一方の機能が変化すれば、その機能的制約によって進化速度が変化してしまうし、さらに遺伝子の機能が途中で失われると機能的制約が解かれて進化速度は急激に速くなるからである。つまり、遺伝子の進化速度の変化を無視した単純な配列比較では正しい結論が引き出せないのである。遺伝子の重複の順番を解明するためには、個々の遺伝子の機能的変化や機能消失を考慮した分子系統学的な手法を、新たに開発し解析することが必要である。

相同な遺伝子群の系統関係と、ゲノム上での互いの遺伝子の位置関係から、種間におけるゲノム構造の変異の歴史的順番を数値的最適化の手法を適用することによって再構築することを試みる。分子系統学とゲノム比較を高度に組み合わせることによって、ある遺伝子の進化の引き金となった突然変異が、ゲノム上で何時どのように起こったかを推定することができるようになる。こうして個々の突然変異が定着してきた歴史を解明することは、進化のメカニズムを知るための第一歩となるであろう。

クジラ目におけるカワイルカの進化

曹　纓

現在地球上には、細菌類からわれわれ人類に至るまで実に多種多様な生物種が生息しており、これらの生物種の見かけ上の多様性にも拘らず、DNA を基本とした遺伝的な仕組みが、調べられた範囲ではあらゆる生物で共通であることから、これら全ての生物は、もとを辿れば一つの共通祖先から由来したものであると考えられる。従って、地球上のあらゆる生物は、一本の巨大な系統樹の中に位置づけられる。このような作業をおこなう研究が、生物系統学である。従来、生物系統学は主に現存生物やすでに絶滅した生物の化石などの形態を比較することによって行われてきたが、近年分子生物学の発展に伴い、DNA や蛋白質などの解析から生物の系統進化を探る分子系統学研究が盛んになってきた。形態レベルでは、系統関係とは無関係に似たような環境に住む生物の形が似てくる“収斂進化”という現象がよく起ることであり、形態だけに基づいた系統樹推定は不十分である。そのために、形態とは独立に、分子データに基づいて、統計的モデルを用いる解析による系統樹推定法が望まれている。分子系統学の研究によって、従来の比較形態学からは思いもかけなかったような新しい説が近年次々に提唱されている、そのうち最も有名になったのはクジラ類が系統的には偶蹄類のなかに入ってしまい、カバと最も近縁だということである。クジラ目では歯クジラ亜目１０科６４種とヒゲクジラ亜目４科１４種がある。ヒゲクジラは発生の過程で歯が吸収され、替りに口腔の上顎からくじらひげを生やしていて、オキアミを主食しているグループであり、歯クジラ亜目はマッコクジラ、アカボウクジラや、イルカ類など口腔に歯を持ち、魚類を主食にしているグループである。更に歯クジラグループの中、淡水に生息するイルカ類で、現在４種それぞれ独立した科として分類されている。アジアにはヨウスコウカワイルカとインドカワイルカが生息しており、また南米ではアマゾンカワイルカとラプラタカワイルカがいる。形の非常に似た口ばしを持つカワイルカは単系統であるか或いは多系統であるかは興味深い問題である。現在野生の揚子江カワイルカは中国にしかいない、絶滅の危機に迫られている。中国科学院水生生物研究所の協力を得て、その貴重な揚子江カワイルカのミトコンドリア全ゲノムを決定することに成功し、最尤法を用いて解析した結果、ガンジスカワイルカは他のカワイルカや海洋性イルカから先に分岐したカワイルカ類に見られる形態的類似性は系統を反映されない、つまりカワイルカは多系統であることが分った。ガンジスカワイルカはマッコウクジラや海洋性のイルカなどほかの歯クジラ類とのあいだを埋める原始的なグループの生き残りであり、このグループはかつて世界中の海で繁栄し、その子孫がガンジス川流域で生き延びたインドカワイルカであると考えられる。



日常生活のマトリックス化計画

モデリング研究系＆予測発見戦略研究センター　樋口知之

唐突ではあるが，私は，キアヌ・リーブス主演の映画「マトリックス」（1999 年）が大好きである．何回見たか分からないぐらい気に入っている．これ以上によく見た映画は，ハリソン・フォード主演の「ブレードランナー」(1982 年)くらいであろう．両方とも近未来映画である点は共通しているが，私が講演や講義で引き合いに出すのはマトリックスの方のみである．どうして私の専門領域である統計科学とマトリックスが繋がるのか，不思議に思われる方も多いと思う．

まずは分かりやすい，両者間の技術的な関連性から説明したい．映画を見ていない方には申し訳ないが，「マトリックス」で最初に採用され，その後の映画や CM で度々使われるようになった撮影法をご記憶であろうか？天井から複数本のワイヤーで吊るした俳優の動作を，３次元的に配置した膨大なカメラ群でもって同時に撮影する．撮像データはすべてコンピュータに取り込まれ，コンピュータの中では“データにもとづく人体モデル”が構成される．こうすると，一瞬の動作もさまざまなアングルから自由自在に見ることが可能だ．これと同じアイデアは，ロボティクスの権威である金出カーネギーメロン大学教授の，バーチャライズド・リアリティプロジェクト，特に３次元ビジョンシステムにみることができる．データからの情報が不足していれば，人体に関する先験的な知識をデジタル情報として人体モデルの構成に利用する．このような，動的かつ複雑な対象を理解するために，同時多点計測（観測）データと先見情報を組み合わせ，コンピュータの中に近似モデルを構成するアプローチ，これは私の専門であるベイジアンモデリングそのものである．

研究対象として地球を考えてみよう．毎日，膨大な数の人工衛星，航空機，船舶，ブイ，地上観測点からの超大量のデータが天気予報機関に届く．スーパーコンピュータ上では，物理・化学プロセスを数値表現したシミュレーションモデルが常時活躍している．それでも気象予報，特に長期予報や局所予報は難しい．データからの情報だけでは高精度予報には全然力不足，一方，シミュレーションモデルは所詮近似モデルであって，未来永劫現実を忠実に表現することはできない．有効な解決策は，先験的情報，この場合シミュレーションによる数値演算結果と，超大量のデータからの両方の情報を活用すること．このデータ同化と呼ばれる情報統合作業も，ベイジアンモデリングの一つの具体事例である．

今，地球をとりあげて説明したが，ぐっと身近な社会生活の問題，例えばマーケティング研究でも，同様の研究スタイルが時代の潮流だ．POS データ，各種会員カード，電子マネー，IC タグ，インターネット調査等々，人々の諸々の日常生活をとらえるデジタルデータの集積は加速するいっぽう．これら膨大なデータと既存の経験など，ありとあらゆる情報にもとづく人間行動のモデル化，そして個人の嗜好にあわせたマーケティング戦略立案がマーケティング研究最前線の姿である．

映画「マトリックス」では日常生活すべてがデジタル情報としてコンピュータの中に埋め込まれ，その中で構成された近似モデル，『マトリックス』が情報と情報の“会話”する場である．ベイジアンモデリングによって築かれるコンピュータの中の情報空間，それはもう『マトリックス』のミニチュア版だとは言えないだろうか．

データを覗いて楽になろう：シミュレーションからデータ同化へ

上野玄太（モデリング研究系 / 予測発見戦略研究センター）

高校時代に習った、「未来を予言するのが物理だ」。舞台を設定して登場人物を紹介し、この初期条件でどんなドラマが展開するか。しかし現実的には、物体に働く力を正確に数え上げることなどもちろんできないし、初期条件や境界条件を適切に与えることすら不可能。そのため、理想的な状況では厳密に成り立っていた関係式が、近似的にしか成り立たなくなってしまう。私の研究課題であるデータ同化とは、データをこそっと覗いてその近似関係式に魂を吹き込む方法。近似式からステップアップして、合理的な予測が可能になる。

そういったデータ同化とはいったい何をする作業かというと、要するに観測データにモデルをあてはめることである。理科の実験で行った、直線のデータへのあてはめ。直線のかわりにシミュレーションモデルを用意してデータにあてはめる手法のことをデータ同化と言っている。

一方で、「未来を予言」を引き継いだ理念がシミュレーション科学にいまだ息づいている。すなわち、シミュレーションとは、確立された物理学の基本法則にのっとって物理の素過程を明らかにするための道具であり、観測データとは独立に進めるべきものである、というものだ。素過程の「素」は素粒子の「素」でもあることに関連してか、この理念には理論家としてのシミュレーション科学者のプライドが感じられる。また、データとつき合わせるにしても、シミュレーションは未来を予言できるはずのものであるから、シミュレーションの計算はすべて完了してから行うべきものともいわれる。ところが、データ同化とは、そういった理念とは反する姿勢の手法である。

理念に反してまでデータ同化を行うのにはそれなりの理由がある。その理由とは煎じ詰めると、シミュレーション結果は観測データを「正確に」再現していない点である。その原因は当然シミュレーションモデルの不備にある。シミュレーションを走らせる際の初期条件、境界条件、モデリングの際に無視した異スケールの物理、１、２次元性などの仮定、グリッドの大きさ、経験的公式、その他の使用が不備の内訳である。ところが、「正確に」再現しないと価値がないシミュレーションも存在する。天気予報が好例だ。そこで、従来のシミュレーションの理念には反するが、データを参考にしながらシミュレーションモデルを修正し、現象の「正確な」再現を図るというのがデータ同化の狙いである。本来は参照するべきでないデータを参考にするのだから、データ同化とはデータのカンニングプロセスが挿入されたシミュレーションといってよいだろう。シミュレーションモデルの不備の調整に血道をあげるよりも、カンニングできるところは潔く覗いてしまうことが、予測や解析といった本来の目的の達成の早道ではないだろうか。

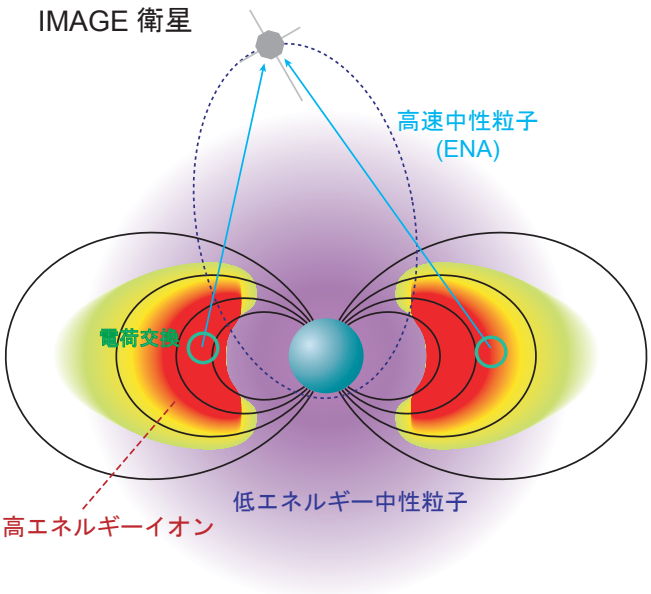
データ同化の宇宙環境科学への応用

中野 慎也（JST CREST 研究員）

地球周辺の宇宙空間は希薄なプラズマで満たされているが，地上数千kmから数万kmくらいのプラズマの運動が主として地球の磁場に支配される領域を地球磁気圏と呼んでいる．磁気圏中のプラズマの分布やダイナミクスは，主に人工衛星の直接観測から得られたデータを用いて研究がなされているのだが，限られた数の衛星による直接観測でグローバルなプラズマの動きを把握することは難しい．

しかし，NASA の IMAGE 衛星により，高速中性粒子を遠隔観測することで，磁気圏における高エネルギーイオンの空間分布に関する情報を2分に1回という時間分解能で得ることができるようになった．下図のように，高速中性粒子は，磁気圏に捕捉された高エネルギーイオンと地球近傍の低エネルギーの中性粒子との電荷交換によって生成されるので，低エネルギーの中性粒子の分布をモデルによって仮定すれば，高速中性粒子から高エネルギーイオンの分布が推定できるというわけである．

さて，磁気圏中の高エネルギーイオンの動きは，磁場・電場に強く支配されているため，高エネルギーイオンの分布は，磁場・電場分布を反映していると考えられる．そこで我々は，高速中性粒子のデータを数値モデルに取り込むことで，高エネルギーイオン分布に加えて磁場・電場分布も同時に推定しようと試みている．データを数値モデルに取り込むというやり方は，データ同化と呼ばれ，天気予報などの目的で発展してきた考え方であるが，その考え方を磁気圏環境のモデリングに応用するというアプローチで研究を進めている．



ターミエ・アレックサンドル研究紹介

データ・マイニング：スーパーから遺伝子まで

データ・マイニングは，データから新しい知識を見つけることである．データ・マイニングの一つの大切な課題は，データからパターンを発見することである．実世界の例をあげる：
データを，全国のライフスーパーの先月の領収書としよう．パターンは，顧客がよく同時に買う商品である．各顧客の領収書を観ると，顧客は違う商品を買っている．私もいつも違う商品を買っている！しかし，すべての領収書を見たら，顧客が買う商品のパターンが出る・・・以下の図を見てください．

領収書 A	領収書 B	領収書 C	領収書 D
パン ☆	**生したけ ※**	**肉 ※**	ラーメン
生したけ ※	**タレ ※**	ウインナ	ビール
ジャム ☆	鮭	**生したけ ※**	**パン ☆**
肉 ※	納豆	**タレ ※**	**ジャム ☆**
タレ ※	**肉 ※**	醤油	

※｛生したけ，タレ，肉｝はよく同時に買う

☆｛パン，ジャム｝　　もよく同時に買う

肉と生したけとタレはよく一緒に売っている．もちろん！焼肉パーティの時、その三つは必要だからである．パンとジャムもよく一緒に売っている：朝食でパンを食べる顧客向けである．その情報で，スーパーのマネジャーは売り物の置き場を考えられる．また，某店のマネジャーはタレを割引するが，肉の値段を少し上げる．すると，顧客はお買得と思って買う，しかし実は合計の値段が変わらない・・・
そんな背景があってデータ・マイニングは大人気になった.

しかし違う利用もある．例えば，最近バイオインフォマティクスは生き物のゲノムを課題にした．生物研究員はたくさんデータを集めた．データの数は多くて難しく人間では分析が出来ない．でもデータ・マイニングでデータの面白い所を見けることができる，そこから関係者は分析が出来る．そうすると，人間の遺伝子の知識が進む.
私の仕事は，特定のデータからそのデータ・マイニングツールを発見することである.

点過程の統計解析： 研究紹介と招待

尾形良彦 （統計数理研究所，総合研究大学院大学）

時間とともに変化するデータの統計解析は時系列解析と呼ばれ，統計数理研究所は世界をリードする研究成果を積み重ねてきました。私自身の専門は時系列解析に比べると研究者数でも極めて少数派で馴染みが薄いのですが，「点過程」と呼ばれるモデルによる統計解析です。

点過程は，突然に発生する事象を数学的に抽象した「点」の発生の確率メカニズムを記述する確率過程です。たとえば，

- インターネットなどでの顧客等のアクセス（サービス工学）
- 神経発火や心脈パルスのスパイク波列（生理学・脳科学）
- 損害・災害・事故・事件発生などの経済・自然・社会現象（保険数学）
- ハードやソフトシステムの故障・バグ（信頼性工学）や疾病発症・出生・死亡（疫学）
- 自然林の樹木，宇宙における銀河，銀河における恒星，の配置（自然，社会，環境）

などです。事象の発生時や位置を示す「点」に，発生規模（スカラー値）や諸特性（ベクトルや図形など）が付加された対象をマーク付き点過程と呼ばれています。データの解析を通して発生時間や位置，何らかの外因性の変量との因果関係を探し，将来の発生を予測することを目指します。点過程のシミュレーション法，推定法，モデル選択法，モデル診断法に関する私達の研究は先駆的なものがあり，国際的に高い評価をうけています。

地震国日本における豊富な地震カタログが利用可能な恵まれた研究環境のもとで，私の研究生涯の後半では主に地震活動のモデルと解析法に関して取り組んできました。地震カタログは発生時刻，震源座標，大きさ（マグニチュード），断層機構テンソルなどを編集した膨大なデータで，これは点過程モデルの研究の独創性の源です。この分野で，伝統的な地震活動解析に加えて統計モデルによる解析法の提案を積み重ね，統計地震学とも言うべき新しい局面を拓きました。これまでETASモデルなどの地震活動の各種点過程モデルを提案し，そのソフトウェアは国際地震学地球内部物理学会やアメリカ地震学会などを通して世界中に提供されています。また政府の地震調査委員会で検討された余震の確率予測や活断層データに基づく直下型地震の長期確率予測の実用化にあたっては，私達が提案した点過程解析法も採用され，日本（気象庁）やカルフォルニア（合衆国地質調査所）で実施または参照されています。

データが膨大であればあるほど，その隠れた本質的な情報を十分汲み取るために時間的・空間的に非定常・非均質なモデルを考慮する必要があり，大規模な統計モデルの研究が避けられない様になってきました。逆問題や時空間モデルなどの大量のパラメタを必要とする大規模モデルはベイズ法と不可分になり，計算機をフルに利用する最適化法やマルコフ連鎖モンテカルロ法などの応用技術開発，推定されたベイズモデルを表示する多次元動画像解析法などの情報学・数値解析などとの境界分野に及ぶ研究の比重も高くなってきました。必要上取り組んだこれらの技術的研究でも世界をリードするものと評価されました。

最初に述べたように点過程モデルの応用分野は多岐に亘っています。点過程の統計解析に関して，統計地震学でも他の応用分野や基礎理論でも，独自の新境地を拓きたいと考える学生諸君は私の研究室を訪れてください。

研究紹介：月齢と地震発生の相関について

予測発見戦略研究センター・地震予測解析グループ　プロジェクト研究員　　岩田貴樹

潮汐とは，太陽や月などの引力により，海水が周期的に満ちたり引いたりすることであり，多くの人に馴染みのある自然現象である．しかし，同様の現象が地球の固体部分（固体地球）にも起きていることはあまり知られていない．海水同様，固体地球も周期的に変形しており，例えば，我々の立っている地面は，潮汐により数十cm程度上下変動している．

とはいえ，潮汐が地球内部に引き起こす力（応力）は，地震を引き起こすのに必要と考えられる応力に比べればごく僅かなものである．しかし，地球内部に蓄積された応力が地震発生の限界に近い状態に達していれば，潮汐による応力が「最後の一押し」として作用し，地震発生を誘発する可能性がある．このような考えに基づき，潮汐と地震発生の関係を調べる研究がいくつか行われている．

潮汐の周期としてよく知られているのは，海水の満ち引き（満潮・干潮）に対応する約半日または1日のものであるが，満ち引きの変化の大小（大潮・小潮）に対応する約半月または1月のものも存在する．これは月齢と関連しているので，月齢と地震発生の相関を調べることで，潮汐の大小と地震発生の関係を調べることが出来る．

丹波山地は近畿地方では有数の，微小地震（マグニチュード3程度以下の，殆ど体で感じることのない地震）活動が盛んな地域である．この地域の近傍では，1995年に兵庫県南部地震（阪神・淡路大震災）が起きている．その兵庫県南部地震発生後，約2年間に起きた丹波山地での微小地震の発生時刻を月齢に直し，それを度数分布にしたものが図(a)である．分布には偏りが見られ，新月または満月（横軸の0˚または180˚）の後，度数すなわち地震の発生数が増えているように見える．この偏りが統計的に有意なものであるかどうか，「点過程モデル」と呼ばれる手法を用いて解析を行った．モデルとして，月齢と地震発生の相関が「ある」とするものと「ない」とするものの2つを考え，赤池情報量基準（AIC）という統計的指標を用いて，どちらのモデルが実際の地震発生時系列によく合うかを調べた．その結果，月齢と相関が「ある」とするモデルの方がよいモデルであり，「ない」としたモデルとの差が有意であることが分かった．また，点過程モデルから推定された月齢に対する1日当たりの地震の期待発生数の増減値を図(b)に示してある．図(a)と似通った様相を示しており，妥当なモデル化を行ったことが分かる．

また，比較のために，兵庫県南部地震が起きる前，約2年間について同様の解析を行ったが，月齢と地震発生の間に有意な相関は見られなかった．兵庫県南部地震を境にして，月齢と地震発生の相関の有無に違いがあることから，丹波山地の応力状態も兵庫県南部地震前後で違いがあると推測出来る．実際，他のデータや研究から，兵庫県南部地震を起こした断層運動によって，丹波山地の応力が高められたことが確かめられている．

このようにして月齢と地震発生の相関を調べることは，地球内部の応力状態を知る一助となり得る．様々な地域に対して解析を行うことで，応力状態の高い地域を判別し，近い将来地震の起きそうな地域を予測する可能性を秘めている．

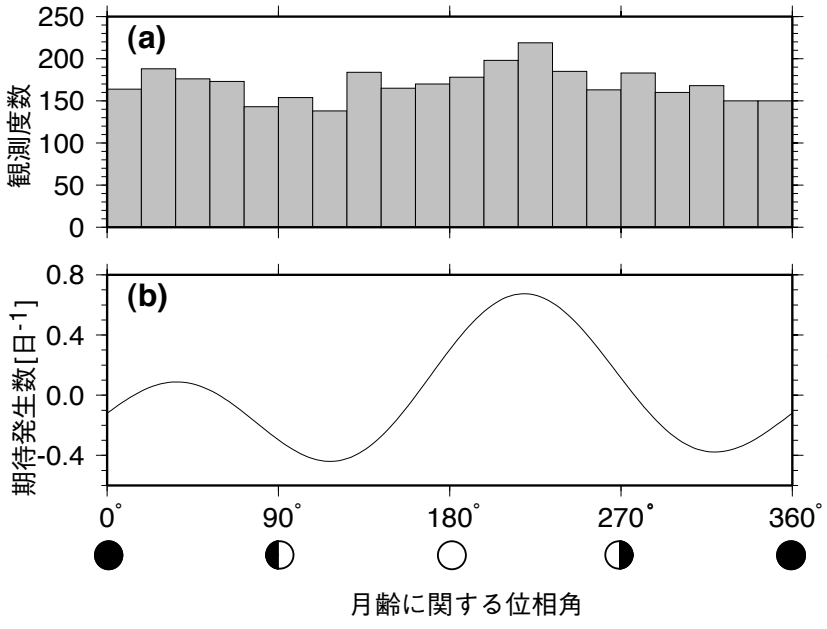


図:(a)兵庫県南部地震後約2年間（1995年1月31日から1996年12月11日）での，丹波山地における微小地震の発生時刻に関する月齢別ヒストグラム．(b)点過程モデルから推定された，月齢によって影響される1日当たりの地震の期待発生数の増減値．横軸は新月が0˚または360˚，満月が180˚となるように月齢を位相角に直したものとして示してある．

パターンインフォマティクスを用いて

将来の地震の発生場所を予測する研究

統計数理研究所・地震予測研究グループ

プロジェクト研究員

楠城一嘉

私は，地震の確率予測をする新しい手法の開発と，それを日本に適用して地震予測をする研究を行っています．パターンインフォマティクス（Pattern Informatics）法（または, PI 法）と呼ばれるこの手法を用いて，過去の地震活動を解析すると，将来の 10 年以内に大きい地震が起こる確率の高い地域（“ホットスポット”と呼びます）を示す予測地図を作成することが出来ます．そこでこの手法が実際の地震予測に有効かどうかを調べるために, 2000 年以前に起きた中部日本の地震活動を解析し，ホットスポットを示す予測地図を作成しました．次に，その地図と，2000 年以降に大地震が発生した場所を示す空間分布を比較しました．その結果，高い割合で，大地震の発生場所とホットスポットが対応していることが分かりました．これまでの研究では，活断層調査等のデータを用い，確率的な手法で長期的な地震発生予測がなされてきたのに対して, 10 年という中期的なタイムスケールで地震の発生場所の予測を試みている本研究は，大変野心的な試みとして位置づけられています．

学習推論グループの紹介

数理・推論研究系 学習推論グループ
江口 真透

私たちの研究所が改組して早 2 年目となった．私は**学習推論グループ**に参加し，南さん，池田さん，藤澤さん，伏木さんたちと共に研究活動を行っている．理論的な考察が得意なメンバーが集まったので基本的には個人研究が主体であるが，その個人研究の「相互作用」が最も期待されるところである．色々議論もあり必ずしも全く同じ考えでまとまるわけではないが，そこが未知の可能性を残すところであり，興味深いところでもある．私が考えている学習推論グループとは，データから学び，そこから新しいことを知り，それを検証する「良い方法」を作ることであり，そのための研究グループである．

現在の学習推論グループの主な活動は，毎週開かれている文献紹介セミナーである．これには，栗木さん，川喜田さんも参加され，かなりの力が注がれている．当番になった人は，選んだテーマの最新の文献の動向をかなり詳細に紹介する．結局は広範な紹介というよりは，高度な数理の展開を含む，かなり突っ込んだディスカッションがなされている．時には，非常に批判的な意見も飛び交うが，そのプロセスを通して自分たちの思い違いや気がつかなかった考えに遭遇して，理解の増進に役立つ．

また，昨年の改組の際，予測発見戦略センターの 4 番目のプロジェクトとして「**遺伝子多様性解析**」を開始した．これは，現在驚異的な速度で生産されているゲノムデータに対して，学習推論の立場から，このデータに特化した新しい統計方法を開発することを目指している．ゲノムデータの持つ新奇さに苦悩し，常に研究興味が更新される．果たして，より普遍的な「学習推論」としてどういう形態に昇華されるのだろうか．いやきっと近い将来，その新しい概念とそれを具体化する方法が私たちの研究グループの中から提案されるに違いないという夢に向かって，自分に課せられ日々の努力を継続したい．

研究紹介「変化点問題と幾何学的方法」

数理・推論研究系　栗木哲

〔連鎖解析と変化点問題〕

図 1 は，実験交配されたマウスにおいて高血圧の原因となる遺伝子の場所を探すための図（ロッドスコア）である．この図の横軸は，19 本の常染色体と 1 本の性染色体上の遺伝子の場所（遺伝子座）を，また縦軸はその位置の遺伝子（あるいはその位置の近くの遺伝子）が高血圧症の原因遺伝子である “もっともらしさ”（尤度）を意味する．この図において染色体 4 の中央付近に第 1 のピークが，また染色体 1 の中央付近に第 2 のピークが見てとれるが，これら両者のピークがデータのランダム性によるものではなく真に原因遺伝子を意味するものかどうかは，十分な吟味が必要である．統計解析の立場からは，ピークが原因遺伝子によるものと判定するための閾値の設定や，またそのピークが原因遺伝子によるものであると判定されたときには，その原因遺伝子の位置の信頼区間を与えることが必要となる．

ここで説明した統計的問題は変化点問題とよばれるものの典型的な例である．同種の問題として，画像データからノイズに埋もれた信号を検出する問題や，経済時系列データにおいて背後にひそむ構造変化を検出する問題がある．

〔積分幾何学の利用〕

このように変化点問題はいろいろな分野であらわれる．しかし変化点問題が必要とする統計モデルは，特異モデルとよばれる範疇の取り扱いの難しい統計モデルである．特異モデルに対しては，正則モデルに基づく通常の統計理論が正しく働かないため，データ解析の処方箋は十分には与えられていない．例えば，モデル選択のための AIC も用いることができないことが知られている．

この特異モデルについては，近年になって主として理論的な立場からの研究が進められている．そのなかには代数的なアプローチやチューブ法（オイラー標数法）とよばれる積分幾何学的なアプローチがある．後者のアプローチは，脳画像データ（fMRI データ）のピーク同定にも用いられている（マッギル大ワースレー教授）．現在筆者はチューブ法などを用いて，ロッドスコアのピークの検出などの変化点問題に関する研究を行っている．

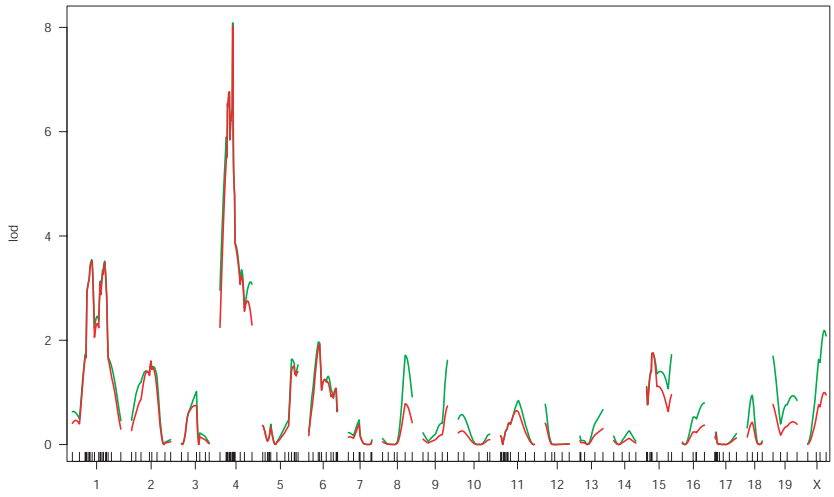


図 1. マウス高血圧症データ（ロッドスコア）

「理論研究とデータ解析」

数理・推論研究系　南　美穂子

ここ数年、マグロ漁による混獲データの解析に関連した研究を主に行っている。これまでもデータ解析に根付いた研究を行いたいと考えてきたが、自分の研究対象となるデータとの出会いがなかなかなく、多くの場合は理論的な研究を中心にその解析例として適当なデータを探して結果を示すという形で研究を進めてきた。研究テーマは、線形混合モデル、欠測データの解析、ラグランジュ分布族(カウンターで客がいなくなるまでに来た客の数の分布として特徴付けられる)、多変量逆正規分布、独立成分分析などで、分野の一貫性はなく、そのときどきで興味を持ったテーマを数年研究するという、移り気な研究スタイルを取ってきた。

混獲データの研究は、ヒストグラムの出力を壁いっぱいにいくつも貼った、国際組織の水産研究所に勤める友人の研究室を訪ねたことから始まった。そのヒストグラムはゼロの値が飛びぬけて多いという、日頃よく見るヒストグラムとは違った特徴をもつものであった。データの説明を聞くうちに、これは混獲が起こりえない状態と起こりうる状態の２つがあって、この２つの状態の分布が混じったものであると思った。そこでこのことを陽に表現するモデルを用いた解析を提案した。
提案した方法と従来のいくつかのモデルとの解析結果を比較するうち、従来のモデルでゼロの多いデータを解析すると、混獲数の減少傾向を過大に示す推定結果が出ることに気付いた。これはそれまで自分の持っていた知識、あるいは思い込みに反することであり、なかなか理由が考え付かなかったのだが、データを眺め推定方程式を眺めているうちに、急に視界が開けるように、なぜそうなるのかがわかった。
データ解析をしてはじめて気付く問題点がある。また、データからどのようなことを知りたいか考えることによって有用な解析方法を新しく考え出すことができるのだろう。逆に、以前、あの理論研究をしていたから、この方法や説明が考え付いたのだと思うこともしばしばある。

混獲データの研究は、今後しばらく続けたいと思っている。次の課題は混獲組成の解析である。解析方法を検討していたら、これまで行ってきた分布論や独立成分分析の研究との関連が出てきた。それぞれの研究をしているときには、この２つの研究テーマに接点があるとは考えたこともなかった。これまで半ば気まぐれで研究テーマを選んできたが、混獲データの解析をきっかけに関連性を再認識することになり面白いと思う。

池田思朗 (数理・推論研究系 学習推論グループ)

携帯電話や無線放送などの通信，あるいは医療で用いられる生体計測機では，しばしば雑音や他の信号が混ざることにより情報がうまく伝わらないことがあります．具体的な要因は様々ですが，完全には取り除けないことが多いでしょう．このような場合に，雑音が確率的な振舞いをするとみなして確率に基づく推論を行ない，その結果の応用を目指す研究を中心に行なっています．以下ではこれまで得られている 2 つの事柄について以下に示します．

1. 脳磁計による脳活動の計測について

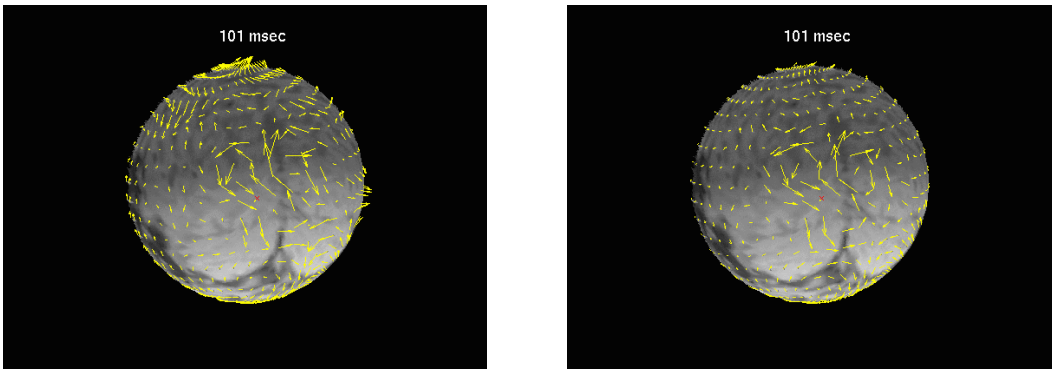


図 1　脳磁計データの改善 (左: 処理前，右: 処理後)

脳磁計 (MEG) は脳の活動による微弱な磁場の変化を計測します．信号が微弱なため，雑音が非常に大きくなります．我々は独立成分解析を行なうことで雑音を軽減し，脳活動を推定するための手法を提案しました．

2. 自動車における衛星放送の受信について

本研究は，NHK に所属し，総合研究大学院大学の博士課程学生として本研究所に所属している浜田正稔 (D2) さんとの共同研究です．

図 2　自動車における NHK 衛星放送受信信号の改善の様子

図 2 は NHK 衛星放送の受信信号を確率推論によって改善した例です．

左は車載システムによって受信した画像をそのまま表示したものです．画像が大変乱れています．これは車の速度によるドップラーシフトの影響と，衛星からの信号が建物などにより反射した影響によるものであると考えられます．ドップラーシフトの大きさと反射の影響を推定し，その推定結果を基に確率推論をしたのが右の図です．非常に鮮明な画像が得られていることがわかります．

現在は計測した結果を研究計算機上で再現し結果を示しています．しかしこの結果を得るための計算量は少なく，実際の車載システムとして実現可能な技術です．

外れ値への対処・ハプロタイプブロック同定

数理・推論研究系　藤澤 洋徳

私は理論にも応用にもどちらにも興味を持っています．その興味の中から，理論研究と応用研究を，一つずつ挙げて紹介したいと思います．

[外れ値の割合が多い場合にもバイアスが小さいロバスト推定]

統計モデルにおいては，しばしば，何らかの意味のパラメータを推定します．代表的で汎用的に使われているパラメータ推定方法の一つが最尤推定です．ところが，最尤推定は，一般的には外れ値に弱く（図 1），最尤推定の代替として，外れ値に強い，様々なロバスト推定が提案されています．ただし，過去のロバスト推定は，外れ値の割合が多くなると，外れ値に引きずられにくいという利点はあるものの，ある程度のバイアスは常に内在している，という欠点を持ち続けていました．

しかしながら，「外れ値とは何であろうか」ということをきうまく定式化すると，ダイバージェンスの上でロバスト推定を議論することができて，さらに，これまできちんと議論できなかった「外れ値の割合が多い場合」も自然に議論できるようになるのです．そして外れ値の割合が多い場合にもバイアスが小さいロバスト推定が提案できます（図 1）．

[ハプロタイプブロック同定]

現在，人のゲノムデータが着々と蓄積されています．そして，ゲノムデータを観察することで，薬効・副作用・疾患，などとの関連が分かり，治療方法の改善などが期待されています．そのような目的を達成するための一つの方法が相関解析です．相関解析において使われるゲノムデータの代表例の一つが一塩基多型（SNP）です．

一塩基多型に基づいた相関解析が行われているうちに，単純に一塩基多型に着目するよりも，連続した一塩基多型から作られるハプロタイプ，さらには，ハプロタイプを適当なブロック単位に分割したハプロタイプブロックに基づいた相関解析がより有益であると考えられてきています．我々のグループと癌研究会のグループは，遺伝統計学の知見と数理統計学の知見を組み合わせて，新しい同定方法を作りました（図 2）．

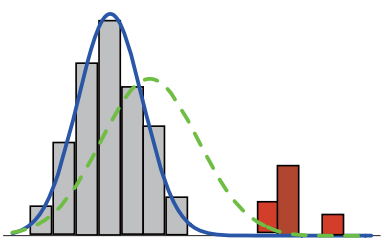


図 1: 最尤推定（点線）と提案しているロバスト推定（実線）の比較．ヒストグラムは得られているデータに対応している．本来は薄い灰色の部分のヒストグラムに対応する滑らかな密度関数推定が欲しい．ところが現実には濃い灰色のような外れ値が存在していた．点線は外れ値に引きずられている．実線はうまく薄い灰色部分だけを捉えている．

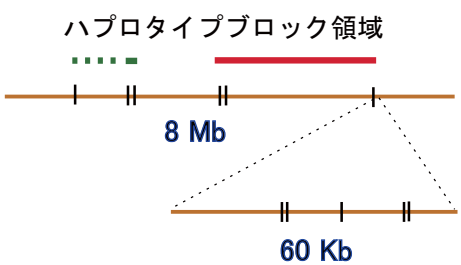


図 2: TAP2 遺伝子領域．縦棒は 11 箇所の一塩基多型の場所．ハプロタイプブロック領域は，実線部分は生物学的に検証されている．我々の方法はデータ解析の観点からこの構造をうまく同定できた．

データ科学雑感

数理・推論研究系　伏木 忠義

本稿では，従来の統計学を含むデータを扱う科学全般をデータ科学とよぶことにして，データ科学について私が普段思っていることをまとめてみたいと思う．

これを読まれている方が「統計学」についてどのような印象をもたれているかわからないが，私は，大学学部生のころ，統計学というと古くからあり，もう完成された学問という印象をもっていた．しかし，データ科学を研究するようになってみて，そのような「思い込み」は間違いだったと感じている．私の知る限りでもデータ科学は 1990 年代以降も大きな発展を遂げている．

1990 年代の統計学における大きな発展のひとつは Bayes 法の各方面への応用と考えられる．Bayes 法は，有益な事前情報がある場合や複雑な現象をモデル化する際に有用な方法論であるが，これまで計算不可能だった事後分布の計算が計算機の進歩により Markov Chain Monte Carlo 法を用いることで実現可能となり，大規模モデルでも Bayes 法を用いることができるようになった．

機械学習において，1990 年代に大きく発展し，現在も発展しつづけているテーマのひとつは，カーネル法である．従来の統計学においては，多変量解析の方法として，線形モデルを用いた方法論が提案されていたが，現実のデータは，多くの場合，非線形構造を含む．カーネル法は，データを高次元の特徴空間に射影して解析を行う方法であり，判別におけるサポートベクターマシンやカーネル主成分分析など多方面で用いられている．その他にも機械学習の分野では，独立成分分析やデータ多様体を用いた非線形次元削減法など多数の興味深い手法が提案されている．

また，現代のデータ科学の特徴のひとつは扱うデータそのものがこれまでのものと大きく異なっているということである．現代のデータ科学では，これまで考えられなかったような高次元データの解析が重要になっている．特に，ゲノムデータに代表されるように高次元だがサンプル数が少ない状況での解析を求められることが多くなっている．このような状況では，伝統的な統計学の方法では有効な結果が得られず，さまざまな新しい方法が提案されている．

ここで，私自身の研究についても少し触れたい．私は，モデル混合を用いた予測についての研究を行っている．パラメータ数に比べてデータ数が少ない場合など，推定量や選ばれたモデルが，データの微小な変化によって大きく変化する場合がある．モデル混合を用いる予測は，推定量や選択されたモデルの不確かさを減らすため，モデルを平均化する方法であり，計算機の発達により 1990 年代ごろから統計科学や機械学習の分野で研究されており，現実のさまざまな例でその有効性が示されている．

このように，私の知る限りでも，今日のデータ科学では，これまでなかったような新しいデータ，計算機性能の格段の向上，半正定値計画法などの計算可能な最適化手法の発展などにより，新たなモデル，新たな方法論が提案され，既存の統計学では扱えなかったものが解析されるようになってきているのである．私自身,「これからの」データ科学に少しでも寄与できる研究を行っていきたいと考えている．

ブースティング法とカーネル法: 統一的枠組みからの研究

川喜田 雅則

統計的判別問題について研究を行なっている. 統計的判別問題とは例えば今日の温度, 湿度などの情報から明日の天気を予測するように, 関連があると思われる情報 (特徴量) が与えられた時にそれから興味のある情報を予測する問題である. この問題は広範囲な応用が可能であり古くから盛んに研究されており, かつ現在でも重要な研究課題である.

統計的判別法としては古くには線形回帰を応用したものや, Fisher の線形判別法, ロジスティック回帰などがある. これらに対して近年統計的学習理論の枠組みから全く新しい手法であるバギング法, ブースティング法, サポートベクタマシンに代表されるカーネル法などが提案された. いずれも従来の統計的判別法とは異なるコンセプトに基づいた興味深い方法となっている.

最初にバギング法とブースティング法を簡単に紹介する. Breiman (1996b) では現存するいくつかの強力な判別手法はデータのわずかな揺らぎに対して過敏に反応して構築する予測方式がまるで異なるものになること (不安定さ) を指摘した. このことはノイズが高いときやサンプル数が少ない時には得られたデータに過剰適合 (オーバーフィット) して汎化性能が悪くなることを示唆している. Breiman (1996a) は与えられた学習データからブートストラップにより疑似的に独立な学習データを複数作成し, それぞれのデータから構築された予測方式の予測を平均化することで汎化性能をあげることができるバギング法を提案した. さらに Freund and Schapire (1997) はさらに上の方法で bootstrap の代わりに判別機が予測に失敗した標本を高い確率で resampling する (reweighting) ことにより作成したデータセットを用いて同様の予測方式を構成するブースティング法を提案した. この名前は元の予測方式の能力を飛躍的に向上 (boost) させることに由来する.

次にカーネル法を紹介する. 一般に判別境界が線形のような単純な形をしているとは限らない. しかし生の特徴量空間で複雑な形をしている判別境界でも, 各特徴量を適当な可算無限個の写像により高次元に持ち上げるとしばしば線形判別境界で識別できる. ただし直接これを実行しようとすると無限個の特徴量を扱う必要があり実際には不可能である. ここでいくつかの線形判別法は特徴量そのものの値ではなく, 特徴量ベクトルの内積がわかれば十分であることに着目する. カーネル法ではそのことを利用して無限個の写像を全く意識することなくその内積を表す簡単なカーネル関数を与えることによって容易に計算を可能にしている. これをカーネルトリックという.

上述の二つの方法は全く異なる方法に見えるが, その考え方には多くの共通部分がある. 例えば両手法共に複雑な判別境界を達成するために, 特徴量を高次元空間に射影した上で線形識別を行なっている. また一般に近似能力を高め過ぎるとオーバーフィットにより汎化能力が落ちることが知られているが, これを避けるためにカーネル法ではしばしばパラメータに中心 0 の正規分布を事前分布として設定し事後確率最大化を行なう. またブースティング法では高次元への射影を一度に行なうのではなく, 逐次的にとってくることでほとんどの射影の係数を 0 に押えている. これはパラメータの事前分布として 0 のまわりの両側指数分布を設定した正則化と関連が深い (Efron et al., 2002). 本研究ではこれらを統一的な視点でまとめなおすことで互いの長所を持つような方法を提案すること, またそれぞれについて得られている理論的知見を互いに輸入して新たな知見を得ること, また漸近的にこれらの手法がどの程度良いかなどを示すことなどを目指している.

References

Breiman, L., 1996a. Bagging predictors. *Mac e Lea* 26, 123–140.

Breiman, L., 1996b. The heuristics of instability in model selection. *A a S a c* 24, 2350–2383.

Efron, B., Johnstone, I., Hastie, T., Tibshirani, R., 2002. Least angle regression. *A a S a c* 32 (2), 407–499.

Freund, Y., Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1), 119–139.

個人情報の保護と活用推進：そのバランス

藤田利治

個人情報保護法（「個人情報の保護に関する法律」）が 2005 年 4 月から全面施行されたが、その後も官民のさまざまな機関からの Winny などを通じた個人情報流出が相次ぐ発覚が続いており、社会的問題となっている。その一方で、「個人の権利利益の保護」を盾にした不適切な匿名化や活用制限といった過剰反応に対する批判もある。

個人情報保護法は、第 1 条（目的条項）の通り、「個人情報の有用性への配慮」（活用）と「個人の権利利益の保護」のバランスを図ることを目的としている。しかし、一般的に想像されがちな保護の対象とは、目先の「現在」に力点が置かれたものであり、またプライバシーなどの必要以上の「私益」のみに限定されたものとなりがちのように思われる。これに対して、「現在」と「私益」のそれぞれについて視野を広げ、「将来」と「公益」の要素を取り入れようという見直し論が熱を帯びてきている。

健康にかかわる領域における情報の蓄積・活用の制限は、どのような事態を招くのであろうか。健康に対して有害な問題が発生した際にも、問題にかかわる実態を迅速に把握して有効な対策を効率的に実施することができずに、国民が長期にわたって有害な状態にさらされたり、取り返しのつかない事態を招いたりする危険が高まる。あるいは、サービスや予防対策についても科学的根拠なしに実際に広く用いてしまうと、無効であったり有害ですらあったりするサービス・対策に多くの国民がさらされる危険が増すことになる。情報の活用制限の代償は、こうした不利益を甘んじて受け入れることを意味しているであろう。

医薬品については、欧米のいくつかの国では数十万、数百万人規模の経時的な医療情報を持つデータベースが構築されて、市販後医薬品の安全性確保などの研究に活用されている。たとえば、米国のMedicaidやHMO（Health Maintenance Organization）データベース、カナダのSaskatchewan州のほぼ全住民をカバーするデータベース、英国の General Practice Research Database 、副作用自発報告のデータベースなど枚挙に遑がない。（http://www.pharmacoepi.org/resources/summary_databases.pdfなど参照）。国際薬剤疫学会が 1989 年に発足し、医薬品のリスク評価にかかわる薬剤疫学が本格化した初期においては、「薬剤疫学の研究に携わるということは大規模データベースを開発すること」と言われていた。その後、欧米の現状は、データベース開発から実際的な活用によるリスク評価・社会的責務の履行に移っている。一方、残念ながらわが国には医薬品の安全性問題に対する迅速なリスク評価を可能にするデータベースは存在しない。情報基盤が未成熟であり、情報活用の有用性が社会的に認識されるに至っていない。

統計数理研究所リスク解析戦略研究センターでの重点研究分野の１つとして医薬品・食品があり、データベース構築やそれに基づく高度な統計的データ処理を通じて、医薬品・食品など人が直接摂取する物質の健康影響についてのリスク研究の基本的枠組みを創設することが目指されている。「将来」と「公益」の要素を取り入れた個人情報の保護と活用推進のバランスが社会的に検討されている中、個人情報の厳格な保護の下で社会の安心・安全のために科学的リスク評価を推進したいものである。

「例外の重み」

数理・推論研究系　志村　隆彰

２つのグループ「金メダル、新記録、最高気温」と「月間日照時間、打率、平均寿命」の違いは何だろうか。一般に多くのもの（データ）の集合は複雑であり、そのまま理解するのは難しい。だから、集まりの特色を簡単に現すことが求められる。文頭のふたつのグループもそのようなものであるが、両者には大きな違いがある。前者は多くのものの中の（極端な）一つであるのに対し、後者は全部のデータを総合して決まる点である。多くの4位入賞よりたった一つの金メダルが世間の注目を一手に集めるように、極端なものは刺激的であり、扱いも容易に見える。一方、全てのデータを見ることは堅実ではあるが、手間がかかりそうでやっかいな印象を与えるかもしれない。しかしながら、学術的には、たとえたった一つであっても極端なものを予測したり、制御したりすることは大変難しく、かつしばしば重要な問題である。

近年、現実的な要請が増してきたことも手伝って、確率分布の裾の研究が盛んになってきた。分布の裾とは　100歳以上の人の割合とか、年収いくら以上の人の割合といった、ある値以上（或いは以下）をとる確率のことであり、稀ではあるが極端なこと、言わば、例外の起こり方の数学的表現である。重要なのは裾が0に近づく速さで、速ければ、極端なことが起こる確率は小さく、無視できるため、平均的に起こることが重要になる。逆に、遅い場合は極端なことが意味を持ってくる。大数の法則と中心極限定理はよく知られた法則であるが、これらが成り立つには、必ず裾が速く0に近づくことを意味する条件がつく。そのような条件がなければ、ランダムな値の和の挙動が正規分布以外へ近づいたり、値の和がその中の最大のものただひとつとほとんど変わらないことも起こる。

裾の挙動が主役の分野として極値統計学を挙げておこう。大地震や大雨などのように、数多く起こる小さなものの蓄積よりも稀に起こる極端に大きいことが現実的意味をもつ現象を扱う分野である。他にも、建築工学、信頼工学、保険数学、ファイナンスなどの安全性やリスクを扱う工学の分野では極値理論が大きな役割を果たしている。

環境データ解析のためのベイズ的方法の開発とその応用

データ科学研究系　柏木　宣久

環境問題は注目を集めてから未だ日が浅く、次々と新たな問題が浮上してくるため、データの整備に手がまわらない状況にあります。それでも、問題解決に向け迅速に決断を下さなければなりません。こうした状況に対処するため、不完全データに基づく推論を実現する統計的方法を開発しています。同時に、推論の精度向上に不可欠なデータを指摘し、一部についてはデータの整備も実施しています。以下では、いくつかの例題について説明します。

★微量化学物質測定における要因分析

環境中の微量な化学物質を測定するため、特別な測定法が用いられています。その精度を管理するため、測定条件の要因効果の推定が望まれています。ところが、測定条件の数は極めて多く、測定にかなりの費用が掛かるため、推定に必要なデータを取得できないでいます。そこで、不完全なデータからでも要因効果を推定できるベイズ的方法を開発しています。

研究組織： 統計数理研究所、国立環境研究所、(財)日本環境衛生センター、(株)NTTデータ

★ダイオキシン類汚染における発生源解析

特定のダイオキシンの毒性はサリンの毒性より強く、ダイオキシン類の健康への影響が懸念されています。主な発生源は農薬、漂白、燃焼、PCB製品等の人間活動です。近年、ダイオキシン類による高濃度汚染が各地で頻繁に発見されています。高濃度汚染の解決には発生源の特定が欠かせません。ところが、ダイオキシン類にはデータが無い未確認発生源が数多く存在するため、発生源の特定は容易ではありません。そこで、未確認発生源についても推論できるベイズ的方法を開発しています。加えて、推論の精度を向上させるため、ダイオキシン類データの充実を図っています。

研究組織： 統計数理研究所、国立環境研究所、北海道環境科学研究センター、宮城県保健環境センター、茨城県公害技術センター、千葉県環境研究センター、東京都環境科学研究所、新潟県保健環境科学研究所、長野県衛生公害研究所、岐阜県保健環境研究所、広島県保健環境センター

★東京湾内水質の時空間予測

我々の研究により、近年の東京湾内で、水温の上昇、塩分濃度の上層における低下と底層における上昇、湾奥底層における貧酸素水塊の増加など、新たな問題が次々と生じているのが明らかになり、将来予測が不可欠な状況になってきました。そこで、将来予測が可能なベイズ的方法を開発しています。

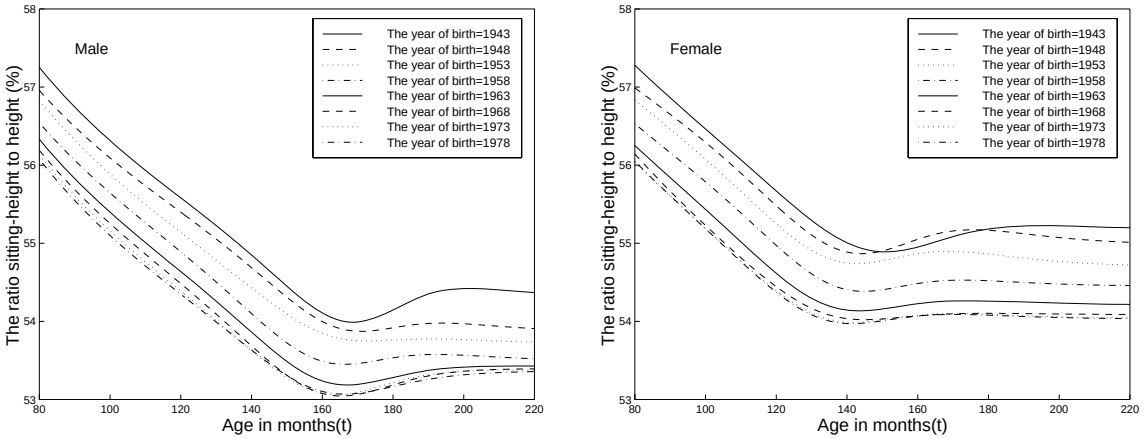
研究組織： 統計数理研究所、東京都環境科学研究所、横浜市環境科学研究所、千葉県環境研究センター

日本人の身体的変化をとらえるために

データ科学研究系：金藤浩司

日本人の食生活が欧米化し、それに伴って日本人の体格のスタイルもよくなったと世間では言われてきました。果たしてこの説は正しいのでしょうか。それとも外部からの情報によってそう思っているだけなのでしょうか。この仮説を科学的に検証するためには、正しいデータに基づいてそれを評価する必要があります。

文部科学省は、児童、生徒および幼児の発育状況及び健康状態を明らかにするため、国の法律に基づいて［統計法による指定統計（第 15 号)］、毎年「学校保健統計調査」を行っています。もしかしたら、今この文章を読まれている方の小・中学校生時代の測定値もその中に入っているかもしれません。そこで公開されている平均値のデータをよく眺めてみると、第二次世界大戦後に生まれた児童生徒の身長やその他の平均値は、確かに増加傾向を示しています。ただ、ここ１０年ぐらいは各年齢での平均値の変化がほとんどなくなっています。そこで、この調査を利用して、時間tでの身長を$y(t)$として、成長曲線（具体的な式は、研究者によっていろいろ提案されている）を$H(\theta;t)$とし、成長モデル$y(t)=H(\theta;t)+\varepsilon(t)$を導入して別の観点から統計的解析を行ってみました。ここで、θは成長曲線の形を決定する未知な母数であり、$\varepsilon(t)$は誤差を表しています。以下の２つの図は、身長と座高のデータを出生年ごとに並び替え、男性(Male)、女性(Female)毎に成長モデルに当てはめ、それを出生年に関して 5 年ごとに座高比曲線＝座高（月齢）／身長（月齢）で表示しています。この図から日本人のスタイルの観点から相対的な足の長さが長くなっていることがわかります（身長に対する座高の比の値が小さくなっている)。この傾向は、1968 年生まれ以降は、平均値でみると男女ともほぼ安定していることが読み取れます。



科学的に物事を述べるためには、正しくとられたデータとそのデータを解析するための統計的モデルと統計的方法論が必要となります。ここでは、国民の共有財産である統計を使うことによって、はじめに示された問題に関する科学的な判断が行えることがわかります。実際、「学校保健統計調査」の統計を用いることで、現在の状況が把握できるとともに、将来の日本人の体型の予想が可能となります。また。小・中学校生の皆さんが毎日使っている机や椅子の適正な配置等の学習環境を整備する基礎的な資料として非常に有効になります。同時に、このような統計は、身近な問題として衣料品のサイズの決め方、自動車の座席の高さや快適な自動車内の空間の設計等の身近な生活環境を改善する大切な基礎データとなっています。

タグチメソッドにおける統計的手法の開発と実践面への応用

データ科学研究系・リスク解析戦略研究センター
河村敏彦

タグチメソッドとは、日本では品質工学ともよばれ、田口玄一博士がほぼ独力で半世紀をかけて構築した“品質”に関する計量的工学手法です([1])。統計的手法の一つである(タグチ流)“実験計画法”を製品開発や工程設計などの様々な問題に応用し、我が国の工業製品の品質向上に多大な貢献をもたらしました。日本は暗黙知(経験や勘に基づく知識のことで、言葉などで表現が難しいもの)が強く、ノウハウを形式知化することはできないと言われていますが、研究開発における従来の試行錯誤的プロセスを効率的にシステム化したタグチメソッドは近年、特に注目されています。また 1980 年のベル研究所での超 LSI のばらつき低減という大きな成功を収めて以来、米国の製造業界(フォード社やゼロックス社など)においても活用されています。

品質工学とは別に統計品質管理という管理手法があります。さらに統計的品質管理には、観察研究において QC ストーリーを活用した問題解決法と、実験計画を利用した改善のための伝統的実験計画法があります。ここでは、従来の実験計画法の活用とその限界について簡単に説明します。伝統的実験計画法は分散分析(平均値の差の検定)を用いられますが、この方法を“ばらつき”の低減研究に用いると、どうしても無理が生じてきます。そもそも、水準間の等分散性を仮定する分散分析では、“ばらつき”の小さい水準を見出すという発想がもともとないからです。そこで、タグチ流実験計画では、品質特性の“ばらつき”を低減する、さらには機能性の向上を図るといった目的に対し、それに適した方法を与えています([2])。

近年、タグチメソッドは統計的品質管理手法の一つ延長線上であるものとして、特に統計的な観点から積極的に研究されています。田口玄一博士の 3 大業績の一つである SN 比(その他は直交表と損失関数)は、製品の品質や工程の機能性を定量的に評価するための評価測度として知られています。しかし、従来の物理量である SN 比は入出力系の単位が異なる場合には性能性の優劣比較をすることはできません。そこで、Nagata et al. は単位に依存しない SN 比を提案し、その SN 比に関する一様性の検定方式の構築を行っています。また同じ問題に対して河村 他 は正値データの場合の SN 比を提案し、いくつかの統計的性質を考察しています。また、宮川はタグチメソッドでは統計的仮説検定は重視されてないことを指摘し、2 値入出力系の SN 比に関する有意差検定を示しています。

参考文献

[1] 田口玄一 (1999): 「タグチメソッド　わが発想法」, 経済界.
[2] 宮川雅巳 (2000): 「品質を獲得する技術」, 日科技連.

次期地球環境観測衛星による温室効果ガス濃度観測精度の評価

リスク解析戦略研究センター　プロジェクト研究員　　友定　充洋

18 世紀末に始まった産業革命以降、人間は石炭や石油といった化石燃料を燃やし、大量の二酸化炭素を排出してきた。そのため、産業革命以前の二酸化炭素濃度は 280ppm ほどであったのが、現在では 380ppm ほどと 35%も上昇した。二酸化炭素には、太陽から放出された光は良く透過するものの、太陽で温められた地表からの放射は透過しにくいといった特徴がある。そのため、二酸化炭素が増加すると太陽から入った熱で温まった地球が冷めにくくなり、地球が温暖化する。地球が温暖化することによって、海面上昇、異常気象の多発等の問題が生じる。ここで、特筆すべき事実は、これまで地球は、暖かくなったり寒くなったりを繰り返してきた。しかし、この数十年間は、これまでの地球では経験のない速さで気温が上昇している。そのため、この速度で気温が上昇し続けると、2100 年には現在より 1.4 度から 5.8 度上昇すると見積もられている。地球温暖化の問題は、ある一国の問題ではなく、全世界の共通の問題、また本文を読んでいただいている方にとっての問題でもある。そのため、1997 年に京都で開催された地球温暖化防止会議では、全世界が取り組む問題として、各国の二酸化炭素排出量の削減目標が決定された。ところが、ところがである。現在、地表で二酸化炭素の観測を定期的に行っている地は少ない。そのため、各国が二酸化炭素排出量の削減目標に向かって勤めても、実際にどれだけ二酸化炭素が減少したかを評価することが困難なのである。そこで！登場するのがテレビの天気予報でお馴染みの衛星である。衛星は、ほぼ全地表を、人がほとんど住んでいない山奥や、海面といった未開地を探査することが可能である。そのため、衛星から二酸化炭素の量を計測することが可能であれば、地球の広い範囲を定期的に観測することが可能となる。日本では、環境省、国立環境研究所、宇宙航空研究開発機構が共同で、2008 年に温室効果ガス観測技術衛星 GOSAT(Greenhouse gases Observing SATellite)を打ち上げ、温室効果ガスを観測する計画がある。計画では、GOSAT から二酸化炭素濃度を 1%の誤差で計測することを目標としている。現在、宇宙航研究空開発機構ではセンサとロケットの開発が、また国立環境研究所では衛星が観測した信号から二酸化炭素濃度を求める方法の開発が行われている。我々は、GOSAT から二酸化炭素濃度が実際に 1%の精度で求めることが可能であるかを問題として研究している。GOSAT では、地表および大気で反射した太陽光を観測し、信号に変換して二酸化炭素濃度が求められる。しかし、信号に変換する際に電気的な雑音(ノイズ)等が入り、求められた二酸化炭素濃度に誤差が生じる。そこで、GOSAT に搭載されたセンサに光が入ってから信号に変換されるまでの流れで、雑音になるものを抽出し、雑音の大きさの計算が可能となるようにモデル化して、GOSAT から二酸化炭素濃度が 1%の精度で求めることが可能であるかを検証している。2008 年まで、1 年以上ありますが、打ち上げが上手くいき、世界全体の二酸化炭素濃度が 1%内の精度で公表され、GOSAT が地球温暖化防止の一役を担う日が来ることを祈る日々である。

金融リスクの統計的計測

山下智志

金融のリスク

金融工学の主たるテーマはリスクの計測と管理である。1960年以降、金融工学は少ないリスクで大きなリターンを得ることを最終的な目的とし、投資リスクを計測・管理する手段として発展した。もちろん収益（リターン）をあげるための技術も数多くあるが、収益をあげるための技術は成功例が少ない。対して、リスクに対する金融工学の効果はめざましいものがあり、少なくとも現在の金融機関の投資戦略では、金融工学的なリスク管理は欠かせない。金融のリスクは大きく分類して、信用リスク、市場リスク、オペレーショナルリスク、法的リスクがある。このうち統計的アプローチが有効なものは、信用リスクと市場リスクである。

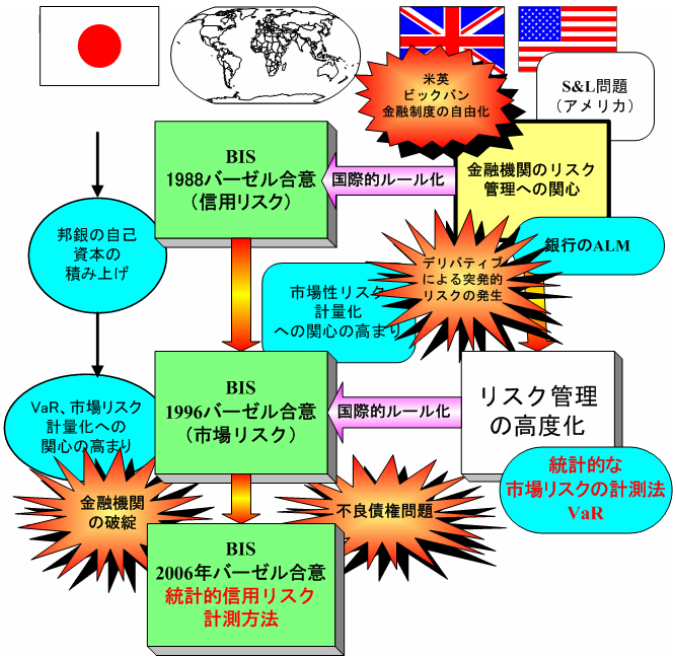
信用リスクの計測

信用リスクとは、国債、社債などの債券や、貸付、ローンなどの債権が、債務者の都合によりデフォルト（債務不履行）となるリスクである。信用リスクに対する取り組みは1980年代に制度的に構築された（BIS規制）が、近年デフォルトデータと統計的モデルより計測することが一般的となっている。従来、信用リスクの統計モデルは判別関数分析が中心であったが、ここ数年確率過程を内挿したモデルや、一般化線形モデルを利用したより精緻な統計モデルが相次いで開発されている。本研究所でも、潜在変数を用いた一般化線形モデルにより、独自の信用リスク測定モデルを考案し、実用化を進めている。

また、多くの信用リスク計測モデルが提案されていると、それぞれのモデルの優劣を測る規準が必要となる。しかし、現時点に置いて認知されている評価基準はない。行政による金融機関の監督や国際規約による銀行監査などの社会的要請を勘案すると、この評価基準の作成は緊急の課題であるといえる。この問題ついては本研究所の過去のノウハウである情報量統計学を応用したアプローチが有効となりうる。そのため、現在関係当局との議論・作業を経て、問題解決に協力している。

市場リスクの計測

市場リスクは、市場価格の変動によって起因する変動リスクと、市場における売却に何らかの障害があることによって起こる流動性リスクに分類される。このうち統計的な分析手法が確立しているのは変動リスクであり、時系列モデルや確率過程などの統計手法を応用することによってリスクの計量化を行っている。このような市場リスクの統計的測定モデルについては９０年代後半にほぼ確立され実用化されている。しかし、昨今の超低金利に対応した金利変動リスクのモデル化や、複雑化したオプションなどの派生商品を対象としたモデルが依然として確立されておらず、さらなる研究が求められている。本研究所では９０年後半に市場リスク計量化モデルの検証実験を行っており、関係諸機関の協力を得て、行政面や企業などへの実務的な貢献をしている。



統計数字の見方

データ科学研究系　佐藤　整尚

世の中にいろいろな統計数字があふれています。また、それらをどのように眺めるかもいろいろなやり方があります。ここでは、数字の眺め方について、すこし、書いてみましょう。巷にあふれている統計数字の中で皆さんは何が一番、気になるでしょうか？今は、少子化問題がさかんに議論されていますから、日本の人口かもしれません。127,756,815 人というのが 2005 年の国勢調査での結果です。この数字だけでもいろいろ議論できそうですが、実際に関心があるのはどのくらい増えているのかだと思います。この場合は前の調査からの増減数が必要になりますので、調べてみますと、2000 年調査と比べると 830,972 人の増加となります。これでもなんかピンとこない場合はさらに、その前の調査の数字と比べる必要があるかもしれません。このように、統計数字は、それ単独ではなくて、何かと比較することでより、その意味がはっきりすることが多いのです。

もっと特徴的なものとして、GDP（国内総生産）の数字が挙げられます。これも、統計数字としては約 542 兆円とかという数字で出てくるのですが、この数字だけで何かわかることは少ないでしょう。通常は前年や前期との比較によって、O.X%成長などと発表されることが多いのです。これも時系列的な比較によって、その統計数字を眺めているといえます。特に GDP の場合は増減の額ではなく増減率が重要になります。率で見ることにより、ほかの統計数字と比較しやすくしていると考えられます。このようにして、過去の値と比較してその伸び率を見るというやり方は特に経済統計などでは多く行われてきています。しかしながら、たとえばある年の 7 月の平均気温を 6 月の平均気温と比べて上がっているからといって、この年の 7 月は暑いという風に結論付けるのは間違いですね。この場合は季節性を考慮すべきところを無視してしまった点が問題であると考えられます。したがって、月次データや四半期データの場合は、単純に前月比、前期比を取るのではなく、季節調整が必要であれば、それを行ってから増減比を取る必要があります。もちろん、季節調整のやり方はいろいろあり、統計学的にも面白い話題なのですが、ここでは割愛したいと思います。

ところで、月次データの場合、前月比でみるのか、前年同月比でみるのかによって、多少なりとも、見方が変わってくるという場合があります。日本の経済統計の場合、季節調整値を前月比でみるというのが主流になっているようです。これは高度成長期のような年率で 2 桁の増加率があった時代であれば、有効であると思いますが、現在のような成長率がマイナスにもなりうる低成長時代では、かえって、振れが大きく出てしまう可能性があります。そこで、近年、著者は、前年同月比の改良版である「前年同月比平均伸び率」を提案しています。これはどういうものかというと、前年の同じ月のトレンド（傾向）の値を求めておいて、これから、今の月の季節調整値の伸び率を計算するというものであります。これによって、前年のたまたまの増減には左右されない伸び率が求められ、また、新たな統計数字の見方ができたといえます。ぜひお試しください。

統計科学の光と影

川崎能典（モデリング研究系／リスク解析戦略研究センター 助教授）

私は高校生の頃は歴史が好きで、特に経済史に興味を持っていた。東大経済学部を志望する高校生としては、ある意味平凡だったのだが、大塚久雄先生の「社会科学の方法」などを読んで、発展段階論を下敷きにしたような経済史のとらえ方に大きな魅力を感じていた。

ところが、いざ入学してみると、こうした動機に関連する講義というのはあまりおもしろく感じられなかった。大学教養課程で使用するレベルの専門書を入学前に読みすぎたので、それで新鮮みにかけたのかもしれない。ともかく、私の心は経済史から離れてしまった。

一方では、文科に入学してしまったことで、近代経済学を学ぶときを除けば、数理的な科目に接する機会が殆どないのを残念に思っていた。ただ、それは漠然として興味のレベルであって、自分に数理的な素養が十分あったかというとむしろなかった。そんな中で、自分が統計学を意識し始めたのは、アメリカの著名ジャーナリスト、デヴィッド・ハルバースタム著「ベスト・アンド・ブライテスト」を読んだときだった。

この本は、ケネディ政権下とそれを引き継いだジョンソン政権下で、まばゆいばかりの俊英をブレーンとして集めながら、なぜアメリカが泥沼のベトナム戦争へと突入して行ったかを様々な記録と証言から克明に描いたルポルタージュである。その登場人物の一人が、国防長官ロバート・マクナマラである。

マクナマラは大学時代に統計学やオペレーションズ・リサーチを専攻しており、1940 年にはハーバード大学ビジネススクールで教鞭をとるようになる。元々企業経営に用いる数値解析手法を陸軍航空軍将校に教えていた彼は、戦争が激化した 1943 年には陸軍航空軍に入隊し、戦略爆撃の解析および立案の仕事に従事する。彼は東京大空襲の計画にも少なからず関わっている。

そうした経歴的なエピソードだけでなく、同書に散見されるさまざまな逸話の中には、マクナマラの統計家としての合理的な判断を伺わせるものが沢山ある。最終的にはベトナム戦争は合理的なものとは言えなかったかもしれないが、人間の認識をうまく抽象化した統計科学の魅力を、自分はそれらのエピソードの中に感じていたのかもしれない。

マクナマラは軍事作戦におけるロジスティックスの重要性を知悉した人だったが、彼が陸軍航空軍に入隊した大戦末期、奇しくも日本においても、軍事作戦における統計科学の重要性を意識した動きがあった。1944 年の統計数理研究所の創設である。自分が統計学を勉強しようと思ったひとつの動機がロバート・マクナマラという人物について知ったことだと思うと、同じ時期に同じような目的で創立された研究所にこうしていま自分が身を置いて研究できていることが、なんだか不思議に思われるのである。

信用デリバティブ：Credit Default Swapの市場構造の解析

リスク解析戦略研究センター　　田野倉葉子
（共同研究者：津田博史　　佐藤整尚　　北川源四郎）

インターネットなどの情報伝達の高速化や金融システムのグローバル化により海外の資本市場で発生した信用不安や金融危機から国内資本市場が深刻な打撃を受けるニュースを最近よく耳にする．こういった経済危機の伝播・拡大を回避して金融リスクをグローバルにコントロールすることは急務となってきている．このためには，リスクの伝播する経路を解析することが必要不可欠である．企業にとって生命線ともいえる信用力に関わるリスクを取り扱う信用デリバティブ取引，Credit Default Swap（CDS）は欧米で活発に取引され，最近日本においても市場が急速に拡大し，注目を集めてきている．実際，ボーダフォン日本法人の買収に名乗りを挙げたソフトバンクのCDSスプレッド（後述）は急拡大し，ソフトバンクに対する信用リスクが大幅に高まったのは記憶に新しい．ある企業に，破産や負債のリストラといった企業の信用力に関わる出来事（信用事由）が発生するとその企業が発行した社債等は通常大幅な額面割れとなり，その保有者は経済的な損失を被ることになる．CDSは将来起こりうる企業の信用事由によって生じる損失をカバーするプロテクションの売買で，売り手が買い手から定期的に一定のプレミアムを受け取る代わりに，信用事由が発生した際に債務の元本金額相当分を支払う契約である．プレミアムのレートはCDSスプレッドと呼ばれ，取引価格である（図1参照）．CDSのメリットは社債等の参照債務から信用リスクのみを切り離して安いコストでリスク移転できることである．また，信用リスクに対して多様な動機を持つ投資家のニーズにあわせて組成する複雑な合成信用デリバティブの基本構成要素であるため，その流動性から信用リスクの新しい指標として注目されている．

CDSは店頭市場で取引されるため，同一の銘柄に対して複数の取引業者による価格情報が存在し，その市場構造は明らかでない．実際，CDSスプレッドはかかる企業の格付，業種，地域性，取引関係等の属性との連動性，また銘柄間の相関が指摘されている．

本研究は，業種および格付間のCDSスプレッドに焦点を当て，一般化パワー寄与率 (Tanokura and Kitagawa (2004)) を適用することにより市場の変動特性を明らかにすることでCDSの市場構造の解析を行った．その結果，業種間の変動の伝播の経路を明確に表現し，輸送用機器と食料品の両業種の重要性が新しい知見として検出できた．また，パワー寄与率の大きい業種は変動の周期に関わらず安定して他の業種の変動に共通して影響を与えることがわかった．

参考文献

Tanokura, Y. and G. Kitagawa (2004), “Modeling influential correlated noise sources in multi-variate dynamic systems,” in: M. H. Hamza (Ed.), *The 15th IASTED International Conference on Modelling and Simulation*, ACTA Press, Marina del Rey, CA, USA, 19-24.

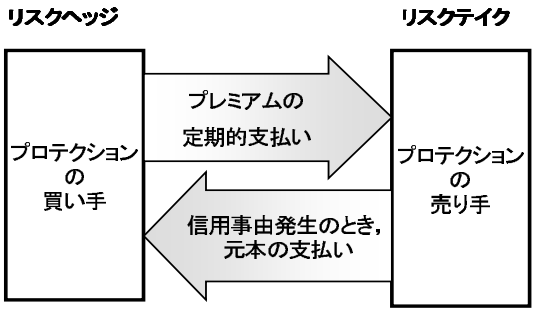


図 1: CDS取引のしくみ

リスク解析戦略研究センター プロジェクト研究員 公文雅之

研究テーマ：ゲーム論的確率およびファイナンスの最適戦略的研究[@]

Glenn Shafer と Vladimir Vovk によって提唱されたゲーム論的確率論では，測度論を使うことなしに大数の法則，中心極限定理，数理ファイナンスにおける価格付けの公式等が比較的簡明に証明されている．また竹内 啓教授は「賭けゲームにおける最適戦略の構造」という視点から，確率論やファイナンスを包括する新たな体系を展開されている．この研究では標記のテーマ，特に「ゲーム論的確率論とカルバック情報量」について研究を行っている．

具体的には「賭けをする人」を表す Skeptic と「自然」あるいは「現実」を表わす Reality との間の交互手順・逐次完全情報ゲームを定式化し，特にコイン投げゲームを含めた「結果当てゲーム」において Reality の行動が大数の法則に従わないとき，Skeptic の資金を限りなく増やすためのベイズ的最適戦略を研究した．

この Skeptic の最適戦略および最適資金過程は陽に書き表すことができる．そして資金過程を漸近評価することにより，最適資金過程が Reality の行動平均とリスク中立確率との間のカルバック情報量に支配されていることが導かれ，コイン投げに対するゲーム論的大数の強法則が収束率，収束因子を含めた精度のよい形で求められる．

この研究は通常何らかの確率的構造を仮定して導かれる確率論の極限定理や，金融工学の Black-Scholes 公式といったオプション価格付け法に対して，自然あるいは現実市場を相手とする賭けゲームを定式化しそこでの最適戦略を考察することの一つの帰結として，これらを捉えようという視点に立っている．

従ってこの研究においては，社会・経済システムが持っている不確実性を最初に何らかの確率的構造を仮定することなくモデル化することができ，確率的な前提に依存しない頑健なアプローチが可能である．

そしてこのように開放的なモデルにおける様々な行動に伴うリスク解析を最適戦略の立場から行うことによって，計量的かつ客観的なリスクの評価・管理方法を導くことができる．特にこの研究の最適戦略はカルバック情報量という統計科学一般における基本的な情報量を内包していることから，従来の統計解析手法やモデリング方法にも新たな視点を提示し得る可能性を秘めている．

[@] この研究は竹内 啓 東京大学名誉教授，竹村 彰通 東京大学教授との共同研究です．

リスク解析戦略研究センター プロジェクト研究員　　河合 研一

研究テーマ　　オプション評価モデルに関する実証分析*

オプション等の金融派生商品の価格は，その原資産となる商品（例えば，株式や債券など）の価格に依存する．そのため派生商品の価格を定量的に評価する際には，原資産価格の挙動をできる限り正確に表現することが重要である．オプションの価格評価モデルとしてよく知られている Black-Scholes モデルでは，株価過程は幾何ブラウン運動に従うと仮定され，収益率は正規分布に従うことになる．しかしながら，実際に観察される収益率の分布は正規分布よりも裾が厚く，左側に裾が長い非対称な分布となることがよく知られている．従って，本研究では原資産の収益率分布に見られる上記のような非正規性を十分に説明可能な Kou (2002) モデルに依拠しつつ Black-Scholes 流の仮定を排したよりデータ指向型のモデルを用いた実証分析を行う．そこでは，株価過程を幾何ブラウン運動として捉えるのではなく，ジャンプ拡散過程を利用し，非完備市場の枠組みの下でオプションを評価している．このように，より現実に即した想定の下で価格付け評価を行うことによって，一層精度の高いリスクの計測・管理が可能になると思われる．

本研究ではもっとも単純なヨーロピアン・オプションを対象とし，その価格付けモデルの日本市場における現実妥当性を実証分析を通じて検証した．まず株価過程に幾何ブラウン運動を仮定することが適切であるかどうかを，つまり，対数収益に対して裾の厚い分布によるモデリングが適当であるかどうかを，株価過程におけるジャンプの有無の検定という形で検証し，ジャンプ拡散モデルを利用することの妥当性を確認した．次いで Black-Scholes モデル と Kou モデルの価格付けパフォーマンスを比べるために，両モデルからそれぞれ計算される理論価格が，市場で観察されるオプション価格をどの程度説明できるかを数量的に計測しその比較を行った．その結果，原資産の挙動をより正確に表現できる Kou モデルの方が Black-Scholes モデルに比べ，優れたパフォーマンスを示すことが確認された．

参考文献

Kou, S. G. (2002) A jump diffusion model for option pricing. *Management Science* 48, 1086--1101.

* 本研究は前川功一教授(広島大学)，Sangyeol Lee教授(ソウル大学)，森本孝之氏(名古屋大学)との共同研究である．

独立成分分析による線形逐次モデルの探索

清水昌平
日本学術振興会特別研究員 PD (統計数理研究所)

因果を吟味するための効果的な方法は, 無作為割付を伴う実験を行うことですが、実験を行うことが困難な状況がしばしばあります．実験を行うためには, 因果に関する事前仮説 (例えば、因果の向き) が必要ですが、因果に関する仮説を構築するだけの事前情報が十分でないことがしばしばあるからです．また, たとえ事前仮説が十分であっても, 応用研究の性格上, 実験が行えない事も多いのです (例えば, 犯罪心理学)．そこで、観察データから因果に関する仮説を構築するための統計的手法が必要になってきます．

図 1 左のような線形逐次モデル (Linear acyclic model) は観察データからの因果推論に最も頻繁に用いられている統計モデルの１つです．従来法では、データからの情報のみを用いて、元のネットワーク (図 1 左) を復元することはできなかったのですが、私たちは「独立成分分析」という信号処理の分野で生まれたデータ解析法のアイデアを用いて、元のネットワークを一意に復元する (図 1 右) ことができるアルゴリズムを開発しました (Shimizu et al., UAI 2005)。

ただし注意しなければならないのは、統計モデルによるデータ解析「のみ」によって、因果に関する「最終的な」結論を導くことはできないということです。しかし、よい因果仮説を事前に構築できない時に、データの助けを借りて、因果に関する「初期」仮説を探索することには意味があるでしょう．

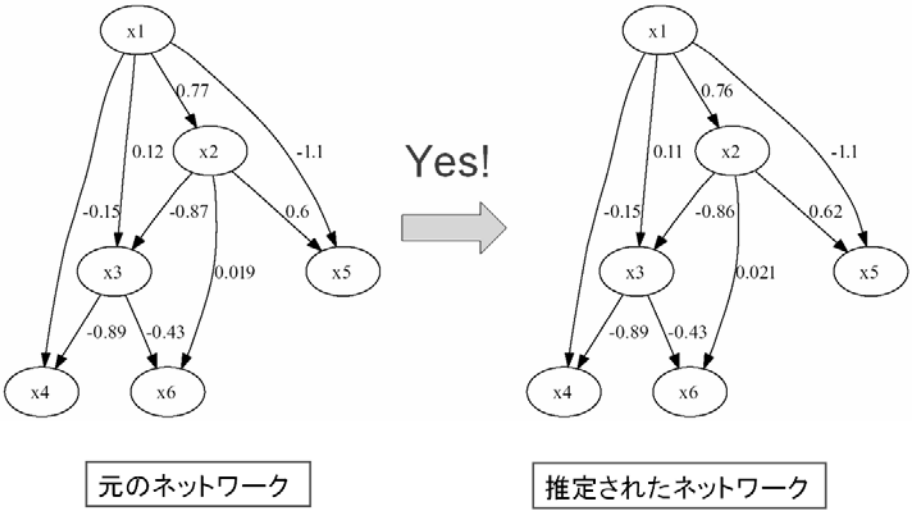


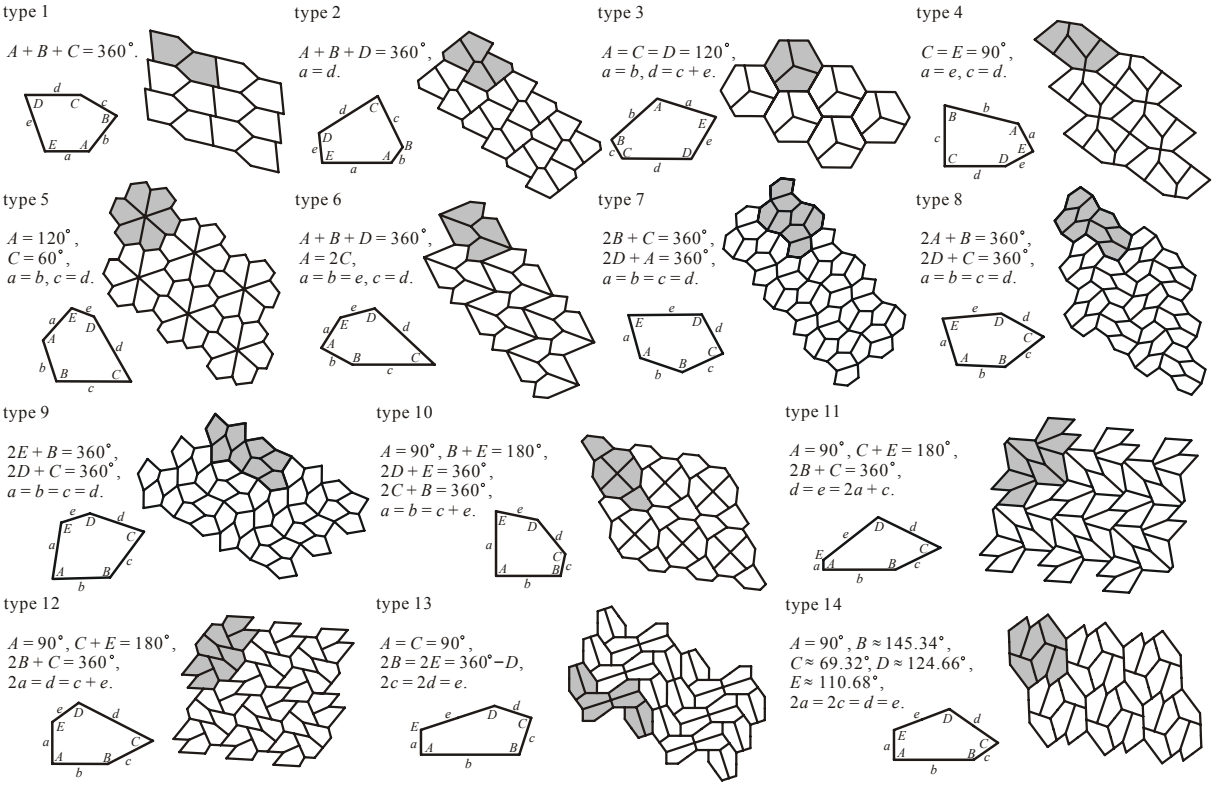
図 1: 独立成分分析で元のネットワークを復元できる？

タイル張り可能な凸多角形はどのようなものがあるか？

統計数理研究所モデリング系プロジェクト研究員
杉本晃久

平面を 1 種類の合同な凸多角形を使ってタイル張りしてみよう．なお，1 種類の合同図形のみで平面を隙間なく充填できる図形を，平面充填形と呼ぶ．古代ギリシャ時代から知られているように，正多角形の場合は正三角形と正方形と正六角形のみが平面充填形である．正多角形以外の凸多角形の場合はどうだろうか？実は，平面充填凸多角形は無数に存在する．なぜならば，すべての三角形と四角形は，内角和が 360 度を整除（整数を他の整数で割るとき，商が整数で余りがないこと）できるので対応する辺どうしが会して全種類の頂点が一点に集結することができ，タイル張り可能だからである．しかし，他の凸多角形は必ずしも充填形とは限らない．例えば，360 度を埋めるような内角の組み合わせがない凸多角形ではタイル張り不可能である．凸六角形の場合は平面充填形が 3 種類に表現でき，7 辺以上の凸多角形には充填形が存在しないことが証明されている．凸五角形の充填形は，現在までに 14 種類に表現されているが（下図参照），これで網羅という証明はなく未解決問題となっている．

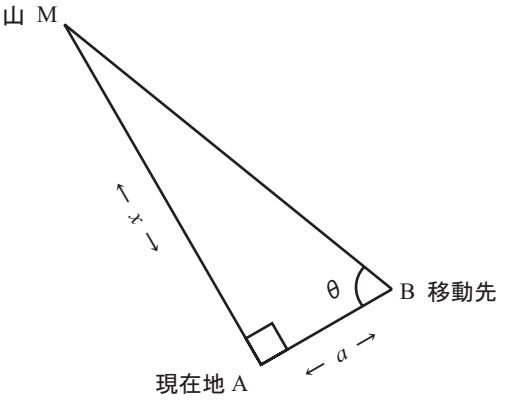
一見簡単そうに思える「タイル張り可能な凸多角形はどのようなものがあるか？」という平面充填凸多角形の網羅問題が，まだ解決していないことに驚かれるかもしれない．本問題以外にもタイル張りには，わからない問題がまだまだある．タイル張りは，三十年ぐらい前までは数学の中で曖昧な分野であったが，現在は物質の配分や回路の設計のような仕事への応用などが開発され，注目を集める分野となっている．



三角測量と量子推定

数理・推論研究系プロジェクト研究員　津田美幸

遠くに見える山までの距離は, 三角測量で知ることができる. 図のように, 現在地 A から山 M までの距離 x は, 移動先 B までの距離 a と, そこから山を見る角度 θ により

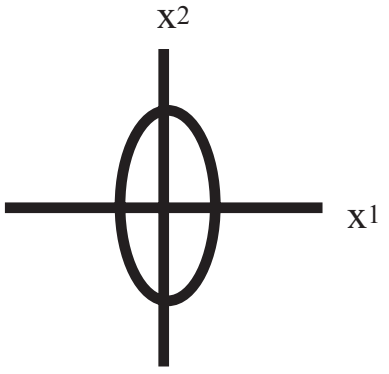


$$x = a\tan\theta$$

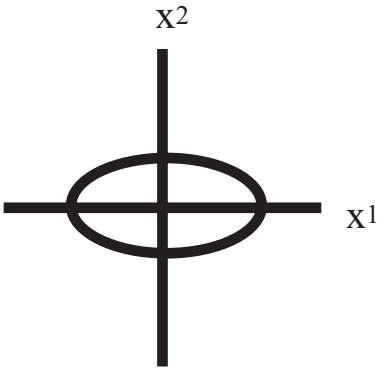
と表される. θ は通常, 直角に非常に近いので, 真の値と少しでも異なる数値を代入してしまうと, 得られる x は本当の値と随分違ってしまう. このような誤差をすることであるが, 移動量はなるべく少なくしたい. (a をどんどん大きくすれば良い, というのも真実だが, 移動距離が増えると大変なので a は固定して考える.) ここで角度 $\angle MAB$ は直角としたのは. θ をなるべく小さくして x の誤差を少なくするためである. (理想的には, $\triangle MAB$ が二等辺三角形になるようにすると一番良い.) もしも $\angle MAB$ がゼロか二直角ならば三角測量はできない.

都心からだと富士山は西南方向なので, 北西か南東に数キロ移動して三角測量すると富士山までの距離 x_1 が分かる. 北東か南西に動くと x_1 は測れない. 一方, 東京湾をはさんで見える房総の山は南東にあるので, そこまでの距離 x_2 を効率よく測るには北東または南西に移動して測量するのがベストで, 北西か南東に行くのは無意味である. この x_1, x_2 のように, 両者を同時に最適に測定できないことがある.

量子力学は, 電子や光子など, 極微の物質世界を調べる分野であるから, 富士山の三角測量なんて全く関係ないが, 少しは似ている面もある. 量子の世界では, 位置と運動量の関係は x_1 と x_2 の関係と同じである. 複素数の実部を位置, 虚数部を運動量に対応させることで, この性質を端的に表される.



x_1(位置) の誤差を少なくすると x_2(運動量) の誤差が大きくなる.



x_2(運動量) の誤差を少なくすると x_1(位置) の誤差が大きくなる.

The Institute of Statistical Mathematics

2006 統計数理研究所 オープンハウス

統計数理の世界

— 研究紹介とエッセイ —

大学共同利用機関法人
情報・システム研究機構
統計数理研究所
http://www.ism.ac.jp/

2006年7月発行