統計数理(2019)
第 67 巻 第 2 号 299–395

創立 75 周年記念号
[統計数理研究所研究活動]

統計数理研究所75周年記念
研究業績紹介

目次

1 樋口知之 モデリング変革の四半世紀を振り返って
2 川崎能典 スパース正則化法によるリスク要因の探索
3 宮里義彦 統計科学における制御理論の研究
4 吉本敦 森林資源管理のための数理最適化モデリング
5 庄建倉 地震間の対話の理解と地震活動確率予測の展望
6 瀧澤由美 神経系の動作の解明と時空間推定システムへの適用
7 三分一史和 時空間データ解析による生体メカニズムの解明
8 伊庭幸人 マルコフ連鎖モンテカルロ法の世界を拡げる
–高次元の多峰性分布からのサンプリング–
9 福水健次 正定値カーネルによる統計的機械学習
10 南和宏 位置情報軌跡の匿名化技術
11 小山慎介 神経スパイク時系列パターンの特徴解析
12 日野英逸 ノンパラメトリックエントロピー推定とその応用
13 早水桃子 組合せ論的系統学における最近の話題–グラフ理論が拓く系統解析の新展開–
14 中野純司 統計数理研究所で開発された R パッケージ
15 上野玄太 混合分布モデルとデータ同化
16 中野慎也 地球磁気圏の撮像観測とそのデータ同化
17 野村俊一 前震識別とその予測可能性
18 吉野諒三 「日本人の国民性調査」から「文化多様体解析」へ
19 前田忠彦 社会調査方法論の実践的研究
20 朴堯星 組織規範継承を可能にする目的指向型成果評価と相互依存性の計量分析
21 廣瀬雅代 立川市町丁目別住民意識調査分析追記–小地域推定モデル活用に向けて–
22 山下智志 制度的制約下におけるデータベース構造化、モデリング、モデル評価
23 金藤浩司 極値分布と指数逆ガウス型分布に関するある一般化について
24 伊藤陽一 医師主導治験による医療機器開発の経験
25 船渡川伊久子 経時データ解析と健康指標の長期推移
26 野間久史 先端医学研究の発展を支える統計数理とデータサイエンス
27 清水信夫 集約的シンボリックデータ解析
28 松井知子 ツイートを利用した都市気温の解析–都市インテリジェンス向上をめざして–
29 吉田亮 機械学習による新物質の発掘
30 島谷健一郎 方向統計学と動物移動生態学
31 Stephen Wu 機械学習に基づく高熱伝導率ポリマーの設計事例
32 村上大輔 大規模空間データのための空間可変係数モデリング
33 栗木哲 オイラー標数法によるウィシャート行列の最大固有値分布の近似
34 二宮嘉行 変化点モデルのための AIC
35 間野修平 抽出法と計算代数
36 加藤昇吾 角度の観測を含むデータのための統計モデル
37 志村隆彰 極値分布の吸引領域と離散性
38 荻原哲平 拡散過程の非同期・ノイズ付観測データに対する最尤型推定法
39 江口真透 一般化平均による統計モデル
40 藤澤洋徳 ガンマ・ダイバージェンスに基づいたロバスト統計

41　持橋大地　統計的自然言語処理と統計学
42　逸見昌之　メタアナリシスにおける公表バイアスの最悪評価に基づく感度解析
43　坂田綾香　スピングラス理論による制限等長定数評価
44　伊藤聡　多層整数計画に基づくクリンチ／エリミネーション数の計算
45　池田思朗　天文学とデータ科学
46　田中未来　制約付き非凸スパース最適化問題に対する DC アルゴリズム
47　今泉允聡　関数推定の理論に基づく深層学習の原理解析

モデリング変革の四半世紀を振り返って

25 years of a paradigm shift in a statistical modeling technology

前所長 樋口 知之 （Tomoyuki Higuchi）*

キーワード：ベイズモデリング、粒子フィルタ、カーネル法、深層学習

私が研究所に入ったのは博士号取得の直後である平成元年 4 月、また所長を退任すると同時に研究所を退職するのが平成が終わる直前の平成 31 年 3 月であるため、平成の 30 年間は私にとって研究者人生そのものである。私が入所した当時は、その 10 年ほど前から研究が活発化していたベイズモデリングの成熟期にあった。事前分布および尤度関数が線形・ガウス型の理論は確立し、応用の観点からも、いくつかの課題では本研究所の研究成果が標準的方法として認知されるなど、研究進展も著しかった。現象や既存知識のモデリングにおいては、計算上からくる線形・ガウス性の制約は、研究者の自然な思考を妨げる。そのため、その制約を“人手”でもって解析的に緩める論文が数多く出版されていた。その流れと平行して、計算機集約的な方法の開発も活発化していた。高い次元の潜在変数がつくる超高次元の分布から、マルコフ連鎖でもって直接的に実現値を得るマルコフ連鎖モンテカルロ法や、その分布を低次元の分布に分解し（いわゆる逐次ベイズフィルタ）、低次元の分布の逐次更新を数値的に得る「非線形フィルタ」などである。私は学位を宇宙プラズマ物理の領域で得たため、統計力学の自由エネルギーや転送積分とほぼ同一の発想にもとづく両手法を自然に理解でき、大いに魅惑された。

平成にはいってしばらくして北川元所長は、「非線形フィルタ」の適用範囲を大幅に拡張するモンテカルロフィルタを提案された（1993,1996 年）。この手法は今日において、粒子フィルタの手法群の中で、原始的ではあるがレガシーとして高く評価されている。北川先生のすぐ近くで研究していた自分は、そのアルゴリズムの大規模並列コンピュータへの実装容易性に衝撃を受け、さまざまな非線性・非ガウス型の時系列モデルへの適用を試みた。また、世界同時的に多くの研究者が、アルゴリズムの高度化と現実問題への応用をすすめ、21 世紀早々（2001 年）にシュプリンガーから書籍を出版できた。私も著者の一人として参加するチャンスを得、当時新進気鋭の海外の若い研究者と交流できたことは、その後の自分の研究者人生にとって大きな糧となった。この友人の多くは現在、後述する深層学習に研究の軸足を移しているが、そのことは偶然でなく、むしろモデリング技術の発展を考えれば極めて自然なことと私は考える。粒子フィルタは、時系列モデルが所与であれば、パラメータや状態変数の推定に関して原理的には万能である。もちろん、モンテカルロ誤差（有限サンプルによる表現限界）からくる、分布の表現能力の喪失（いわゆる退化問題）や、尤度値の不安定性など、数値的問題は避けられないが、問題や経験に即して“人”が対象を自由にモデリングできるようになった点は、モデリング技術の発展において大きな革新であった。

2000 年代にはいって、非線形モデリングの観点からは、別の形で大きな進歩があった。それはカーネル法の登場である。1998 年に Google が、また 2004 年には Facebook が創業されるなど、ビッグデータを新しい情報サービスという価値に転換する企業が当時、続々と誕生していた。ビッグデータを用いた容易なタスクは、識別関数の構築を通した判別や分別である。カーネル法は、データ空間で複雑な識別境界面を同定するのではなく、あえて超高次元の特徴

* 中央大学理理工学部：〒112-8551 東京都文京区春日 1-13-27

ベクトル空間を構成し、そこで豊富に蓄積された線形のモデリング技術を適用するものである。もちろん、データ間の類似度を規定するカーネル関数は“人”が設定しなければならないが、データ空間での非線形モデリングに求められる巧みの技の多くを、カーネルトリックに押しつけられた利点は大きい。また、パターン認識手法の多くはデータどうしの内積計算を含むため、カーネルトリックによって既存の線形モデリングの諸手法が非線形版に自然に拡張できた。2000 年代はビッグデータの登場により、統計的モデリングの主たる興味である生成モデルの構築から、カーネル法と最適化の活用による、複雑な識別関数（識別モデル）の自動構築に、研究のトレンドがシフトした。それと同時に、機械学習の言葉で代表される研究者コミュニティが育っていったのもこの時代の特徴である。

2000 年代半ばから 2010 年代にかけて、非ガウス型モデリングにおいても大きな発展がスパースモデリングの普及によってもたらされた。応用現場における統計的モデリングの成否は、膨大な説明変数群の中からタスクの解決に有効な特徴ベクトルの構築にあると断言できる。なお、画像、音声、自然言語の処理においては、現在、その課題は深層学習によって大幅に自動化されている。カーネル法は、データ空間から（暗に）高次元の特徴ベクトルを構成する方法なので、情報圧縮の文脈では逆センスの手法である。今、興味ある対象を膨大な説明変数の中から少数個で線形回帰表現する問題を考える。何次の回帰モデルとするのか、さらにどの変数を使うかで、膨大な数の回帰モデルが存在する。モデル数は組み合わせ爆発しており、AIC などの情報量規準による最適モデルの探索（変数選択）は機能しない。一方スパースモデリングでは、回帰係数に L1 正則化を加えた上で推定を行う。最適化関数の形が L1 であることから、L1 正則化は非ガウスモデリングと言える。この最適化の結果として、重要な説明変数のセットが自動的に浮きあがる。このように、説明変数の選択といった、限定されたモデリング技術に関しては、スパースモデリングにより線形・非ガウス型モデリングが実現された。

2010 年代にはいっての最大の衝撃は、特定のタスクでの深層学習の圧倒的パーフォーマンスである。その性能の高さは、入出国時の自動顔画像判別、スマートスピーカーでの音声認識、多言語自動翻訳など、私たちの生活に身近な製品として既に具現化している。これらの技術は、人の働き方や社会の構造そのものにも直接的に影響を与えていることから、深層学習はこの四半世紀の情報科学技術における最大のブレイクスルーであると言わざるを得ない。深層学習で使われるニューラルネットワークは、層数が大幅に増えた以外、第二次ニューロブームの時のものと違いは無く、そのパラメータ学習アルゴリズムも、Back propagation を基本とする以前のものと大差はない。層数が大幅に増えた結果、パラメータ数も爆発的に増え、学習アルゴリズムもさほど賢くなっているわけでもないので、必然と計算リソースはこれまでとは桁違いに必要となる。ただし、上述したように、それまでの特徴ベクトル構成法は機械学習の『匠の技』と言え、“機械”学習にもかかわらず、人間の判断が最も性能向上に大切という羊頭狗肉の面もあったが、その問題を特定領域では基本的に解決した点は画期的である。深層学習では、通常、最適化関数は L2 であるため、誤差にガウス性を仮定している。よって、非線形・ガウス型モデリング技術は、深層学習により、少なくとも予測・判別性能の観点からはほぼ極みに達したと言える。今後は、深層学習の大きな成功事例である、入力データが画像、音声、テキストなどに、一般的な問題での入力データをあえて変換すれば、特徴ベクトルの選択問題を回避可能となる。

非ガウス・非線形モデリングを自動化する試みは、この四半世紀に大きな飛躍を遂げた。その一方で、ビッグデータ× AI（人工知能）を土台に大きな存在感を示す米中のプラットフォーマーに対する風当たりも強くなってきている。今後は、モデリング技術の向上を計る上で、モデルの“人”による解釈可能性、帰納法で構築された意思決定システムの説明責任、データのバイアスと社会的偏見の分離など、人に寄り添った視点が極めて重要になってくる。

スパース正則化法によるリスク要因の探索

Identifying Risk Factors by Sparse Regularization

モデリング研究系 川崎 能典 （Yoshinori Kawasaki）

1. 円滑閾値型推定方程式による変数選択・グルーピング

情報通信技術やデータ計測技術等の発展に伴い，大規模なデータが蓄積されるに至って久しい．統計科学においては高次元データとしばしば言及される状況は，典型的には個体に付随して観測される属性が多数に及んでいる場合を指し，しばしば個体数に比べて属性数が遙かに上回る．こうした状況下で回帰分析を行う場合，「説明変数候補は多数得られているが，目的変数と関連性のある要因はごく少数である」という制約を置くのが現実的である．それを実現したのが LASSO (Least Absolute Shrinkage and Selection Operator) とその変種であり，一般にスパース正則化，L^1 正則化，スパース推定法などと言われる．

ペナルティつき損失関数を一般に $L(\theta)+\sum_{j=1}^{d}\rho_j(|\theta_j|)$ と記す．ここで ρ_j は j 番目のパラメータ θ_j に関する非負のペナルティ関数であり，$\rho_j(|\theta_j|)=\lambda_j|\theta_j|$ なら adapive LASSO である．ここで $\rho_j(|\theta_j|)=w_j\theta_j^2/2$ としつつも，$\delta_j\in[0,1]$ によって $w_j=\delta_j/(1-\delta_j)$ と取り直すことで，解くべき推定方程式は $(1-\delta_j)\partial L(\theta)/\partial\theta_j+\delta_j\theta_j=0$ $(j=1,\ldots,d)$ となる．$\delta_j=1$ は $w_j=\infty$ に対応し，このとき $\theta_j=0$ に帰着する．これを円滑閾値型推定方程式 (Smooth-Threshold Estimating Equation, Ueki (2009)) と呼び，以下 STEE と略す．

Ueki and Kawasaki (2011) は，STEE の方法論を変数のグルーピングにも拡張し，スパース変数選択とグルーピングを半自動的に行う方法を提案した．応用例としては，気管支疾患の判別や与信スコアリングの問題を取り上げている．Kawasaki and Ueki (2015) は，電話によるマーケティングデータを例に，STEE 法の性能を他のスパース正則化法と比較している．

2. 多重共線性と標準化更新度

STEE 法は，初期推定量の良さに依存している．初期推定量は飽和重回帰やリッジ回帰から構成するが，データ間に多重共線性が強ければ，有意な変数は偶発的に生き残っているに過ぎない可能性がある．そこで，ひとつのモデルを信頼するのではなく，たまたま選ばれなかったかもしれない真の因果変数も拾い上げる方法の構築が必要となる．

Ueki and Kawasaki (2013) は，線形回帰モデルの枠組みにおいて，飽和回帰モデルからの相対差で適合度基準を設定し，その基準を満たすモデルを前進選択法で探索することで，複数の「説明力同等」なモデルを手元に残す方法を提案した．手順は以下の通り．1) まず各変数を 1 つだけ含んだ p 個の単回帰モデルから出発し並列的に探索，2) 適合度基準を満たしたモデルにはお墨付きを与えて終了，3) 満たさないモデルについては，それ以外の変数を各ステップで 1 つずつ取り込み，適合度基準を満たすまで深掘り，4) 現在のモデルの変数添字集合を C とし，いま変数 k を加えたときのあてはまりの改善度を標準化更新度 (Standardized Update)

$$SU_{k,C}=\frac{||y-X_C\hat{\beta}_C||^2-||y-X_{C\cup\{k\}}\hat{\beta}_{C\cup\{k\}}||^2}{||y-\bar{y}1||^2-||y-X\hat{\beta}||^2},k\notin C$$

で測り，SU がある閾値を超えた時に変数 k を採用する．

数値実験の結果，完全多重共線性の下でも SU を使う提案手法は偽陽性・偽陰性ともに小さく，真の回帰関数に関連している変数の組合せを高精度で発見できた．一方 Elastic Net は全般的に偽陽性率が高く，完全多重共線性の下では偽陰性も高いことが示された．

3. 効果がマスクされた変数の探索

多重共線性が深刻な説明変数群を使って推定された線形飽和重回帰モデルでは，有意な説明変数は偶発的に有意になっているに過ぎない可能性がある．一方，飽和重回帰モデルで有意にならず，かつ目的変数との周辺相関がないと思われる説明変数でも，特定の説明変数集合を伴って推定されれば有意になることがある．

Ueki, Kawasaki and Tamiya (2017) では，効果が他の変数にマスクされている説明変数を効率的に探索する方法を提案している．飽和回帰モデルが推定可能な状況であっても，部分回帰的全探索は計算負荷が高い．提案手法は，説明変数・被説明変数を合わせた変数群内の全てのペアで計算した相関係数を基に変数間の接続性を双方向グラフで表現し，特定の説明変数から目的変数までの最短経路をダイクストラ法で求めることで，効果を浮かび上がらせる共変量集合を特定する．この方法は，ケースの次元より説明変数の次元が大きい状況下で，かつ説明変数群に多重共線性が潜んでいる状況でも適用可能である．(ただし上掲論文中の応用例では，対象に関する知見から，変数群を独立なブロックに分けることが妥当性を持つので，各ブロックごとに計算している．)

謝　辞

本稿で紹介した一連の研究は，植木優夫博士 (執筆時点では理化学研究所革新知能統合研究センター所属) との共同研究である．リスク解析戦略研究センター草創期の 2008 年に特任研究員として着任した同氏から，Ueki (2009) の草稿について討論させて頂いたのを契機に，その後は公募型共同利用 (25–共研–1018，26–共研–1014，27–共研–1013，28–共研–1011，29–共研–1009，30–共研–1013) が一連の研究成果につながっていった．ここに記して感謝申し上げたい．

参　考　文　献

Ueki, M. (2009). A note on automatic variable selection using smooththreshold estimationg equations, *Biometirka*, **96**, 1005–1011.

Ueki, M. and Kawasaki, Y. (2011). Automatic grouping using smooth-threshold estimating equations, *Electronic Journal of Statistics*, **5**, 309–328.

Ueki, M. and Kawasaki, Y. (2013). Multiple choice from competing regression models under multicollinearity based on standardized update, *Computational Statistics and Data Analysis*, **63**, 31–41.

Kawasaki, Y. and Ueki, M. (2015). Sparse predictive modeling for bank telemarketing success using smooth-threshold estimating equations, *Journal of Japanese Society of Computational Statistics*, **28**, 53–66.

Ueki, M., Kawasaki, Y. and Tamiya, G. (2017). Detecting genetic association through shortest paths in a bi-directed graph, *Genetic Epidemiology*, **41**, 481–497.

統計科学における制御理論の研究

Research on Control Theory in Statistical Science

モデリング研究系 宮里 義彦（Yoshihiko Miyasato）

1. 制御理論の背景

制御理論の歴史は伝達関数と周波数領域の評価に基づく古典制御に始まり，状態空間と時間領域の評価に基づく現代制御の時代を通過し，周波数領域と時間領域，伝達関数と状態空間の双方の評価方式と表現方式を統合したポスト現代制御の時代を経て現在に至っている．いずれの手法においても制御系を設計するに当たって，制御対象の適正なモデルを求める必要があるが，制御対象をどのようにモデリングするか，またそのモデルに含まれる不確定要因をどのように評価するかによって，適用される制御手法や達成される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リングやモデルの不確定要因を考慮した制御という視点が重要で，モデリングと制御は切り離して考えられない．

2. 制御理論の研究

このような制御理論とモデリングの関係を考慮して，モデルリングと制御を同時に行う適応学習制御を中心とした研究を行っている．特に適応制御は制御器の実時間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が，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て，現在は適応制御に関連する非線形制御と線形制御の立場から，およびそれらの多体系（マルチエージェント系）への適用の立場から研究に従事している．また応用研究の分野でも，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）の走行試験から良好な結果を得ることが出来た．

3. 研究事例 I：非線形制御の立場

未知の対象を制御する場合，対象のパラメトリックモデルを求めて，モデルのパラメータの推定値を正しいと見なして制御器の設計をする（certainty equivalence の立場）．ところがその推定値は常に正しいとは限らず，またパラメトリックモデル自体も対象の近似表現の一つにすぎないので，所望の制御性能を達成するためには，対象のさまざまな不確定要因も考慮に入れて，制御系を構成しなければならない．このような問題に対して，パラメータの推定誤差を制御問題の外乱と見なして，外乱の影響を抑制する非線形適応制御系を設計する手法を開発し，パラメータが時間的に変動したり不確定要素が存在する場合でも，所望の制御性能を達成できるようになった．この手法は非線形パラメトリックモデルの一種であるニューラルネットを含む制御系の設計にも適用が可能である．同じ考え方は高次振動モードを有する複雑なシステム（非線形プロセス，柔軟構造物，弾性アームなど）のモデリングと制御にも適用可能で，低次元モデルに含まれるモデリング誤差（スピルオーバー）の影響を低減化して，実用的な有限次元（低次元）の制御器でシステムの振動抑制と制御を実現することができる．

4. 研究事例 II：線形制御の立場

パラメータの上下限が規定された線形システムとして記述されるプロセスは，その上下限を端点とするポリトープの内点として表現することができる（ポリトピックモデル）．そのような対象についてはポリトープの内点を定めるパラメータをスケジューリングパラメータ（SP）と見なして，SP に応じて制御器の特性を変化させることで，より高性能の制御結果が得られる．この制御方式はゲインスケジューリング（GS）制御と呼ばれていて，制御理論においては線形行列不等式に基づく制約下における設計手順として定式化される．GS 制御は SP が正確に求められれば，ポリトープ内の変動に対してシステムの安定化と外乱抑制が達成されるが，SP が正確でないと安定性も保証されない．このような場合に，プロセスの操業データを用いて制御誤差を観測して SP を実時間で再調整する適応型の GS 制御方式を開発した．ポリトピックモデルに基づく適応制御は，制御のためのモデリングにも多大の影響を与えると思われる．

5. 研究事例 III：多体系（マルチエージェント系）への適用の立場

未知パラメータを含む複数の動的システムを個々のエージェントとするマルチエージェント系に対して，適応的に速度追従型の群生行動を実現するフォーメーション制御やリーダーフォローワー型の追従行動を実現するコンセンサス（合意形成）制御の研究を行っている．フォーメーション制御においては，未知パラメータの推定誤差と群生行動に関係するポテンシャル関数の誤差を等価的な外乱と見なした $\mathcal{H}_\infty$ 制御問題の解としてフォーメーション制御機構を導出し，コンセンサス制御においては限定された通信構造に対応するネットワークグラフに着目し，未知パラメータの推定誤差とネットワーク密度に関する不確定性を等価的な外乱と見なした $\mathcal{H}_\infty$ 制御問題の解としてコンセンサス制御機構を導出している．実際の問題としては高速道路における自動車の群制御（スマートハイウェイ）や複数のロボットマニピュレータによる協調動作，ドローンの群制御や宇宙機のランデブー問題などが該当し，これらの調和行動の実現のための基本原理を解明する研究である．

6. 最後に

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，既存モデルの統計的な当てはめに留まらず対象の物理的特性や原理にも目配りを行い，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不可欠である．また制御理論がこれまで取り扱ってきた時間軸上の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，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．

参 考 文 献

宮里義彦 (2013). 適応制御の回顧と展望, 計測と制御, **52** (4), 361-367.
宮里義彦 (2014). 分布定数系の適応制御, システム／制御／情報，**58** (9) 365-370.
宮里義彦 (2017). 無限次元系の協調制御, 計測と制御, **56** (12), 925-930.
宮里義彦 (2018).『適応制御』，コロナ社，東京．

森林資源管理のための数理最適化モデリング

Mathematical Optimization for Forest Resource Management

モデリング研究系 吉本 敦（Atsushi Yoshimoto）

キーワード：森林資源管理, 数理最適化, 最適制御, 動的計画法, 0–1 整数計画法.

森林資源管理において，森林を「いつ，どこから，どのくらい」伐採するかを決めることは，古くから取り組まれてきた主要な課題の一つである．このような意思決定をサポートするツールとして，資源の伐採量あるいはそこから得られる利益の最大化と言った資源管理の目的に対して，与えられた条件下で最適な伐採量・方法の在り方を探索できる最適化モデルが構築されてきた．そして，地域的な政策や経済的な要求に対応すべく広くその開発・応用がすすめられてきた．特に，最適な伐採の時空間的配置を考慮した管理の探索ができる整数計画法は，大規模な資源開発，あるいは土地利用の変化に伴う環境への負荷を評価するアプローチとして注目され，例えば，空間的に隣接し合う場所での伐採を同時期に行うことができないという条件を加えることで，伐採が空間的に分散され，大規模な伐採域の創出を避けることができる．森林資源管理は基本的に植林と伐採という単純な制御で対応できるが，それらの時空間的な組み合わせの違いは，生態系のみならず，地域的な社会経済へ異なる影響を及ぼす．森林資源管理問題の研究では，個々の管理ユニット（森林経営上の単位）である林分を対象とする林分レベルと，それらの集合体である森林全体を対象とする森林レベルに大きく分類することができ，それぞれに対し効率的な数理最適化モデルの開発が進められている．

林分レベルでは主に最適な間伐戦略の探求が課題となり，その解の探求では，まず経営から得られる総利益などの最適化を最適制御あるいは変分法の枠組みで連続的に捉え，目的関数を定式化する．次に間伐行動が離散的であることから，動的計画法のフレームワークに変換し，繰り返し演算を行うことにより最適解を探求するというものである．仮に $x(t)$ を森林の状態を表す時間依存の状態変数，$u(t)$ を森林の成長に影響を与える間伐などの程度を表す制御変数 (control variable) とする．ここで現時点の森林の状態および制御により得られるであろう微小時間間隔の利益の現在価値あるいはパフォーマンス指標を $\dot{I}(x(t), u(t))$ とすることで，下記のように目的を時間 t_0 から t_n までの積分値の最大化として，最適間伐戦略を探求することができる．

$$\begin{aligned} &\max_{\{u(t)\}} J = \int_{t_0}^{t_n} \dot{I}(x(t), u(t))dt \\ &subject\ to \\ &\dot{x}(t) = f(x(t), u(t)), \quad x(t_0) = x_0. \end{aligned}$$

なお，$f(\cdot)$ は森林の状態変化を表す $x(t)$ および $u(t)$ の関数であり，x_0 は初期状態を表す．上記のように最適制御の枠組みにて定式化できれば，離散的に発生する間伐制御 $u(t)$ による動的計画法への変換が可能になり，林分レベルの数理最適化モデルが構築できる (Yoshimoto *et al.*, 2016) .

一方，森林レベルの数理最適化モデルの構築は 1970 年代に線形計画法の応用から始まり，Johnson and Scheurman (1977) により最適な伐採時期・伐採量を探求するモデルとして Model I および Model II が開発され，伐採計画問題の基本構造が定式化された．Model I が各林分に対し計画期間内で実施可能な伐採パターン（施業）を決定変数 (decision variable) として定式

化するのに対し，Model II は各林分の各期における伐採量を決定変数として定式化している．Model I を例に，x_{ij} を決定変数とし，第 i 番目の林分において第 j 番目の施業を実施する面積比率を表すこととする．また，c_{ij} を x_{ij} を実施することにより得られる利益の現在価値とし，森林全体から得られる総利益の現在価値の最大化を目的とすれば，下記のように面積比率制約と各期伐採量制約を伴う伐採計画問題が定式化できる．

$$
\begin{aligned}
& Z = \max_{\{x_{ij}\}} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \cdot x_{ij} \\
\text{(Model I)} \quad & \textit{subject to} \\
& \sum_{j=0}^{n} x_{ij} = 1 \quad \forall i, \quad (1-\alpha) \cdot v_0 \leq \sum_{i=1}^{m} \sum_{j=1}^{n} v_{ij}^{p} \cdot x_{ij} \leq (1+\alpha) \cdot v_0 \quad p = 1, \cdots, T.
\end{aligned}
$$

なお，m は林分数，n は 1 つの林分に対する施業数，α は毎期伐採量変化の許容率，v_{ij}^{p} は x_{ij} の実行に伴う期間 p で得られる伐採量，T は計画期間，そして v_0 は伐採量変化の基準となる基準伐採量である．ここで x_{ij} を 0–1 決定変数とすれば上記問題は伐採を施す林分も特定できる数理最適化モデルとなる．

森林レベルの数理最適化モデルの応用はさらに拡大し，1980 年代後半から広大な皆伐域を回避することによる野鳥獣の生息域などの保護，森林景観の確保等の環境保全といった環境配慮型志向に伴い，新たに「隣接しあった林分を同時期に伐採することはできない」という空間的制約条件（隣接林分制約条件）を課すようになってきた．ここで A^S を空間的な林分同士の隣接行列，A^V を施業間の同時期伐採関係を表す伐採隣接行列とし，$vec(X') = (x_{11}, \cdots, x_{1n}, \cdots, x_{m1}, \cdots, x_{mn})'$ とすれば下記により隣接林分制約条件を定式化することができ，上記の Model I に加えることにより解の探求が可能になる (Yoshimoto and Konoshima, 2016)．

$$
[\tilde{A} + \mathrm{diag}(\tilde{A} \cdot 1_{mn})] \cdot \mathrm{vec}(X') \leq \tilde{A} \cdot 1_{mn} \quad \text{ただし} \quad \tilde{A} = A^S \otimes A^V.
$$

昨今ではこれらのモデルの拡張により，林分の集約を伴う最大伐採域許容問題 (Yoshimoto and Asante, 2018a) や老齢林バッファーゾーン形成問題 (Yoshimoto and Asante, 2018b) など様々な空間的制約条件を伴う森林資源管理の数理最適化モデルの開発が盛んに行われている．また生態系保護の観点から病害虫・外来種などの拡散防止に向けた数理最適化モデルも開発されている (Yoshimoto *et al.*, 2017)．

参 考 文 献

Johnson, K.N. and Scheurman, H.L. (1977). Techniques for prescribing optimal timber harvest 572 and investment under different objectives–discussion and synthesis. *Forest Sci.* Mono. **573**:18p.

Yoshimoto, A. and Asante, P. (2018a). A new optimization model for spatially constrained harvest scheduling under area restrictions through maximum flow problem. *Forest Sci.* **64**(4):392–406.

Yoshimoto, A. and Asante, P. (2018b). Focal-point aggregation under area restrictions through spatially constrained optimal harvest scheduling. *Forest Sci.*, in press, DOI:10.1093/forsci/fxy044.

Yoshimoto, A., Asante, P. and Konoshima, M. (2016). Stand-level forest management planning approaches. *Curr. Forestry. Rep.* **2**:163–176.

Yoshimoto, A., Asante, P., Konoshima, M., Surový, P. (2017). Integer programming approach to control invasive species spread based on cellular automaton model, *Nat. Resour. Model.* **30**(2), DOI:10.1111/nrm.12101.

Yoshimoto, A. and Konoshima, M. (2016). Spatially constrained harvest scheduling for multiple harvests by exact formulation with common matrix algebra. *J. Forest Res.* **21**:15–22.

地震間の対話の理解と地震活動確率予測の展望

Understanding the conversations among earthquakes and developing probability earthquake forecasting models

モデリング研究系 庄 建倉（Jiancang Zhuang）

地震は地震波形記録図の中では観測地での地震動を示す連続なパルス波形である。地震カタログの中では、各地震は地震波形を解析して生成された地震の位置、深さ、大きさ、震源機構、記録精度などの情報を記録した一列の数字で表示される。しかし、現実ではこれらの地震、特に陸上または沖合の大地震は人類に生命や財産の莫大な損失をもたらす。 正確な地震予知を達成することは長年にわたる人類の願いである。

現代の地震学的研究結果によれば、地震の発生は自己組織化された臨界現象でああり、単一の地震の規模とタイミングを正確に予測することは非常に困難である [1]。しかし、個々の地震は完全に独立していない。 地震と地震の間には「対話」がある。地震の前後にはしばしば前震と余震がある。 同時に、これらの大地震は元の断層のコロンボ応力分布を変化させ、余震の分布に影響を与え、その後の大地震の発生を遅らせたり、早めたりする。 同時に、地震断層では高応力状態にあるとき様々な異常が発生する。したがって、地震活動は完全にランダム、予測不可能ではない。これらの予測可能なコンポーネントの解明は、将来の地震リスクの推定、防災政策の制定、震災後救助に大きな影響を与える。

地震活動において、どの成分を予測でき、どの成分を予測できないのか？地震活動の予測可能な最大の成分は地震のクラスタリングである。地震学者はこれらのクラスターの特徴について深く研究し、多く経験則を得ることができた。 ETAS モデルは、これらの経験則の統合に基づいて提案した [2]。 このモデルでは各地震は特定の確率ルールに従って自らのクラスターを誘発する。これらの確率ルールは既存の経験則である [2,3]。 今日既に、地震活動の最も大きな予測成分を定量化する確率モデル、特に地震活動の仮説検定および他のモデルの予測評価の標準相場モデルはかなり成功している。同時に ETAS モデルは地震予測の確率モデルでもあり、カリフォルニア地震予知プログラム（UCERF3）で採択された。

ETAS モデルの予測レベルを改善する、すなわちその予測の確率利得を改善させるには多く方法が存在する。 例えば、本来の ETAS モデルを 2 次元震央から 3 次元震源版 [4,5] に，点震源から有限断層源 [6,7] に拡張すると、モデルはより詳細になり、より高い確率利得予測を得ることができる。 地震活動データの以外の GPS 変形データ、ULF 地電位信号、地磁気信号など他の地球物理観測データと組み合わせると、より高い確率利得予測が得られる。これらの現象の大部分が地震の前兆であるかどうかは論争があり、客観的な評価が非常に重要である。ETAS モデルは現在までの地震活動の最良の定量化モデルであることが証明され、外部前兆の励起効果を組み合わせた新しい ETAS モデルが提案されている [8,9]。

ETAS モデルの推定手法も大きく発展した。 確率除群法、確率再構築法、ベイズノンパラメトリック推定手法の開発は、犯罪、テロ事件、森林火災などの地震以外の自然現象や社会現象の研究のため、より広い Hawkes モデルにも適用される [10]。 他の応用分野における予測理論や手法の開発は、地震の確率予測に応用することもできる。

参 考 文 献

[1]Zhuang, J., D. Wang, and M. Matsu'ura (2016) Features of the earthquake source process simulated by Vere-Jones' branching crack model. Bulletin of the Seismological Society of America. Volume 106. doi:10.1785/0120150337.

[2]Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes, J. Am. Stat. Assoc., 83(401), 9-27.

[3]Zhuang J., Ogata Y. and Vere-Jones D. (2002). Stochastic declustering of space-time earthquake occurrences. Journal of the American Statistical Association, 97: 369-380.

[4]Guo, Y., Zhuang, J., and Zhou, S. (2015) A hypocentral version of the space-time ETAS model. Geophysical Journal International, 203: 366-372. doi: 10.1093/gji/ggv319.

[5]Guo, Y., J. Zhuang, and N. Hirata (2018) Modeling and forecasting 3D-hypocenter seismicity in the Kanto region. Geophysical Journal International, 214: 520-530. doi:10.1093/gji/ggy154.

[6] Guo, Y., Zhuang J., Hirata N., Zhou S. (2017) Heterogeneity of direct aftershock productivity of the main shock rupture. Journal of Geophysical Research Solid Earth, 122, 5288-5305 doi:10.1002/2017JB014064.

[7] Zhuang, J., M. Murru, G. Falcone, Y. Guo (2019) An extensive study of clustering features of seismicity in Italy from 2005 to 2016.Geophysical Journal International. 216:302-318. doi:10.1093/gji/ggy428.

[8] Zhuang J., (2011) , Next-day earthquake forecasts for the Japan region generated by the ETAS model. Earth Planets Space, 63, 207-216. doi:10.5047/eps.2010.12.010. [pdf] doi:10.1029/2003JB002879.

[9] Zhuang, J., M. Matsu' ura, P. Han (2019) Critical zone of the branching crack model for earthquakes: inherent randomness, earthquake predictability, and precursor modelling. Submitted to Europhysics, in Revision

[10] Zhuang, J., J. Mateu (2018) A semi-parametric spatiotemporal Hawkes-type point process model with periodic background for crime data.Accepted by Journal of the Royal Statistical Society, Ser. A.

神経系の動作の解明と時空間推定システムへの適用

Study in Neural Operation and its Application to Time-Space Estimation

モデリング系 瀧澤 由美（Yumi Takizawai）

1. 神経系の動作の研究

生物の神経系は，実際に生起した事象の時刻と場所（時間・空間）を知覚する能力を有する．脳は情報を知覚する広範で高度な機能をもつが，それらの原理は現在においてもほとんど明らかではない．筆者は多くの機能の中で特に基本的な機能として，時間・空間知覚能力に着目し，その動作原理を電気物理学的に解明することを試みた．脳の基本動作は多数の神経細胞とその結合である神経システムよりなる．従来の研究では，神経システムの動作の電気物理的観測，心理学による考察などの機能面からの解明，または刺激（入力）と応答（出力）を模擬する人工システムとしてとらえ，数理的データ的な面からの解明が試みられてきている．しかし，この取り組みは人工物に焦点をあて，生体としての神経細胞，神経システムの実態をとらえるものとは異なる．本研究では神経細胞（ニューロン）単体を能動的な電気信号（パルスまたはプラトー）の発振器としてとらえる．次に，神経細胞群（ニューロン集合体）をニューロン群により自律的同期システムとしてとらえる．安定な同期信号の発生は，正確

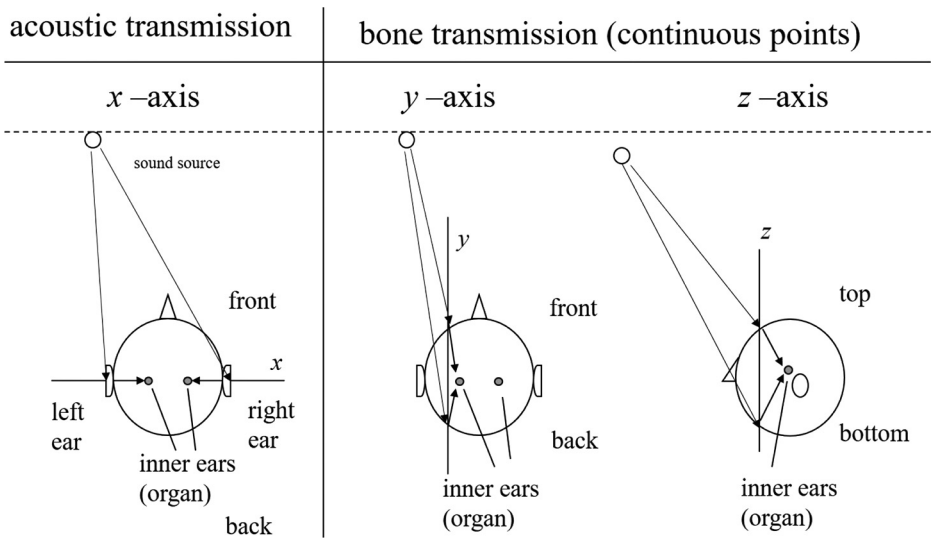


図 1．両耳と骨伝導による 3 次元位置の推定．

な時刻の知覚を可能にし，同時に場所（空間）の知覚を実現する．脳の高度な情報処理は，この同期パルスに基づくディジタル処理と緩やかで安定なプラトー信号に基づくアナログ処理とにより達成される．脳内における時空間知覚と処理は脳内マッピングとして動物および人間の脳において観測されている．筆者はこれの電気物理的モデルとその解析により，それが実現されることを示した．筆者による上記の研究成果は国際学会において注目され，2012 年学会より Best Paper Award を授与された．

2. 時空間推定の実システムへの応用研究

上記の研究成果の適用例として，電磁波を用いた位置計測システムを研究し，実用化を進めている．このシステムでは電磁波の送信と受信時刻差から，対称物までの距離を精密に測定する．この方式は，液化天然ガス（LNG）タンカーにおいて電磁波（マイクロ波）を用いた積載量計量システムを実現し，資源の貯蔵，輸送に用いられる．一方，航空機の搭載により，資源探査，植生観測による農業管理への適用が進められている．この研究では特にマイクロ波を用いた小型高性能な円偏波アンテナ（送信／受信）を実現し，実用化を進めている．

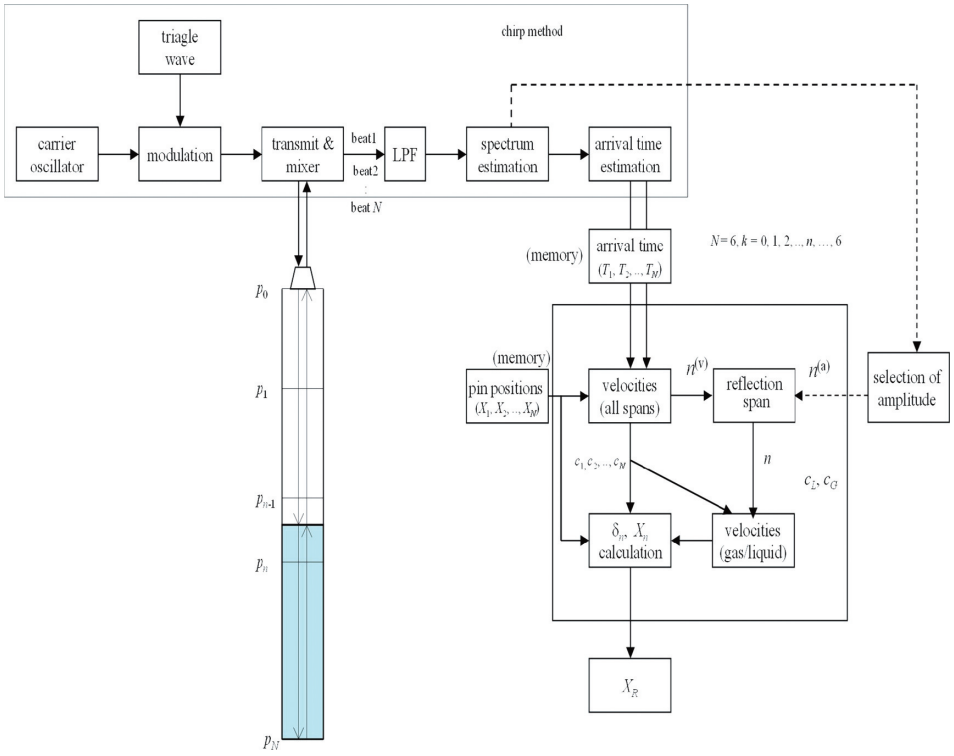


図 2．液面計測システムの一例．

時空間データ解析による生体メカニズムの解明

Elucidation of biological mechanism by spatiotemporal data analysis

モデリング研究系　三分一 史和（Fumikazu Miwakeichi）

1. 時空間解析によるニューロン間の因果的結合性とネットワーク構造の推定

このプロジェクトでは、兵庫医科大学、ドイツのゲッティンゲン大学との共同研究として呼吸リズムを形成するニューロンネットワーク機構の解明を目指している。具体的には、マウスの脳幹のスライス標本に蛍光色素を添加し、ニューロンの活動による細胞内カルシウム濃度の変化をイメージングデータ（動画）で記録するカルシウムイメージング法を用いて研究を行っている。自発的な呼吸を引き起こすニューロンの活性化タイミングを特定するために、標本表面に留置された電極により数十〜数百個の細胞の局所場電位データも同時計測しており、呼吸性のバースト波形が出現したときの画像を解析すれば、呼吸に関係するニューロンの検出と活性化タイミングを調べることができる（図 1）。

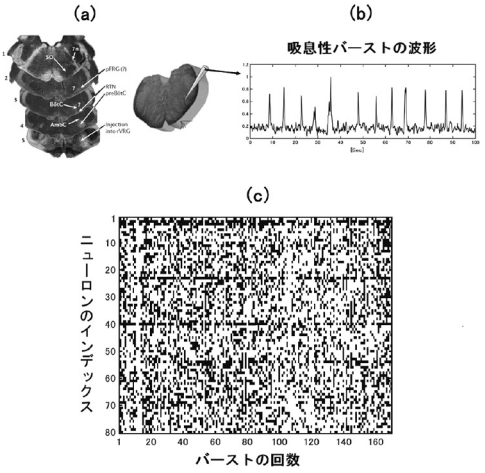


図 1. (a) ラットの脳幹のスライスの模式図
(b) 計測される吸息性バーストの波形 (局所場電位)
(c) ニューロンの活性化タイミングを記録したラスタープロット

これまでの研究において、動画のブレ補正法、ノイズ低減のための時空間フィルタリング法、などの事前処理法を開発した。そして、事前処理を行ったイメージングデータに遅延時間を考慮した相互相関解析を適用することにより、呼吸に関連する興奮性ニューロン、グリシン抑制性ニューロン、GABA 抑制性ニューロンの 3 種類のニューロンを特定することが可能となり、呼吸性ニューロン間の活性化順序は呼吸サイクルごとに変化するが、同じニューロンタイプで

は活性化順序に一定の規則性が存在するという重要な発見をした。

さらに進んで、呼吸リズムを生成するニューロンネットワーク構造の推定と同調メカニズムの解明を目指すには時空間解析の方法による因果的結合性の推定を行う必要がある。しかし、相互相関解析では、呼吸と関係するニューロンのと活性化タイミングの検出は可能であるが、ニューロン同士の因果的結合性を推定することができない。そこで、ニューロンの自励活動のみを考慮した自己回帰モデルと他のニューロンからの入力を考慮した外生変数型自己回帰モデルをカルシウムイメージングデータに適用し、これら 2 つのモデルの赤池情報量規準量の差を調べることにより、特定の興奮性ニューロン Y(図 2 ☆印) と因果的結合性を持つニューロンの推定を試みた。例えば、図 2(b) と (c) の両方に検出されているニューロン A と C はニューロン Y と双方向に結合しているが、ニューロン B は図 2(c) では検出されていないので、ニューロン Y からの因果性はあるが、ニューロン Y へ因果性はないという非対称的な結合性であるということが分かった。現在は、薬剤の添加や物理的な方法で脳領域やニューロン間の結合を離断させる阻害実験の前と後のデータ時系列モデルを適用し、パラメータを定量比較することにより、ニューロンネットワークの実在性の検証を行っている。

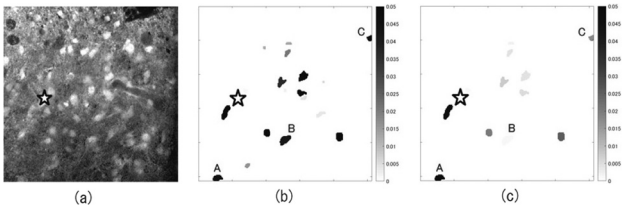


図 2. (a) 活動性ニューロンの空間分布
(b) ☆印の位置にある興奮性ニューロン (Y)”が”影響を及ぼしているニューロン
(c) ☆印の位置にある興奮性ニューロン (Y)”に”影響を及ぼしているニューロン

2. 時系列データにおける変化点の検出

人や動物の健康状態の把握や異常状態の検出のために生体システムのモニタリングが必要な場合が数多くある。複数のセンサで生体信号をリアルタイム計測し、信号間の相互相関解析あるいは回帰分析などでデータを逐次処理し、パラメータの変化から異常値を検出するというアプローチが一般的であるが、異常状態にあるときはデータの定常性が崩れていたり、非線形性が生じていたりして、一般的な処理方法の適用条件を満たさなくなり、解析そのものが困難になる。本研究では正常状態にある生体信号を用いて時系列モデルを同定し、その同定したモデルでデータを逐次フィルタリングを行うイノベーションアプローチに基づいた異常値検出方法の開発を行っている。具体的には、牛の体温データを用いた出産時刻の予測法の開発を目指している。体温はサーカディアンリズムに従った変動をしているが、出産直前に体温が低下することが知られており、この現象を利用して出産時刻の予測を行うことは原理的には可能である。しかし、体温データには気温や給餌タイミングなどの外部要因の影響も重畳しており、これらの影響を考慮する必要がある。これまでに、季節調整法に外生変数型自己回帰モデルを組み込んだ時系列モデルを用いたモニタリングシステムを構築し、実データを用いた検証を行っている。

マルコフ連鎖モンテカルロ法の世界を拡げる

― 高次元の多峰性分布からのサンプリング―

Expanding world of Markov chain Monte Carlo

– Methods for sampling high-dimensional multimodal distributions–

モデリング研究系 伊庭 幸人 （Yukito Iba）

要 旨

多峰性分布に強いマルコフ連鎖モンテカルロ法とその応用について述べた．

キーワード：レプリカ交換モンテカルロ法，マルチカノニカル法，多峰性分布，レアイベントサンプリング

1. はじめに

マルコフ連鎖モンテカルロ法 (MCMC) は高次元・多変量の確率分布からの乱数生成（サンプリング）の手法である．当初は統計物理の世界で使われていたのが，1990 年代以降に統計学や機械学習の世界で利用されるようになり，大きな影響をもたらしたことはよく知られている．著者は長年マルコフ連鎖モンテカルロ法の応用に興味を持っているが，その間，以下の 2 つの点にこだわりを持ち続けてきた．

1. 通常の MCMC ではサンプリングしにくい多峰性のある分布（確率の高い領域が複数に分かれている分布）に対して有効な手法の開発．
2. 統計データ解析でも統計物理でもない分野への応用．

MCMC というと，ベイズ統計やそのソフトウェアを思い浮かべる方が多いかもしれないが，それとはかなり違う問題意識で研究を進めてきたといえると思う．以下，その軌跡を紹介したい．

2. 拡張アンサンブル MCMC とその応用

多峰性の分布に対して MCMC を適用する手法として Simulated Annealing 法が知られている．Simulated Annealing 法では「温度」に対応するパラメータを最初は「高温」に設定し次第に「低温」にすることで，重要でない分布の山（局所的極大）にトラップされることを防ぐ．しかし，これはあくまで最適化の範疇であって，多峰性の分布を正しくサンプルすることができるとは限らない．統計物理での有限温度の分布，ベイズ統計での事後分布からのサンプリングには別のアイディアが必要である．1993 年ごろにこの問題に興味を持った著者は，今日「レプリカ交換モンテカルロ法」（パラレルテンパリング）と呼ばれている方法を独自に思いついたが，あまり知られていない先行研究（木村・瀧（1990）及び Geyer（1991））の存在に気づいたため，査読つき論文を出すには至らなかった（短報は「統計数理」の研究報告会要旨に掲載され，オンラインでも見られる． 伊庭（1993））．それから数年たってから考えたのは，温度に

限らず，収束を悪くしている特定の制約条件を緩めることでよりよい結果が得られるということで，この考え方をタンパク質の格子モデルに適用して開発した手法（MSOE 法）は当該分野で高く評価されている（第 2 報の論文 Chikenji et al.（1999）は web of science で引用数 74）．また，それまでに考えたことを踏まえて，関連する一連の手法を「拡張アンサンブル法」としてまとめたレビュー論文 Iba（2001）は 2018 年現在 web of science で 142 件引用されている．

3. さまざまな分野での高次元分布からのサンプリング

MCMC が有用な分野としては統計物理とベイズ統計が代表格であるが，これらの分野では，それぞれの主役である「ギブス分布（カノニカル分布）」「事後分布」が高次元の確率分布であるという点が共通である．そう考えると，統計物理やベイズ統計以外でも MCMC によって面白いことができる分野はもっとあるのではないか．「最適化からサンプリングへ」というスローガンは，単に MCMC の応用ということを超えて，いろいろな学問の世界で「確率構造」「測度構造」を考える糸口になるかもしれない．過去 10 年ほど行ってきた研究は，こうした発想に基づくものである．

具体的な応用としては，ランダム行列やランダムグラフの大偏差の数値計算，力学系の珍しい軌道のサンプリング，ランダム系の不純物に関する平均（クエンチ平均）の効率的計算，ラテン方陣や分割表の数の計算，複雑な帰無仮説下での検定統計量の分布の計算（サロゲーション法への応用）などがある．これらの例の多くでは，与えられた分布の極端な裾からのサンプリングが必要とされるが，そこで多峰性分布からのサンプリングの手法が重要となる．マルチカノニカル法，レプリカ交換モンテカルロ法などを利用することで，与えられた確率分布のもとでの生起確率が 10^{-200} といった極端なレアイベント（大偏差事象）のサンプリング及び確率計算ができることが示されている．興味を持たれた方は，解説 Iba et al.（2014）の中の事例と引用文献を参照されたい．

4. 今後の展望

マルチカノニカル法やレプリカ交換モンテカルロ法（パラレル・テンパリング）など多峰性分布に対応した MCMC 手法は，ベイズ統計にもとづくデータ解析でも有用なはずであり，研究レベルではレプリカ交換モンテカルロ法を中心にいろいろな適用例がある．しかし，現場での応用が MCMC ソフトウェア中心に展開しているため，十分活用されていないのが現状である．これらの手法に対応する統計ソフトウェアが登場することで，より自由な統計モデリングが行えるようになることを望みたい．

参 考 文 献

Chikenji, G., Kikuchi, M. and Iba, Y. (1999). Multi-self-overlap ensemble for protein folding: ground state search and thermodynamics, *Physical Review Letters*, **83** (9), 1886–1889.

Geyer, C. J. (1991). *Computing science and statistics: Proceedings of 23rd Symposium on the Interface* (ed. E. Keramidas), , 156–163, Interface Foundation, Fairfax Station.

Iba, Y. (2001). Extended ensemble Monte Carlo, *International Journal of Modern Physics C*, **12** (05), 623–656.

Iba, Y., Saito, N. and Kitajima, A. (2014). Multicanonical MCMC for sampling rare events: an illustrative review, *Annals of the Institute of Statistical Mathematics*, **66** (3), 611–645.

伊庭幸人 (1993). メトロポリス的モンテカルロ法の緩和について (統計数理研究所研究活動 (平成 4 年度 研究報告会要旨)), 統計数理, **41** (1), 65–67.

木村宏一, 瀧 和男 (1990). 時間的一様な並列アニーリングアルゴリズム, 電子情報通信学会 NC-90-1, 1–8.

正定値カーネルによる統計的機械学習

Statistical Machine Learning by Positive Definite Kernels

数理・推論研究系 福水 健次 （Kenji Fukumizu）

要　　旨

正定値カーネルとそれが定める再生核ヒルベルト空間を用いて，確率分布を表現する方法論を確立し，さまざまな統計的問題に適用してきた．その一連の研究に関して概説する．

キーワード：正定値カーネル，再生核ヒルベルト空間，統計的推論，検定，ベイズ推定

1. はじめに

カーネル法は，データを（非線形）写像することによってデータの高次モーメントを扱う方法論であり，サポートベクターマシンの提案以来，機械学習の主要技術の一つとして発展してきた（福水, 2010）．データに変換を施してから解析する手法は古くから存在するが，カーネル法の特徴は，特殊な内積を持つ関数空間への写像を用いることにより，写像後のデータに対する線形の処理が効率的に行える点にある．

正定値カーネルとは，集合Ω（データが存在する空間） 上に定義された対称な 2 変数関数 $k : \Omega \times \Omega \to \mathbb{R}$ で，任意の点 $x_1, \ldots, x_n \in \Omega$ に対しグラム行列 $(k(x_i, x_j))$ が半正定値性を満たすものである．Ω 上の正定値カーネル k に対し，Ω 上の関数からなるヒルベルト空間 H が定まり，カーネル法ではこれを「特徴空間」と呼ぶ．このヒルベルト空間は特別な内積を有しており，第 2 変数を $x \in \Omega$ に固定して第 1 変数に関する関数とみなした $k(\cdot, x)$ と任意の関数 $f \in H$ の内積が，$\langle f, k(\cdot, x)\rangle_H = f(x)$ と関数値に一致する．この性質を再生性といい，ヒルベルト空間 H を再生核ヒルベルト空間と呼ぶ．

データ解析に正定値カーネルを用いる際には，データの存在する空間 Ω に正定値カーネル k を定め，次の「特徴写像」によって特徴ベクトル $\phi(x) \in H$ を仮想的に作成する．

$$\phi : \Omega \to H, \qquad x \mapsto k(\cdot, x).$$

再生性を用いると，

$$\langle \phi(x), \phi(y)\rangle_H = k(x, y)$$

が得られるが，これは 2 つの特徴ベクトル $\phi(x), \phi(y)$ のヒルベルト空間 H における内積が，正定値カーネルの値の評価によって容易に計算されることを意味しており，カーネル法の鍵となる．データ $x_1 \ldots, x_n$ に対して特徴ベクトル $\phi(x_1), \ldots, \phi(x_n)$ を想定し，これらに線形回帰，主成分分析など様々な既存のデータ解析手法を適用したものがカーネル法として総称されている．

2. 分布の表現とその応用

Ω 上の確率分布 P に対し $\mu_P := \int k(\cdot, x) dP(x) \in H$ という関数を，分布 P の表現として

用いることが可能である．これを**カーネル平均**と呼ぶ．k が非線形カーネルの場合，カーネル平均は分布 P の高次モーメントの情報を有している．特にガウスカーネルなどは，分布 P から再生核ヒルベルト空間 H への 1 対 1 写像 $P \mapsto \mu_P \in H$ を定めることが知られており，確率分布の特性関数と同様の役割を持つ．このようなカーネルを特性的カーネルと呼ぶ（Fukumizu et al., 2004）．確率 P に従う有限個の i.i.d. データ $X_1, \ldots, X_n$ が与えられた場合，μ_P を $(1/n)\sum_{i=1}^{n} k(\cdot, X_i)$ により推定することが可能である．

同時分布 P を持つ確率変数 $(X.Y)$ に対して（非心）**共分散作用素**を $C_{YX} = E_P[k_X(\cdot, X)k_Y(\cdot, Y)]$ により定める．ここで，k_X, k_Y はそれぞれ X, Y が値をとる空間に定義された正定値カーネルである．C_{YX} は有限次元確率ベクトルの共分散行列の一般化であり，X と Y の統計的関係を表現している．これも，グラム行列を用いて有限サンプルから容易に推定可能である．

近年，筆者を含むグループらにより，カーネル平均と共分散作用素を様々な統計的問題へ適用する研究がなされてきた．例えば，Fukumizu et al.（2004）では，特性的なカーネルを導入し，確率変数の条件付き独立性を共分散作用素によって特徴づけ，それを回帰問題における次元削減に応用している．また，２つの分布の距離をカーネル平均の距離により定義し，２標本問題に応用する研究や，独立性，条件付き独立性の検定にカーネル法を応用する研究などが発展している．これらカーネル平均を用いた方法に関しては Song et al.（2013）や Muandet et al.（2017）などを見ていただくとよい．

また，Fukumizu et al.（2013）は，カーネル平均と共分散作用素を用いてベイズ事後確率を推定する方法を提案した．この方法は状態空間モデルにおけるフィルタリングの問題などに応用されている．

カーネル平均や共分散作用素を用いたデータ解析は一般的なノンパラメトリック推論の方法論であり，最近では，近似ベイズ計算への応用，因果推論への応用，さらに深層学習と組み合わせた生成モデルへの応用など，大きな広がりを見せている．本稿によって興味を持たれた方がさらに研究を発展させていくことを期待している

参 考 文 献

福水健次 (2010) 『カーネル法入門 – 正定値カーネルによるデータ解析』, 朝倉書店, 東京.

Fukumizu, K., Bach, F.R. and Jordan, M. I.（2004）. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, *Journal of Machine Learning Research*, 5, 73–99.

Song, L., Gretton, A., and Fukumizu, K. (2013) Kernel Embeddings of Conditional Distributions. *IEEE Signal Processing Magazine* 30(4), pp. 98–111.

Muandet, K., Fukumizu, K., Sriperumbudur, B. and Sch ́olkopf, B. (2017), Kernel Mean Embedding of Distributions: A Review and Beyond, *Foundations and Trends in Machine Learning*, Vol. 10: No. 1–2, pp 1–141.

Fukumizu, K., Song, L., and Gretton, A. (2013) Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels. *Journal of Machine Learning Research*, 14, 3753–4783.

位置情報軌跡の匿名化技術
Anonymization Techniques for Trajectory Location Datasets

モデリング研究系 南 和宏（Kazuhiro Minami）

1. 位置情報の匿名化における課題

スマートフォンの普及に伴い，我々の位置情報の取得が容易になり，多くのユーザーの移動履歴は，交通情報の提供，都市設計といった社会サービス，また商圏分析等の商用ビジネスにも活用されている．その一方，位置情報から，個人の興味に関するプライバシーな情報が漏洩する危険性が懸念される．よって位置情報の安全な２次利用には，匿名化と呼ばれる個人の識別情報を取り除くデータ処理が不可欠である．一般に，位置情報の匿名化処理では，氏名等の個人の識別子を削除するだけでは不十分である．なぜなら目撃情報，名簿等の外部知識から特定の日時，場所の位置情報が識別され，その結果，その位置情報を含む軌跡全体が特定されるリスクが存在するからである．したがって，位置情報から k 未満のユーザーへの絞り込みを防ぐための k-匿名化処理（Sweeney, 2002）が必要となる．

しかし通常の k-匿名化の手法を位置情報軌跡に適用する場合，２つの課題が存在する．１つは，位置情報軌跡のような時系列データの場合，k-匿名化を実施するとデータの有用性が著しく劣化する問題である．長期の位置情報軌跡を匿名化する場合，k-匿名化の前提となる軌跡群へのグループ化が困難である．そのため，k-匿名化を実現するための一般化処理による情報損失は大きくなり，有益なデータ分析に堪えなくなる．２つめは，位置情報軌跡のデータ間に時空間の相関性が存在し，匿名化した位置情報から統計的推論により元の軌跡情報が復元される問題である．位置情報軌跡には，人の移動に関する物理的制約が反映し，車，電車といった交通手段により移動経路は限定される．また長期的な移動軌跡には通勤，病院への通院といった個人の生活習慣を反映した特徴的なパターンが現れる．そのような移動パターンに関する外部知識を用いると匿名化された位置情報から元の位置情報が復元される危険性がある．

近年，著者はこれらの課題を解決するための２つの匿名化技術に取り組んできた．１つは位置情報軌跡を複数のセグメントに分割する動的仮名交換手法（Tanjo et al., 2014）であり，ミックスゾーンと呼ばれる複数ユーザーの集積点でのランダムな仮名の再割当により移動先の不確定性を確保する手法である．もう１つは，ユーザーの移動パターンをマルコフ過程でモデル化し，隠れマルコフモデルにおける内部状態の推定問題として匿名化データの安全性の評価を行う手法（Minami, 2014）である．

2. ミックスゾーンにおける動的仮名割当

個人の行動パターンが顕著に現れる位置情報軌跡の場合，その中の幾つかの点に過ぎない外部知識を用いて軌跡全体の識別が可能であり，情報漏えいリスクが非常に高い．位置情報軌跡の開示リスクを局所するため，Mano et al.（2013）では位置情報軌跡に紐付けられる仮名を動的に更新し長期間の軌跡データを複数の軌跡セグメントに分割する方式を提案した．この仮名の更新は複数のユーザーが同一の時間，場所に存在する「ミックスゾーン」と呼ばれる領域でラ

ンダムな仮名交換の形式で実施し，ミックスゾーン前後の軌跡セグメント間の関連性を分断する．個人の位置情報はミックスゾーンを経由することで代替経路が増大するので，その不確定性に着目して位置情報軌跡の仮名化データに対するプライバシー指標を定式化した．ただし，攻撃者の外部知識と整合性を保持する代替経路の列挙には，ミックスゾーンを頂点，位置情報の軌跡セグメントを辺とするグラフにおける排他的辺素パス問題を解く必要がある．一般の排他的辺素パス問題は NP 困難であるため，(Tanjo et al., 2014)では排他的辺素パス問題を制約充足問題に変換する効率的な安全性評価手法を開発した．

3. 隠れマルコフモデルに基づく匿名化処理の安全性評価

仮名更新による位置情報軌跡の分割は位置情報軌跡の識別リスクを局所化する手法である．しかし位置情報が識別されると同じ軌跡セグメント内の位置情報は依然として漏洩してしまう．したがって分割して次元を削減した軌跡セグメント単位に k-匿名化を実施することが望ましい．ただし，位置情報軌跡には時空間の相関性が存在するので，通常の k-匿名化では不十分な場合が多い．

Minami (2014)では，ユーザーの移動パターンをマルコフ過程でモデル化し，匿名化データの安全性を隠れマルコフモデルにおける観測情報から内部状態の推定問題として定式化した．モデルにおける観測情報は匿名化データ，内部状態遷移は秘匿すべき元の位置情報に対応し，匿名化アルゴリズムは，内部状態から観測情報への確率的な変換を定義する記号出力行列として記述される．さらに匿名化データの安全性は，観測情報，記号出力行列，内部状態のマルコフ過程が与えられたときに正しく内部状態を推定する条件付き確率として定式化した．実データを用いた評価実験では，通常の k-匿名化処理では想定した安全性が確保できず，追加の秘匿処理の必要性を明らかにし，統計モデルに基づく安全性評価の有用性を実証的に示すことができた．

参 考 文 献

Mano, K., Minami, K. and Maruyama, H. (2013). Protecting Location Privacy with K-Confusing Paths Based on Dynamic Pseudonyms, *5th IEEE International Workshop on SEcurity and SOCial Networking*, March.

Minami, K. (2014). Preventing denial-of-request inference attacks in location-sharing services, *2014 Seventh International Conference on Mobile Computing and Ubiquitous Networking*, 50–55, January.

Sweeney, L. (2002). k-anonymity: a model for protecting privacy, *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, **10** (5), 557–570.

Tanjo, T., Minami, K., Mano, K. and Maruyama, H. (2014). Evaluating data utility of privacy-preserving pseudonymized location datasets, *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, **5** (3), 63–78, September.

神経スパイク時系列パターンの特徴解析

Statistical analysis of neural spike trains

モデリング研究系 小山 慎介（Shinsuke Koyama）

要　　旨

本研究では，神経スパイク時系列の統計パターンを記述する方法を提案する．提案方法の新規性は，スパイク発火時刻のゆらぎ（分散）を平均のべき関数で表すところにある．べき関数のパラメータを調整することで，脳内で観測されるさまざまなスパイク発火ゆらぎを再現することができる．この枠組みに基づきスパイク時系列の点過程モデルを構築し，データからパラメータを推定する方法を提案する．

キーワード：スパイク時系列，点過程，ゆらぎのスケーリング則

1. はじめに

高度な情報処理を行う脳は，異なる働きを持つ部分が有機的に組み合わされたヘテロ構造体である．このような見方は，およそ 100 年前にブロードマンが解剖学的・細胞構築学的観点から大脳皮質を区分して「脳地図」を描いたことに始まる．今日，その区分が脳の機能と密接に関係していることも明らかになっている．

近年 Shinomoto et al.(2009) は、神経スパイク発火パターンの大脳皮質全体にわたる非一様性に着目した．スパイク発火の不規則性を測るために「局所変動係数 (local variation) L_V」を提案し，大脳皮質のさまざまな領野のデータから L_V を求めたところ，細胞ごとに固有の発火パターンがあり，それが大脳皮質の機能とも相関している，という事実を見いだした．この発見は，大脳皮質が機能的に区分できるという事実が神経細胞が用いる信号レベルにも反映されているということを示唆している．本研究の目的は，スパイク発火パターンをより系統的に特徴付けるための統計的方法を構築することである．

2. アイデアのスケッチ

Shinomoto らが提案した局所変動係数 L_V は，時系列の非定常性の影響を取り除いたスパイク間隔 (ISI) の変動係数 (coefficient of variation C_V = 標準偏差/平均) とみなせる．したがって彼らが見いだしたことは，局所的に定常と見なせる短い時間スケールで

$$\mathrm{Var(ISI)} = \phi \mathrm{E(ISI)}^2 \tag{2.1}$$

という関係を持つことを仮定すると，領野・皮質層ごとに固有な ϕ の値をとり，神経細胞を ϕ の値で大まかに分類できると理解できる．

一方で Troy and Robson(1992) は，定常な光輝度刺激に対する網膜ガングリオン細胞のスパイク間隔の C_V を推定したところ，おおよそ $C_V^2 \propto \mathrm{E(ISI)}$ という関係があることを発見した．これを書き換えると

$$\mathrm{Var}(\mathrm{ISI}) = \phi \mathrm{E}(\mathrm{ISI})^3 \tag{2.2}$$

という平均と分散の関係式が得られる．ここで式 (2.1) と比べてべき指数が異なることに着目しよう．

式 (2.1) と (2.2) を一般化してべき指数 α を導入すれば，少なくとも局所的に定常と見なせる時間スケールで，スパイク間隔の平均と分散の間に

$$\mathrm{Var}(\mathrm{ISI}) = \phi \mathrm{E}(\mathrm{ISI})^\alpha \tag{2.3}$$

というスケーリング則が得られる．これが提案する方法の根幹をなす仮定である．本研究の目的は，式 (2.3) を仮定してスケール因子 ϕ とべき指数 α を用いてスパイク発火パターンを特徴付けるための統計的方法を構築することである．

まず定常リニューアル過程を考えよう．ISI の平均を $\mu = \mathrm{E}(\mathrm{ISI})$ として $\mathrm{ISI} \to \mu^{-1}\mathrm{ISI}$ とスケール変換すると，式 (2.3) より $\mathrm{E}(\mathrm{ISI}) \to 1$, $\phi \to \mu^{\alpha-2}\phi$ とリスケールされる．したがって，平均と分散がそれぞれ μ と $\phi\mu^\alpha$ である確率密度関数 $f(x;\mu,\phi)$ でこのスケール変換に対して不変なものは

$$f(x;\mu,\phi) = \mu^{-1} f(\mu^{-1}x; \mu^{\alpha-2}\phi) \tag{2.4}$$

を満たす．ここで $f(x;\phi) := f(x;1,\phi)$ である．すなわち，平均が 1 で分散が ϕ である任意の確率密度関数を式 (2.4) でスケール変換することによって式 (2.3) を満たす確率密度関数を作ることができる．

Koyama (2015) では，このアイデアを非定常リニューアル過程に拡張することで，ゆらぎのスケーリング則をもつスパイク時系列モデルを提案した．また，このモデルに基づいてパラメータを推定する方法を提案し，実験データに適用して有効性を確認した．

3. まとめと展望

本研究では，スパイク発火時刻のゆらぎを記述するためのスケーリング則を定式化し，これに基づいてスパイク時系列の統計モデルとパラメータ推定方法を提案した．この方法を脳の幅広い部位から記録したデータに適用してスパイク発火パターンを特徴付け，脳機能との関連を調べることが今後の課題である．

参 考 文 献

Koyama, S. (2015). On the spike train variability characterized by variance-to-mean power relationship, *Neural Computation*, **27**, 1530–1548.

Shinomoto. S. et al. (2009). Relating neuronal firing patterns to functional differentiation of cerebral cortex, *PLoS Computational Biology*, **5**, e1000433.

Troy, J. B. and Robson, J. G. (1992). Steady discharges of X and Y retinal ganglion cells of cat under photopic illiminance. it Visual Neuroscience, **9**, 535–553.

ノンパラメトリックエントロピー推定とその応用
Non-parametric Entropy Estimation

モデリング研究系 日野 英逸 （Hideitsu Hino）

情報理論において最も基本的な量の一つである情報量は $I_f(x) = -\log f(x)$ で定義される．ここで $f(x)$ はデータ $x \in \mathbb{R}^d$ が従う分布の確率密度関数である．情報量の期待値はエントロピーと呼ばれる: $H(f) = E_f[I_f(X)] = -\int f(x)\log f(x)\mathrm{d}x$. 情報量，エントロピーあるいはこれらを用いて導出できる KL ダイバージェンス及び相互情報量は統計学や機械学習など非常に多くの分野で重要な役割を果たしている．

微分エントロピーの推定方法として最も良く利用されている手法の一つが，k-近傍法と確率密度関数の 1 次の展開に基づく推定法である．最近傍法に基づくエントロピー推定量は, Kozachenko and Leonenko（1987）により提案され，任意の次元の確率変数に対して mean square consistency を持つことが示されている．この結果は一般の k-近傍に基づく推定量に拡張され（Goria et al., 2005），その後も各種の拡張と理論的解析がなされている（Beirlant97, 1997; Paninski, 2003）．確率密度関数 $f(z)$ の検査点 $z \in \mathbb{R}^p$ における値を観測データ集合 $\mathcal{D} = \{x_i\}_{i=1}^n$ を用いて推定する問題を考える．検査点 z を中心とする半径 ε の p 次元超球を $b(z;\varepsilon) = \{x \in \mathbb{R}^p | \|z - x\| < \varepsilon\}$ で表す．この超球の体積は $|b(z;\varepsilon)| = c_p\varepsilon^p$ である．ただし，$c_p = \pi^{p/2}/\Gamma(p/2+1)$ である．中心 z の ε 球に含まれる確率質量を $q_z(\varepsilon) = \int_{x\in b(z;\varepsilon)} f(x)\mathrm{d}x$ で定義する．この定義式の被積分関数を Taylor 展開すると $q_z(\varepsilon) = |b(z;\varepsilon)|\left(f(z) + O(\varepsilon^2)\right) = c_p\varepsilon^p f(z) + O(\varepsilon^{p+2})$ を得る．超球の半径 ε を十分小さいと仮定してその 2 次以上の項を無視し，確率質量を全観測データに占める超球内の点の割合で近似することで，確率密度関数の推定量 $\hat{f}(z;\varepsilon) = \frac{k_\varepsilon}{nc_p\varepsilon^p}$ を得る．ここで k_ε は観測データ集合 $\mathcal{D}$ の中で半径 ε の超球の中に含まれるものの個数である．一方，ε の代わりに超球に含まれるサンプル数 k を固定した場合，$\hat{f}(z;\varepsilon)$ は $\hat{f}^{nn}(z;k) = k/(nc_p\varepsilon_k^p)$ のようにかける．ここで超球の半径 ε_k は検査点 z からその k 番目に近い点までの距離で決定されることになる．$\hat{f}_i^{nn}(x_i;k)$ を，データ集合 $\mathcal{D}\backslash\{x_i\}$ を用いて k-近傍法により推定した推定量として，$-\ln\hat{f}_i^{nn}(x_i;k)$ の経験期待値を計算することで，k-近傍エントロピー推定量 $\hat{H}^{nn}(\mathcal{D};k) = -\sum_{i=1}^n \ln\hat{f}_i^{nn}(x_i;k)$ を得る（Goria et al., 2005）．この方法は確率質量関数の一次展開に基づく方法であるが，筆者らはより高次の展開に基づく手法を提案した（Hino et al., 2015）．検査点 z を中心とした半径 ε の超球内の確率質量 $q_z(\varepsilon)$ は ε に関する二次の Taylor 展開をすると，$q_z(\varepsilon) = c_p f(z)\varepsilon^p + \frac{n}{4(p/2+1)}c_p\mathrm{Tr}\nabla^2 f(z)\varepsilon^{p+2} + O(\varepsilon^{p+4})$ の形で表わされる．上式左辺の $q_z(\varepsilon)$ を比 k_ε/n で近似し，両辺を $c_p\varepsilon^p$ で割ることで，$\frac{k_\varepsilon}{nc_p\varepsilon^p} = f(z) + C\varepsilon^2 + O(\varepsilon^4)$ を得る．ここで $C = n\mathrm{Tr}\nabla^2 f(z)/4(p/2+1)$ である．さらに，$Y_\varepsilon = \frac{k_\varepsilon}{nc_p\varepsilon^p}$ と $X_\varepsilon = \varepsilon^2$ を導入し，ε に関する 4 次以上の項を無視することで，応答変数 Y の説明変数 X に関する一次式 $Y_\varepsilon \simeq f(z) + CX_\varepsilon$ が得られる．この式は説明変数 X による応答変数 Y の線形回帰式とみなせる．複数の半径 $\mathcal{E} = \{\varepsilon_i\}_{i=1}^m$ と置き，$\mathcal{E}$ に含まれる各 ε で定まる $\{(X_\varepsilon, Y_\varepsilon)\}_{\varepsilon\in\mathcal{E}}$ の組を観測データとみなして，二乗誤差 $R = \frac{1}{m}\sum_{\varepsilon\in\mathcal{E}}(Y_\varepsilon - f(z) - CX_\varepsilon)^2$ を最小化するように $f(z)$ と C を求める．これは単回帰に他ならず，この回帰によって得られた切片が，$f(z)$ の推定量 $\hat{f}(z)$ である．以上より検査点 z における密度の推定量 $\hat{f}^s(z)$ が得られたので，leave-one-out 推定量とし

てエントロピーの推定量 $\hat{H}^s(\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_i^s(x_i)$, を得る．ここで $\hat{f}_i^s(x_i)$ は x_i を用いずに求めた密度の推定値である．筆者らは，この推定量を始めとして，より直接的に一度の回帰問題を解くことでエントロピーを推定する手法，誤差構造として Poisson 分布を仮定した手法（Hino et al., 2016），及び確率質量関数の高次展開に基づく局所フラクタル次元推定量（Hino et al., 2017）を提案し，ノンパラメトリックなアプローチによるデータ分布の特徴付けの方法論の開発を進めている．

また，関連して筆者らは一般にデータに重みが与えられている場合に適用可能な情報量推定手法を提案している（Hino and Murata, 2013）．例えば観測されたデータそれぞれに対して，その信頼度が与えられている場合や，あるいは同一の事象が重複して観測されるとして個々の事象を観測した頻度を重みとして表現する場合のように，データ $\mathcal{D}_x$ が与えられた上で，各データ点 $x_i \in \mathcal{D}_x$ の重要さとして重みが付与されることが考えられる．過去のデータの重みが小さくなっていくような忘却係数付きのオンライン観測データもこうした重み付きデータの一例である．この重み付き情報量推定量の応用として情報論的クラスタリング（Hino and Murata, 2014），変化点検知（Koshijima et al., 2015）などの方法を開発した．情報論的クラスタリング手法は同種の手法と比較してより正確なクラスタリングを実現し，変化点検知の手法は従来の手法では捉えられなかったデータの背後にある構造的な変化を抽出することに成功している．

参 考 文 献

Beirlant, J., Dudewicz, E. J. , Györfi, L. and Meulen, E. C. (1997). Nonparametric Entropy Estimation: An Overview, *International Journal of the Mathematical Statistics Sciences*, **6**, 17–39.

Goria, M. N. and Leonenko, N. N. and Mergel, V. V. and Novi Inverardi, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses, *Journal of Nonparametric Statistics*, **17**(3), 277–297.

Hino, H., Fujiki, J., Akaho, S. and Murata, N. (2017). Local Intrinsic Dimension Estimation by Generalized Linear Modeling, *Neural Computation*, **29**(7).

Hino, H.,and Murata, N. (2013). Information estimators for weighted observations, *Neural Networks* , **46**(0), 260–275.

Hino, H.,and Murata, N. (2014). A Non-parametric information theoretic clustering algorithm based on Quantile-based entropy estimator, *Neural Computation* , **26**(9), 2074–2101.

Hino, H., Koshijima, K. and Murata, N. (2014). Non-parametric entropy estimators based on simple linear regression, *Computational Statistics & Data Analysis*, **89**(0), 72 – 84.

Hino, H., Akaho, S. and Murata, N. (2016). An Entropy Estimator Based on Polynomial Regression with Poisson Error Structure, *Neural Information Processing - 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16-21, 2016, Proceedings, Part II*, 11 – 19.

Koshijima, K., Hino, H. and Murata, N. (2015). Change-Point Detection in a Sequence of Bags-of-Data, *Knowledge and Data Engineering, IEEE Transactions on*, **27**(10), 2632–2644, Oct.

Kozachenko, L. F. and Leonenko, N. N. (1987). Sample estimate of entropy of a random vector, *Problems of Information Transmission*, **23**, 95–101.

Paninski, L. (2003). Estimation of entropy and mutual information, *Neural Comput.*, **15**, 1191–1253, June.

組合せ論的系統学における最近の話題

－グラフ理論が拓く系統解析の新展開－

Recent topics in combinatorial phylogenetics

—From graph theory to statistical analysis—

モデリング研究系　早水 桃子（Momoko Hayamizu）

要　　旨

生物の進化の道筋を解明する系統学的なデータ解析において，今日では系統樹を拡張した「系統ネットワーク」というグラフ構造が広く用いられるようになっているが，記述能力が高い汎用的なモデルが必ずしも優れたモデルであるとは限らないため，系統ネットワークのサブクラスの中で生物学的な妥当さと数学的な性質の良さを兼ね備えたものを見出すことは重要である．特に Francis and Steel (2015)が定義した「系統樹ベースのネットワーク」(tree-based phylogenetic network; TBN）は系統樹に辺を追加する単純な操作で得られる系統ネットワークのサブクラスで，TBN の数学的性質や計算複雑性に関する未解決問題は理論生物学分野のホットトピックスになっている．Hayamizu (2018)は，Francis and Steel が取り上げた決定／探索問題や数え上げ問題だけでなく，TBN に関する列挙問題や最適化問題にもスポットライトを当て，これらの問題を高速に解くアルゴリズムを統一的な視点で生み出す「TBN の構造定理」を証明し，多様な統計学的な応用を可能にした．本稿では，その研究成果の一端を紹介する．

キーワード：系統樹推定，系統ネットワーク，細分系統樹，離散アルゴリズム

1. 研究の背景

生物の進化は古くから系統樹（phylogenetic tree）を用いて記述されてきたが，例えば植物，菌類，細菌類が進化する過程では異種交雑（hybridization）や遺伝子の水平伝播（horizontal gene transfer; HGT）といった木構造で記述しきれない現象が起きうるため，あらゆる種の進化を系統樹だけで描写することはできないといわれている．また，仮にそのような現象を考慮しなくてもよい種を対象にした系統解析を行う場合でも，現実のデータを扱い，その情報を忠実に描写したいなら，木構造よりも融通の利くグラフ構造が欲しいと考えるのは自然であろう．

このようなニーズに動機づけられ，**組合せ論的系統学**（*combinatorial phylogenetics*）という理論生物学の一領域では，系統樹を拡張した**系統ネットワーク**（*phylogenetic network*）とその様々なサブクラスに関する研究がこれまでに多数行われている（Huson *et al.* (2010); Steel (2016)）．その研究成果は既に実際のデータ解析に応用されており，例えば，SplitTree などのソフトウェアは系統ネットワークを使ってデータを可視化するツールとして広く使われている（Bryant and Moulton (2004)）．ただし，このトレンドは系統ネットワークが系統樹に取って代わることを意味しているのではなく，系統樹は依然として進化を記述するファンダメンタルなモデルであることを強調しておく．

2. 系統樹ベースのネットワーク（TBN）と細分系統樹

興味のある現存種の集合 X を葉とする根付き二分系統ネットワーク N が与えられているとき，これらの種が辿った進化の道筋を系統樹モデルで記述するとなれば，N の中に X を葉とする何らかの系統樹 T を見出したくなる．そこで Francis and Steel (2015)は，系統樹に余分な辺を加えてできる**系統樹ベースのネットワーク**（*tree-based phylogenetic network*; *TBN*）を定義した．TBN は**細分系統樹**（*subdivision tree*）という全域木を持つ系統ネットワークと定義することもできるため，TBN を論じるうえで細分系統樹という概念は本質的な役割を果たす．

3. TBN の構造定理が導く一連のアルゴリズムと統計学的な意義

Hayamizu (2018)は，根付き二分系統樹 N の細分系統樹の集まり $\{T_1, \ldots, T_{\alpha(N)}\}$ を特徴づける構造定理を示し，次の一連の問題を解く高速なアルゴリズムを統一的な視点で記述した．

(1)**決定／探索問題**： 系統ネットワーク N が与えられたとき，N が TBN か（すなわち細分系統樹が存在するか）否かを決定し，存在するならば一つ見つける問題．Francis and Steel (2015)はこの問題を解く線形時間アルゴリズムを与えたが，次のように数え上げ問題に拡張すると，多項式時間では解けないかもしれないと予想していた．

(2)**数え上げ問題**: 系統ネットワーク N が与えられたとき，N の細分系統樹の個数 $\alpha(N) \in \mathbb{Z}_{\geq 0}$ を求める問題．Hayamizu (2018)は，これを解く線形時間アルゴリズムを与え，N が TBN のとき，$\alpha(N)$ は N の複雑さを評価する尺度になるため，モデル選択の文脈に関連する．

(3)**列挙問題**： 系統ネットワーク N が与えられたとき，N の全ての細分系統樹 $T_1, \ldots, T_{\alpha(N)}$ を列挙する問題．入力 N のサイズに関する多項式時間でこの問題を解くアルゴリズムが存在しないことはすぐに分かる（列挙したい解の個数 $\alpha(N)$ 自体が N のサイズに関する指数関数で表される場合があるため）．しかし，Hayamizu (2018)は，これを高速に解く線形時間遅延アルゴリズム（列挙アルゴリズムの中で最も効率的なクラスに属すもの）を与えた．全ての解ではなく指定の個数 $k \in \mathbb{N}$ の解のみを列挙するには，$O(k|V(N)|)$ 時間で十分である．これにより細分系統樹の一様サンプリングなどの応用が可能になる．

(4)**最適化問題**: 系統ネットワーク N と辺の重みづけ関数 $w \geq 0$ が与えられたとき，ある目的関数の値 $f(T)$ を最大化（または最小化）する細分系統樹 T を求める問題．力まかせ探索では指数時間を要するが，Hayamizu (2018)の構造定理は最適解を線形時間で求めるアルゴリズムを導く．この最適化問題は，N の各辺が存在するか否かの不確かさに応じた確率 w が与えられ，尤度または対数尤度 $f(T)$ を最大化するベストな細分系統樹 T を求めるという最尤推定の文脈で現れる問題である．

参 考 文 献

Bryant, D. and Moulton, V. (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks, *Molecular Biology and Evolution*, **21**, 255–265.

Francis, A. and Steel, M. (2015). Which phylogenetic networks are merely trees with additional arcs?, *Systematic Biology*, **64**, 768–777.

Hayamizu, M. (2018). A structural theorem for tree-based phylogenetic networks, preprint available at `arXiv:1811.05849[math.CO]`.

Huson, D. H. and Rupp, R. and Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*, Cambridge University Press.

Steel, M. (2016). *Phylogeny: Discrete and random processes in evolution*, SIAM.

統計数理研究所で開発されたRパッケージ
R packages developed at the Institute of Statistical Mathematics

モデリング研究系　中野 純司（Junji Nakano）*

1. はじめに

統計数理研究所は計算機の出現以来、その時代の最新の統計計算環境を維持してきた。そしてその上で種々のソフトウエアが開発・公開されてきている。ただ計算機技術は急速に発展しており、過去のソフトウェアをソースコードでそのまま配布しても使える人は限られてしまう。そのため統計科学技術センターを中心として、それらのソフトウェアをその時代の計算機環境に合うように保守し、さらに利用しやすくすることが試みられてきた。われわれは、統計科学分野でデファクトスタンダードになっている統計解析ソフトウェア R からそれらのソフトウェアを利用できるように、R のパッケージを作成して公開している。本稿ではそれらを紹介するが、多くは CRAN(`https://cran.r-project.org/` またはそのミラーサイト、例えば `https://cran.ism.ac.jp/`) から入手できる。

2. timsac

TIMSAC(TIMe Series Analysis and Control program) は、赤池弘次氏を中心として開発された時系列データの解析、予測、制御のための総合的プログラムパッケージである。オリジナル TIMSAC(TIMSAC-72) は 1972 年に発表され、その後、TIMSAC シリーズとして TIMSAC-74,TIMSAC-78,TIMSAC-84 が統計数理研発行の Computer Science Monograph に発表された。工業プロセスの最適制御、経済変動の分析など広い分野で現在でも利用されている。R パッケージ `timsac` は、FORTRAN で書かれているオリジナルプログラムの計算処理機能の多くをサブルーチン化し、R 関数を通して入出力を行い、必要であればその解析結果等を R でグラフィック表示することにより時系列データ解析を容易にしたものである。

3. SAPP

尾形良彦氏を中心として開発された SASeis (Statistical Analysis of Seismicity) は、地震活動などの統計的解析とモデリングのためのプログラムパッケージである。そこでは大森・宇津の公式と点過程 ETAS (Epidemic Type Aftershock Sequence) モデルを扱っているが、それは地震活動の標準化モデルとして世界各国で使用されている。そこで FORTRAN で書かれたこれらの計算処理機能を R パッケージにしたものが `SAPP` である。

4. TSSS

R パッケージ `TSSS` は、北川源四郎氏による書籍「FORTRAN 77 時系列解析プログラミング」(岩波書店、1993) に掲載されていたプログラムを基に作成された時系列データ解析のための関数群である。代表的な時系列のモデリングに必要な最小二乗法、最尤法、カルマンフィ

* 中央大学国際経営学部：〒192-0393 東京都八王子市東中野 742-1

ルタによる推定の方法、情報量規準 AIC を用いたモデルの評価・選択の方法が実現されている。TSSS は本書のデータをデータセットとして組み込んでおり、関数のドキュメントにおける例題の一部ではこれらのデータセットを用いている。なお、時変係数 AR モデルの時変分散と時変 AR 係数を推定する関数 (tvvar, tvar) については、OpenMP を使った拡張パッケージ tvvarOMP を作成しており、それにより並列処理も行えるようになっている。このパッケージは http://jasp.ism.ac.jp/ism/TSSS/ から入手できるが、近々、CRAN で公開予定である。

5. catdap

CATDAP (CATegorical Data Analysis Program) は、坂元慶行氏を中心に開発された最適な分割表（クロス表）の探索のためのプログラムである。最適な説明変数の選択に AIC (Akaike Information Criterion) が使われている。R パッケージ catdap は、FORTRAN で書かれた CATDAP の計算処理機能をサブルーチン化することにより、R からこれらを関数として利用できるようにした。最近、主として石黒真木夫氏によりいくつかの拡張が行われた。例えば現在の関数 catdap2 では連続値目的変数に適用できたり、目的変数、説明変数に欠測が含まれるデータに適用できたりするようになっている。

なお、関数 catdap2 の機能をさらに使い易くするため、パッケージ R commander (Rcmdr) を使ったメニューインタフェース (RcmdrPlugin.catdap) も利用可能になっている。(http://jasp.ism.ac.jp/ism/RcmdrPlugin.catdap/)

6. NScluster

R パッケージ NScluster は、ネイマン・スコット型空間クラスターモデルのシミュレーションとパラメータ推定のための関数群である。 これらの関数は U.Tanaka, Y. Ogata and K. Katsura, Simulation and estimation of the Neyman-Scott type spatial cluster models (Computer Science Monographs, No.34, 1-44, The Institute of Statistical Mathematics, 2008) の FORTRAN プログラムをもとに開発された。

パラメータ推定のためにシンプレックス法を用いているが、モデルによってはかなりの計算時間がかかる。そのため、この時間のかかる計算処理の部分を OpenMP を使って並列化している。OpenMP が利用可能な環境であれば、実行時間の大幅な短縮を図ることができる。

7. Rhpc

R パッケージ Rhpc は基本的な並列計算用パッケージ snow の流れをくむ R の並列化のためのパッケージである。その特徴は、並列化のために MPI ライブラリを用いるが 2GB 以上のデータ処理に対応している、多くの部分を C でプログラムして実行速度を上げている、R から MPI 外部プログラムを利用し易くなっている、などである。特に最近のスーパーコンピュータ上での利用を念頭において開発されている。

8. Rmpenv

R パッケージ Rmpenv は任意精度による実数と複素数の四則計算および基本的な数学関数、さらに行列積や逆行列を求める関数などを実現するパッケージである。現在機能拡張中であり、まだ公開にはいたっていない。

謝　辞

以上のパッケージの作成は嵯峨優美氏、中間栄治氏との共同研究によるところが大きい。また、オリジナルの FORTRAN プログラム作成者の方々に深く感謝したい。

混合分布モデルとデータ同化

Finite mixture models and data assimilation

モデリング研究系 上野 玄太 （Genta Ueno）

1. 混合分布モデルを用いたプラズマデータ解析

超高層物理学を専攻した大学院時から継続している研究課題である。宇宙空間のプラズマの速度分布は正規分布から離れた形状をとることが多く、複数のピークを持ったり分散の異なる分布の重ね合わせが見られたりする場合が多い。そのような複雑な分布の理解は、地上を優雅に照らすオーロラ現象のメカニズムの理解につながる。プラズマの速度分布を詳細に観測できるようになったのは、1992 年に日米共同で打ち上げられた人工衛星ジオテイルが最初であり、現在に至るまで 12 秒ごとに速度分布データを取得している。

Ueno et al.（2001）は、プラズマ速度分布データに対して正規混合分布モデルを当てはめ、複数の成分に分離を可能としたものである。分離のために推定した各成分のパラメータ (混合比、平均ベクトル、分散共分散行列) は、プラズマの密度が急変する境界層と呼ばれる領域のデータ解析における基礎的な物理量としてそのまま利用できる。このモデルの適用により、25 年間余りに取得・蓄積されている大量の速度分布データを対象にした統計解析の道を開いた。この手法により、磁気圏境界層での応用研究を進めた（Lui et al., 2005; Nishino et al., 2007a, d, c, b; Nakai and Ueno, 2011）。かつては高温プラズマに埋もれて取り扱いが困難であった低温プラズマの解析を実施したものである。このモデルはノイズ除去にも応用でき、電子観測データからの光電子成分の除去への応用が Ueno et al.（2001）にある。このモデルはもともとジオテイル衛星に搭載されたプラズマ観測器に対して開発したものであるが、その後、2007 年に NASA により打ち上げられた人工衛星テミスによるプラズマ観測データへ展開している（Chaston et al., 2013）。

応用研究と並行して、新しい方法論の提案を行った。中村 他（2005）は、プラズマ観測装置に検出不可能な速度があることに動機を得て、データ欠損領域が存在する場合の混合分布モデルを提案したものである。

2. データ同化の方法論への展開

2005 年度からは、大気海洋結合モデルを軸としたデータ同化手法の研究を集中的に進めた。Ueno et al.（2007）は大気海洋結合モデルにアンサンブルデータ同化手法を用いた初の研究である。つづいて、推定精度を上げることを目的として、システムノイズ・観測ノイズのパラメータの最適化を行うこととした。状態ベクトルの確率分布をアンサンブルによる近似表現すると、時系列モデルの尤度関数の表式が混合分布モデルのそれと同等の形になるところが面白い。この表式の類似性に着目することで、時系列モデルにおいても EM アルゴリズムを導出が可能であり、特に観測ノイズの分散共分散行列の推定に有効である（Ueno and Nakamura, 2014）。最尤法によるパラメータ推定を大気海洋結合モデルに対して行った結果、予測精度の向上を実現すると同時に、データ同化による状態推定の限界を明らかにした（Ueno et al., 2010）。限界の一つは、静穏時の大気海洋の状態とエルニーニョなど変動時の状態で、結合モデルによる再現

性に違いが見られたことである。その状況を解決するため、パラメータに時変性を許し、観測状況に適応的にフィルタのゲインを推定するベイズ法を開発した(Ueno and Nakamura, 2016; Nakabayashi and Ueno, 2017)。

参 考 文 献

Chaston, C. C., Yao, Y., Lin, N., Salem, C. and Ueno, G. (2013). Ion heating by broadband electromagnetic waves in the magnetosheath and across the magnetopause, *Journal of Geophysical Research: Space Physics*, **118**, DOI: http://dx.doi.org/10.1002/jgra.50506.

Lui, A. T. Y., Hori, T., Ueno, G. and Mukai, T. (2005). Plasma transport from multicomponent approach, *Geophysical Research Letters*, **32**, 1, DOI: http://dx.doi.org/10.1029/2004GL021891.

Nakabayashi, A. and Ueno, G. (2017). An extension of the ensemble Kalman Filter for estimating the observation errorcovariance matrix based on the variational Bayes's Method, *Monthly Weather Review*, **145**, 199–213, DOI: http://dx.doi.org/10.1175/MWR-D-16-0139.1.

Nakai, H. and Ueno, G. (2011). Plasma structures of Kelvin-Helmholtz billows at the dusk-side flank of the magnetotail, *Journal of Geophysical Research*, **116**, DOI: http://dx.doi.org/10.1029/2010JA016286.

Nishino, M. N., Fujimoto, M., Terasawa, T., Ueno, G., Maezawa, K., Mukai, T. and Saito, Y. (2007a). Geotail observations of temperature anisotropy of the two-component protons in the dusk plasma sheet, *Annales Geophysicae*, **25**, 769–777.

Nishino, M. N., Fujimoto, M., Ueno, G., Maezawa, K., Mukai, T. and Saito, Y. (2007b). Geotail observations of two-component protons in the midnight plasma sheet, *Annales Geophysicae*, **25**, 2229–2245.

Nishino, M. N., Fujimoto, M., Ueno, G., Mukai, T. and Saito, Y. (2007c). Origin of temperature anisotropies in the cold plasma sheet: Geotail observations around the Kelvin-Helmholtz vortices, *Annales Geophysicae*, **25**, 2069–2086.

Nishino, M. N., Fujimoto, M., Terasawa, T., Ueno, G., Mukai, T. and Saito, Y. (2007d). Temperature anisotropies of electrons and two-component protons in the dusk plasma sheet, *Annales Geophysicae*, **25**, 1417–1432.

Ueno, G., Nakamura, N. and Higuchi, T. (2001). Separation of photoelectrons via multivariate Maxwellian mixture model, *Discovery Science, Proceedings of the 4th International Conference, DS 2001* (eds. K. P. Jantke and A. Shinohara), Lecture Notes in Computer Science, **2226**, 470–475, Springer, Washington, D.C. 11.

Ueno, G. and Nakamura, N. (2014). Iterative algorithm for maximum-likelihood estimation of the observation-error covariance matrix for ensemble-based filters, *Q. J. R. Meteorol. Soc.*, **140**, 295–315, 1, DOI: http://dx.doi.org/10.1002/qj.2134.

Ueno, G. and Nakamura, N. (2016). Bayesian estimation of the observation-error covariance matrix in ensemble-based filters, *Quarterly Journal of the Royal Meteorological Society*, **142**, 2055–2080.

Ueno, G., Nakamura, N., Higuchi, T., Tsuchiya, T., Machida, S., Araki, T., Saito, Y. and Mukai, T. (2001). Application of multivariate Maxwellian mixture model to plasma velocity distribution function, *Journal of Geophysical Research*, **106**, 25655–25672, 1.

Ueno, G., Higuchi, T., Kagimoto, T. and Hirose, N. (2007). Application of the ensemble Kalman filter and smoother to a coupled atmosphere-ocean model, *SOLA*, **3** (1), 5–8, 1.

Ueno, G., Higuchi, T., Kagimoto, T. and Hirose, N. (2010). Maximum likelihood estimation of error covariances in ensemble-based filters and its application to a coupled atmosphere-ocean model, *Q. J. R. Meteorol. Soc.*, **136**, 1316–1343, 7, DOI: http://dx.doi.org/10.1002/qj.654.

中村永友, 上野玄太, 樋口知之, 小西貞則 (2005). 欠損混合分布モデルとその応用, 応用統計学, **34**, 57–73.

地球磁気圏の撮像観測とそのデータ同化

Data Assimilation of Imaging Observation in Earth's magnetosphere

モデリング研究系 中野 慎也（Shin'ya Nakano）

1. はじめに

地球磁気圏は，宇宙空間の中でも地球の持つ磁場の影響が及ぶ範囲を指す．地球磁気圏は，太陽風と呼ばれる太陽からのプラズマ (電荷を持った粒子で構成されるガス) の流れの影響で非対称な形状をしており，太陽側は地上 6 万 km 程度まで，太陽と反対側には地上数百万 km 以上まで広がっている．通常，地球磁気圏の研究では，人工衛星によって直接その場所の環境を観測して得たデータを用いる．しかし，現在運用されている衛星の情報を集めても，広い磁気圏中の数点の情報が得られるに過ぎず，磁気圏の大域的な現象の全体像をつかむのが難しい．

地球磁気圏のプラズマの空間分布を，遠隔から 2 次元的に捉える撮像観測は，衛星による直接観測の欠点を解決する有用な方法である．特に，2000 年から 2005 年に運用されていた人工衛星 IMAGE は，様々な手段による撮像観測を実現した衛星であり，有用なデータが取得されている．しかし撮像観測は，プラズマ密度以外の物理量について情報を得るのが難しいという欠点がある．そこで我々は，データ同化技術を活用することにより，人工衛星 IMAGE による撮像観測データから磁気圏の大域的な現象の全体像を捉える手法の開発を進めてきた．本稿では，これまでに行ってきた撮像観測のデータ同化について紹介する．

2. 高速中性粒子データ同化

地球磁気圏の荷電粒子の中でも，比較的エネルギーの高い 1keV–100keV (eV は荷電粒子のエネルギーを表す単位．電子を 1 ボルトの電圧で加速して得られるエネルギーが 1eV となる) の陽イオン (主に H^+) は，磁気嵐と呼ばれる地上の全球的な地磁気変動を引き起こす他，オーロラ嵐をはじめとする高緯度域の電離圏現象とも密接に関係している．この 1keV–100keV 程度の陽イオンの空間分布に関する情報を得る手段として，高速中性粒子撮像観測がある．高速中性粒子は，高エネルギーの陽イオンが，磁気圏中に漂う地球起源の中性粒子から電子を受け取ることによって生成される中性の粒子である．磁気圏中の陽イオンは，地球磁場によるローレンツ力の影響で自由に動くことができないが，中性に変化すると力を受けずに高速で直線運動する．これを遠隔で観測することにより，陽イオンの空間分布に関する情報が得られる．我々は，IMAGE 衛星で観測された高速中性粒子のデータを磁気圏荷電粒子モデルに同化し，荷電粒子分布の時間発展を推定する手法の開発を進めてきた Nakano et al.（2008）．ここで問題となるのは，荷電粒子の動きを決める電場と磁場のうち，地球磁気圏においては電場について十分な情報がないということである．そこで，電場についてはデータ同化の過程で推定することにより，全体の感電粒子分布の時間発展を推定することを実現した．高エネルギーの荷電粒子は，運動にエネルギー依存性があるため，磁力線方向の運動を平均化した Boltzmann 方程式で扱っている（Fok et al., 2001）．データ同化には，当初，粒子フィルタを再帰的に適用する手法を用いていたが，開発の過程で，融合粒子フィルタ Nakano et al.（2007），アンサンブル変換カルマンフィルタ Bishop et al.（2001）に変更され，現在は次に述べる極端紫外光データ同化と

の統合を進めている．

3. 極端紫外光データ同化

IMAGE 衛星では，30.4nm の波長の極端紫外光による撮像観測も行っていた．太陽から来る紫外光のうち，30.4nm の波長のものはヘリウムイオン (He^+) に散乱されるため，これを遠隔から観測することにより，磁気圏の He^+ の分布について情報が得られる．極端紫外光の撮像観測で得られる情報は，磁気圏の中でもエネルギーが低い 1eV–10eV 程度の荷電粒子の情報である．この低エネルギーの荷電粒子は，地球の高度 20000–30000km 以下の領域に高密度で分布しており，プラズマ圏と呼ばれる．プラズマ圏の低エネルギー He^+ は，水素イオン (陽子; H^+) など，他の低エネルギー荷電粒子と同じ方程式にしたがって動くため，He^+ のデータから，プラズマ圏中のプラズマを構成する荷電粒子全体の動きを推定することができる．我々は，アンサンブル変換カルマンフィルタを用いて，プラズマ圏の物理モデルに，極端紫外光撮像観測データを同化し，プラズマ圏のプラズマ密度分布の時間発展を推定する手法を開発した Nakano et al.（2014）．低エネルギーの荷電粒子の動きも電場と磁場の影響を受けるが，電場について十分な情報がないため，高速中性粒子データ同化と同様にデータ同化の過程で電場の推定を行っている．また，ここで用いているプラズマ圏モデルは，地球磁力線方向の密度構造についてある関数系を仮定した 2 次元モデルであるが，磁力線方向の密度構造に関するパラメータを周辺尤度最大化で推定することにも成功した．

4. おわりに

現在我々は，高速中性粒子データと極端紫外光データの両方を磁気圏統合モデルに同化する新たなデータ同化システム開発を進めている．高速中性粒子の起源となる高エネルギー荷電粒子は，極端紫外光で観測される低エネルギー荷電粒子の分布するプラズマ圏と比較して，磁気圏のやや外側に分布しており，磁気圏内の電場を推定する上で，高速中性粒子データと極端紫外光データは，互いの情報を保管する役割を果たすと考えられる．最初に述べたように，地球磁気圏に関して観測から得られる情報は非常に限られている．我々は，データ同化を活用し，物理法則の知見を活用して直接観測できない物理量を推定しながら，地球磁気圏で起こる様々な現象の全体像を包括的に捉えることを目指している．

参 考 文 献

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001., Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects, *Mon. Wea. Rev.*, *129*, 420–436.

Fok, M.-C., R. A. Wolf, R. W. Spiro, and T. E. Moore, 2001., Comprehensive computational model of Earth's ring current, *J. Geophys. Res.*, *104*, 8417.

Nakano, S., G. Ueno, and T. Higuchi, 2007., Merging particle filter for sequential data assimilation, *Nonlin. Process. Geophys.*, *14*, 395–408.

Nakano, S., G. Ueno, Y. Ebihara, M.-C. Fok, S. Ohtani, P. C. Brandt, D. G. Mitchell, K. Keika, and T. Higuchi, 2008., A method for estimating the ring current structure and the electric potential distribution using ENA data assimilation, *J. Geophys. Res.*, *113*, A05208, doi:10.1029/2006JA011853.

Nakano, S., M.-C. Fok, P. C. Brandt, and T. Higuchi, 2014., Estimation of temporal evolution of the helium plasmasphere based on a sequence of IMAGE/EUV images, *J. Geophys. Res.*, *119*, 3708–3723, doi:10.1002/2013JA019734.

前震識別とその予測可能性

Discrimination of Foreshocks and Its Predictability

モデリング研究系 野村 俊一（Shunichi Nomura）

要　　旨

一連した地震の群れの中の最大地震すなわち本震には，しばしば前震と呼ばれる先行活動が見られ，本震を事前予測する重要な手がかりとなる．本稿では，地震カタログから構成したクラスター（地震群）の時空間的配置および地震規模推移を特徴量として，地震群が前震である確率および一定期間に一定規模の本震が起こる確率を評価する研究について解説する．

キーワード：地震予測，前震，クラスタリング，ロジスティック回帰，スプライン関数

1. はじめに

大地震の後には多数の余震が発生するが，ときに大地震に先駆けて地震が集中して起こることがあり，これを前震活動という．もしも大地震が実際に起こるより前にその前震活動を特定することができれば，大地震を短期的に予測することが可能となる．前震活動とその他の地震活動との完全な事前判別は困難であるが，Ogata et al.(1995) によると前震活動とその他の地震活動では，地震間の時空間的距離やマグニチュード差の傾向に違いがある．以降では，野村・尾形 (2018) を例に前震識別の方法論を 3 つの段階に分けて解説する．

2. 地震群（前震群候補）の構成

まず，地震活動を地震の群れとして捉えるために，地震活動の点群に対してクラスタリング手法を用いて地震群を構成する．ここでは Ogata et al.(1995) に倣い Single-link 法を採用し，日本のマグニチュード 4 以上の地震について，震央間距離 Δd(km) と時間距離 Δt(日) が $\sqrt{(\Delta d)^2 + (c\Delta t)^2} \leq 33.33$ を満たす地震同士を連結していくことでクラスター（地震群）を構成した．ただし，$c = 1.11$(km/日) とおいた．

3. 前震群候補からの特徴量抽出と前震群の定義

次に，構成した地震群内の各地震に対して，その時点までに発生した地震のみからなる部分群を作り，前震群候補とする．さらに，実際の前震群を定義付けた上で，前震群候補である部分群から，前震群の識別に有効な特徴量を抽出する．ここでは，図 1 のように部分群の最後の地震発生時点から 30 日以内に，部分群内の最大マグニチュードを超える地震が起きたときに前震群であると定義し，その事前識別のため各部分群から次の特徴量を抽出した．

- 群内の地震数：$N \geq 2$
- 群内の一番目・二番目に大きいマグニチュード：M_1, M_2
- 群内の期間長：T（日）

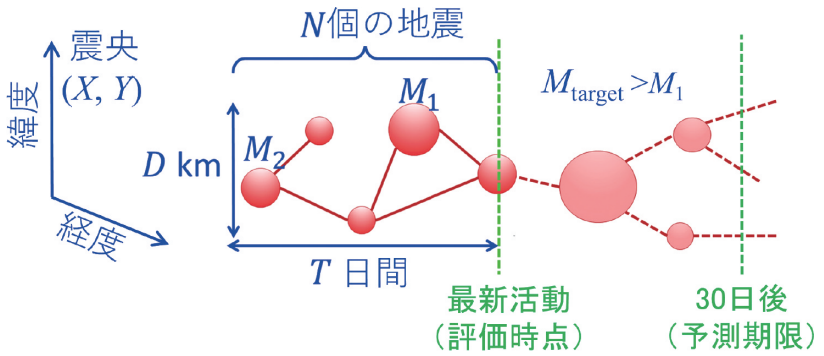


図 1．前震群候補から抽出する特徴量と前震群の定義

- 群内の平均震央間距離：D（km）
- 群内の中心経度・緯度：(X, Y)（度）

4. 前震群候補の前震確率評価

最後に，前震群候補である部分群について抽出した特徴量に基づき，前震群であるかどうかを確率で評価することにより識別する．ここでは，ロジスティック回帰を用いて前震確率 p を

$$\mathrm{logit}p = \log \frac{p}{1-p} = g(X,Y) + f_1(N, M_1, M_2) + f_2(N, M_1, T) + f_3(N, M_1, D) + \varepsilon_i$$

のように評価した．右辺第 1 項 $g(X, Y)$ は地域による前震確率の変化を表しており，薄板スプライン関数を用いて推定する．$f_1(N, M_1, M_2), f_2(N, M_1, T), f_3(N, M_1, D)$ は相関の強い特徴量の組合せごとに，特徴量による前震確率への影響を表しており，それぞれ 3 次 B-スプライン関数を用いて推定する．最後の項は，上記には含まれない前震確率に影響する特徴量の効果を，地震群 i ごとの変量効果として取り入れたものである．

1926 年から 1999 年までの気象庁カタログから，前震確率の評価式を学習し，2000 年から 2017 年 10 月までの同カタログに適用して前震確率を評価した結果，最大で 80% 前後の前震確率となった地震群が 2 つあり，そのいずれもが実際にマグニチュード 6 以上の本震を起こした．

5. まとめ

本稿では，進行中の地震群が大地震前の前震活動であるかについて，地震群の特徴量に基づいて確率的に識別する方法を解説した．前震識別に基づく地震予測手法は，短期間での大規模地震の予測について比較的高い発生確率を与えうる数少ない手法であり，まだ改善の余地は大きいものの，将来的には大地震のリアルタイム予測へと実装されることを期待する．

参 考 文 献

野村俊一, 尾形良彦 (2018). マグニチュードと時空間情報に基づく前震確率評価モデルの識別性能, 「地震予知連絡会会報」,**99**, 1-6

Ogata, Y., Utsu, T. and Katsura, K. (1995). Statistical features of foreshocks in comparison with other earthquake clusters, *Geophysical Journal International*, **121**, 233-254.

「日本人の国民性調査」から「文化多様体解析」へ

From “Japanese National Character Survey” to “Cultural Manifold Analysis (CULMAN)”

データ科学研究系 吉野 諒三（Ryozo Yoshino）*

「日本人の国民性」調査は昭和 28 年より開始され, 戦後民主主義の発展のための「科学的世論調査」の基盤となってきた。これは, 世界で希有の長期継続統計的無作為標本抽出調査であり、内閣府「社会意識に関する調査」, NHK「日本人の意識調査」とともに日本の三大意識調査と呼ばれている。その後、これを追従し、米国 GSS, ドイツ ALLBUS,EU の Eurobarometer 等々の一般社会調査が遂行されるようになり, 近年ではアジア諸国でも同様の調査が開始されている。

1970 年頃より国民性調査は国際比較に拡大され、筆者は平成元年に入所以来、「日本人の国民性調査」と「意識の国際比較」に携わってきた。「国際比較は意識調査の宝庫である」とは, 戦後の科学的世論調査を含む社会調査の発展に大きな足跡を残した林知己夫が到達した認識であった。様々な国を比較する時は, 翻訳の問題, 各国固有の調査方法の違いなど, 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」を追求しているのであり, また「データの科学」を計量的文明論のために試行錯誤しているのである。

林知己夫と鈴木達三らにより構築された「連鎖的比較方法論 CLA を, 筆者らは「文化多様体解析 CULMAN」へと発展させてきた。特に, この 20 年ほどは一連の大型科研費を獲得して,CULMAN の具現化のため, 米国西海岸日系人調査（1999-2001）, ハワイ日系人調査（2000-2002）, 東アジア価値観調査（2002-2005）, 環太平洋価値観調査（2004-2009）, アジア太平洋価値観国際比較調査（2010-2015）等々を展開してきた。

これらの調査研究が進む中で、共同研究者のハワイ大学黒田安昌名誉教授（現）の論文はハワイ教育委員会に認められ、初等中等学校の教科書に掲載されるようになった。また、ワシントン大学カシマ・テツデン名誉教授（現）は、日系人社会への貢献で 2018 年に日本政府から旭日小勲章が授与された。少数者差別問題が世界に渦巻く中で、多様な民族・人種の共存共栄モデルの提示に尽力してきた貢献が認証されたのである。

近年の米大統領選挙やイギリスの EU 離脱国民投票では、事前の世論調査は大きな失敗を見せた。欧米では調査方法の質が著しく下がり、信頼性を失っている。また, 急速な IT 技術の進展に対し, 法律の規制が追いつかぬ中で, Cambrdige Analytica 社や Facebook 社の個人情報の扱い等、企業が法律や倫理を踏みにじるような事件もある.

世論と選挙に関する Galbraith の次の文章は、今日でも我々に示唆するものは重い。「・・・

* 同志社大学 文化情報学部：〒610-0321 京都府京田辺市多々羅都谷 1-3

経済的に豊かな階層の人々からなる勢力には資金と影響力がある。そして, 彼らは投票する。一方, 貧しい階層に対する支援者からなる勢力は, 人数的には多いが, 貧しい人々の多くは残念ながら投票しない。民主主義は存在するが, それは恵まれた人々のための民主主義と言えなくもないのである。・・・」(John Kenneth Galbraith,1996「よい世の中」佐々木直彦・純子訳,p.26, 日本能率協会マネジメントセンター.])

民衆の感情を掴みながらも,「感情の世論」に流されず, 建設的な目標を支える「理性の世論」を掴む調査のために「日本人の国民性調査と意識の国際比較」が貢献し続けていくことを強く望む。

(注) 本研究所の主要な調査は、多数の書籍、研究リポート、WEB サイト上で公刊されている。「国際比較データの解析」(吉野・林・山岡,2010, 朝倉書店) 参照.

社会調査方法論の実践的研究

Practical Studies on Social Survey Methodology

データ科学研究系 前田 忠彦（Tadahiko Maeda）

1. はじめに

社会調査法は社会現象に関するデータ取得のための中心的な手段の一つである。筆者はそのような社会調査を企画・実践するとともに調査データを素材とした調査方法論研究を主なテーマとしている。実践と研究が一体を成すスタイルで研究を続けており，現実問題に合わせた調査設計法自体が研究の重要な一面である。社会調査には様々なプロセスが含まれるが，その全ての段階に調査方法論上の研究課題が潜んでおり，次節に紹介するような具体的な調査を研究材料として研究を進めてきた。実施した社会調査には多くの共同研究も含まれる。

2. 調査プロジェクトの例

2.1 日本人の国民性調査および関連調査

中心的な調査研究の一つが，統計数理研究所が 1953 年以来 5 年に一度実施している「日本人の国民性調査」である (最新調査は 2018 年実施の第 14 次全国調査) 。これは，同じ調査手法（訪問面接法），同じ調査項目で横断調査を繰り返すことを基本とした継続社会調査で，5 年に一度の本調査実施の他に中間年には様々な関連研究を行っている。多くの先輩方の努力により続けられてきたプロジェクトであるから，その資産を活用しながら，近年は自身のアイデアを調査設計等に生かすようにしている。

2.2 共同調査研究

2010 年には「2010 年格差と社会意識に関する全国調査」(略称 SSP-I2010 調査) を実施した。吉川徹客員教授と共に多数の研究者を含む共同研究体制を組み，調査の企画・実施から解析まで，大阪大学人間科学研究科との緊密な連携の下で多数の成果を生んだ調査研究である。

2011 年には国立国語研究所と「第 4 回鶴岡市における言語調査」を実施した。1950 年に第 1 回調査が両研究所の協力で実施されて以来，1972 年第 2 回，1991 年第 3 回と約 20 年間隔で行われてきた。山形県鶴岡市における共通語化の進行を，各回のクロスセクション調査と，パネル調査を組み合わせたデザインで研究する，恐らく世界最長の言語に関する継続調査である。

3. 具体的な研究テーマ例

2 節で紹介したプロジェクトでの中心的調査は，調査員が調査対象者を訪ね，面接で回答を取得する「個別訪問面接法」を採用している。最も「正統的」でよく用いられてきた調査手法であるが，そのような調査手法についても未だ様々な研究テーマが残されている。

3.1 調査員効果に関する研究

このような調査員が介在する調査では，調査員の持つ何らかの特徴が調査結果に影響を与える。最近はこの「調査員効果」について二つの面からの研究を行っている。第一は調査員特性

が，回収・非回収に与える影響という面であり，例えば松岡亮二氏（早稲田大学）との共同研究では，「日本人の国民性第 13 次全国調査」について，調査地点の特徴や調査員特性を含めたマルチレベル分析により，対象者・地点・調査員の特性と回収状況の関連を総合的に検討した．第二は調査員属性が回答内容に与える効果についてのもので，特定テーマの調査項目に対して，調査員の属性が与える効果の可能性等を検討している（金沢大学小林大祐氏との共同研究）。

3.2 調査パラデータの解析－訪問記録を例として

調査パラデータとは，質問への回答という最も中心的な調査データに加えて，調査の実施プロセスに付随して得られる様々な情報を指す。最近では，面接調査における調査員の活動（訪問記録）を分析した。この分析を通じて，面接調査での回収・不能がどのような経緯を経て決まるのか，その経緯は対象者の住む住居特性と関係するか，等を分析することによって，調査員の活動に資する知見を得ることを目的とした研究である。

3.3 調査モード間の比較研究

測定プロセスのうち対象者から回答を取得する手段を「調査モード」と呼ぶ。調査員が回答を面接で聴き取る「他記式」に対し，対象者が自ら調査票を読み回答を記入する「自記式」のモードもあり，回答結果にこの調査モードが大きく影響することがある。尾崎幸謙氏（筑波大学）との共同研究では，留置法 (自記式）と面接法（他記式）で同時に行われた実験的調査を比較分析した。留置法では，面接法に比べて，暮らし向きの満足度が低いとか，選挙で投票に行く頻度が低いといった違いが明瞭に見られ，この違いは傾向スコアを用いて回答者の属性（共変量）の分布が二つのモードで異なることを調整した上でも消えない。このことから，「社会的望ましさ」への対象者の敏感さが両モード間で異なる反応を引き出す可能性等を推察することができる。モード間の差が生ずるメカニズムも一様ではなく，研究課題が残されている。

3.4 調査不能バイアスの調整

このテーマの背景・動機となっているのは，近年の社会調査特に面接調査全般における回収率の低下傾向である。例えば 2013 年実施の「日本人の国民性第 13 次全国調査」では回収率が 50% と，半数近くの人が調査に協力しない状態での結果が得られている。この協力率の低さで母集団の推定を正確に行えているのかという点（調査不能バイアスの問題）が懸念される事態と言える。伏木忠義氏 (新潟大学) との共同研究を進め，例えば 2 節で紹介した SSP-I2010 調査を題材としてこの問題を検討した成果を発表した。

3.5 標本設計・サンプリングの精度とその一貫性に関する検討

サンプリングの設計と，その設計下での標準誤差の大きさ等の調査精度の評価は，社会調査設計上の重要な論点の一つである。各調査プロジェクトでの標本設計を担当し続ける中で，継続社会調査でのそうした精度の一貫性の有無や，市区町村合併のような社会制度の変化が調査設計に与える影響も無視できない論点であることに気づき，こうした点の検討も続けている。

4. 社会調査法研究のこれから

「はじめに」に述べたように，社会調査のプロセスの全てにわたって，調査方法論上の研究課題が含まれており，最も伝統的な調査手法である面接調査法に限っても，様々な研究テーマが残されているのが現状である。他方で，回収率が低下し続ける訪問面接法による調査研究には限界が見え始めていることも感じている。ランダムサンプリングと訪問面接法の組み合せという伝統的な手段に代わる，現代社会にふさわしい調査方法の研究も必須である。

組織規範継承を可能にする目的指向型成果評価と相互依存性の計量分析

Does Goal-Oriented Managerial Behaviour Applying Performance Management Improve Interpersonal Facilitation Among Public Officials in Japan? A Multilevel SEM Analysis Focusing on Division Level Interdependence

データ科学研究系 朴 堯星（Yoo Sung Park）

1. 研究の背景と目的

現在，多くの自治体が抱える問題の一つは，行政職員の協力行動をいかに引き出すかである．バブル崩壊以降，日本の行政組織は慢性的な赤字財政を抱えている．そのため，2000 年代ごろから成果評価と分権化を軸とする組織運営が取り組まれている．しかし，多くの自治体では，成果評価の導入によって業務遂行の個人化が進み，人材育成，チェック体制が弱まり（村林 2004)， これまで緩やかに維持されてきた組織としての機能が揺らいでいる状況である．元来，成果評価の導入は，官僚制型組織運営の弊害を克服することを狙いとする．すなわち，組織機能の向上のために導入されたものであるにもかかわらず，その実態は，当初の期待とはかけ離れているアイロニックな状況である．総務省 (2017) によれば，都道府県・市区町村において 977 団体 (54.4 %) がすでに行政評価を導入済みである．導入団体数は毎年，増加していることが知られており，従来の組織運営体制への後戻りはできないだろう．むしろ成果評価をめぐる弊害を払拭するための，組織規範の継承を可能とする職場での協力体制を強化させる組織心理学的要因を計量的に探ることが重要と考える．そこで本研究では，組織の長期的な発展に向けて組織が機能するために必要とされる職員の行動として Borman and Motowidlo(1993) が提唱した task performance（TP) と Contextual performance（CP) に着目して，職場での協力体制を強化させる制度設計の条件としての「相互依存性」がもたらす文脈効果を明らかにする．具体的には，個人-市町 -県の多層同時比較調査を遂行し，マルチレベル構造方程式モデリングを用い，目的指向型成果評価のもと，相互依存性が行政職員の協力行動に及ぼす影響を明らかにする．

2. 研究の方法

本研究は以下の手順で進めている．まず，自治体行政に適した課業相互依存尺度を開発するため，自治体行政の業務内容と法律で定まっていない自治体行政の業務内容を整理し，インタビュー調査を行い，自治体行政組織における課業相互依存の実態を把握したうえで，つぎに，「多層的相互依存尺度」の開発を行った．さらに，その後，成果評価と分権化を軸とした行政改革にいち早く取り組んできた三重県庁職員 279 名を対象とした．調査対象者は，多段集落抽出法に基づき，三重県本庁 26 課中，業務内容に応じてより明確に相互依存性が表れると考えられる事業執行部門を対象とし，業務特性に偏りが出ないように 17 課を選出し，これらの課に所属する常勤職員全員 (課によっては，7 名－ 24 名程度が所属) に質問紙調査を実施した．ま

た，三重県庁の職員との比較を試みるため，三重県所在の 3 市（津市，松阪市，尾鷲市）の職員を対象とした同様の調査を実施している．最後に，調査から得られたデータをもとに，課レベル変数が個人レベル変数に及ぼす影響を確かめるため，マルチレベル分析を行った．調査・分析を終え，相互依存性が，行政職員の協力行動が促進するジョブデザインの条件になりうることについての総括的な考察を行った．

3. 研究成果

まず，新しく開発した「相互依存性」尺度の信頼性および妥当性を検討したうえで，マルチレベル SEM を行い，課の違いがもたらす影響を確かめた．マルチレベル SEM は，within レベル (個人レベル) における random slope や random intercept が，1 つの値ではなく between レベル (課レベル) で値がバラつく因子として捉える．これが，between level の変数によって推定されることで，レベル内の関係 (within) と各レベル間の関係 (between) を同時に検討することができる (Snijders and Bosker，2012)．本研究では課の主効果の検討を目的としているため，切片のみにランダム効果（random intercept）を仮定し，切片の課レベル残差の程度から推定されたモデル間の比較を行い，切片の課レベルのばらつきを，課レベル変数である業務相互依存性と目標相互依存性で説明するモデルを検討した．その結果，課レベルで相互依存性がうまく取れているほど，個人の対人的促進が高まっていることを確認している (表 1)．具体的には，課レベルの業務相互依存性には，直接，個人の対人的促進を高める効果があるが，同時に課レベルの目標

相互依存性を媒介して個人の対人的促進を高める効果もあることが明らかになった．このことは，個人の所属する課の働き方を相互依存的なものに変えることで，職場での協力行動を促すのが可能になることを意味する．

表 1. マルチレベル SEM を用いた多層レベル解析結果

	モデル1	モデル2	モデル3	モデル4
固定効果				
個人レベル				
対人的促進 ← 職務充実度	0.498 ***	0.490 ***	0.486 ***	0.490 ***
職務充実度 ← 目的指向型経営管理行動	0.540 ***	0.540 ***	0.540 ***	0.540 ***
課レベル				
目標相互依存性 ← 業務相互依存性	0.544 ***	0.544 ***	0.544 ***	0.544 ***
対人的促進 ← 業務相互依存性		0.715 ***		0.285 †
対人的促進 ← 目標相互依存性			0.973 ***	0.791 ***
変量効果				
切片の課レベル残差	0.026	0.021	0.002	0.002
AIC	1978.80	1962.80	1962.20	1962.80
CFI	0.833	0.940	0.910	0.949

個人レベル：n=414, 課レベル：N=35, *** p < .001, ** p < .01, * p<.05, † p<.10.

これまで多くの自治体組織では，成果評価の導入を期に，職場での知の継承が疎かになってきていることに悩まされていた．これに対し，本研究では，職場での協力体制を構築するための，個人レベルと課レベルでの組織心理学的メカニズムを解明したものである．そもそも仕事を個人で遂行するのではなく，課の全員で相互依存的に行うことによって，課全体の目標に対する認識の共有度合いが高まる．そのような過程で，自然に組織運営にかかわる知の継承が緩やかに促されることになるのではないか．言い換えれば，職場での組織規範の継承には，職員個々人が属する部署での仕事のやり方も相互依存的なものへと変えることが重要であると考えられる．今後は，部署内の職員同士のパーソナルネットワークを組み込んだモデリングへの拡張を検討していきたい．

参 考 文 献

朴堯星・坂野達郎（2015），“自治体職員の対人的促進に関するマルチレベル分析：課レベルの相互依存性に着目して”，『計画行政』, 38(3), 55-64.

YOOSUNG PARK “Performance Evaluation System and Interpersonal Facilitation: An Empirical Evidence of Public Officials in Mie Prefectural Governments of Japan”, The spring 2015 conference of The Korea Association for Survey Research, Seoul.2015.6.15.

立川市町丁目別住民意識調査分析追記 ―小地域推定モデル活用に向けて―

Addemdum to “Hirose et al. (2018)”: Applicability of Small Area Explicit Model to Japanese Survey Data

データ科学研究系　廣瀬　雅代（Masayo Y. Hirose）*

要　　旨

小地域推定モデルに基づく統計的推測法は, わが国の調査データ分析にも大いに貢献し得る. 本論文では, そのような手法の有用性を示す資料の説明力を高めるべく, 廣瀬ら (2018) の分析結果に補足する形で, 手法の有用性の解釈を容易にする資料を提示する. このような資料を通して, わが国の小区分別調査データ分析の可能性がさらに広がることを期待する.

キーワード：EBPM, 小区分別調査データ分析, 小地域推定

1. わが国の小地域推定モデル活用に向けた資料の追記

Evidence Based Policy Making (EBPM) の重要性が高まっている今日, 細かな政策やサービスを計画する際, 区分ごとの実態を効率よく把握することは重要な課題のひとつになり得る. しかし, 区分数が多くなるにしたがって, 慣習的に用いられている区分ごとの推定法は, 信頼性を大幅に低下させる懸念があり時には計画遂行にも大きく支障を与えかねない. このような問題に対して, 小地域推定モデルに基づく統計的推測法の需要が, 理論面及び応用面において急速に高まっている (Rao and Molina, 2015; 久保川, 2016). わが国の調査データも例外ではない. 立川市住民意識調査データ (朴・土屋, 2017) でも, 町丁目別に区切ることにより同様の問題が懸念される. 廣瀬ら (2018) はその調査データに小地域推定モデルを適用するべく, 国勢調査と共通の項目を用いて, 国勢調査小地域集計結果と各推定値からの乖離を絶対相対誤差によって測ることで, 小地域推定モデルに基づく推測法 (以後 MBA 法と呼ぶことにする) の有用性を評価した. しかし, そのような統計的推測法の活用推進の為には, より解釈が容易となる指標を用いた方が適用手法の有用性を説きやすい. 例えば, 政策設計やサービス計画を立てる側に MBA 法の有用性を説く場面では, 単純であるが解釈が容易になる絶対誤差の指標に基づく資料が大いに役立つと考えられる.

そこで, 本論文では, 廣瀬ら (2018) の資料に追記する形で, 国勢調査の集計結果と各推定法 (慣習的な推定法と MBA 法) に基づく推定値の乖離を絶対誤差 ($AER = |\hat{P}_i - p_i| \times 100$) で測り, その結果を図 1 に示す. ここで, 廣瀬ら (2018) の定義と同様に, $\hat{P}_i$ は, 立川市住民意識調査データ (朴・土屋, 2017) から慣習的な推定法と MBA 法によって推定された第 i 町丁目の男性割合に対する各推定値を表しており, 平成 27 年度国勢調査小地域集計結果 (http://www.e-stat.go.jp)

* 九州大学マス・フォア・インダストリ研究所：〒819-0395　福岡市西区元岡 7447

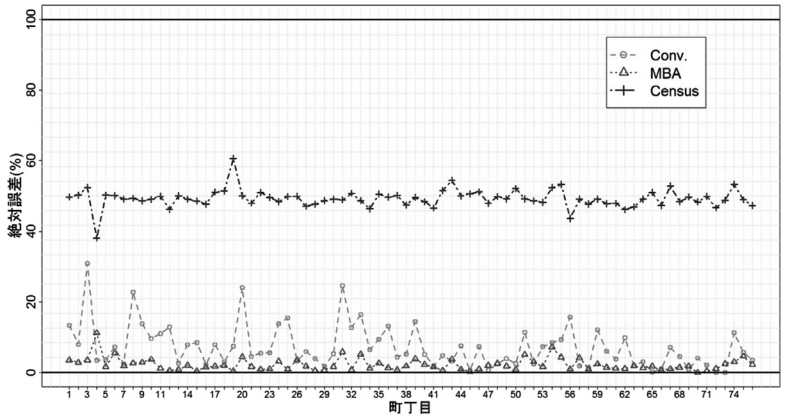


図 1. 立川市 76 町丁目における国勢調査の男性割合からの絶対誤差の比較 (横軸は有効回答サンプルサイズの昇順に町丁目を並べている): 慣習的な推定値の AER(Conv.), MBA 法による推定値の AER(MBA), 国勢調査の各町丁目の男性割合 p_i(Census)

から計算された第 i 町丁目における男性割合 p_i を真としてみなすことにする. この図によって, 指標として絶対相対誤差を用いた場合よりも, 慣習的な方法と比べて MBA 法がどの程度誤差を抑えているのか解釈しやすくなったように見える. 特に, 有効回答サンプルサイズが 3 番目に小さい町丁目では, 慣習的な推定法によって 31 ポイントもの絶対誤差が引き起こしているのに対し, MBA 法を用いると 3.6 ポイントまで誤差が小さくなっている様子も確認することができる. また, 絶対誤差の平均は, 慣習的な推定法が 7.3 ポイント, MBA 法は 2.3 ポイントであった. すなわち, 国勢調査小地域集計結果 p_i を真として考えると, この結果は, 絶対誤差の平均が小地域推定モデルの活用により慣習的な推定法の 1/3 以下に抑えられたことを意味している.

小地域推定モデルに基づく推測法は, 慣習的な推定法より取り扱いにくい. しかし, それでもこの資料の補足によって, わが国の小区分別調査データ分析での小地域推定モデル活用の機会が広がることを期待したい.

謝　辞

横浜市立大学の土屋隆裕氏と統計数理研究所の朴堯星氏には, 図 1 を作成する為に, 廣瀬ら (2018) の論文で用いた立川市住民意識調査データを引き続き使用する許可をいただいた. また, 千葉大学の佐野晋平氏, 川久保友超氏, 東京大学の菅澤翔之助氏には, 今回の資料補足のきっかけとなる, 図に関する重要な助言をいただいた. この場をお借りして御礼申し上げたい.

参　考　文　献

久保川達也 (2016). 推定における縮小法の展開－高次元解析と小地域推定－日本統計学会誌, 46, 43-67

朴堯星・土屋隆裕 (2017). 多摩地域 住民意識調査—立川市郵送調査 (2016)—, 統計数理研究所調査研究リポート No.120

廣瀬雅代・朴堯星・土屋隆裕 (2018). 小地域集計を活用したモデルに基づくアプローチによる防災に関する立川市町丁目別住民意識調査分析, 日本統計学会誌, 48, 49-70.

Rao, J.N.K. and Molina, I. (2015). Small Area Estimation, 2nd Ed., Wiley, New York

制度的制約下におけるデータベース構造化、モデリング、モデル評価

Database structuring, modeling, and model evaluation under institutional constraints

データ科学研究系 山下智志（Satoshi Yamashita）

1. はじめに

私個人の今世紀に入って以降の研究活動の特徴は、社会ニーズを吸上げ、データの取得、理論構築、モデリング、モデル評価、社会実装まで一気通貫的に行い、各ステップの整合性を重視することにあった。特に、秘匿性データや統合データなどのデータベースの構築を長期間にわたって実行してきた。また、社会実装を想定しているため、法律、条約、規制、会計ルールなどのコンプライアンス面を考慮しながらモデリングを行うところに特徴がある。以下、研究活動実績を紹介したい。

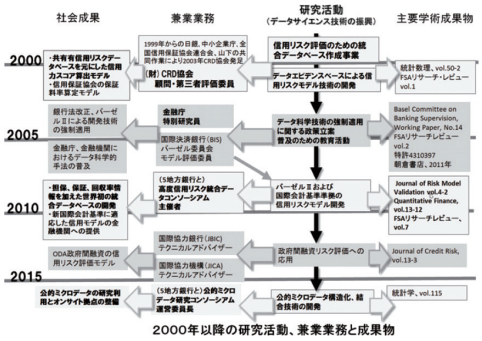
2. 統計モデル、データベース、政策の整合性と効率性

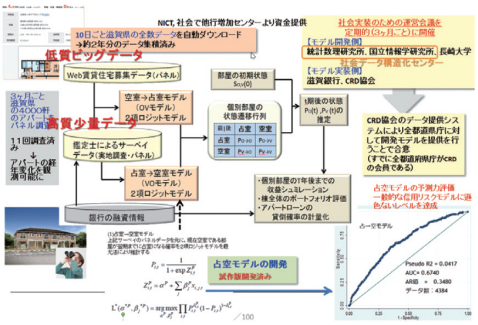
公開された研究業績は、いわゆる数理ファイナンスの分野や企業データ分析の方法論に関する著作が多い [1]。これは数理ファイナンス系方法論の研究が学術論文執筆と親和性が高く、書きやすいもしくは審査に通りやすいと言うことが原因である。実態としては方法論構築に費やしたエフォートは、データベースの作成、高度化やニーズの引き上げや政策化に費やしたエフォートよりも少ないのが実感である。

特に、金融庁特別研究員を 10 年近く併任していたため、全世界の銀行の共通ルールであるバーゼル II（新 BIS 規制）[2] や銀行法改正に関する統計学的根拠の整理を長年続けていた。また、一般的な信用リスクモデルを改良したカントリーリスク計量化モデルは、国際協力機構（JICA）や国際協力銀行（JBIC）における政府間金融（いわゆる ODA）の意思決定に利用されている。

データベース構築に関わる活動としては、日本最大の信用リスクデータベースである CRD 協会を日銀、中小企業庁、全国信用保証協会連合会との連携により 2003 年に設立し、現在も顧問を行っている。そこでは日本の企業の過半数である１６０万社の財務データを蓄積しているだけでなく、そのまた地方銀行とのデータ共有コンソーシアムを立ち上げ、世界で初めて秘匿性の高い与信データの統合を統計数理研究所内で実現した。これにより銀行間で与信構造に差異があることが確認された。現在は総務省との共同で、公的ミクロデータの統合化に取り組んでいる。以下、現在運営管理中のデータベースである。

- 公的ミクロデータ（オンサイト拠点）：　政府統計のミクロデータ（法人統計データ、労働統計データ）を利用するため、統数研内に政府データの窓口である (財) 統計センターと直接つながるオンサイト拠点
- 国際ミクロデータ：　アジア 6 カ国の政府ミクロデータ（主として労働統計）





- 高度信用リスク統合データベース： 地方銀行５行（滋賀、群馬、八十二、伊予、北陸）の貸出先法人の全数データの統合データベース。与信情報と回収情報があるのは世界唯一。お互いに秘匿情報のため、分析方法に工夫が必要。
- CRD 協会法人データベース：1996 年より収集を開始した中小企業財務データベース (法人としては 2003 年から)。毎年約 160 万社、与信データベースとして日本最大である。
- 民間の購入データ（帝国データバンク、日経 Needs Company）2017 年から開始した政府企業統計ミクロデータと民間信用データとの結合実験として帝国データバンクの企業情報（約３万件）がある。その他、民間の企業情報データとして日経 Needs と契約を続けている。

3. 賃貸住宅データを用いた、質の異なるデータベースの接続実験

様々なデータベースが増えるとともに、複数のデータベースを用いた統計モデルの構築方法が問題となる。特に質の異なるデータを利用するための方法論が必要とされている。具体的な研究としては、Web データを定時観測（楽天不動産、SUUMO から 10 日ごとにデータをダウンロード）することによって大規模データベースを作成し、賃貸住宅の入居化要因を分析する。一方、Web データの精度を補完するため、不動産鑑定士による現地パネル調査（賃貸住宅定点観測データ：約 4400 戸）を行った。それぞれのデータを統合することにより、より正確な賃貸住宅の収益予測を行う。２つの質の異なるデータベースを接続することにより、これまでリスク計量化モデルが考案されてこなかったアパートローン与信リスクについて、より正確な賃貸住宅の収益予測を行うことを目指している。現在、モデル精度は個人ローンのデフォルト予測と同等であり、実用化への具体的準備が進んでいる。

参 考 文 献

[1] Satoshi Yamashita, Toshinao Yoshiba,“ A Collateralized Loan’s Loss Under a Quadratic Gaussian Default Intensity Process”, Quantitative Finance, vol.13-12, p.1935-1946, 2013 年 6 月

[2]Satoshi Yamashita, 他 “Studies on the Validation of Internal Rating Systems”, Basel Committee on Banking Supervision, Working Paper, No.14, 2005 年 5 月

極値分布と指数逆ガウス型分布に関するある一般化について

Generalization for the Extreme Value and Exponential Inverse Gaussian Distributions

データ科学研究系 金藤 浩司 （Koji Kanefuji）

要　　旨

本報告では実数上で定義される二つの確率分布の一般化について紹介する。この一般化によって、一般化極値分布と一般化指数逆ガウス型分布を導出している。本報告とは異なり分布関数によって定義された一般化極値分布もある。そこでは母数の値により、三つのタイプの極値分布（タイプ I:ガンベル型、タイプ II:フレシェ型、タイプ III : ワイブル型）を表現している。

キーワード：母変動係数; 寿命分布；確率素分

1. はじめに

本稿では、タイプ I の極値分布の一般化を検討する。この確率分布の母歪度は零ではなく、非対称の性質を有している。同様の性質を有する実数上の分布として、指数逆ガウス型分布 (Kanefuji and Iwase; 1996) がある。この確率分布は、タイプ I の極値分布が利用される場面においてデータ解析上の別の候補となる一つの確率分布である。指数逆ガウス型分布に関しても同様な一般化を行う。極値分布は指数変換によるガンマ分布との関連性はよく知られている。同様の関連性は、逆ガウス型分布と指数逆ガウス型分布の間にも見られる。この関連性がタイプ I の極値分布と指数逆ガウス型分布の一般化の元となるアイデアである。さらに、3 母数ガンマ分布や 3 母数逆ガウス型分布は、本稿での一般化手法と直接的な関連性を有している。

実際のデータ解析において、タイプ I の極値分布が用いられる場合において、データから計算される標本歪度がその母歪度から大きく外れている場合が多々存在する。これらの変動に対応するため、本報告で定義するような一般化分布が有用となる。

また、Jenkinson(1955) により、一般化極値分布として、三つのタイプの極値分布を包含する分布が提案されている。

2. 二つの確率分布の一般化

タイプ I の極値分布 $EV(\mu, \sigma^2)$ は、次のように定義される。

$$\exp\left(\frac{X-\mu}{\sigma}\right) \sim Ga(1, 1^2),$$

ここで、$0 < \sigma < \infty$, $-\infty < \mu < \infty$, $-\infty < X < \infty$ であり、記号 $Ga(m, c^2)$ は、母平均 m、母変動係数 c であるガンマ分布を表している。

定義 1. タイプ I の一般化極値分布 $X \sim GEV(\mu, \sigma^2, \lambda)$ を以下で定義する。

$$\exp\left(\lambda \frac{X-\mu}{\sigma}\right) \sim Ga(1, \lambda^2),$$

ここで、$0 < \sigma < \infty,\ -\infty < \mu < \infty,\ |\lambda| < \infty,\ -\infty < X < \infty$ であり、λ は無次元量である。$GEV(\mu, \sigma^2, \lambda)$ に従う確率変数の確率素分 $f(x)dx$ は以下である。

$$f(x)dx = \frac{\left(\frac{1}{\lambda^2}\right)^{\frac{1}{\lambda^2}-\frac{1}{2}}}{\Gamma\left(\frac{1}{\lambda^2}\right)} \exp\left(\frac{1}{\lambda}\cdot\frac{x-\mu}{\sigma} - \frac{1}{2\lambda^2}\left\{2\exp\left(\lambda\frac{x-\mu}{\sigma}\right)\right\}\right)\frac{dx}{\sigma},$$

ここで、$-\infty < x < \infty, -\infty < \mu < \infty, 0 < \sigma < \infty, |\lambda| < \infty$ であり、$\Gamma(x)$ はガンマ関数である。

指数逆ガウス型分布 $EIG(\mu, \sigma^2)$ は以下の様に定義される。

$$\exp\left(\frac{X-\mu}{\sigma}\right) \sim IG(1, 1^2),$$

ここで、$0 < \sigma < \infty,\ -\infty < \mu < \infty,\ -\infty < X < \infty,$ であり、記号 $IG(m, c^2)$ は、母平均 m、母変動係数 c である逆ガウス型分布を表している。

定義 2. 一般化指数逆ガウス型分布 $X \sim GEIG(\mu, \sigma^2, \lambda)$ を以下で定義する。

$$\exp\left(\lambda\frac{X-\mu}{\sigma}\right) \sim IG(1, \lambda^2),$$

ここで、$0 < \sigma < \infty,\ -\infty < \mu < \infty,\ \lambda < \infty, -\infty < X < \infty$ であり、λ は無次元量である。$GEIG(\mu, \sigma^2, \lambda)$ に従う確率変数の確率素分 $f(x)dx$ は以下である。

$$f(x)dx = \frac{1}{\sqrt{2\pi}}\left[\exp\left(\lambda\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\lambda^2}\left(2\sinh\left(\lambda\frac{x-\mu}{2\sigma}\right)\right)^2\right\}\frac{dx}{\sigma},$$

ここで、$-\infty < x < \infty, -\infty < \mu < \infty, 0 < \sigma < \infty, |\lambda| < \infty$ であり、$\sinh(x)$ は双曲線正弦関数である。

謝　辞

本稿の内容は、岩瀬晃盛広島大学名誉教授との共同研究の成果である。

参 考 文 献

Jenkinson, A. F. (1955). The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements, *Quarterly Journal of the Royal Meteorological Society*, **81** (348), 158–171.

Kanefuji K. and Iwase K. (1996). Exponential Inverse Gaussian Distribution, *Computational Statistics*, **11** (*), 315–326.

医師主導治験による医療機器開発の経験

An experience of developing medical device by investigator-initiated clinical trials

データ科学研究系 伊藤 陽一（Yoichi M. Ito）

要　　旨

本稿では、アカデミア発の医療機器開発において、規制当局のどのような話し合いの上で、臨床試験 (治験) が立案され、承認に至ったということに関して記述を行う。医療機器に関する承認審査の規制が改正され、開発中断となっていた医療機器に関して、医師主導治験を実施することによって、承認申請を行うことができた。医師主導治験の実施にあたっては、主に有効性の判断基準、対象施設と症例数について、規制当局 (PMDA) との間で話し合いが行われ、当初の計画が変更された。治験の結果として、対象者 20 名中 17 名で有効性が確認され、新規の医療機器として承認された。

1. 開発の経緯

平成 14 年 (2002 年) に薬事法が改正され (平成 17 年 (2005 年) 施行)、医療機器に係る安全対策の見直しが行われた。本改正によって、リスクに応じたクラス分類制度が導入され、体内に留置して不具合が生じた場合に生命に危険を及ぼす可能性が高いものをクラス IV（高度管理医療機器）とし、医薬品と同様の Good Clinical Practice（GCP）基準に基づく治験が必要とされた。一方、平成 15 年 (2003) に「医薬品の臨床試験の実施の基準に関する省令」（GCP）の改正により、従来、医師自ら実施することはできなかった薬事法上未承認・適応外の医薬品を用いた治験を医師主導治験として実施することができるようになった。このような流れを受けて、アカデミア発の医薬品・医療機器開発を推進するために、平成 19 年 (2007) 年に、文科省の橋渡し研究支援推進プログラムにより、北海道大学、札幌医科大学、旭川医科大学の 3 大学共同で、北海道臨床開発機構という組織が作られた。

北海道大学病院整形外科三浪明男教授は、平成 16〜18 年度厚生労働省科学研究費補助金/免疫とアレルギー疾患予防・治療研究事業「関節リウマチ上肢人工関節に関する研究」によって、新規の人工関節の開発を行った。当該研究を通して、人工関節の有効性および安全性を確保するためには、インプラントの設置後のアライメント、コンポーネントの固定、軟部組織とのバランスが重要であり、また、正常手関節の運動解析によって、矢状面での伸展、屈曲より Dart thrower(投げ矢) 面での伸展、屈曲運動がより生理的であり、手関節周囲軟部組織に加わるストレスが少ないことが確認された。これらの条件を満たすべく、摺動面が卵円形をした橈尺屈半拘束型のモデルが採用され、原材料に関してはシリコンではなく、他関節において最も多くの実績のあるチタン合金および超高分子量ポリエチレンを使用することとされた。本製品を製造したナカシマプロペラ株式会社 (現ナカシマメディカル株式会社) が、2007 年 3 月に医療機器製造販売承認申請した。しかし、承認審査の過程において、審査側から、本開発品目に用いら

れている材料は人工膝関節や人工股関節において多くの使用実績があるものの適用部位が手関節の同等製品が存在しないことから、安全性および有効性を確認した臨床データが必要との見解が示された。しかし、ナカシマプロペラ株式会社は人を対象とした治験の経験を有していなかったことから、申請を取り下げ、開発は中断されていた。このような経緯から、北海道臨床開発機構が医師主導治験を支援することにより、開発が継続されることとなった。

2. 有効性指標の選択と有効性の判断基準

ICH-E9 統計ガイドラインでは、「主要変数の選択には、関連した研究領域で一般に認められている規範と標準を反映させるべきである。先行研究や公表論文で使用された実績のある、信頼性及び妥当性の確立した変数を使用することが薦められる」と述べられている。そこで、本治験では、Figgie らによって用いられた人工手関節に関する評価尺度 (Wrist Scoring System by Figgie) を用いることとした。Wrist Scoring System by Figgie では、除痛 (Pain relief) の程度について 50 点、関節を何度動かすことができるかという関節の可動性 (Range of motion) について 20 点、関節を様々な角度で止めることができるかどうかという関節の安定性 (Function) について 30 点の計 100 点満点で評価される。(Figgie et al. 1990).

当初、観察期間は人工手関節の埋植後 12 ヶ月、Wrist Scoring System by Figgie 70 点以上を有効性の判断基準として提案したが、PMDA との相談の結果、観察期間は 12 ヶ月から 18 ヶ月に延長され、実際に関節が動くことを評価するため、Range of Motion のスコアが 10 点以上という条件が追加された。

3. 対象施設と症例数

有効性指標に基づき、統計学的に有効性を証明できる必要症例数は 8 例と推定されたため、北海道大学病院において、目標症例数 10 例での治験実施を提案した。しかし、PMDA との相談の結果、どこの施設でも手術できることを確認するために、実施施設は 2 施設以上、また、有害事象の発現割合の推定の観点から、20 例以上の実施が要求され、2 施設、20 例で治験を実施することとなった。

4. 対象施設と症例数治験成績および考察

有効性については、主要評価項目である埋植後 18 か月時点における有効性の判断基準を達成した被験者の割合は 85 % (17 例)、安全性については、有害事象・不具合等が評価され、本品との因果関係が否定できない有害事象が 1 例、2 事象、不具合が 2 件、X 線学的評価で 18 か月に緩みありと判断された被験者が 4 例であった。本成績を持って、新規の医療機器として承認された。治験デザイン上の PMDA との相談のポイントとしては、有効性指標の明確化、安全性については長期の安全性を重視したものであった。レギュラトリーサイエンスの観点から、有効性と安全性のバランスが考慮された結果であると思われる。

参 考 文 献

Figgie, M. P., Ranawat, C. S., Inglis, A. E., Sobel, M., & Figgie, H. E., 3rd. (1990). Trispherical total wrist arthroplasty in rheumatoid arthritis. J Hand Surg Am, 15(2), 217-223.

医薬品医療機器総合機構 (2016). 審査報告書. http://www.pmda.go.jp/medical_devices/2016/M20161117001/510462000_22800BZX00385_A100_2.pdf

経時データ解析と健康指標の長期推移

Longitudinal Data Analysis and Long Term Trends of Health Related Measures

データ科学研究系 船渡川 伊久子（Ikuko Funatogawa）

1. 経時データ解析のための自己回帰線形混合効果モデル

生物統計学分野において，経時データ解析の手法は線形混合効果モデルの発表から，大きく発展しましたが，その多くは静学的内容です．線形混合効果モデルと時系列解析で用いられる自己回帰モデルを拡張した自己回帰線形混合効果モデルを提案しました（Funatogawa et al., 2007）．このモデルは，反応を直前の反応と固定効果および変量効果の共変量に回帰し，誤差の分散共分散構造を拡張しています．反応は漸近値に向かい推移し，変量効果は漸近値の個体間差を表し，非線形混合効果モデルでの monomolecular 成長曲線に対応します．周辺モデルとして表現することで，従来提案されていなかった倹約的で汎用性の高い分散共分散構造を導出し，特に欠測のあるデータで活用できます（Funatogawa et al., 2008b）．複数の反応変数への拡張も容易です（Funatogawa et al., 2008a）．反応を一時点前の反応に回帰すると，時間依存性共変量の影響を線形および非線形混合効果モデルとは異なる方法で扱い，過去の共変量の影響を取り入れるダイナミックなモデルとなります．投与量が時間依存性共変量で，観測間隔が一定でない場合に，状態空間表現を用いて最尤推定を行う方法を提案しました（Funatogawa and Funatogawa, 2012a）．また，投与量が時間依存性共変量で，観測された反応の値によって変更される場合，モデルが正しければ最尤推定値に偏りはありません（Funatogawa and Funatogawa, 2012b）．モデルのメカニスティックな側面に着目し，和書（船渡川・船渡川, 2015）および英文書籍（Funatogawa and Funatogawa, in press）を出版しました．一方，経済学の分野では，動学的パネルデータ分析と呼ばれ，反応を以前の反応に回帰するモデルが個体間差を考慮する形で観察研究に用いられています．

2. 健康指標の長期推移

喫煙は，喫煙開始から死亡までが非常に長いことや出生コホート間の喫煙習慣の複雑な違いが，その影響の大きさを分かりにくくしています．喫煙開始年齢は喫煙期間も表す重要な指標ですが，各国の長期推移の報告は限られます．喫煙開始割合，喫煙率，肺癌死亡率の加齢変化の出生年による長期推移を，日本および英国についてそれぞれ WHO Bulletin と BMJ Open に発表しました（Funatogawa et al., 2013; Funatogawa et al., 2012）．喫煙と肺癌の研究は，関心の高い分野で，しばしば医学の主要ジャーナルに掲載されますが，喫煙の早期中止でリスクが減少するという主張が目立ち，若年期の早期喫煙開始の危険性が十分に伝えられていません．背景には，統計学や疫学の方法論上の問題があると考えています．米国や英国女性の喫煙開始，喫煙率，肺癌死亡率の出生年による変化と関連した内容を，N Engl J Med（レター）や Lancet（レター）で報告しました（Funatogawa, 2018; Funatogawa, 2013）．

Body Mass Index（BMI）は重要な健康関連指標ですが，数十年単位の経年的加齢変化を計量的に評価した報告はなく，横断調査から出生コホートを考慮せずに求めた加齢変化を用いて

いました．そこで，日本の代表的な繰り返し横断調査で，無作為抽出で行われ，60 年以上の記録が存在する国民健康・栄養調査のデータを利用し，出生コホートを考慮した BMI の加齢変化を，0〜25 歳女性の結果を BMJ，20〜60 歳代成人男女の結果を Int J Epidemiol に発表しました（Funatogawa et al., 2008c; Funatogawa et al., 2009）．日本人女性は，より最近の出生コホートほど，子供の頃はより過体重ですが，成人するとより痩せていること，横断調査と出生コホート別では加齢変化パターン自体が異なること等を示しました．喫煙や肥満に関する日本語での解説を発表しています（船渡川, 2014a; 船渡川, 2014b; 船渡川・船渡川, 2015）．

参 考 文 献

Funatogawa, I. (2013). The first generation in which many women began smoking, *Lancet*, **381**(9876), 1455.

Funatogawa, I. (2018). Incidence of lung cancer among young women, *The New England Journal of Medicine*, **379**(10), 988.

Funatogawa, I. and Funatogawa, T. (2012a). An autoregressive linear mixed effects model for the analysis of unequally spaced longitudinal data with dose-modification, *Statistics in Medicine*, **31**(6), 589–599.

Funatogawa, I. and Funatogawa, T. (2012b). Dose-response relationship from longitudinal data with response-dependent dose modification using likelihood methods, *Biometrical Journal*, **54**(4), 494–506.

Funatogawa, I. and Funatogawa, T. (in press). *Longitudinal Data Analysis: Autoregressive Linear Mixed Effects Modesl*, Springer, Singapore.

Funatogawa, I., Funatogawa, T., Nakao, M., Karita, K. and Yano, E. (2009). Changes in body mass index by birth cohort in Japanese adults: results from the National Nutrition Survey of Japan 1956-2005, *International Journal of Epidemiology*, **38**(1), 83–92.

Funatogawa, I., Funatogawa, T. and Ohashi, Y. (2007). An autoregressive linear mixed effects model for the analysis of longitudinal data which show profiles approaching asymptotes. *Statistics in Medicine*, **26**(2113-30), 2113–2130.

Funatogawa, I., Funatogawa, T. and Ohashi, Y. (2008a). A bivariate autoregressive linear mixed effects model for the analysis of longitudinal data, *Statistics in Medicine*, **27**(6367-78), 6367–6378.

Funatogawa, T., Funatogawa, I. and Takeuchi, M. (2008b). An autoregressive linear mixed effects model for the analysis of longitudinal data which include dropouts and show profiles approaching asymptotes. *Statistics in Medicine*, **27**, 6351–6366.

Funatogawa, I., Funatogawa, T. and Yano, E. (2008c). Do overweight children necessarily make overweight adults? Repeated cross sectional annual nationwide survey of Japanese girls and women over nearly six decades, *British Medical Journal*, **337**(a802).

Funatogawa, I., Funatogawa, T. and Yano, E. (2012). Impacts of early smoking initiation: long-term trends of lung cancer mortality and smoking initiation from repeated cross-sectional surveys in Great Britain, *BMJ Open*, **2**(5).

Funatogawa, I., Funatogawa, T. and Yano, E. (2013). Trends in smoking and lung cancer mortality in Japan, by birth cohort, 1949-2010, *Bulletin of the World Health Organization*, **91**(5), 332–340.

船渡川伊久子 (2014a). 肺の健康とタバコ 近年の日本における肺癌発生の推移と関連因子, 健康管理, **61**(8), 19–25.

船渡川伊久子 (2014b). 思春期の栄養と運動を考える 小児・思春期の発育についての疫学的検討, 思春期学, **32**(1), 145–149.

船渡川伊久子, 船渡川隆 (2015). 『経時データ解析』, 朝倉書店, 東京.

先端医学研究の発展を支える統計数理とデータサイエンス

Statistical Mathematics and Data Science for the Developments of Advanced Medical Researches

データ科学研究系 野間久史（Hisashi Noma）

1. はじめに

世界規模で進む社会の高齢化により、医療費・医療資源の効率的な配分は、高水準の医療・福祉を維持するべく、先進諸国において重要な問題となっている。特に日本では、既に 65 歳以上の高齢者の割合が人口の 4 分の 1 を超えており、WHO が定める「超高齢社会」となっている。年間の医療費も 40 兆円を超えており、深刻な状況にある。このような中で、将来に向けて高水準の医療・福祉の持続、および、医療技術のさらなる発展を図るためには、その基盤となる信頼できる科学的根拠が不可欠であり、そのために、統計学・データサイエンスは極めて重要な役割を果たしている。統計数理研究所においても、2018 年 4 月に、医療・健康科学領域におけるデータサイエンスの最先端の研究および人材育成の拠点として、医療健康データ科学研究センターが設立されている。本稿では、著者らが取り組む研究プロジェクトの一部を紹介する

2. ネットワークメタアナリシス

医療費・医療資源の効率的な配分のために、医療政策・診療ガイドラインの策定において重要になるのが、既に多く存在する医薬品・医療技術のいずれが最も高い有効性・安全性を持ち、経済的であるか、ということである。しかしながら、多くの治療法をすべて比較して、それらの優劣を比べるための十分な検出力を達成する臨床試験を行うためには、一般的に、数万人〜数十万人以上の規模の試験を行う必要があり、現実的には不可能である。この問題を解決するために、近年の医療統計学の研究から、ネットワークメタアナリシスという新しい方法論が開発された。ネットワークメタアナリシスは、過去に行われた臨床試験の結果を統合し、対象となる治療法間の比較評価を行ったエビデンスを提供してくれる新しい方法として、近年、先端的な臨床医学・医療技術評価で急速に普及している。本邦からも、優れた先進的な研究成果がいくつか報告されており、例えば、著者の野間も参加した、双極性障害の 17 種類の薬物療法のネットワークメタアナリシス（Miura et al., 2014）などがある。

ネットワークメタアナリシスにおいて、統計科学の方法論は、そのエビデンスの科学的な妥当性および精確性の根幹を支えるための中心的な役割を果たしている。ネットワークメタアナリシスは、複数の異なる情報源から得られるエビデンスを統合するため、その異質性を適切に考慮した複雑な構造を持つマルチレベルモデルを用いる必要がある。これらの統計的推測手法には、最尤法やベイズ法を基礎とした方法が一般的に用いられるが、Noma et al. (2018) などによって、一般的なネットワークメタアナリシスが行われる条件下におけるこれらの推測手法の不正確性が明らかにされ、それを解決するための方法論の開発・整備が活発に進められてい

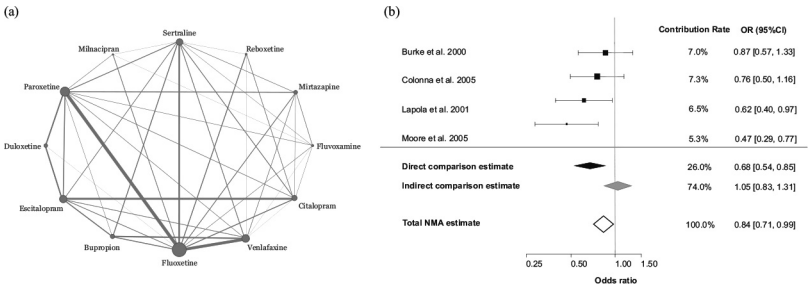


図 1. (a) 新世代抗うつ薬の臨床試験のエビデンスのネットワーク。(b) Escitalopram vs. Citalopram の比較における複合尤度法による推定量の分解の結果（Noma et al. (2017) より転載）。

る。Noma et al. (2018) では、リサンプリング法を用いて、この不正確性の問題を解決した推定・検定の手法を提案している。また、特に、近年、試験間の異質性を考慮した治療効果の評価方法として、将来の集団において期待される治療効果の「予測」も重要な問題となっているが、Nagashima et al. (2018) は、現在、国際的なスタンダードとなっている予測区間の構成方法の不正確性を明らかにし、基礎的な 2 治療の比較試験の統合解析において、これより大幅に優れた性能を持つ正確な予測区間の構成方法を提案した。この正確な予測手法のネットワークメタアナリシスへの拡張に関する研究も現在進行中である。加えて、治療効果の評価におけるバイアスの評価において重要となる、ネットワーク上でのエビデンスの不整合性を評価するための方法の開発研究も活発に行われており、Noma et al. (2017) は、複合尤度法を用いた新しい不整合性の評価方法を提案している。Oxford 大学，京都大学の研究グループと協同し、精神医学領域のネットワークメタアナリシスで史上最大規模の研究であった新世代抗うつ薬の臨床試験のネットワークにこれを応用し、興味深いスポンサーシップバイアスの可能性を示唆する結果を報告している（図 1）。

参 考 文 献

Miura, T., Noma, H., Furukawa, T. A., et al. (2014). Comparative efficacy and tolerability of pharmacological treatments in the maintenance treatment of bipolar disorder: a systematic review and network meta-analysis. Lancet Psychiatry 1, 351-359.

Nagashima, K., Noma, H., and Furukawa, T. A. (2018). Prediction intervals for random-effects meta-analysis: A confidence distribution approach. Statistical Methods in Medical Research, doi: 10.1177/0962280218773520.

Noma, H., Nagashima, K., Maruo, K., Gosho, M., and Furukawa, T. A. (2018). Bartlett-type corrections and bootstrap adjustments of likelihood-based inference methods for network meta-analysis. Statistics in Medicine 37, 1178-1190.

Noma, H., Tanaka, S., Matsui, S., Cipriani, A., and Furukawa, T. A. (2017). Quantifying indirect evidence in network meta-analysis. Statistics in Medicine 36, 917-927.

集約的シンボリックデータ解析

Aggregated Symbolic Data Analysis

データ科学研究系　清水信夫（Nobuo Shimizu）

近年，様々な分野において，Web システムを用いたデータの収集が多用されており，各分野における活動の詳細なデータが計算機上に連続的に蓄積されるようになっている．それらのデータは連続変数とカテゴリー変数が混在した多次元データであることが多く，データの個体数についても非常に大きな場合が多々存在する．

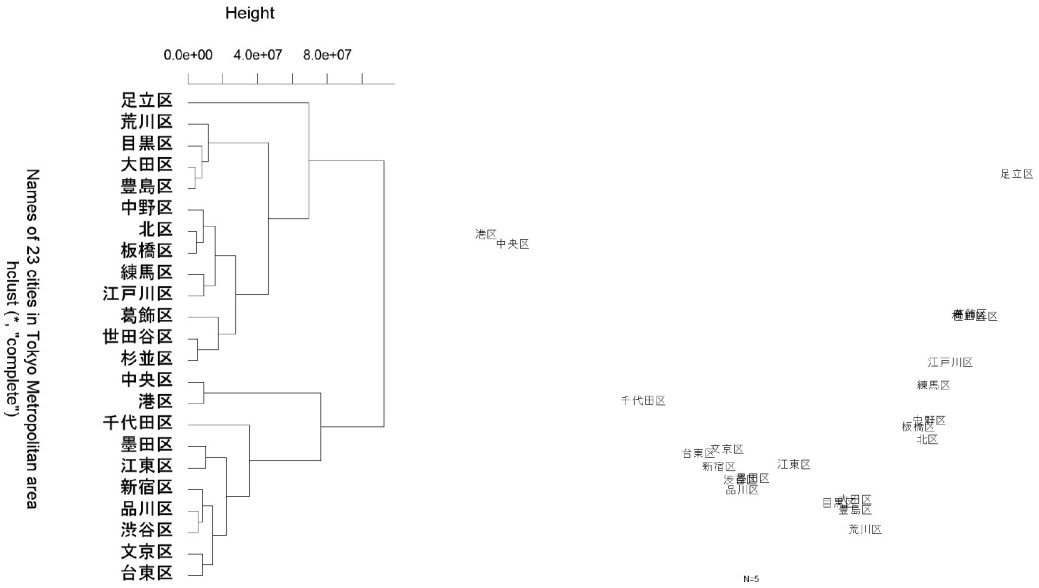
このようなデータは，全体像を見たり詳しい解析を行ったりするための，一般的なデータ管理・処理ソフトウェアによる取り扱いが困難なほどの巨大かつ複雑なデータ，いわゆる“ビッグデータ”の代表例と考えられ，従来多用されていた個体データに着目する方法以外の解析手法の開発が求められる．ただし，そのようなときは，個体データが意味のある自然な比較的少数のグループに分かれる場合が少なからず存在する．したがって，個体データそのものではなく，個体がまとめられたグループに関心を向けた手法が必要である．その方法の一つとして，Diday(1988) によりシンボリックデータ（Symbolic Data, SD）という概念が提案され，データの中で個体がまとめられたグループを SD として解析するシンボリックデータ解析（Symbolic Data Analysis, SDA）が提唱されている．

SDA においては，データとして各連続変数ごとに 1 つの値ではなく，ある値を中心としてばらつきをもつデータ（区間データや分布値データ）などで表されるものが考えられ，それらを SD と考えた場合の解析として従来の各種多変量解析手法を拡張する研究が，Bock and Diday(2000)，Billard and Diday(2005)，Diday and Noirhomme-Fraiture(2008) などにまとめられている．それら以外にも，SD に関するクラスタリングに関しては Verde(2004) や Irpino and Verde(2006) など，多次元尺度構成法に関しても Denœux and Masson(2000) や Groenen et al.(2006) などの研究がある．これまでの SDA においては，データは最初から区間や集合のような形で与えられている場合が多く，そこではグループ内の複数の変数間の関係は無視される．例えば，2 つの連続変数間の相関関係は考慮されない．

しかしながら，現代のビッグデータにおいては、元の個体データは保持されている．個体数や変数が極めて多いデータの場合は移動や計算に困難を伴うが，どうしても必要ならばグループに関するいかなる記述統計量も計算することは可能である．そこで，グループにおける多次元データの情報を可能な限り簡潔な形で持つために，複数の記述統計量を考えることにし，それを集約的シンボリックデータ（Aggregated symbolic data，ASD）と呼ぶこととする．

ASD は，グループ内の個体データのそれぞれの変数および複数の変数の組み合わせに関し，情報の脱落を可能な限り抑えつつ保持すべき値を可能な限り少なくして取り扱いを容易にするために，2 次までのモーメントについて求められる統計量の集合として表される．ASD に含まれる統計量の例としては連続変数における平均および分散共分散行列，カテゴリー変数における分割表などがある．これらを用いて，ASD 間の非類似度を尤度比検定統計量やカイ 2 乗統計量などを用いて定義し，それらを用いてクラスタリングや多次元尺度構成法を行う方法を提案した．この方法を東京都区部の不動産情報データに適用し，23 の特別区をそれぞれグループとして考え，データの各変数から各グループ間の非類似度を ASD を用いて求めた上で階層的ク

ラスタリングおよび多次元尺度構成法を行った例を図 1 に示す.



(a) 階層的クラスタリング (b) 多次元尺度構成法

図 1. 集約的シンボリックデータを用いた不動産情報データの解析例
(清水・中野・山本 (2018))

参 考 文 献

Billard, L. and Diday, E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*, John Wiley & Sons Ltd, Chichester, UK.

Bock, H.-H. and Diday, E. (2000). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*, Springer-Verlag, Berlin.

Denœux, T., Masson, M. (2000). Multidimensional scaling of interval-valued dissimilarity data, *Pattern Recognition Letters*, **21**, (1), 83–92.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis, *Classification and Related Methods of Data Analysis*, 673–684.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, John Wiley & Sons Ltd, Chichester, UK.

Groenen, P. J. F., Winsberg, S., Rodriguez, O. and Diday, E. (2006). I-Scal: Multidimensional scaling of interval dissimilarities, *Computational Statistics and Data Analysis*, Elsevier, **51**, (1), 360–378.

Irpino, A. and Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data, *Data Science and Classification*, Springer, Berlin, 185–192.

Verde, R. (2004). Clustering methods in symbolic data analysis, *Data Science and Classification*, Springer, Berlin, 299–317.

清水信夫, 中野純司, 山本由和 (2018). 集約的シンボリックデータのカイ 2 乗統計量を用いた非類似度とその不動産情報データへの適用, 統計数理, **66** (2), 279–294.

ツイートを利用した都市気温の解析 –都市インテリジェンス向上をめざして–

Urban Temperature Analysis Using Participatory Sensing Data "Tweets"

–Toward Improvement of Urban Intelligence–

モデリング研究系 松井 知子（Tomoko Matsui）

1. はじめに

近年、地球温暖化によって異常気象リスクが顕在化している。特に都市での熱波リスクは深刻である。このリスクに対処するには各地点、各時刻での気温状況を正確に把握することが重要である。気象庁などは都市内にいくつかの観測拠点を設けて、定期的に気温を測定している。しかし、観測拠点数や時間間隔はスパースである。そこで本研究では、都市の各所で絶え間なくつぶやかれる「暑い」「だるい」などの暑さに関係すると想定するツイートデータ（ヒートツイート）を利用して、上記のスパースな気温観測データを空間的、時間的に補間する統計・機械学習の方法について研究を行った。将来的には本方法を熱波対策に役立て、都市インテリジェンス向上を目指したいと考えている。

2. ヒートツイートと気温観測データとの関係

気温補間においてヒートツイートが有用できることを確かめるために、コピュラモデルを用いて各観測拠点におけるヒートツイートと気温観測データの関係を調べた。ここでヒートツイートは、一般化加法モデルとロジスティックモデルにより、主要駅／公園／河川からの距離、日中・夜間人口密度、土地利用の種類などを考慮した確率的な強度で表した。表 1 に気温、気温変化のそれぞれとヒートツイートの lower tail（気温は低い／気温変化は小さい、ヒートツイートは少ない時に相当）と upper tail（気温は高い／気温変化は大きい、ヒートツイートは多い時に相当）における依存性を示す。一般に気温が高い時にヒートツイートは多くなると考えられてきたが、熱波で問題となる upper tail では特に気温変化がヒートツイートに関係していることを新たに見出した。

表 1. 気温／気温変化とヒートツイートの依存性

	Lower tail	Upper tail
気温	0.40	0.00
気温変化	0.20	0.23

3. S-BLUE 法による気温補間

S-BLUE（spatial best linear unbiased estimation）法（Peters, G. W., Nevat, I. and Matsui, T. (2015)）の枠組みを用いて、各観測拠点において正確ではあるがスパースに観測される気温データとヒートツイートを組み合わせ、都市気温の時空間モデルを表す線形汎関数を構成す

る。S-BLUE 法ではガウス過程を内包しており、連続／離散量の異種混合データの複雑な非線形な関係性をうまく表現することができる。また、平均二乗誤差最小化に基づいてパラメータを効率的に推定することができる。図 1 に、都内について上記モデルを用いて、朝 6 時、昼 12 時、夕方 18 時の都内の気温を補間した結果を示す。ヒートツイートを利用することにより、昼 12 時には東京の中心部（山手線内）がヒートアイランド現象によって気温が上昇する様子を捉えることができた。

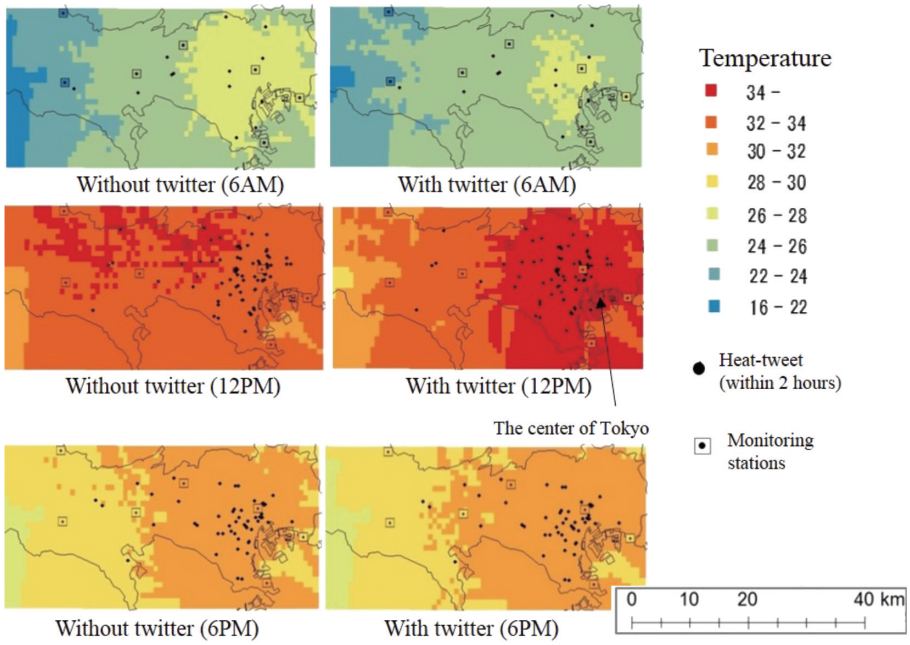


図 1. 朝 6 時、昼 12 時、夕方 18 時の都内の気温を補間
（左側：ヒートツイートを利用しない場合、右側：ヒートツイートを利用する場合）

4. おわりに

本研究は統計的機械学習研究センター「都市インテリジェンス」プロジェクトにおける、村上大輔助教、Gareth W. Peters 教授（Heriot-Watt 大）、山形与志樹博士（国立環境研究所）との共同研究の成果（ Murakami, D., Peters, G. W., Yamagata, Y. and Matsui, T. (2016)）である。今後もこのプロジェクトを基盤として、都市レジリエンス向上を目標とした、統計数理／機械学習に基づく技術とその理論の研究開発を進めていきたい。

参 考 文 献

Murakami, D., Peters, G. W., Yamagata, Y. and Matsui, T. (2016). Participatory Sensing Data Tweets for Micro-Urban Real-Time Resiliency Monitoring and Risk Management, IEEE Access, 4, 347-372, doi:10.1109/ACCESS.2016.2516918.

Peters, G. W., Nevat, I. and Matsui, T. (2015). How to Utilize Sensor Network Data to Efficiently Perform Model Calibration and Spatial Field Reconstruction, *Modern Methodology and Applications in Spatial-Temporal Modeling* (eds. G. W. Peters and T. Matsui), Chap. 2, 25-62, Springer.

機械学習による新物質の発掘

Machine learning for accelerated materials discovery

データ科学研究系 吉田 亮（Ryo Yoshida）

キーワード：マテリアルズインフォマティクス，機械学習，ベイズ推論，転移学習

1. マテリアルズインフォマティクス

MI の問題の多くは，順問題と逆問題の形式に帰着する．順問題の目的は, 系の入力 S に対する出力 Y の予測である．物性予測の文脈では，入力は物質（分子，組成，結晶等），出力は物性値（エネルギー，電子状態等）に相当する．これまでの材料研究では，第一原理計算や分子動力学計算等の理論計算が順方向の予測を担ってきた．このタスクをデータ科学のモデルに代替させることが，MI の中心的課題のひとつである．これに対し，逆問題では文字通り逆方向の予測を行う．すなわち，出力 Y の値（例えば目標物性）を設定した上で，それを達成する入力 S の状態（構造）を予測する．データ科学の観点からみると，これらの計算は，物質構造の“表現・学習・生成”を行うことに相当する．記述子と呼ばれる特徴ベクトルを通して物質の構造を“表現”し，データのパターンから構造から物性の数学的写像を“学習”する．さらに，計算機を用いて所望の物性値を有する物質を“生成”し，有望な候補物質を炙り出す．対象となる入力 S は，分子，組成，結晶，混合物，プロセス，合成経路等，問題に応じて多様な形式をとりうる．

2. 機械学習によるハイスループットスクリーニング

構造と物性の関係を表す実験や理論計算のデータから，物質 S の物性 Y の予測モデル f(S) を導くことが目的である．記述子と呼ばれる特徴ベクトルを通して物質の構造を“表現”し，データのパターンから構造から物性の数学的写像を“学習”する．記述子は MI における最も基本的な要素技術である．入力 S の形式が多様であることから，対象領域ごとに独自に研究が展開している．膨大な数の候補物質のライブラリを作製した上で，訓練済みモデルを用いてスクリーニング実験を実施する．実験や理論計算に比べて機械学習のモデルは圧倒的に計算コストが低いため，膨大な数の候補物質を対象とする物性評価を行うことができる．

3. 転移学習

機械学習の他の応用領域に比べて，材料研究のデータ数は圧倒的に少ない．データ科学が本格的に導入されて間もないこともあり，データベースの整備は発展途上の段階にある．とりわけ，研究対象が最先端に近づくにつれて，スモールデータの傾向はより顕著になる．スモールデータに対する解決策として，転移学習と呼ばれるアプローチが有望視されている．転移学習では，あるタスクの訓練済みモデルを別のタスクに再利用する．我々は，XenonPy.MDL という訓練済みモデルライブラリを開発している（Yamada et al. (2019)）．低分子化合物，高分子，無機結晶等，様々な物質に対する>105 の物性推算モデルが収録されている．図 1 は，ニューラルネットワークの転移学習に基づくポリマー定圧熱容量（Cp）の予測を例示したものであ

る．第一原理計算で低分子化合物の化学構造と定容熱容量（Cv）の関係を表す 133,805 個のデータを取得し（Ramakrishnan et al. (2014)），ソースモデルを導いた．ソースモデルの部分ネットワークを用いて特徴量の計算を行い，高分子材料データベース PoLyInfo（Otsuka et al. (2011)）に登録されている 58 個の Cp のデータを用いて予測モデルを構築した．ソースタスクの学習過程で Cv と Cp に共通する縮約特徴量を獲得し，これを用いることでたった 58 個のデータからポリマー Cp の予測モデルを導くことに成功した．

4. ベイズ推論による物質構造の設計

我々は，ベイズ推論や機械学習を方法論の基軸とし，新物質の創製を目的に設計と合成を対象とする機械学習の手法とソフトウェアを開発してきた（Ikebata et al. (2017)）．実験やシミュレーションから得られるデータを用いて，物質の構造から物性の順方向の予測モデルを構築する．これに条件付き確率のベイズ則を適用し，物性から構造の逆方向のモデルを導き，このモデルから仮説物質を発生させることで，所望の物性を有する埋蔵物質を炙り出す．確率的言語モデルに基づく構造生成器や機械学習の様々な技術を結集させて構築した確率推論のアルゴリズムである．現在，この手法を用いて様々な材料系を対象に実証研究を進めている（Wu et al. (2019) など）．機械学習で物質構造の設計図が描き，大量の埋蔵物質を発掘する．これが本研究のグランドチャレンジである．

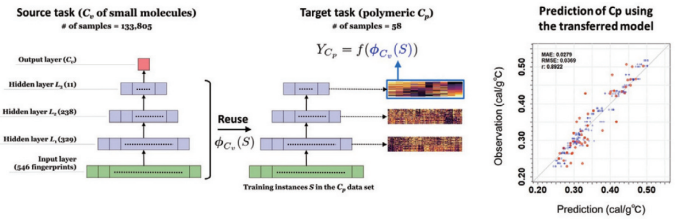


図 1．転移学習によるポリマー定圧熱容量（Cp）の予測

参 考 文 献

Ramakrishnan, R., Dral, P. O., Rupp, M. and von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules, Scientific Data, 1, 140022.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. and Yamazaki, M. (2011). PolyInfo: Polymer database for Polymeric Materials Design, 2011 International Conference on Emerging Intelligent Data and Web Technologies, 22-29.

Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Transfer learning: accelerated materials discovery with small data.

Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, Journal of Computer-Aided Molecular Design, 31(4), 379-391.

Wu, S., Kondo, Y., Kakimoto, M., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Schick, C., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm.

方向統計学と動物移動生態学

Circular statistics and movement ecology

データ科学研究系 島谷 健一郎（Kenichiro Shimatani）

動物の移動軌跡データ

動物装着型 GPS を用いる研究は、水中・空中・陸上を問わず、様々な動物について世界各地で盛んに行われている。GPS データには、動物がいつどこにいたのかという情報に加え、各時点における速さと方向を求めることで、どのように動いていたかという情報も含まれる。それは、いつどこで何をしていたかという推定につながる。

円周自己回帰モデル

速さに特徴的なパターンが見られなくても、方向には見られる場合がある。そこでは、方向とその変化に関する時系列モデルが基本的なデータ解析の道具となる。そこで基本となるのは、各時点で方向はランダムに変動するランダムウォークである。しかし、実際の動物の動きで多く見られるのは、巣や餌場などへ向かっていく動きや、ある場所の周辺をうろついているようなパターンである。こうした移動軌跡を表現するには、360 度で元の 0 度に戻るという方向データの特性を鑑みた、円周上の統計モデルが適している。そこで、Shimatani et al. (2012) では、Kato (2010) の円周自己回帰モデル（図 1a）の適用を提唱した。

この自己回帰モデルでは、目的変数は特定の方向 α に集中する傾向を示す。シミュレーションで順に方向を生成し、速さを一定にして軌跡データを作ると、ある方向に向う軌跡を描く。ランダムな変動には、von-Mises 分布などの円周上の確率分布を用いる。確率密度関数のグラフが尖がっていてランダムな変動が小さいと、目標方向へ向かって進む軌跡が得られる（図 1b ①）。変動の大きい確率分布を用いると、軌跡は震えながらゆっくり目標方向に進んでいく（図 1b ②）。言い換えると、確率分布の尖り具合を決めるパラメータは、行くべき方向へ向かう調整力という解釈を伴う。一方、自己回帰モデルの回帰係数 w の値が 1 に近いと、目標方向への集中が弱いため方向修正する効果が弱く、何かの拍子にあらぬ方向に向かうとしばらく進んでから元の方向に戻る（図 1b ③）。つまり、回帰係数 w は、その方向に向かおうという動物の意志の強さという解釈を伴う。なお、$w = 1$ とするとランダムウオーク (動物行動では correlated random walk と呼ばれる) となる。

実際のデータが与えられたときは、最尤法などでパラメータを最適化することで、動物の意志の強さ、目標方向、方向の調整力に関する定量的評価を与えられる。また、$w = 1$ としたランダムウオークと AIC などの情報量規準で相対評価することにより、動物が進みたい方向を持っていたのかどうかを検証できる。

基本統計モデルが満たすべき条件

一般に、ある現象についてのある型のデータに関する統計モデルは、次の 3 つの条件を満た

しているべきである (Shimatani et al. 2012)。

1. 未知パラメータにその現象に関する定量的解釈が伴う。
2. 数学的あるいは数値計算で扱いやすい数式からなる。
3. 広範な現象に適用でき、状況に応じ適宜拡張できる。

上述したように、Kato (2010) の円周自己回帰モデルを動物軌跡に適用すると動物の行動に関する解釈を伴っており、1 を満たす。2 もよく知られた関数で尤度が書けるなど、数学的扱いも容易である。説明変数を、1 時点だけでなく 2 時点や 3 時点前まで含め、それらの線形和（方向なので 1 度複素平面の単位ベクトルに直して線形和を取り、再び argument で方向に戻す）を使う式に拡張できる。すると、左に回り始めたらしばらく左に回るなどの、滑らかな軌跡を描く。さらに、集中させる方向を、特定の地点と各時点における地点を結ぶ方向にする（α は時間とともに変わる）と、特定地点の周辺をうろつきまわる軌跡を描ける（図 1c）。そこでは回帰係数 w は、その地点に留まりたい意志の強さといった解釈ができる。さらに、3 次元空間における軌跡へも、若干の修正で可能である（島谷 2015）など、Kato (2010) の円周自己回帰モデルは豊かな拡張性を有する。すなわち、3 も満たし、動物行動軌跡の方向に関する基本統計モデルとしての役割を担える。このように、方向統計学は動物の移動生態学 (movement ecology) に関する基本モデルを与え、様々な発展をもたらす可能性を有する。逆に、動物行動に関するデータは、方向統計学の発展を促す。両者は表裏一体となって相性よく発展していくと期待できる。

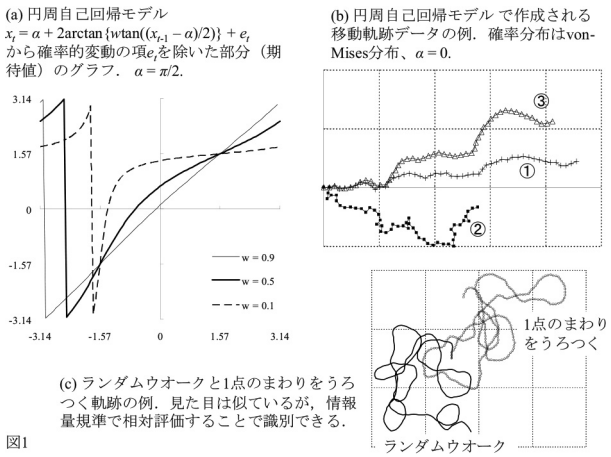


図1

参 考 文 献

Kato, S. 2010. A Markov process for circular data. Journal of the Royal Statistical Society B 72:655-672.

Shimatani I.K., Yoda, K., Katsumata, N., Sato, K. (2012). Toward the Quantification of a Conceptual Framework for Movement Ecology Using Circular Statistical Modeling, PLoS One, doi:10.1371/journal.pone.0050309.

島谷健一郎 (2015) 3 次元軌跡データの基本モデルとその限界. 数理解析研究所講究録 1940: 95-100.

機械学習に基づく高熱伝導率ポリマーの設計事例

A case study on high thermal conductivity polymer design based on machine learning

データ科学研究系 ウ ステファン (Stephen Wu)

キーワード：機械学習，マテリアルズ・インフォマティクス，ポリマー

1. はじめに

新しい材料を開発する従来の方法は，専門家の知識を使用して特定の物性要件を持つ少数の候補者を提案することです．このような方法は，膨大な化学空間（およそ 10^{60} 候補者（Bohacek et al., 1996））のために非効率的であることが知られている．人工知能技術の急速な発展のおかげで，機械学習に基づくより効率的な候補生成方法は，化学フロンティアの拡大ペースを加速すると期待されている．しかし，機械学習はどれほど強力であっても，単独で材料開発の挑戦を克服することはできません．計算分子設計法の開発の長い歴史にもかかわらず，合成までたどり着いたアルゴリズム的に設計された新規分子の例は限られている．これは，合成の難しさ，専門家の知識と学習された機械の知能の不一致，実用的なアプリケーションで厳しい要件などによるものであると推測しています．本研究では，Bayesian 分子設計という機械学習技術によって設計された高熱伝導率の新規ポリマーの発見を成功した．

2. 手法

ポリマーの熱伝導率（λ）は実用的重要性のために多くの注目を集めている（Anderson, 1966）． しかし，設計のためにポリマーのメカニズムを完全には理解していない．本研究はこれまでに提案されているいくつかの計算設計法（Zhong et al., 2001）と異なり，iqspr と呼ばれるベイジアン分子設計アルゴリズムに基づくデータ駆動アプローチを使用している（Ikebata et al., 2017）．データのソースは PolyInfo データベース（Otsuka et al., 2011）である．

R パッケージの iqspr は，ユーザーが指定した材料特性が与えられた探索空間を表す事後分布に対してサンプリングを実行します．ベイズの定理により，事後分布は尤度と事前分布との積として計算することができる．事前分布は，材料の物性に何の制約もなく化合物らしいものを表し，尤度は材料が与えられた場合に所望の材料物性を得る確率である．この問題設定の大きな課題は，信頼性の高い尤度モデルを訓練するためのデータが不十分であり，事後分布サンプリング・ステップ中に欠陥が生じることです．小さなデータの問題を克服するために，iqspr で元の目標の λ ではなく，λ に関連する物性を選択する．図 1 は，PoLyInfo におけるいくつかの関連するポリマー物性のデータを示す． 最終的に，高い λ の元のターゲットに代えて，高いガラス転移温度（Tg）および溶融温度（Tm）を目標とした．その後，iqspr から生成された提案候補は転移学習と専門知識を備えたポストスクリーニングを経た．我々の問題における転移学習は，λ に相関し，十分に大きなデータセットを有する物性のモデルを構築して，これらの

モデルの一部は，λのモデル構築を助けるために利用される．そして，専門知識による産業応用の観点から設計の要素を追加する．

3. 結果

iqspr パッケージでは，Tg が> 250 ℃，Tm が> 350 ℃になるようにターゲット領域を設定する． 回帰モデルにはデフォルト値が使用され，Tg 値と Tm 値が記録されたすべてのホモポリマーを含む訓練データが使用されます．ホモポリマーらしい分子を生成するために使用された従来のモデルでは，PoLyInfo データベースで利用可能な 14,424 ホモポリマーのすべて使用した． その後，生成された分子はすべてポストスクリーニングを経て，三つの候補が専門家によって選択され，うまく合成された． 図 2 は，我々の転移学習モデルからの予測値と比較した，3 つの候補のλを示す．

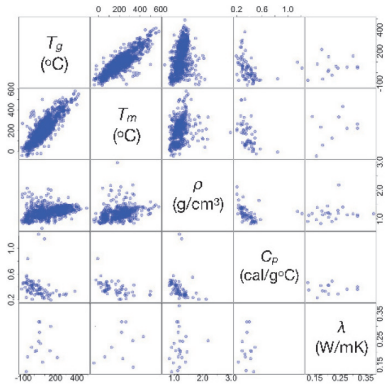


図 1．PoLyInfo 材料物性データの散布図
（ρ: 密度, Cp: 定圧比熱）

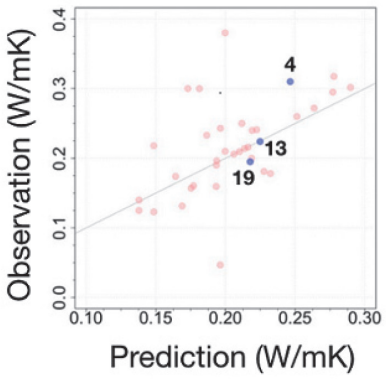


図 2．PoLyInfo のλ観測値と転移学習モデルからの予測値との比較（赤い点：訓練データ，青い点：新しく合成したポリマー）

参 考 文 献

Anderson, D.R. (1966). Thermal Conductivity of Polymers, Chemical Reviews, 66(6), 677-690.

Bohacek, R.S., McMartin, C. and Guida, W.C. (1996). The art and practice of structure-based drug design: A molecular modeling perspective, Medicinal Research Reviews, 16(1), 3-50.

Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, Journal of computer-aided molecular design, 31(4), 379-391.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. and Yamazaki, M. (2011). PolyInfo: Polymer database for Polymeric Materials Design, 2011 International Conference on Emerging Intelligent Data and Web Technologies, 22-29.

Zhong, C., Yang, Q. and Wang, W. (2001). A group contribution model for the prediction of the thermal conductivity of polymer melts, Industrial & Engineering Chemistry Research, 40, 4151-4153.

大規模空間データのための空間可変係数モデリング

Spatially varying coefficient modeling for large dataset

データ科学研究系 村上 大輔 （Daisuke Murakami）

1. 背景

センサ技術の発展に伴い，地理空間データが急速に大規模化してきている．例えば土地被覆や気象関連の情報は高解像度の衛星画像として提供されており，世帯構成や産業構成のような社会経済統計は街区のような空間詳細な単位毎に整備されつつある．それら標本数 N の大きな地理空間データを柔軟にモデル化する方法が研究・実務で求められている．

Gaussian process (GP) は，地理空間データの背後にある空間過程をモデル化するために幅広く用いられてきた．GP の共分散が距離減衰関数に従うことを仮定することで，地点間の空間的な従属関係を柔軟にモデル化することができる．一方で，GP を推定するためには $N \times N$ 次元の共分散行列の逆行列を評価する必要があり，標本数 N が数百万あるいはそれ以上になりうる昨今，その計算負荷は実用上の課題となってきている．そのような中，GP を高速に推定するための近似手法が近年活発に議論されてきた．

本節では，GP に基づいた空間モデルの高速化の一環といて取り組んでいる Spatially varying coefficient (SVC) モデルの高速化に関する研究を紹介する．

2. SVC モデルとその高速化

SVC モデルとは，各回帰係数の背後に GP を仮定することで回帰係数を場所毎に推定しようというモデルである．例えば都市部と郊外部の違いなどを柔軟に捉えることができるなど，SVC モデルは実用上便利である．しかしながら，回帰係数毎の GP を推定しようという同モデルの計算量は極めて大きいことが知られている．そこで本研究では，大規模データへの適用を見据え，SVC モデルの推定を次の手順で高速化した：(i) 各 GP を主成分のみを残して低ランク近似する；(ii) サイズが N に依存する行列・ベクトルを予め処理する（内積をとる）ことで，N に依存しない計算量で評価できるように事後確率を書き直す；(iii) 同事後確率を逐次的に最大化していくことで回帰係数の空間分布を決める各 GP(低ランク) を推定する．手順 (i) は上述の逆行列の計算量を削減するための近似，手順 (ii) と (iii) は各 GP を特徴づけるパラメータの推定を高速化するための処理である．

以上で高速化した SVC モデルの計算時間を，ベイズ SVC モデル（提案手法はこれを近似）および地理的加重回帰モデル（従来手法）と比較した．計算はすべて統計ソフトウェア R 上で行った．計算時間の比較結果は図 1 に示すとおりである．同図より，ベイズ SVC モデルは計算負荷が極めて大きく，実用上難があることを確認した．従来手法の計算時間もまた標本数 N が大きくなるにしたがって急激に増加しており，大規模データのモデリングの観点からは課題が残されているとの示唆を得た．

対照的に，提案手法の計算時間は N に関して線形にしか増加しておらず，例え SVC 数 K が 8 の（8 個の GP を同時推定する）場合であってもその増加は極めて遅い．例えば $N = 5,000$

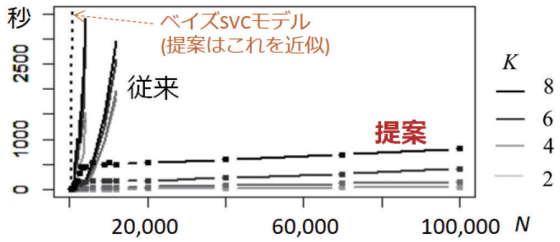


図 1．計算時間の比較（N は標本数，K は SVC の数）．

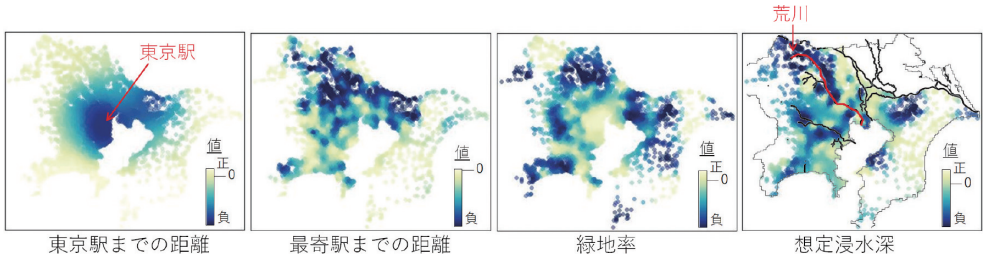


図 2．推定された場所毎の回帰係数

の場合の提案モデルの推定時間は平均で 454 秒となり，従来モデルが適用困難な $N = 100,000$ の場合でも，その推定時間はわずか 836 秒（平均値）となった．提案手法を用いることで，計算時間が大幅に短縮されることを確認した．なお，多くの場合に提案手法の SVC 推定精度が従来手法を上回ることもまたモンテカルロシミュレーションにより確認している．

3. 住宅地価分析への応用

提案モデルを東京都市圏の住宅地公示地価 (2010) 年の分析に適用した．説明変数は東京駅までの距離，最寄駅までの距離，1km グリッド内の緑地率，想定浸水深である．各データは国土数値情報ダウンロードサービス（http://nlftp.mlit.go.jp/ksj/）で公開されている．

推定された回帰係数（SVC）を図 2 にプロットした．この図から東京駅までの距離は都心を中心とした広域的な影響パターンを持つことが確認された．対照的に，最寄り駅までの距離は郊外部での影響力が大きく，特に鉄道網が比較的疎な北部ではその影響力が強まっていることが確認できる．緑地率もまた郊外部で強い効果を持つが，その影響は負であり，緑地の多さは地価を低下させるという結果が得られた．これは緑地よりも都市施設の多い土地のほうが好まれる傾向があるためである可能性がある．なお，例外的に都心部と横浜市中心部では緑地率は正に有意となっており，それら地域では緑地が好まれているの示唆を得た．想定浸水深は荒川沿岸の地価をより強く低下させていると推定された．この結果は，荒川沿岸の地価が水害リスクが適切に反映された値付けになっており，水害リスクにより適応した都市パターンであることを意味するものである．以上の結果は直感に整合する．

オイラー標数法によるウィシャート行列の最大固有値分布の近似

The Euler characteristic method for approximating the distributions of the largest eigenvalues of Wishart matrices

数理・推論研究系 栗木 哲 (Satoshi Kuriki)

要 旨

実対称ランダム行列の最大固有値は，行列の 2 次形式として定義される確率場の最大値であるため，オイラー標数法（チューブ法）によってその近似分布を与えることができる．本稿では，ウィシャート行列の行列サイズと自由度パラメータが無限大に発散する状況でも，オイラー標数法近似が真の極限分布の裾確率を精確に近似することを確認する．

キーワード：チューブ法, オイラー標数法, ランダム行列, Tracy-Widom 分布

1. 最大固有値とエクスカーション集合のオイラー標数

$n \times n$ 実対称行列 A が自由度 N のウィシャート分布 $W_n(N, I_n)$ に従うとする．その最大固有値は，M をランク 1 の $n \times n$ 直交射影行列の全体とするとき

$$\lambda_1(A) = \max_{\|h\|=1} h^\top A h = \max_{U \in M} \mathrm{tr}(UA),$$

すなわち M を添字集合とする確率場 $\{\mathrm{tr}(UA)\}_{U \in M}$ の最大値である．これより，エクスカージョン集合 $M_x = \{U \in M \mid \mathrm{tr}(UA) \geq x\}$ が定義される．このランダム集合のオイラー標数 $\chi(M_x)$ の期待値が，最大固有値のオイラー標数法近似である（栗木, 2019）：

$$\Pr(\lambda_1(A) \geq x) \approx E\big[\chi(M_x)\big] \quad (x \text{ が大きいとき}).$$

$\widetilde{A}$ が自由度 N の $n \times n$ 複素ウィシャート分布に従う場合も，添字集合 $\widetilde{M}$ を適当にとることにより同じ議論ができる．ラゲール多項式を $L_n^{(\alpha)}(x) = x^{-\alpha} e^x (d/dx)^n (x^{n+\alpha} e^{-x})/n!$ とおく．

定理 1. (i) $A \sim W_n(N, I_n)$, $\alpha = N - n$.

$$E[\chi(M_x)] = \frac{\sqrt{\pi}(-1)^{n-1}(n-1)!}{2^{\frac{N+n-1}{2}} \Gamma(\frac{N}{2}) \Gamma(\frac{n}{2})} \int_x^\infty \lambda^{\frac{N-n-1}{2}} e^{-\frac{\lambda}{2}} L_{n-1}^{(\alpha)}(\lambda)\, d\lambda.$$

(ii) $\widetilde{A} \sim CW_n(N, I_n)$, $\alpha = N - n$.

$$E\big[\chi(\widetilde{M}_x)\big] = \frac{n!}{\Gamma(N)} \int_x^\infty \lambda^{N-n} e^{-\lambda} \Big\{ L_{n-1}^{(\alpha)}(\lambda) L_{n-1}^{(\alpha+1)}(\lambda) - L_n^{(\alpha)}(\lambda) L_{n-2}^{(\alpha+1)}(\lambda) \Big\} d\lambda.$$

(i) と同等な表現は，Kuriki and Takemura (2001, 2008)で与えられている．

2. エッジ極限

Marchenko-Pastur 則より，ウィシャート行列の固有値の極限分布は適当なスケーリングのもとで有限サポートをもつ．サポートの上界 μ_+ を拡大するスケーリング（エッジ極限）を行う．

$$x \mapsto s = \frac{x-\mu_+}{\sigma} \quad \text{ただし} \quad \frac{N}{n} \to \gamma, \quad \mu_+ = (1+\sqrt{\gamma})^2 n, \quad \sigma = (1+\sqrt{\gamma})\left(1+\frac{1}{\sqrt{\gamma}}\right)^{\frac{1}{3}} n^{\frac{1}{3}}.$$

ラゲール多項式の極限定理（Johnstone, 2001）によって次を得る．

定理 2. (i) $A \sim W_n(N, I_n)$, $N, n \to \infty$ s.t. $N/n \to \gamma$ のとき

$$E[\chi(M_x)]\Big|_{x=\mu_+ + \sigma s} \to \frac{1}{2}\int_s^\infty \mathrm{Ai}(x)\, dx.$$

ここで Ai は第一種エアリー関数．

(ii) $\widetilde{A} \sim CW_n(N, I_n)$, $N, n \to \infty$ s.t. $N/n \to \gamma$ のとき

$$E\big[\chi(\widetilde{M}_x)\big]\Big|_{x=\mu_+ + \sigma s} \to \int_s^\infty \big\{\mathrm{Ai}'(x)^2 - \mathrm{Ai}(x)^2\big\} dx.$$

図 1 より，オイラー標数法近似は，真の極限分布である Tracy-Widom 分布の上側裾確率を精確に近似することがわかる．実際それらの相対誤差は，$s \to \infty$ のとき

$$\Delta(s) \sim -2^{-5}\pi^{-\frac{1}{2}} s^{-\frac{9}{4}} e^{-\frac{2}{3}s^{3/2}} \quad \text{(実の場合)}, \qquad 2^{-10}\pi^{-1} s^{-\frac{9}{2}} e^{-\frac{4}{3}s^{3/2}} \quad \text{(複素の場合)}.$$

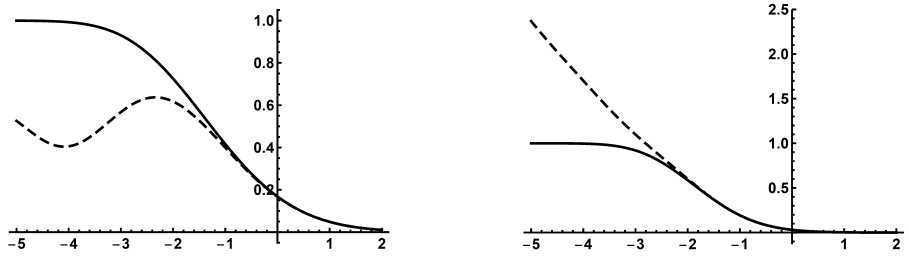


図 1．極限分布（Tracy-Widom 分布）の上側確率（実線）とそのオイラー標数法近似（破線）．
左：実ウィシャートの場合，右：複素ウィシャートの場合．

定理 2 で扱う極限では，オイラー標数法の確率場の添字集合の次元と体積が無限に発散するため，オイラー標数法が有効に働くかどうかは自明ではない．無限次元の添字集合を直接扱うことのできるオイラー標数法の開発は，これからの大きな研究課題である．

参 考 文 献

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis, *The Annals of Statistics*, **29** (2), 295–327.

栗木哲 (2019). チューブ法の理論・応用とその周辺, 「統計数理」, 投稿中.

Kuriki, S. and Takemura, A. (2001). Tail probabilities of the maxima of multilinear forms and their applications, *The Annals of Statistics*, **29** (2), 328–371.

Kuriki, S. and Takemura, A. (2008). Euler characteristic heuristic for approximating the distribution of the largest eigenvalue of an orthogonally invariant random matrix, *Journal of Statistical Planning and Inference*, **138** (11), 3357–3378.

変化点モデルのための AIC

AIC for change-point models

数理・推論研究系 二宮 嘉行 （Yoshiyuki Ninomiya）

要　　旨

変化点問題は長い間議論されているが，その理由の一つは，変化点モデルに通常の統計的漸近理論を満たさせない非正則性が存在することにある．本研究の目的は，そのような変化点モデルの AIC を導出することである．AIC の罰則項は最大対数尤度の漸近バイアスの二倍であり，正則条件を満たすモデルであればそれはパラメータの数の二倍となる．一方変化点モデルでは，その非正則性がゆえにそれは $2m+2p_m$ とならない．ここで m と p_m はそれぞれ変化点の数と変化点以外のパラメータの数である．本稿では，その漸近バイアスが $6m+2p_m$ で評価できることを示した Ninomiya (2015) の結果を紹介する．

キーワード：構造変化，情報量規準，非正則性，Brown 運動，ランダムウォーク

1. モデル

独立観測系列 $\{\boldsymbol{x}_i,\ 1\le i\le n\}$ に対し，m 個の変化点 $k^{(1)},\ldots,k^{(m)}$ をもつモデルを考える．簡単のため，$\boldsymbol{x}_i$ の分布は指数型分布族に属する，つまり確率関数は

$$k^{(j-1)}+1\le i\le k^{(j)}\quad\text{のとき}\quad \exp\{\boldsymbol{\theta}^{(j)\mathrm{T}}\boldsymbol{T}(\cdot)+S(\cdot)-A(\boldsymbol{\theta}^{(j)})\} \tag{1.1}$$

と書けるものとする（$1\le j\le m+1$）．ここで $k^{(0)}=0$ かつ $k^{(m+1)}=n$ としている．また，$\boldsymbol{\theta}^*=(\boldsymbol{\theta}^{*(1)\mathrm{T}},\ldots,\boldsymbol{\theta}^{*(m+1)\mathrm{T}})^{\mathrm{T}}$ と $\boldsymbol{k}^*=(k^{*(1)},\ldots,k^{*(m)})^{\mathrm{T}}$ を，$\boldsymbol{\theta}=(\boldsymbol{\theta}^{(1)\mathrm{T}},\ldots,\boldsymbol{\theta}^{(m+1)\mathrm{T}})^{\mathrm{T}}$ と $\boldsymbol{k}=(k^{(1)},\ldots,k^{(m)})^{\mathrm{T}}$ の真値とする．いま，

$$\boldsymbol{\theta}^{*(1)}\ne\boldsymbol{\theta}^{*(2)}\ne\cdots\ne\boldsymbol{\theta}^{*(m+1)}, \tag{1.2}$$

を仮定し，$\boldsymbol{\theta}^*$ と $\boldsymbol{k}^*$ は未知であるとする．加えて，$\boldsymbol{\theta}^{*(1)},\ldots,\boldsymbol{\theta}^{*(m+1)}$ は自然パラメータ空間に含まれるパラメータ集合の内点であり，そのパラメータ集合上で $\partial^2A(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}$ は正定値であるとする．さらに，後の漸近論のため，$1\le j\le m$ に対して $\lim_{n\to\infty}k^{*(j)}/n=\kappa^{(j)}$ であることを仮定する．ただし $0<\kappa^{(1)}<\cdots<\kappa^{(m)}<1$ としている．

2. 結果

$\hat{\boldsymbol{k}}_{\boldsymbol{x}}$ と $\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}$ を $\boldsymbol{k}^*$ と $\boldsymbol{\theta}^*$ の $\boldsymbol{x}=(\boldsymbol{x}_1^{\mathrm{T}},\ldots,\boldsymbol{x}_n^{\mathrm{T}})^{\mathrm{T}}$ に基づく最尤推定量とし，$f(\boldsymbol{x}|\boldsymbol{k}^*,\boldsymbol{\theta}^*)$ を $\boldsymbol{x}$ の同時確率関数とする．モデル選択は $f(\boldsymbol{y}|\boldsymbol{k}^*,\boldsymbol{\theta}^*)$ と $f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$ の間の Kullback-Leibler ダイバージェンスの二倍

$$2\mathrm{KL}\{f(\boldsymbol{y}|\boldsymbol{k}^*,\boldsymbol{\theta}^*),f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}=2\mathrm{E}_{\boldsymbol{y}}\{\log f(\boldsymbol{y}|\boldsymbol{k}^*,\boldsymbol{\theta}^*)\}-2\mathrm{E}_{\boldsymbol{y}}\{\log f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\},$$

を小さくしようとすること行うことができる．ここで，$\boldsymbol{y}$ は $\boldsymbol{x}$ のコピーであり，$\mathrm{E}_{\boldsymbol{y}}$ はその $\boldsymbol{y}$

に関する期待値を表すとする．右辺の第一項はモデルに依らないため，第二項を考えさえすればよい．そのシンプルな推定量は $-2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$ であるが，これは過小評価する．そこで，AIC 型の情報量規準では，それをバイアス補正した

$$\begin{aligned}&-2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})+2\mathrm{E}_{\boldsymbol{x}}[\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})-\mathrm{E}_{\boldsymbol{y}}\{\log f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}]\\ &=-2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})+2\mathrm{E}\Big[\sup_{(\boldsymbol{k},\boldsymbol{\theta})}L_{\boldsymbol{x}}(\boldsymbol{k},\boldsymbol{\theta})-L_{\boldsymbol{x}}\Big\{\operatorname*{argsup}_{(\boldsymbol{k},\boldsymbol{\theta})}L_{\boldsymbol{y}}(\boldsymbol{k},\boldsymbol{\theta})\Big\}\Big],\end{aligned}\tag{2.1}$$

を考える．ここで，E は $\boldsymbol{x}$ と $\boldsymbol{y}$ の両方での期待値を表し，また $L_{\boldsymbol{x}}(\boldsymbol{k},\boldsymbol{\theta})=\log f(\boldsymbol{x}|\boldsymbol{k},\boldsymbol{\theta})-\log f(\boldsymbol{x}|\boldsymbol{k}^*,\boldsymbol{\theta}^*)$ である．しかし，(2.1) における期待値は陽に求められないため，通常の AIC と同じようにその漸近評価を用いることを考える．つまり，(2.1) の代わりに

$$-2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})+2\mathrm{E}\{b(\boldsymbol{k}^*,\boldsymbol{\theta}^*)\}\tag{2.2}$$

を考える．ここで，$b(\boldsymbol{k}^*,\boldsymbol{\theta}^*)$ は $\sup_{(\boldsymbol{k},\boldsymbol{\theta})}L_{\boldsymbol{x}}(\boldsymbol{k},\boldsymbol{\theta})-L_{\boldsymbol{x}}\{\operatorname{argsup}_{(\boldsymbol{k},\boldsymbol{\theta})}L_{\boldsymbol{y}}(\boldsymbol{k},\boldsymbol{\theta})\}$ の弱極限であるとする．ただし，$\sup_{(\boldsymbol{k},\boldsymbol{\theta})}$ と $\operatorname{argsup}_{(\boldsymbol{k},\boldsymbol{\theta})}$ は $L_{\boldsymbol{x}}(\boldsymbol{k},\boldsymbol{\theta})$ が $\mathrm{O_P}(1)$ あるいは正の値となるような $(\boldsymbol{k},\boldsymbol{\theta})$ の集合上でとるものとする．

$A'(\boldsymbol{\theta}^{*(j)})$ を $\partial A(\boldsymbol{\theta}^{(j)})/\partial\boldsymbol{\theta}^{(j)}|_{\boldsymbol{\theta}^{(j)}=\boldsymbol{\theta}^{*(j)}}$，$B_1^{(j)}(\boldsymbol{\theta}^*)=A(\boldsymbol{\theta}^{*(j+1)})-A(\boldsymbol{\theta}^{*(j)})-(\boldsymbol{\theta}^{*(j+1)}-\boldsymbol{\theta}^{*(j)})^{\mathrm{T}}A'(\boldsymbol{\theta}^{*(j)})$，$B_2^{(j)}(\boldsymbol{\theta}^*)=A(\boldsymbol{\theta}^{*(j)})-A(\boldsymbol{\theta}^{*(j+1)})-(\boldsymbol{\theta}^{*(j)}-\boldsymbol{\theta}^{*(j+1)})^{\mathrm{T}}A'(\boldsymbol{\theta}^{*(j+1)})$ とし，$Q_{k,\boldsymbol{x}}^{(j)}$ を

$$\begin{aligned}&I_{\{k<k^{*(j)}\}}\sum_{i=k+1}^{k^{*(j)}}[(\boldsymbol{\theta}^{*(j+1)}-\boldsymbol{\theta}^{*(j)})^{\mathrm{T}}\{\boldsymbol{T}(\boldsymbol{x}_i)-A'(\boldsymbol{\theta}^{*(j)})\}-B_1^{(j)}(\boldsymbol{\theta}^*)]\\ &+I_{\{k>k^{*(j)}\}}\sum_{i=k^{*(j)}+1}^{k}[(\boldsymbol{\theta}^{*(j)}-\boldsymbol{\theta}^{*(j+1)})^{\mathrm{T}}\{\boldsymbol{T}(\boldsymbol{x}_i)-A'(\boldsymbol{\theta}^{*(j+1)})\}-B_2^{(j)}(\boldsymbol{\theta}^*)].\end{aligned}$$

とする．すると次の定理が得られる．

定理 1． $\boldsymbol{x}$ が (1.1) にしたがい，また (1.2) が成立すると，(2.2) における漸近バイアスは

$$\mathrm{E}\{b(\boldsymbol{k}^*,\boldsymbol{\theta}^*)\}=\sum_{j=1}^{m}\mathrm{E}\Big(\sup_k Q_{k,\boldsymbol{x}}^{(j)}+Q^{(j)}_{\operatorname{argsup}_k Q_{k,\boldsymbol{y}}^{(j)},\boldsymbol{x}}\Big)+p_m,\tag{2.3}$$

で与えられる．ここで p_m は $\boldsymbol{\theta}$ における異なるパラメータの数である．

情報量規準を使うメリットの一つはその使いやすさであることを鑑み，$1\le j\le m$ に対して

$$\boldsymbol{\theta}^{*(j+1)}-\boldsymbol{\theta}^{*(j)}=\alpha_n^{-1/2}\Delta_{\boldsymbol{\theta}^*}^{(j)}\quad\text{かつ}\quad\mathrm{O}(1)\ne\alpha_n=\mathrm{o}(n)\tag{2.4}$$

なる近接条件を仮定する．ここで $\Delta_{\boldsymbol{\theta}^*}^{(j)}$ は定数ベクトルである．すると変化点推定量の漸近挙動が変わり，定理 1 の代わりに以下が得られる．

定理 2． (2.4) のもとで (2.2) における漸近バイアスは

$$\mathrm{E}\{b(\boldsymbol{k}^*,\boldsymbol{\theta}^*)\}=3m+p_m$$

で与えられる．

参 考 文 献

Ninomiya (2015). Change-point model selection via AIC. *Annals of the Institute of Statistical Mathematics* 67, 943–961.

抽出法と計算代数

Samplers and Computational Algebra

数理・推論研究系 間野 修平（Shuhei Mano）

要　　旨

Diaconis と Sturmfels は，Markov chain Monte Carlo について，Markov 連鎖の推移の基底である Markov 基底を導入し，それが十分統計量が定める多項式環のトーリックイデアルの Gröbner 基底により与えられることを示した．筆者は，微分作用素環の考察により，標本経路が目的の分布に従う Markov 連鎖を構成することで，直接抽出が可能であることを示した．

キーワード：代数統計, 抽出法, 計算代数, Gröbner 基底, 超幾何系

1. はじめに

近年の代数統計の研究は，Pistone と Wynn (1996) による実験計画法における母数の識別性に関する研究と，Diaconis と Sturmfels (1998) による Markov chain Monte Carlo (MCMC) における定常分布を目的の分布とする Markov 連鎖の推移の基底である Markov 基底の導入が源流とされる．後者の周辺の発展については Aoki et al. (2012) を参照．MCMC の長所は正規化定数の計算を要しないこと，短所は定常分布からの抽出を保証し難いことで，正規化定数を効率的に計算できるなら MCMC を使う必要はない．筆者は，微分作用素環の考察により，標本経路が目的とする分布に従う Markov 連鎖を構成し，推移確率に現れる正規化定数を計算すれば，目的の分布からの直接抽出が可能であることを示した (Mano 2017)．本稿では，その概略を紹介する．詳細と参考文献については Mano (2018) を参照されたい．

2. Markov 基底

カウントベクトル $c \in \mathbb{N}_0^m$ が従う分布からの MCMC における Markov 連鎖の状態空間は，十分統計量 $b \in \mathbb{C}^d$ について $\mathcal{F}_b(A) := \{c; Ac = b, c \in \mathbb{N}_0^m\}$ と表せる．A は階数 d の非負整数値 $d \times m$ 配置行列 (行空間に $(1, ..., 1)$ を含む) とする．集合 $\mathcal{M}(A) = \mathrm{Ker}A \cap \mathbb{Z}^m$ を A に関する移動という．Markov 基底 $\mathcal{B} \subset \mathcal{M}(A)$ は $\mathcal{F}_b(A)$ に既約な Markov 連鎖を与える移動の集合である．任意の移動 z は $z = z^+ - z^-$, $z_i^+ := \max\{z_i, 0\}$, $z_i^- := \max\{-z_i, 0\}$ と表され，多項式環の二項式と 1 対 1 に対応づけられる．特に，$I_A := \langle \{x^{z^+} - x^{z^-}; z \in \mathcal{M}(A)\} \rangle$ はトーリックイデアルである．ここで，$x^z := \prod_{i=1}^m x_i^{z_i}$ とした．

定理 1（Diaconis, Sturmfels 1998）．$\mathcal{B} = \{z_i; i \in \{1, ..., s\}\} \subset \mathcal{M}(A)$ が Markov 基底であることは，$\{x^{z_j^+} - x^{z_j^-}; j \in \{1, ..., s\}\}$ が I_A の生成系であることの必要十分条件である．

I_A の Gröbner 基底は生成系だから Markov 基底である．Gröbner 基底を求める一般的なアルゴリズム (Buchberger 1976) があるので，原理的には任意の A に対し Markov 基底が得られる．

3. A 超幾何系

定義 1（Gelfand et al. 1990）. 2 節で導入した行列 A，十分統計量 b に対し，次の消去作用素が定める線形偏微分方程式系を A 超幾何系 $H_A(b)$ とよぶ.

$$\sum_{j=1}^{m} a_{ij} x_j \frac{\partial}{\partial x_j} - b_i, \qquad i \in \{1, ..., d\}, \qquad \partial^{c^+} - \partial^{c^-}, \qquad c \in \mathrm{Ker} A \cap \mathbb{Z}^m.$$

$H_A(b)$ は微分作用素環の左イデアルで，A 超幾何イデアルとよばれる.

A 超幾何系 $H_A(b)$ の原点周りの級数解

$$Z_A(b; x) := \sum_{c \in \mathcal{F}_b(A)} \frac{x^c}{c!}, \qquad c! := \prod_{i=1}^{m} c_i!$$

を A 超幾何級数とよぶ．ただし，$b \notin A\mathbb{N}_0^m$ のときは $Z_A(b; x) = 0$ と規約する．m 個の対数アフィンモデルからの長さ n の多項抽出によるカウントベクトル $c \in \mathbb{N}_0^m$ の分布は離散指数型分布族であり，十分統計量 $b \in \mathbb{N}_0^d$ で $Ac = b$ と条件づけた分布の確率函数は $x^c/c!$ に比例する．正規化定数が A 超幾何多項式になるので，Takayama et al. (2018) は A 超幾何分布とよんだ．周辺度数が所与の分割表の分布など，カウントデータ解析における典型的な統計モデルを含む.

4. 直接抽出法

A 超幾何多項式の性質から，A の第 i 列ベクトルを a_i として，

$$p_A(b; i) := \frac{\mathbb{E}(C_i | AC = b)}{n} = \frac{Z_A(b - a_i; x)}{Z_A(b; x)} \frac{x_i}{n}, \qquad \sum_{i=1}^{m} p_A(b; i) = 1$$

が従う．$p_A(b; i)$ を，A 超幾何多項式を状態空間とし，推移ごとに次数が 1 下がる Markov 連鎖において，$Z_A(b; x)$ から $Z_A(b - a_i; x)$ への推移確率とみなすと，次のアルゴリズムが得られる.

アルゴリズム 1（Mano 2017）. A 超幾何分布からの逐次直接抽出.

1. $t_1 = j$ を確率 $p_A(b; j)$ で抽出.
2. $i = 2, ..., n$ について，$t_i = j$ を確率 $p_A(b - (a_{t_1} + \cdots + a_{t_{i-1}}); j)$ で抽出.

A 超幾何多項式は $H_A(b)$ の標準単項式のベクトル Q が従う Pfaffian 系 $\partial_i Q = P_i Q$ を用いて求める．有理函数の行列 P_i, $i \in \{1, ..., m\}$ は標準単項式の Gröbner 基底による標準形として原理的には任意の A に対して得られる (Saito et al. 2010; 日比ら 2011).

5. おわりに

様々な統計モデルについて MCMC による抽出を直接抽出に置き換えていくことは，代数統計の新しい挑戦のひとつになると期待される.

参 考 文 献

Mano, S.: Partition structure and the A-hypergeometric distribution associated with the rational normal curve. Electron. J. Statist. **11**, 4452–4487 (2017)

Mano, S.: *Partitions, Hypergeometric Systems, and Dirichlet Processes in Statistics.* JSS Research Series in Statistics, SpringerBriefs in Statistics (2018)

角度の観測を含むデータのための統計モデル

Statistical Models for Data Which Include Angular Ones

数理・推論研究系　加藤 昇吾（Shogo Kato）

1. はじめに

様々な学問分野において，角度として表される観測値が得られることがある．例えば，気象学における風向の観測はその一例である．風向は，西を $-\pi$ とし，反時計回りを正の向きとすれば，南を $-\pi/2$，東を 0，北を $\pi/2$ のように角度で表すことができる．つまり，任意の風向は $-\pi$ 以上 π 未満の角度 θ，もしくは円周上の点 $(\cos\theta, \sin\theta)$，として表現できる．他にも，医学・地震学・動物行動学・生命情報学など，多くの分野で角度の観測が存在している．

角度の観測を含むデータには，統計解析をする上で大きな問題がある．それは，このようなデータを解析する上では，統計学が主に対象としている実数値データのための解析手法をそのまま使うことができないという問題である．例えば，平均や分散などの要約統計量や，実数値データのための確率分布・回帰モデル・時系列モデルなどの統計モデルは，角度データに直接応用すると，しばしば不自然な解析結果を与えることにつながってしまう．この問題を解決するため，角度の観測を含むデータの統計的手法を考えることが統計学における重要な研究テーマとなっている．本報告では，このテーマにおける研究の背景と著者の研究結果を概観する．

2. 角度データのための確率分布

方向統計学における中心的なアプローチは，データに何らかの確率分布を仮定したパラメトリックな統計解析法である．著者は，角度データのための確率分布（円周上の確率分布）の 1 つとして知られる「円周上のコーシー分布」に着目し，これに関連したパラメトリックな統計モデルを研究してきた．円周上のコーシー分布は確率密度関数

$$f(\theta) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}, \quad -\pi \le \theta < \pi; \quad -\pi \le \mu < \pi,\ 0 \le \rho < 1,$$

で定義される確率分布である．ここで，μ は分布の位置を定めるパラメータ，ρ は分布の集中度を調節するパラメータである．

方向統計学においては，円周上のコーシー分布は古くから知られていたもののほとんど注目されていない分布であった．この分野で長きにわたり研究の中心だったのは，「フォン・ミーゼス分布」とよばれる円周上の確率分布およびそれを応用したパラメトリックな統計手法である．フォン・ミーゼス分布は，実数上の正規分布からの自然なアナロジーにより導かれることから，「円周上の正規分布」とよばれることもある．一方，正規分布が持ついくつかの扱いやすい性質が成立しない問題点も指摘されており，そのことがフォン・ミーゼス分布およびそれを応用した統計手法の理論的性質を導くことを困難にしている一面もあった．そのような中，McCullagh (1996) などによって開拓された円周上のコーシー分布に著者は興味を持ち，この分布に関連した統計モデルの研究を行うようになった．

3. 角度の観測を含むデータのための統計モデル

著者は，円周上のコーシー分布に関連した統計モデルの論文を今までに複数篇執筆してきたが，ここでは，その中から 2 篇についてある程度詳しく説明し，4 篇について簡潔に紹介する．

Kato (2010) では，円周上のコーシー分布を誤差分布として用いた新たな円周上のマルコフ過程を提案した．この研究は Fisher and Lee (1994) によるマルコフ過程と関連がある．彼らは，回帰曲線としてメビウス変換，誤差分布としてフォン・ミーゼス分布を仮定した円周上のマルコフ過程を提案した．それに対し，Kato (2010) では，誤差分布としてフォン・ミーゼス分布の代わりに円周上のコーシー分布を用いることにより，Fisher and Lee (1994) のモデルでは得られなかった多くの挙動に関する性質を得ることに成功した．例えば，任意時点 t の角度を所与としたときの $t+h$ 時点における角度の条件付分布が円周上のコーシー分布となることや，その条件付分布のパラメータも複素数を用いれば簡潔に表現することができることを示した（t と h は自然数をあらわす）．

Kato and Pewsey (2015) では，円周上のコーシー分布の拡張として，2 次元トーラス $[-\pi,\pi)^2$ 上の分布を提案した．2 次元トーラス上の分布の既存研究としては，フォン・ミーゼス分布を拡張した 2 変量フォン・ミーゼス分布がよく知られている．この分布は，条件付分布がフォン・ミーゼス分布となる利点があるが，正規化定数が特殊関数の無限和で表されることや，パラメータの解釈が困難であること，周辺分布がよく知れていない分布になること，などの問題点がある．それに対して，Kato and Pewsey (2015) で提案した 2 次元トーラス上の分布ではこれらの問題がすべて解決されたモデルとなっている．

その他の研究結果について簡潔に紹介する．Kato *et al.* (2008) では説明変数・被説明変数が共に角度となる回帰モデルを提案し，誤差分布として円周上のコーシー分布を用いることで回帰モデルのいくつかの扱いやすい性質を得られることを示した．Kato and Jones (2010, 2015) では，円周上のコーシー分布を特別な場合として含む柔軟な円周上の確率分布を提案した．Kato and McCullagh (2018) では，円周上のコーシー分布を球面上の分布へと拡張し，メビウス変換に関して閉じていることやパラメータ推定が容易にできることを示した．

参 考 文 献

Fisher, N. I. and Lee, A. J. (1994). Time series analysis of directional data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **56**, 327–339.

Kato, S. (2010). A Markov process for circular data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 655–672.

Kato, S. and Jones, M. C. (2010). A family of distributions on the circle with links to, and applications arising from, Möbius transformation, *Journal of the American Statistical Association*, **102**, 249–262.

Kato, S. and Jones, M. C. (2015). A tractable and interpretable four-parameter family of unimodal distributions on the circle, *Biometrika*, **102**, 181–190.

Kato, S. and McCullagh, P. (2018). Möbius transformation and a Cauchy family on the sphere, *arXiv:1510.07679v2*, 1–30.

Kato, S. and Pewsey, A. (2015). A Möbius transformation-induced distribution on the torus, *Biometrika*, **102**, 359–370.

Kato, S., Shimizu, K. and Shieh, G. S. (2008). A circular–circular regression model, *Statistica Sinica*, **18**, 633–645.

McCullagh, P. (1996). Möbius transformation and Cauchy parameter estimation, *The Annals of Statistics*, **24**, 787–808.

極値分布の吸引領域と離散性

Maximum domain of attraction and Discreteness

数理・推論研究系 志村 隆彰 （Takaaki Shimura）

キーワード：極値理論, 極値分布, 確率分布の裾, 最大値吸引領域, 離散化.

近年，東日本大震災のような巨大地震や西日本豪雨のような豪雨など，極端に大きな被害をもたらす自然災害（激甚災害）が続き，大きな社会問題になっている．このような災害をランダムな現象とみなせば，“起こる頻度（確率）は少ないが，一旦起こったときの被害（量）が極端に大きい”ことが特徴である．このような“稀に起こる極端な事象とその起こり方”を研究するのが極値理論 (Extreme Value Theory) である．極値理論では，極端な事象の起こり方を表す確率分布 F の裾（確率）$\bar{F}(x) \equiv 1 - F(x)$ （x より大きい値が起こる確率を表す）の挙動が重要で，x が F の上限 $x_F \equiv \sup\{x : F(x) < 1\}$ に近づくときの 0 への収束が，正規分布のように収束が速いとき裾が軽い，コーシー分布のように遅いとき裾が重いという（heavy tail は最近広く知られるようになってきている）．

確率統計でもっとも重要な定理は中心極限定理であるが，極値理論でこれに当たるのが多数の中の最大値の振る舞いを表す次の定理である．

定理０． $X_1, X_2, \ldots$ を共通の確率分布 F に従う実数値独立確率変数列とし，X_n までの最大値を $M_n \equiv \max\{X_1, \ldots, X_n\}$ とする．このとき，共通分布 F が適当な条件を満たせば，非退化分布 G が存在して，適当な定数列 $a_n > 0$ と $b_n \in \mathbf{R}$ により，

$$\mathcal{L}\left(\frac{M_n - b_n}{a_n}\right) \to G \quad (n \to \infty)$$

となる ($\mathcal{L}()$ はカッコ内の確率変数の分布を，$\to$ は分布の収束を表す)．このときの極限分布 G は極値分布と呼ばれ，３種類ある（フレシェ分布，グンベル分布，（極値）ワイブル分布）．

各極値分布に対して，上記の収束が成り立つような確率変数列の共通分布 F の集合をその極値分布の（最大値）吸引領域といい，分布の裾 $\bar{F}(x)$ が分布の上限に近づくときの漸近挙動で特徴づけられる．本稿では，F として上限が無限の分布を扱うので，極値分布はフレシェ分布あるいはグンベル分布になる．裾の重さに注目すると，パレート分布のような裾の挙動がべきオーダーで重い（正確には裾が正則変動する）ものはフレシェ分布の吸引領域に，指数分布などの裾がより軽いものはグンベル分布の吸引領域に属する．また，分布の連続性の観点からは，正規分布，指数分布，コーシー分布など様々な裾挙動を持つ連続分布がいずれかの極値分布の吸引領域に属する半面，幾何分布やポアソン分布などの裾が重くない離散分布はどの極値分布の吸引領域にもはいらない（指数分布と幾何分布は連続分布か離散分布かという違いであることに注意）．この事実は，吸引領域への属性と離散性との相性が悪いことを示唆している．分布が吸引領域に入らないということは，理論的に，それに従う独立同分布列の最大値の挙動がとらえきれないだけでなく，応用上も，実データが必然的に離散であるから，統計解析の精度に影響することを意味している．

この背景の元，最大値吸引領域と分布の離散性の関係を考察した結果を述べる．実数上の分

布に対して，区間 $(n-1,n]$ の確率を一点 $\{n\}$ に集中させて離散分布（整数値分布）を作る操作を分布の”離散化”と呼び，離散分布に対して，連続分布でその離散化が元の離散分布と一致するものを離散分布の”連続化”と呼ぶ（連続化は一意ではない）．幾何分布は指数分布の離散化，幾何分布の連続化のひとつが指数分布である．このような場合，指数分布は離散化で吸引領域への属性を失い，幾何分布は連続化で属性が回復可能ということにする．

問題は，第一に，吸引領域に属する分布で，離散化によりその属性を保つか，失うかの判別条件を与えること．第二に，連続化で属性が回復可能な分布の特徴付けである．

定理 A. 吸引領域に属する分布が，離散化により，吸引領域に留まるための必要十分条件は，長い裾を持つことである：$\lim_{x\to\infty}\bar{F}(x+1)/\bar{F}(x)=1$ (long-tailed distribution).

定理 B. 分布 F が吸引領域に回復可能であるための必要十分条件は，そのフォン・ミーゼス関数 $\bar{F}_0$ ($\bar{F}$ を滑らかにしたもの）が次を満たすことである．

$$\lim_{n\to\infty}\left(\log\frac{\bar{F}_0(n+1)}{\bar{F}_0(n+2)}\right)^{-1}-\left(\log\frac{\bar{F}_0(n)}{\bar{F}_0(n+1)}\right)^{-1}=0.$$

定理 A から，パレート分布や対数正規分布は離散化で属性を失わないこと，定理 B から幾何分布の他に，ポアソン分布も回復可能であることがわかる（定理 B の吸引領域は実はグンベル分布の吸引領域）．

次に，幾何的分布と呼ばれる離散分布に対して，吸引領域に属する連続化を与える具体的方法を与える．離散分布 F_1 が幾何的分布であるとは，ある正の指数 γ が存在して，$\lim_{n\to\infty}\bar{F}_1(n+1)/\bar{F}_1(n)=e^{-\gamma}$ となるときをいう．また，F が指数的分布であるとは，ある正の指数 γ が存在して，$\lim_{x\to\infty}\bar{F}(x+k)/\bar{F}(x)=e^{-\gamma k}$ となるときをいい，その全体を $\mathcal{L}(\gamma)$ であらわす．$\mathcal{L}(\gamma)$ はグンベル分布の吸引領域に含まれることが知られている．

定理 C. F_1 を指数 $\gamma(>0)$ の幾何的分布，G を $[0,1]$ 上の分布とする．このとき，$F_1*G\in\mathcal{L}(\gamma)$ となるための必要十分条件は，

$$G(x)=\frac{1-e^{-\gamma x}}{1-e^{-\gamma}}\quad(0\le x\le 1)$$

である．

幾何的分布，指数的分布はそれぞれ幾何分布，指数分布の一般化であり，ガンマ分布などが指数的分布である．この定理は，幾何的分布は指数が共通である限り，同じ $[0,1]$ 上の分布を（確率変数の意味で独立に）足すことで，吸引領域へ回復可能であることを主張している．たとえば，幾何分布に従う実データは定理 0 の極限定理は成り立たないが，データにランダムな補正をすることで成り立つようにできるのである．

参 考 文 献

Shimura, T. (2012). Discretization of distributions in the maximum domain of attraction, Extremes **15**, 299-317.

Shimura, T. (2012). Limit distribution of a roundoff error, Statistics and Probability Letters, **82**, 713-719.

高橋 倫也・志村 隆彰. (2016). 「極値統計学」(ISM シリーズ 進化する統計数理 5), 近代科学社.

拡散過程の非同期・ノイズ付観測データに対する最尤型推定法

Maximum likelihood type estimation for diffusion processes with noisy, nonsynchronous observations

数理・推論研究系 荻原 哲平（Teppei Ogihara）*

近年株式市場における一日内の全取引の情報を記録した「高頻度データ」の利用可能性が高まり，従来のデータに比べて膨大な情報量をもつため，金融市場のミクロ構造のさらなる解明が期待され，このようなデータを用いたデータ解析手法が活発に研究されている．通常，日内において株価が観測されるのは株価の約定時であるため, 異なる株式に対して観測時刻が一致していないという「非同期観測」の問題が必然的に生じ，特にデータの共変動（共分散）の推定が困難になる．データの線形補完や直前データを用いた補完などによるシンプルな「同期化」を行ったデータに対する共変動推定量には深刻なバイアスが存在することが知られている．

確率空間 $(\Omega, \mathcal{F}, P)$ 上の二次元確率過程 $X = \{X_t\}_{0\le t\le T}$ が以下を満たすとする：

$$dX_t = \mu(t, X_t, \sigma_*)dt + b(t, X_t, \sigma_*)dW_t, \quad t \in [0, T].$$

ここで, σ_* はモデルの d 次元パラメータ, $\{W_t\}_{0\le t\le T}$: 二次元標準ブラウン運動, μ は $\mathbb{R}^2$ 値，b は $\mathbb{R}^2 \otimes \mathbb{R}^2$ 値既知関数とする．X の成分を X^1, X^2 と書き, X^1, X^2 の観測時刻をそれぞれランダム時刻 $\{S_i^n\}_{i=0}^{\ell_{1,n}}, \{T_j^n\}_{j=0}^{\ell_{2,n}} \subset [0, T]$ で表す．データが高頻度観測になる極限を扱うため，$n \to \infty$ の時 $\max_{i,j} |S_i^n - S_{i-1}^n| \vee |T_j^n - T_{j-1}^n| \to^p 0$ を仮定する．この時観測データ $\{S_i^n\}_i, \{T_j^n\}_j, \{X^1_{S_i^n}\}_i, \{X^2_{T_j^n}\}_j$ からモデル・パラメータ σ_* を推定する問題を考える．

非同期観測モデルに対する最尤型推定法

Ogihara and Yoshida (2014) において，σ_* に対する疑似対数尤度関数を用いた最尤型推定法が提案されている．$\mu \equiv 0$ かつ b, X_0 と $\{S_i^n\}_i, \{T_j^n\}_j$ が非ランダムの時，$\{X^1_{S_i^n}\}_i, \{X^2_{T_j^n}\}_j$ は多変量正規分布に従い，共分散関数を計算することが可能であるため，対数尤度関数を二次形式で与えることができる．一般の場合でも同様に疑似対数尤度関数 $H_n(\sigma)$ を近似的に構成可能であり，$H_n(\sigma)$ を最大にするパラメータの値として最尤型推定量 $\hat{\sigma}_n$ が定義される．このように定義された $\hat{\sigma}_n$ に対して，μ, b の微分可能性や非退化性等の条件と，観測時刻列 $\{S_i^n\}_i, \{T_j^n\}_j$ の漸近挙動に関する条件の下，漸近混合正規性：

$$\sqrt{n}(\hat{\sigma}_n - \sigma_*) \to^{s\text{-}\mathcal{L}} \Gamma^{-1/2}\mathcal{N} \quad \text{as} \quad n \to \infty \tag{0.1}$$

が示される．ただし，$\to^{s\text{-}\mathcal{L}}$ は stable convergence を表し，∂_σ を σ に関する微分，$\Gamma = \text{P-}\lim_{n\to\infty}(-\partial_\sigma^2 H_n(\sigma_*)/n)$, $\mathcal{N}$ を $(\Omega, \mathcal{F}, P)$ のある拡張上で定義された，$\mathcal{F}$ と独立に $N(0, I_d)$ に従う確率変数とする．ここで I_d は d 次元単位行列である．

* 東京大学 数理・情報教育研究センター：〒113-8656 文京区弥生 2-11-16

非同期観測モデルにおける局所漸近混合正規性

Jeganathan (1983) では，統計モデル $\{P_{\theta,n}\}$ の局所漸近混合正規性の下，任意の推定量の漸近分散の下界を与える minimax 不等式が示された．この下界を達成する推定量は漸近有効推定量と呼ばれる．拡散過程の高頻度観測モデルに対しては，局所漸近混合正規性を示すには拡散過程の推移確率密度関数の漸近挙動を解析する必要があり，非同期性のない規則的な観測モデルに対して Gobet (2001) では Malliavin 解析の技術を用いて局所漸近混合正規性を示した．このモデルにおいては Genon-Catalot and Jacod (1993) の最尤型推定量が漸近有効となる．Ogihara (2015) では，Gobet (2001) の Malliavin 解析を用いた手法を応用し，拡散過程の非同期観測モデルに対する局所漸近混合正規性を証明し，最尤型推定量 $\hat{\sigma}_n$ が漸近有効であることを示した．

非同期・ノイズ付観測モデルに対する最尤型推定法

高頻度データの解析上の問題点として，非同期観測に加えて，拡散過程によるモデリングにおける仮想的な観測ノイズの存在が実証研究から示唆されている．このような観測ノイズは「マーケット・マイクロストラクチャー・ノイズ」と呼ばれる．

観測ノイズ $\{\epsilon_i^k\}_{k=1,2,i\in\mathbb{Z}_+}$ を $(X_t,W_t)_{0\le t\le T}$ と独立で，ある正定数 v_1,v_2 に対して，$E[\epsilon_i^k]=0$，$E[\epsilon_i^k\epsilon_j^l]=v_k\delta_{kl}\delta_{ij}$ を満たす確率変数として，観測が $\{X_{S_i^n}^1+\epsilon_i^1\}_{i=0}^{\ell_{1,n}},\{X_{T_j^n}^2+\epsilon_j^2\}_{j=0}^{\ell_{2,n}}$ で与えられるような統計モデルとして非同期・ノイズ付観測がモデル化される．

この時 Ogihara (2018) において，モデル・パラメータ σ_* の最尤型推定量が以下のように構築されている．まず L_n を正整数列で，$L_n\to\infty$, $(L_n/\sqrt{n})\vee(n^{1/4}/L_n)\to 0$ を満たすものとし，観測区間全体 $[0,T]$ を L_n 個の区間に等分する．等分された各区間において，ϵ_i^k が正規分布に従うと仮定すれば拡散過程 X が局所的に条件付正規分布で近似されることを用いて対数尤度関数を正規近似し，各区間の近似対数尤度関数を足し合わせることで疑似対数尤度関数 $\mathcal{H}_n(\sigma)$ が構築される．最尤型推定量は $\check{\sigma}_n=\mathrm{argmax}_\sigma\mathcal{H}_n(\sigma)$ と計算される．

Ogihara (2018) において，μ,b の微分可能性や非退化性等の条件と，観測時刻列 $\{S_i^n\}_i,\{T_j^n\}_j$ の漸近挙動に関する条件の下，漸近混合正規性：

$$n^{1/4}(\check{\sigma}_n-\sigma_*)\to^{s\text{-}\mathcal{L}}\tilde{\Gamma}^{-1/2}\mathcal{N} \tag{0.2}$$

が示された．ただし，$\tilde{\Gamma}=\text{P-}\lim_{n\to\infty}(-\partial_\sigma^2\mathcal{H}_n(\sigma_*)/\sqrt{n})$ である．特にこの結果は ϵ_i^k が正規分布に従わない場合でも成立する．

参 考 文 献

Genon-Catalot, V., Jacod, J. (1993): On the estimation of the diffusion coefficient for multi-dimensional diffusion processes, Annals of Institute of Henri Poincare, 29, 119-151.

Gobet, E. (2001) Local asymptotic mixed normality property for elliptic diffusion : a Malliavin calculus approach. Bernoulli, 7, 899-912.

Jeganathan, P. (1983) Some asymptotic properties of risk functions when the limit of the experiment is mixed normal. Sankhya Ser. A, 45, 66-87.

Ogihara, T. (2015): Local asymptotic mixed normality property for nonsynchronously observed diffusion processes, Bernoulli, 21, 2024-2072.

Ogihara, T. (2018): Parametric Inference for Nonsynchronously Observed Diffusion Processes in the Presence of Market Microstructure Noise, Bernoulli, 24, 3318–3383.

Ogihara, T. and Yoshida, N. (2014): Quasi-likelihood analysis for nonsynchronously observed diffusion processes, Stochastic Processes and their Applications, 124, 2954-3008.

一般化平均による統計モデル

A statistical model via generalized mean

数理・推論研究系 江口 真透（Shinto Eguchi），

1. 一般化平均と情報幾何

1930 年に，コルモゴロフと南雲によって独立に発表された一般化平均ついて考察する．正の数 x, y に対する一般化平均は

$$\mathrm{GM}_\phi(x,y) = \phi\big((1-\pi)\phi^{-1}(x) + \pi\phi^{-1}(y)\big) \tag{1.1}$$

と定まる．ここで $\phi : \mathbb{R} \to (0,\infty)$ は単調関数とする．同様に x, y, z の一般化平均を考えると，定義から $\mathrm{GM}_\phi(x, \mathrm{GM}(y,z)) = \mathrm{GM}_\phi(\mathrm{GM}(x,y), z)$ がいえる．典型例の算術平均，幾何平均，調和平均などが含まれるが，‘平均’の持つべき公理から特徴付けられた一般化平均は生成関数 ϕ の関数自由度を持つので多様な平均の考えが展開できる．

統計学のために一般化平均を積極的に援用した方法を提案した．はじめに情報幾何との関連について考察し，つぎに統計モデリングのために幾つかの応用を紹介する．ロジスティック回帰や比例ハザードモデルにおける予測関数を一般化平均を使った準線形モデルを提案した．さらにクラスタリングのエネルギ関数やロス関数の混合のために一般化平均がキーになることが示された．

情報幾何は確率密度の関数空間の上に双対リーマン幾何をベースに豊かな直観を与え，確率に関連する全ての分野へ幾何的考えを構築している．その中の基本定理として，この関数空間の上でのピタゴラス定理が挙げられる (Amari-Nagaoka, 2007)．確率密度関数 $p(x)$ と $q(x)$ を混合測地線でつなぎ，一方で，$r(x)$ と $q(x)$ を指数測地線でつないだとき，この 2 つの測地線が $q(x)$ で直交するならば，またその時に限り，

$$D_0(p,r) = D_0(p,q) + D_0(q,r) \tag{1.2}$$

が成立する．ここで D_0 は KL ダンバージェンスとする．この性質から最尤推定と十分統計量の関係，赤池情報量規準の妥当性などが導かれる．この考察において $r(x)$ と $q(x)$ をつなぐ指数測地線とは

$$\mathrm{EG}(q,r) = \exp\big((1-\pi)\log q(x) + \pi \log r(x) - \kappa(\pi)\big) \tag{1.3}$$

と定められる．ここで $\kappa(\pi)$ は正規化定数とする．このように $\mathrm{EG}(q,r)$ は生成関数 $\phi = \exp$ を使って正の数の代わりに密度関数に対する一般化平均と見れる．したがって，一般の ϕ に

$$\mathrm{EG}_\phi(q,r) = \phi\big((1-\pi)\phi^{-1}q(x) + \pi\phi^{-1}r(x) - \kappa_\phi(\pi)\big) \tag{1.4}$$

が定まり，これを一般化指数測地線と呼ぶ (Eguchi-Komori, 2015)．同様な考えから一般化 KL ダイバージェンスを導出すると，ピタゴラス定理が示される．次の節では一般化平均を直接に統計モデリングに応用することを考察する．

2. 準線形ロジスティック回帰モデル

一般化平均 (1.1) は正の数の平均で在ったが，実数 x と y に対しては

$$\mathrm{RGM}_\phi(x,y) = \phi^{-1}\big((1-\pi)\phi(x) + \pi\phi(y)\big) \tag{2.1}$$

と定める．(1.1) の生成関数 ϕ の代わりに ϕ^{-1} を取られていることに注意する．統計の応用としては実数の代わりに，回帰関数，予測関数，エネルギー関数，ロス関数などの実数値関数の一般化平均を考えることができる．例えば，ロジスティック回帰において p 変数の説明変数 $X = x$ を与えたとき 2 値反応変数 y の条件付き確率関数を

$$p(y|x) = \frac{\exp\{yf(x)\}}{1+\exp\{f(x)\}} \qquad (y = 0, 1) \tag{2.2}$$

とする．予測関数を一般化平均によって

$$f_\tau(x,\beta,\pi) = \frac{1}{\tau}\log\big(\textstyle\sum_{k=1}^K \pi_k \exp(\tau\beta_k^\top x_k)\big) \tag{2.3}$$

とし，準線形予測関数と呼ぶ．ここで τ は逆温度パラメータ，$x = (x_1, \cdots, x_K)$ と $\beta = (\beta_1, \cdots, \beta_K)$ は p 変数の同じ K 分割とする．Cf. Omae et al. (2017). 逆温度パラメータ τ を極限 ∞ を取ると $f_\tau(x,\beta,\pi) = \max_{1\le k\le K} \beta_k^\top x_k$ となり，極限 $-\infty$ を取ると $f_\tau(x,\beta,\pi) = \min_{1\le k\le K} \beta_k^\top x_k$ となる．極限 0 を取ると線形予測関数に帰着される．

$Y = 0$ のときの X の条件分布 $p(x|Y=0)$ は正規分布 $N(\mu_0, \Sigma)$ に従っているが，$Y = 1$ のときの X の条件分布 $p(x|Y=1)$ は混合正規分布 $\sum_{k=1}^K \pi_k^* N(\mu_k, \Sigma)$ に従っていると仮定する．これは $Y = 0$ サンプルは均一な母集団から得られたが $Y = 1$ サンプルは非均一な異質な母集団の混合から得らた状況を考えている．このとき，$\tau = 1$ のとき

$$f_\tau(x,\beta,\pi) = \log\frac{p(x|Y=1)}{p(x|Y=0)} \tag{2.4}$$

が成立する．このような仮定のもとでは，準線形予測関数の判別の最適性が示される．

準線形予測関数の定義において説明変数 x の K 分割が必要であるが，これは教師なし学習によって構成できる．典型的にはクラスター分析によって K 分割が求まる．またスパース学習を準線形モデル (2.3) をロジスティック回帰 (2.2) に代入したモデルの対数尤度関数に (2.3) のパラメータ β_k の積の L_1 ペナルティを入れたものを考える方法も有力である．

比例ハザードモデル，分割クラスタリング，混合ロス関数，モデル平均，メタアナリシス，半教師学習などについても一般化平均を使うと興味深い展開ができると思われる．線形モデルから準線形モデルへの拡張や，エネルギー関数やロス関数の混合について，幾つかの新しい知見が得られた．広い実用のためには未だ多くの解決すべき問題があるが，線形モデルを柔軟に結合する統計方法として着実な完成が近い将来になされることが望まれる．

参考文献

Amari, S. I., & Nagaoka, H. (2007). Methods of information geometry (Vol. 191). American Mathematical Soc.

Eguchi, S., & Komori, O. (2015). Path connectedness on a space of probability density functions. In International Conference on Networked Geometric Science of Information (pp. 615-624). Springer, Cham.

Omae, K., Komori, O., & Eguchi, S. (2017). Quasi-linear score for capturing heterogeneous structure in biomarkers. BMC bioinformatics, 18(1), 308.

ガンマ・ダイバージェンスに基づいたロバスト統計

Robust Statistics via Gamma-Divergence

数理・推論研究系　藤澤 洋徳（Hironori Fujisawa）

1. はじめに

外れ値が存在するとき推定値にはバイアスが生じる．たとえば標本平均を考えよう．データの一つが非常に大きな値を取るとき，標本平均は非常に大きな値となってしまう．このバイアス問題を克服する単純な手法は中央値である．外れ値が一つであれば，中央値は大きく変わらない．ただし，外れ値の割合が大きい時には，中央値は大きくずれる．実は，外れ値の割合が大きい場合にも，バイアスを十分に小さくできる手法，というのは，長らくきちんと議論されていなかった．この問題はガンマ・ダイバージェンスによって解決される．**ロバスト統計で 25 年以上未解決だった重要問題が解決されたと考えている (Fujisawa and Eguchi, 2008).**

2. 本研究の特徴

過去の研究との大きな違いは二つある．一つはバイアスの議論の仕方であり，もう一つは外れ値への意識の仕方である．

上述したバイアスと言うのは，厳密には，統計科学で通常使われているバイアスではない．厳密に言えば「潜在バイアス」と呼ばれるものである．この潜在バイアスを小さくすることはロバスト統計の最大の目的の一つである．しかしながら，潜在バイアスを直接に議論することが難しかったため，理論的には，影響関数のような代替指標を使って議論することが主流であった．**これまで使われてきた代替指標を使わずに，潜在バイアスを直接に議論した点が，本研究の独創的な点である.**

もう一つは潜在バイアスを議論するときの前提条件に対する意識の違いである．過去のロバスト統計では，前述した影響関数という，強力で便利な代替指標を使って議論することが主流であった．しかし，この道具を使うためには，外れ値の割合が小さいということを暗に想定している．影響関数を使いながら，外れ値の外れ値らしさが軽く登場してくる．**本研究では意識レベルを逆にしている．外れ値の外れ値らしさの方を重視する．外れ値の割合が小さいという仮定は一切おいていない．この点も本研究の独創的な点である.**

結果的に，外れ値の割合が大きくても潜在バイアスが小さいロバスト推定，が可能になった．それはガンマ・ダイバージェンス $D_\gamma(g,f)$ に基づく手法である：

3. ガンマ・ダイバージェンス

ガンマ・ダイバージェンス $D_\gamma(g,f)$ は以下で与えられる：

$$d_\gamma(g,f) = -\frac{1}{\gamma}\log\int g(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\log\int f(x)^{1+\gamma}dx,$$
$$D_\gamma(g,f) = d_\gamma(g,f) - d_\gamma(g,g).$$

ここで，g と f は密度関数であり，$\gamma > 0$ はロバスト性を調整するパラメータである．実際のパラメータ推定は，g に基づく期待値を経験密度関数 $\bar{g}$ で置き換え，f をパラメトリックモデル f_θ に置き換えて，$d_\gamma(\bar{g}, f_\theta)$ の最小化で行う．

図 1 はガンマ・ダイバージェンスに対して成立する近似的なピタリアン構造である．詳細は省くが，外れ値が外れ値らしいという仮定を置いている．**このピタゴリアン構造は，ガンマ・ダイバージェンスに基づくロバスト推定が，なぜ上手く働くかの直感的な説明を与える．**データ発生分布 g とパラメトリック分布 $h = f_\theta$ のダイバージェンス $D_\gamma(g, h)$ を小さくするのが普通の方法である．ここで次の二点に注意する：(i) 直交性が近似的に成り立っている．(ii) g と f は固定されていて自由に動けるのはパラメトリック分布 h だけである．そうすると，$D_\gamma(g, h)$ を小さくしようとすると，$D_\gamma(f, h)$ が小さくなり，結果的に，パラメトリック分布 h はターゲット分布 f に近づく．これがガンマ・ダイバージェンスに基づいたロバスト推定が上手く働く理由である．

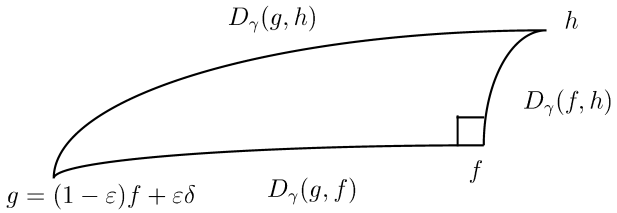


図 1．ピタゴリアン構造．g はデータ発生分布であり，ロバスト統計で慣習的に用いられている汚染分布 $g(x) = (1-\varepsilon)f(x) + \varepsilon\delta(x)$ を想定．ε は外れ値の割合．f はターゲット分布（外れ値に汚染されなかった場合の分布）．δ は外れ値の分布．h はパラメトリック分布 f_θ．

また，ある種の仮定の下では，外れ値の割合が大きくても潜在バイアスが小さくなる手法は，本質的に，ガンマ・ダイバージェンスに基づく手法だけであるという，ある種の唯一性が証明できる．加えて，パラメトリックモデルが指数型分布族に入っている場合には，ピタゴリアン構造を利用して，ロス関数の単調減少性をもつきれいなパラメータ推定アルゴリズムも提案できる．

4. 発展

唯一性の定理の仮定は弱く，ガンマ・ダイバージェンスはある種の決定打である．そのため，研究の発展は難しい状態が続いたが，最近になって研究が進むようになった．パラメトリック分布を，面積を変化させるパラメータ λ をわざと組み込んだ拡張モデル $\lambda f(x;\theta)$ に代えることで，ガンマ・ダイバージェンスでなくても，外れ値の割合が大きくても潜在バイアスが小さくなる手法を構築することができた (Kanamori and Fujisawa, 2015)．この研究と関連研究で，**日本統計学会研究業績賞**を頂いた．当初に提案したパラメータ推定アルゴリズムは，ピタゴリアン構造を利用していてきれいであったが，スパース罰則と組み合わせると，使いやすいパラメータ推定アルゴリズムの構築が容易でなく，ロバスト性とスパース性を同時に併せもつ手法の提案が難しかった．この問題は，Majorization-Minimization アルゴリズムの適用で克服されて，ガウシアン・グラフィカル・モデリングや回帰モデリングに適用されている (Hirose, Fujisawa and Sese, 2015; Kawashima and Fujisawa, 2017)．Google Scholar などで検索することで，その他にも様々に発展をして注目を浴びていることを見て取ることができる．

統計的自然言語処理と統計学

Natural language processing and Statistics

数理・推論研究系 持橋 大地（Daichi Mochihashi）

1. 「静的な統計」から「動的な統計」へ

2011 年に統数研に准教授として着任後, 2016 年の『統計数理』64 巻 2 号において, 特集「統計的言語研究の現在」を企画・担当した. 言語に関しては『統計数理』では 2000 年の 48 巻 2 号で特集「“ことば” 新研究」が組まれており, 二者を比較すると, 16 年の間に大きな変化があったことに気づく. 2000 年の特集では文章のジャンルの分類や編集距離による類似和歌の発見, 多変量解析による文書の因子分析などが主な内容であったが, 2016 年では構文構造の教師なし学習, 言語変化への統計的アプローチ, CRF(条件付確率場) の詳細な解説など, 内容が大きく様変わりしている. 統計的な手法としても, 前者が多変量解析をベースとしているのに対し, 後者では系統樹の統計モデル, ポアソン過程, 条件付確率場 (ロジスティック回帰の隠れマルコフモデル化), 階層ベイズモデルのように最新の幅広い統計あるいは機械学習の手法が取り入れられるようになった.

手法以外にも, 研究の哲学ともよぶべきものが, 「静的な統計」から「動的な統計」へと変わったように感じられる. 前者では, 文や文章は確定した「モノ」であり, それをどう扱うかがテーマとなっていたが, 後者では言語自体も変化するものであり, 内部に多くの文脈依存性を持っていることがフォーカスされている. 例えば CRF は文の構文解析や, 各単語の品詞が互いに依存するマルコフ確率場のモデルであり, 特集に含まれている文の読み時間の動的な推定や Twitter のツイートのモデルも, 時間依存性や空間依存性を含んでいる.

現在, 「動的」とは主に, ある文や文章内の現象を動的にモデル化しているが, 上の言語系統樹の話にあるように, 世代を超えて言語自体が時間的・空間的に変わりうる様相を統計的に明らかにすることにも今後取り組みたく, 国立国語研究所や国立民族学博物館との共同研究を現在行っている.

2. 単語と分節化の理論

英語のように単語に分かれていない[1] 日本語や中国語, タイ語のような言語にとって, 文を「単語」に分けることは最も基礎となる重要な課題である. 従来はこのために, 人手で準備した大量の「正解」の単語列をもとに CRF などを学習することで単語分割や品詞推定がなされてきたが, 「正解」が真に正解であるという保証はなく, また日々無数に生まれる新語には対応できないという限界がある.

一 の 皇子 は 、 右 大臣 の 女御 の 御 腹 に て 、 寄せ 重く 、 疑ひ なき まうけ の 君 と 、 世 に もて かしづき きこゆ れ ど 、 この 御 にほひ に は 並び たまふ べく も あら ざり け れば 、 おほかた の やむごと なき 御思ひ に て 、 この 君 を ば 、 私 ものに 思ほし かしづき たまふ こと 限り なし 。 はじめより おしなべて の 上 宮仕 し たまふ べき 際 に は あら ざり き 。

図 1.『源氏物語』の教師なし形態素解析の結果の一例. 辞書や文法は一切用いていない.

[1] ラテン語や英語も, もともとは単語を分けて書かず一続きに書くのが普通であった.

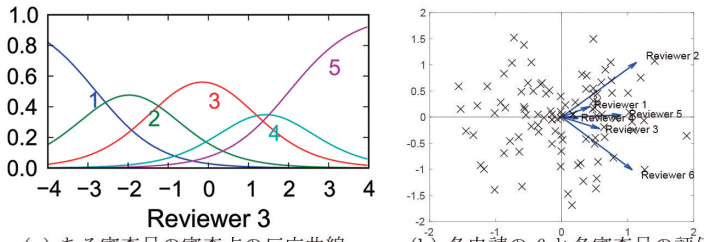
これに対し, ノンパラメトリックベイズ法による文字–単語の階層ベイズモデルを仮定し, 出力された生の文字列のみから「単語」を逆に推定する統計モデルを 2009 年頃に発表した (Mochihashi et al., 2009). これにより, 例えば「源氏物語」の文字列のみから単語を図 1 のように推定できる. その後の統数研での共同研究で, さらに品詞を同時に教師なし推定したり (Uchiumi et al., 2015), 教師データも利用した半教師あり学習も高精度で行えるようになった (Fujii et al., 2017).

統計的には, これはセミマルコフモデルの一種による分節化であると考えられる. したがって, 同様の統計モデルを言語だけでなく, 他の時系列データにも適用することができる. 言語において文字列にあたる出力をガウス過程からの波形とすれば, ロボティクスにおいてロボットの関節角の時系列を分節化して「動作」を取り出すことも可能になった (Nagano et al., 2018). 音声認識においても, 音声から単語を自動認識することが可能になるため, 共同研究を行っている.

3. 離散データと「科学の科学」

離散データを扱う統計的自然言語処理は, 他の多くの分野と繋がりを持っている. 2015 年から現在まで研究員を務めている日本学術振興会の学術情報分析センターでは, 科研費の申請に査読者が付した “5”,“2” などの審査点から, 項目反応理論により図 2 のように正規分布に従う各申請のスコアを統計的に算出する研究を行った. 多次元で行えば, 矢印で示した各審査員の「評価軸」も客観的なデータから数学的に明らかになり, 公正な審査に貢献する. また, 大規模な確率的潜在意味解析 (LDA) を用いて, 学振に登録されている研究者の専門性を計算し, 審査の際に適切な審査員を推薦する試みも行っている.

科学的な発見は論文, つまり言語の形で発表される. 論文の生産自体も離散的なイベントであり, 点過程として捉えることができる. また, そこには複雑な相互依存関係があると考えられる. 自然言語処理の背景を生かし, こうした「科学の科学」へも今後取り組みたいと考えている.



(a) ある審査員の審査点の反応曲線. (b) 各申請の θ と各審査員の評価軸.

図 2. 項目反応理論による潜在的な得点と評価軸の計算

参 考 文 献

Fujii, R., Domoto, R. and Mochihashi, D. (2017). Nonparametric Bayesian Semi-supervised Word Segmentation, *Transactions of ACL*, **5**, 179–189.

Mochihashi, D., Yamada, T. and Ueda, N. (2009). Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling, *Proceedings of ACL-IJCNLP 2009*, 100–108.

Nagano, M., Nakamura, T., Nagai, T., Mochihashi, D., Kobayashi, I. and Kaneko, M. (2018). Sequence Pattern Extraction by Segmenting Time Series Data Using GP-HSMM with Hierarchical Dirichlet Process, *IROS 2018*, 4067–4074.

Uchiumi, K., Tsukahara, H. and Mochihashi, D. (2015). Inducing Word and Part-of-speech with Pitman-Yor Hidden Semi-Markov Models, *ACL-IJCNLP 2015*, 1774–1782.

メタアナリシスにおける公表バイアスの最悪評価に基づく感度解析

Worst-case sensitivity analysis for publication bias in meta-analysis

数理・推論研究系 逸見 昌之（Masayuki Henmi）

1. 医学研究におけるメタアナリシス

同じ目的あるいは何らかの共通点を持った複数の研究から得られた統計解析の結果を統合して、よりエビデンスの高い結果を得るための統計解析のことをメタアナリシスと呼ぶ。医学研究における最も典型的な応用例は、医薬品の効果を検証するための臨床試験や疾患等の原因を探る疫学研究に対するメタアナリシスであり、例えば臨床試験であれば、ある疾患に対する新薬 A と既存薬 B とを比較するランダム化臨床試験の結果が複数ある際に、個々の研究では新薬 A の効果を支持する十分なエビデンスが得られなかったとしても、複数個の結果を統合することによって十分なエビデンスが得られることがある。

2. 公表バイアスの問題

一般に統計解析においては、観測データが関心のある母集団から偏りなく得られていることが重要であるが、メタアナリシスの場合は、観測データとなる複数の研究結果は公表されている論文等からしか得られないので、データに偏りが生じやすくなる。例えば、臨床試験や疫学研究などにおける二群比較では、統計的検定で有意な結果が公表されやすく、それらの結果だけを集めてメタアナリシスを行えば、当然その結果も有意となる。このように、公表データの偏りによりメタアナリシスの結果に生じるバイアスのことを公表バイアス (publication bias) と呼ぶ。図 1 は、ある薬剤を妊婦に投与した場合に早産を予防できるかどうかを検証するために行われた、複数のランダム化臨床試験の結果をプロットしたものである。横軸は各試験での対数オッズ比の推定量、縦軸はその標準誤差の逆数を表している。図の下の方ほど標準誤差が大きくなり、検定結果が有意になりにくくなるが、その部分で片側が欠けていることから、公表バイアスの存在が強く疑われる。このような図は一般にファンネルプロット (funnel plot) と呼ばれ、公表バイアスの存在の可能性を視覚的に検討するためによく用いられる。

3. 最悪評価による感度解析

公表バイアスを調整して妥当な統合結果を得るためには、個々の研究結果の公表のプロセスに関する強い仮定が必要であるが、それを観測される研究結果から検証することは不可能である。そこで、その仮定を有り得る範囲内で変化させて統合結果がどのように変わるかを調べるという感度解析を行うことが推奨されているが、その仮定をどのように変化させるかというのは難しい問題である。この問題に対し我々は、標準誤差が大きい推定量を有する研究ほど結果が公表されにくいという、定性的な弱い仮定のみを用いて、統合結果として得られる信頼区間

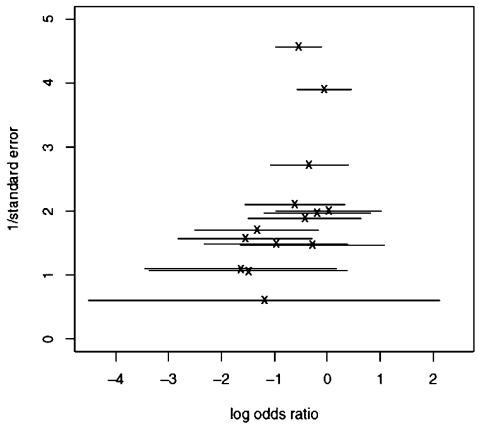


図 1. 臨床試験のファンネルプロット

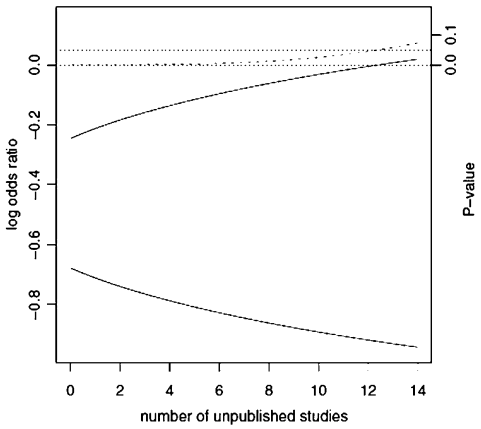


図 2. 信頼区間の存在範囲 (実線) と P-値の上限 (破線)

と P-値の最悪評価を行った。図 2 は、図 1 のデータを用いたメタアナリシスに対して、未公表の研究数ごとに、統合結果から得られる信頼区間の存在範囲 (実線) と P-値の上限 (破線) を表しており、この図から未公表の研究数が 13 を超えると統合結果としての検定の有意性が逆転する可能性があることが分かる。

参 考 文 献

Henmi, M., Copas, J.B. and Eguchi, S. (2007). Confidence Intervals and P-values for Meta-analysis with Publication Bias. *Biometrics* **63**, 475-482.

スピングラス理論による制限等長定数評価
Evaluation of restricted isometry constants using spin-glass theory

数理・推論研究系 坂田 綾香（Ayaka Sakata）

要 旨

制限等長定数は、圧縮センシングにおける完全再構成条件を与えるが、その厳密評価は困難である。我々は統計物理学におけるスピングラス理論を用いて、制限等長定数を制度よく見積もる方法を提案した。

キーワード：圧縮センシング、制限等長性

原信号がゼロ成分を多く持つ (スパース性) という事前知識の下で、次元より少ない観測から信号を復元する枠組みである圧縮センシング (CS) は、様々な分野において利用されている。$\boldsymbol{A} \in \mathbb{R}^{M\times N}$ を観測行列とすると、線形観測 $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ から $S(< N)$ 個の非ゼロ要素を持つ信号 $\boldsymbol{x}(S$ スパースベクトル) を再構成する問題として CS は定式化される。ℓ_0, ℓ_1 再構成法は代表的手法であり、S スパースベクトル $\boldsymbol{x}$ が完全復元される十分条件は制限等長定数により与えられる [Candés and Tao (2005)]。ここでは、観測行列 $\forall \boldsymbol{A} \in \mathbb{R}^{M\times N}$ はコラムが典型的に $(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})_{ii} = 1$ $(i \in \{1, \cdots, N\})$ を満たすように規格化されているとする。全ての S スパースベクトル $\boldsymbol{x}$ に対して次の不等式が成立するとき、行列 $\boldsymbol{A}$ は制限等長性を満たす。

$$(1 - \delta_S^{\min})||\boldsymbol{x}||_F^2 \leq ||\boldsymbol{A}\boldsymbol{x}||_F^2 \leq (1 + \delta_S^{\max})||\boldsymbol{x}||_F^2 \tag{0.1}$$

また $0 < \delta_S^{\min} \leq \delta_S^{\max}$ を制限等長定数 (RIC) と呼ぶ。RIC が与える ℓ_0、ℓ_1 再構成による完全復元条件は [Candés et al. (2006)] 等に示されている。

RIC は $\boldsymbol{A}$ のグラム行列の固有値と関係づけられる。S スパースベクトル $\boldsymbol{x}$ の非ゼロ要素の位置を $T \subseteq V = \{1, \cdots, N\}$, $|T| = S$ として表現し、$\boldsymbol{A}$ の $i \in T$ コラムからなる行列を $\boldsymbol{A}_T$、また $\boldsymbol{x}_T = \{x_i | i \in T\}$ とすると、$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}_T\boldsymbol{x}_T$ である。そして全ての T について、次の不等式が成立する。

$$\lambda_{\min}(\boldsymbol{A}_T^{\mathrm{T}}\boldsymbol{A}_T)||\boldsymbol{x}_T||_F^2 \leq ||\boldsymbol{A}_T\boldsymbol{x}_T||_F^2 \leq \lambda_{\max}(\boldsymbol{A}_T^{\mathrm{T}}\boldsymbol{A}_T)||\boldsymbol{x}_T||_F^2$$

$\lambda_{\min}(\boldsymbol{B})$, $\lambda_{\max}(\boldsymbol{B})$ は $\boldsymbol{B}$ の最小・最大固有値を表し、上付きの T は転置を表す。(0.1) との比較から，$\lambda_{\min}^*(\boldsymbol{A}; S) = \min_{T:T\subseteq V,|T|=S} \lambda_{\min}(\boldsymbol{A}_T^{\mathrm{T}}\boldsymbol{A}_T)$ および $\lambda_{\max}^*(\boldsymbol{A}; S) = \max_{T:T\subseteq V,|T|=S} \lambda_{\max}(\boldsymbol{A}_T^{\mathrm{T}}\boldsymbol{A}_T)$ を用いて、RIC は次のように表現される．

$$\delta_S^{\min} = 1 - \lambda_{\min}^*(\boldsymbol{A}; S), \quad \delta_S^{\max} = \lambda_{\max}^*(\boldsymbol{A}; S) - 1 \tag{0.2}$$

(0.2) を厳密に評価するには、あらゆる T について固有値を評価しなくてはならない。これは計算量的に困難であるため、様々な近似方法が考えられてきた [Bah and Tanner (2010)]。我々は統計物理学におけるスピングラス理論を用いて、RIC の評価を改善することに成功した。そ

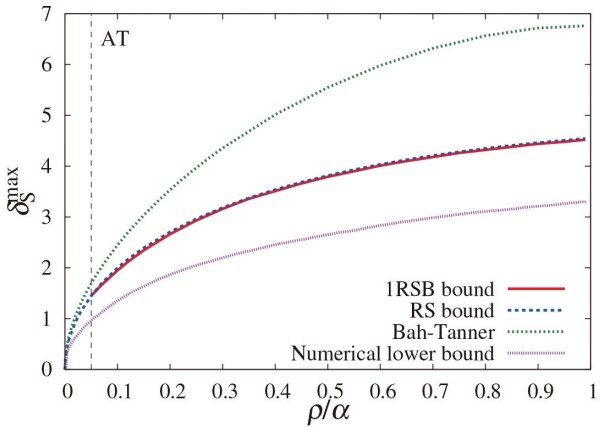


図 1. $\alpha = 0.5$ での RIC の ρ/α 依存性。

の際、$M = N\alpha$, $S = N\rho$ $(\alpha, \rho \sim O(1))$ として $N \to \infty$ 極限を考えた。

図 1 は、観測行列がガウシアンランダム行列の場合の $\delta_S^{\max}$ について、先行研究と我々の評価法を比較したものである。垂直点線の ρ/α において、レプリカ対称性の破れに伴った相転移現象が起きる。"Numerical lower bound"とは数値的に評価された RIC の下界であり、RIC がこれより大きいことを保証している。"Bah-Tanner"は先行研究による RIC の上界評価である。我々の"1RSB bound", "RS bound"も上界評価であるが、先行研究よりも精度の高い上界評価となっている。1RSB bound とは、レプリカ対称性の破れを 1 段階考慮した評価となっており、逐次的に対称性の破れを考慮していくことで上界評価が正しい値に近づいていくことが数学的に保証される。また $\delta_S^{\min}$ に関しても精度の良い評価を達成し、更にランダム一次転移や Gardner 転移と呼ばれる相転移現象が見られることも発見した。具体的評価法は [Sakata and Kabashima (2015)] に示されている。

提案手法は、ランダム直交行列から構成された観測行列に対しても適用可能である。様々な行列に対して RIC を評価することで、効率的観測方法などが提案できると考えられる。

謝　辞

本研究は樺島祥介氏 (東京工業大学) との共同研究である。

参 考 文 献

Bah, B. and Tanner, J. (2010). Improved Bounds on Restricted Isometry Constants for Gaussian Matrices, *SIAM Journal on Matrix Analysis and Applications*, **31**, 2882–2898.

Candés, E. J. and Tao, T. (2005). Decoding by linear programming, *IEEE Transactions on Information Theory*, **51**, 4203–4215.

Candés, E. J., Romberg, J. and Tao, T. (2006). Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information, *IEEE Transactions on Information Theory*, **52**, 489–509.

Sakata, A. and Kabashima, Y. (2015). Replica Symmetric Bound for Restricted Isometry Constant, *Proceedings of IEEE International Symposium on Information Theory*, **2015**, 2006–2010.

多層整数計画に基づくクリンチ／エリミネーション数の計算

Calculation of Clinch and Elimination Numbers Based on Multilayered Integer Programming

数理・推論研究系　伊藤 聡（Satoshi Ito）

1. はじめに

リーグスポーツのシーズン中のどの時点においても，最終的にリーグ優勝やプレーオフ出場権など特定の状況（指標）が達成されることが確定する最小の勝ち試合数（クリンチ数），もしくは逆にその状況（指標）に届かないことが確定する最小の負け試合数（エリミネーション数）が存在する．本研究は，順位決定に係る複数の判定基準が存在する場合のクリンチおよびエリミネーション数の計算を，多層の整数計画問題を解くことにより高速に行う汎用的な枠組みを開発することを目的としている．

2. 整数計画モデルと上下界の多層構造

リーグに属するチームの集合を L とし，そのチーム数を n とする．リーグ L の全チーム間のこれまでの勝敗記録と残り試合数が与えられているとし，w_{ij} をチーム $i \in L$ のチーム $j \in L$ に対する現時点での勝数，g_{ij} をチーム i, j 間の残り試合数とする．チーム $i \in L$ のチーム $j \in L$ に対する今後の勝数を x_{ij} と表すことにすると，

$$X := \left\{ x = (x_{ij}) \in \mathbb{Z}^{n \times n} \;\middle|\; x_{ij} + x_{ji} \leq g_{ij},\ x_{ii} = 0,\ x_{ij} \geq 0 \ (\forall i, j \in L) \right\}$$

は今後起こり得る勝敗に関するシナリオを過不足無く与える（$\mathbb{Z}$ は整数の集合であり，引分がない場合は第 1 式の不等号を等号に変えるものとする）．

順位判定基準が複数あり，そのうち最初の m 個が勝敗数のみに基づく基準（すなわち $m+1$ 番目の基準は得失点差など勝敗数以外に基づくもの）であるとする．このときチーム $a \in L$ の例えば（簡単な例として）第 k 位クリンチ数は，チーム a が今後引き分けることなく，かつ m 個のうちのいずれかの基準でチーム a より上位の成績を持つチームが k 個存在するという制約条件のもとで，チーム a の今後の勝数 $\sum_{j \in L} x_{aj}$ の最大値 $\bar{z}$ を求めることにより得られる（この最大化問題に許容解がなければ既に第 k 位以上が確定していることになるし，$\bar{z}$ が残り試合数と等しければ今後全勝しても第 $k+1$ 位以下の可能性があることになり，そうでなければ第 k 位クリンチ数は $\bar{z}+1$ となる）．ここで注意すべきことは，m 個の順位判定のうち最初の $m-1$ 個は等号なし（$<$）であり，最後の m 番目のみが等号つき（$\leq$）の順位判定となることである．例えば 2016 年に始まった男子プロバスケットボールの B.LEAGUE では，①勝率（勝数）→ ②当該クラブ間勝率（当該クラブ間 1 試合平均勝数）→ ③当該クラブ間得失点差 → ④当該クラブ間 1 試合平均得点 → ⑤得失点差 → ⑥1 試合平均得点（総得点）→ ⑦理事会が必要と判断した場合に抽選と判定基準が続くが，この場合 $m = 2$ であり，判定基準②に関して同成績であっても③以降の基準により順位が変わる可能性が残っているため，チーム a にとってより都合の悪い状況を想定するという意味で②に対する順位判定は等号つきである必要がある．

上記の例のような整数計画モデルは当該チーム間成績が絡む場合などには非線形となり，汎用最適化パッケージを用いて直接解くことは必ずしも容易でない．そこで，最大目的関数値の上界と下界をうまく利用することを考える．図 1 のように，$m-1$ 番目までの順位判定のうちいずれか一つの不等号（$<$）を等号つき（$\leq$）に変えることにより $m-1$ 個のレベルの上界が得られ，また，1 番目を除く順位判定をそれ以降を含めて取り去る（すなわち順位判定を途中で打ち切る）ことにより $m-1$ 個のレベルの下界が得られる．さらに，m 番目の順位判定を等号なし（$<$）に変えて得られる，よりタイトな下界を用いることにより，勝敗数以外の要素に基づく $m+1$ 番目以降の判定基準を考慮する必要性の有無も判定することができる．

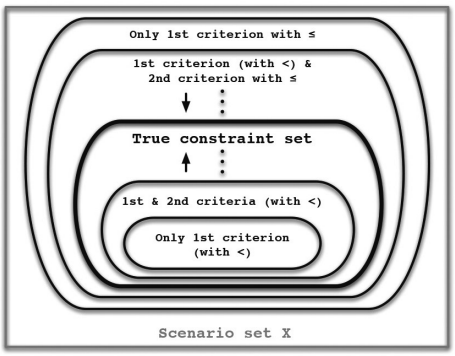
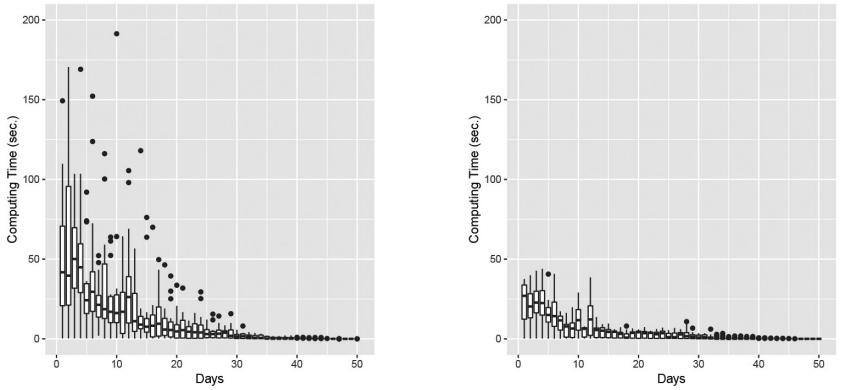


図 1: 上下界の多層構造

3. 数値実験

B.LEAGUE において考えられる指標のうち，B1 チャンピオンシップ・トーナメント進出のエリミネーション数について，2016–2017 シーズン第 15 節以降の 50 日分の勝敗記録に対して数値実験を行った結果を図 2 に示す．全 18 クラブのエリミネーション数を計算するのに要した個々の時間を一日単位で箱ひげ図にプロット（四分位点から 1.5 IQR を超える外れ値を点で表示）しているが，上下界の情報を使わずに解いた (a) に対して，これを活用した (b) では，極めて時間がかかるケースを抑えることにより平均計算時間を短縮することに成功している．



(a) Results from scratch

(b) Results with using upper and lower bounds

図 2: B1 チャンピオンシップ・トーナメント進出エリミネーション数の計算時間

参 考 文 献

Ito, S. and Shinano, Y. (2018). Calculation of clinch and elimination numbers for sports leagues with multiple tiebreaking criteria, ZIB-Report 18-51, urn:nbn:de:0297-zib-70591.

天文学とデータ科学

Astronomy and Data Science

数理・推論研究系 池田 思朗（Shiro Ikeda）

要 旨

深層学習の成功をきっかけとして，データ科学分野へ大きな注目が集まっている．新たなデータ処理の方法は産業を通じた社会への貢献だけでなく，自然科学の分野にも影響を及ぼすはずである．生物学ではこうした流れからバイオインフォマティックスが生まれた．天文学ではまだそうした大きな流れは生まれていないが，今後はデータ科学によって大きく変わっていくはずである．

キーワード：天文学, データ科学, Event Horizen Telescope, すばる望遠鏡.

1. はじめに

天文学では，観測結果がデジタルデータとして保存されるようになって以来，計算機上で統計的な解析が行われてきた．しかし，これからの天文観測では爆発的にデータが増えていくため，新たなデータ解析手法の導入は必要不可欠である．今後の天文学は，データ科学との有機的な共同研究が必要となる．

この数年，天文学者と議論を重ねてきた経験から，データ科学から天文学への貢献の可能性には大きくふたつの方向性があると考えている．ひとつは新たな方法による解析精度の向上だろう．限られたデータからデータ科学の方法によってより多くの情報を得ようという方向性である．もうひとつはビッグデータへの対応である．今後，観測機器の高性能化によって高精度なデータが非常に多く得られるようになると，これまで天文学者が解析に用いてきた方法では対応しきれない．ここにも新しいデータ科学の方法が求められている．

2. 電波干渉計とスパース推定

電波干渉計は天体から発せられる電波を複数のアンテナで受信し，相関処理をしたのちに天体のイメージを得ようというものである．離れた位置にあるアンテナの信号に相関処理を行えば，光の干渉と同じように画像のフーリエ変換に対応する情報を得られる．したがって，この逆問題は理想的には逆フーリエ変換によってイメージングできるはずである．しかし，アンテナ数は限られており，一般にフーリエ空間上の観測点の数はイメージのピクセル数に比べて少ない．このため，電波干渉計のイメージングの問題は不良設定問題である．

条件が足りない不良設定問題に対して，データ科学ではこの 20 年，LASSO や圧縮センシングといったスパース推定の手法が開発された．例えば画像上で局所的に光源が分布しているコンパクトな天体の場合，こうした方法は有効である．我々はブラックホールシャドウの撮像のためのプロジェクト Event Horizon Telescope (EHT) に参加していて（Honma et al., 2014;

Ikeda et al., 2016; Akiyama et al., 2017)，スパース推定は EHT に欠かせない方法のひとつになっている．こうした貢献はまさに，新たな方法による解析精度の向上であろう．

3. すばる望遠鏡の観測データからの超新星の選別

もうひとつ，現在共同研究を行っているのは，すばる望遠鏡の超広視野主焦点カメラ (HSC: Hyper Suprime-Cam) を用いたサーベイ観測「すばる戦略枠プログラム」に関するプロジェクトである．このプログラムは 2014 年に開始され，5 年間で 300 晩の観測を行う．最終的に観測データはペタバイトのオーダーになると見積られている．そこで期待されている成果のひとつは，遠方で発生する Ia 型超新星を数多く検出することである．これにより，宇宙論パラメータの推定精度が向上すると考えられる．

突然現れる超新星探索の発見は，ある日の画像から以前に撮った画像を差し引き，差分イメージに引き残された超新星を見つけ出すことになる．これまでは差分から超新星の候補となる画像パッチを自動的に取り出し，それを人間が本物か偽物か判定を行っていた．しかし，HSC の差分画像では，一晩に観測される超新星候補の数は数万以上に登り，その中にある目的の超新星は高々数十個しかないと見積もられている．このため，人間の目に頼るのではなく，本物・偽物の判定結果を返す関数を機械学習の方法によって作成し，実際の観測に用いている (Morii et al., 2016)．この研究は，これまでの天文学の方法をビッグデータでも引き続き行うための対応である．

4. まとめ

天文学は，歴史の初期から存在する最も古い学問のひとつである．歴史的にみれば，常に最先端の技術が投入されてきた．現在大きく発展しているデータ科学の方法が取り入れられるのも当然の流れである．

ここに挙げたように，天文学からデータ科学への期待は，より多くの情報を引き出し，来たるビッグデータに対応するという 2 点になるだろう．その先，データ科学が主導して天文学へ新たな提案をすることができれば，天文データ科学と呼べる分野が確立するだろう．

これまで天文学に関するさまざまなデータ解析の相談を受けてきた．今後 10 年の間に起きる天文学の変化の力になりたいと考えている．

参 考 文 献

Akiyama, K., Ikeda, S., Pleau, M., Fish, V. L., Tazaki, F., Kuramochi, K., Broderick, A. E., Dexter, J., Mościbrodzka, M., Gowanlock, M., Honma, M. and Doeleman, S. S. (2017). Superresolution Full-polarimetric Imaging for Radio Interferometry with Sparse Modeling, *The Astronomical Journal*, **153** (4), 159(12pages).

Honma, M., Akiyama, K., Uemura, M. and Ikeda, S. (2014). Super-resolution imaging with radio interferometry using sparse modeling, *Publications of Astronomical Society of Japan*, **66** (5), 95(14pages).

Ikeda, S., Tazaki, F., Akiyama, K., Hada, K. and Honma, M. (2016). PRECL: A new method for interferometry imaging from closure phase, *Publications of Astronomical Society of Japan*, **68** (3), 45(9pages).

Morii, M., Ikeda, S., Tominaga, N., Tanaka, M., Morokuma, T., Ishiguro, K., Yamato, J., Ueda, N., Suzuki, N., Yasuda, N. and Yoshida, N. (2016). Machine-learning Selection of Optical Transients in Subaru/Hyper Suprime-Cam Survey, *Publications of Astronomical Society of Japan*, **68** (6), p. 104(8pages).

制約付き非凸スパース最適化問題に対する DC アルゴリズム

DC Algorithm for Constrained Nonconvex Sparse Optimization

数理・推論研究系 田中 未来 (Mirai Tanaka)

1. 制約付き非凸スパース問題

近年, 統計数理のさまざまな分野において Lasso をはじめとするスパース最適化の研究が進められている. 本稿では制約領域 $S \subseteq \mathbb{R}^n$ 上における損失関数 $l(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ とスパース正則化関数 $r(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ の和を最小化する問題

$$\text{最小化} \quad l(\boldsymbol{x}) + r(\boldsymbol{x}) \quad \text{制約条件} \quad \boldsymbol{x} \in S \tag{1.1}$$

に対する効率のよいアルゴリズムを提案した著者らの論文（Tanaka and Takeda, 2018）を紹介する.

以下では l, r, S について以下のような仮定をおく. まず l については連続的微分可能であることと, ある $L \in \mathbb{R}_+$ が存在して $(L/2)\|\boldsymbol{x}\|_2^2 - l(\boldsymbol{x})$ が凸関数となることを仮定する. 次に r については連続であることと, ある $\lambda \in \mathbb{R}_+$ が存在して $\phi(\boldsymbol{x}) := \lambda\|\boldsymbol{x}\|_1 - r(\boldsymbol{x})$ が凸関数となることを仮定する. 後者の仮定は多くのスパース正則化関数が満たすものである. さらに $l + r$ が下に有界であることと S が空でない凸集合であることを仮定する. 以上の仮定をおいたとしても問題（1.1）は符号制約付き回帰, 主成分分析, 標準単体上の最小 2 乗問題, C-SVM の双対問題などを含むさまざまな問題のスパースな解を求める問題に対応する. しかしながら問題（1.1）は制約付き非凸最適化問題であり, 一般に大域的最適解を求めることは難しい. 以下では問題（1.1）の停留点を効率よく求める DC アルゴリズムについて述べる.

2. DC アルゴリズムとその各反復で解く子問題

2 つの凸関数 $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $h : \mathbb{R}^n \to \mathbb{R}$ について, これらの差を最小化する問題

$$\text{最小化} \quad g(\boldsymbol{x}) - h(\boldsymbol{x}) \tag{2.1}$$

を DC 最適化問題と呼ぶ. この問題の停留点を求めるために DC アルゴリズム (アルゴリズム 1) がよく用いられる. このアルゴリズムは緩やかな仮定の下で DC 最適化問題の停留点に収束する点列を生成する.

アルゴリズム 1 問題（2.1）に対する DC アルゴリズム

1: 適当な初期点 $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ をとる.
2: **for** $t = 0, 1, \ldots$ (収束するまで)
3: $\quad h$ の劣勾配 $\boldsymbol{s}^{(t)} \in \partial h(\boldsymbol{x}^{(t)})$ を求める.
4: $\quad \boldsymbol{x}^{(t+1)} \in \operatorname{argmin}_{\boldsymbol{x}}\{g(\boldsymbol{x}) - (\boldsymbol{s}^{(t)})^\top \boldsymbol{x}\}$ と更新する.

問題（1.1）を DC 最適化の枠組みで解くために, 次のように凸関数 g, h を定める:

$$g(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 + \delta(\boldsymbol{x} \mid S), \qquad h(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x}\|_2^2 - l(\boldsymbol{x}) + \phi(\boldsymbol{x}).$$

ここで $\delta(\boldsymbol{x} \mid S)$ は $\boldsymbol{x} \in S$ のときに 0, そうでないときに $+\infty$ をとる関数である. このように g, h を定めると問題（1.1）を DC 最適化問題（2.1）に帰着できる.

問題（1.1）に対する DC アルゴリズムの各反復では次の子問題を解く:

$$\text{最小化} \quad \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + \gamma\|\boldsymbol{x}\|_1 \quad \text{制約条件} \quad \boldsymbol{x} \in S. \tag{2.2}$$

この子問題を高速に解くことができるとき, DC アルゴリズムは全体として高速なものとなる.

著者らは次の 4 つの場合について子問題（2.2）の最適解ないしその近似が効率よく計算できることを示した. 1 つ目は S が超直方体 $\{\boldsymbol{x} : \underline{\boldsymbol{x}} \le \boldsymbol{x} \le \bar{\boldsymbol{x}}\}$ の場合である. この場合は簡単な解析により子問題（2.2）の最適解を陽に書き下すことができる. また, S が ℓ_2 ノルム球 $\{\boldsymbol{x} : \|\boldsymbol{x}\|_2^2 \le 1\}$ の場合についても最適解を陽に書き下すことができる. S が標準単体 $\{\boldsymbol{x} : \mathbf{1}^\top \boldsymbol{x} = 1, \boldsymbol{x} \ge \mathbf{0}\}$ の場合, 子問題（2.2）は $\boldsymbol{v}$ の標準単体への射影の計算となるので, 既存の $O(n)$ 時間アルゴリズムを用いることができる. 著者らが解析した S の中で最も興味深いものは箱型制約に単一線形制約が付加した $\{\boldsymbol{x} : \boldsymbol{a}^\top \boldsymbol{x} = b, \underline{\boldsymbol{x}} \le \boldsymbol{x} \le \bar{\boldsymbol{x}}\}$ である. 著者らは対応する子問題に対する 2 分探索に基づく多項式時間アルゴリズムを構築した.

3. 計算機実験

制約付き非凸スパース最適化問題に対する既存の DC アプローチ（Tono et al., 2017）との比較実験の結果の一部として, ある問題例をそれぞれの手法で解いたときの解パスの比較を図 1 に示す. 既存手法はパラメータ λ の値を大きくとっても厳密にスパース解を出力したとは言えないのに対し, 著者らによる提案手法は $\lambda > 2$ で厳密なスパース解を出力した. また, 詳細は割愛するが, 著者らによる提案手法は既存手法や汎用ソルバと比べて高速に優れた解を出力した.

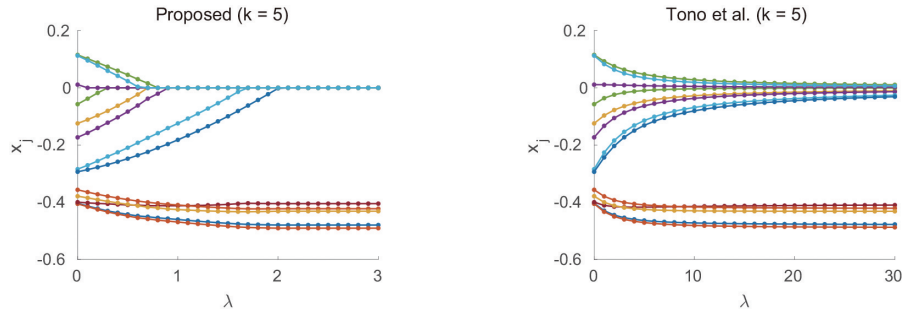


図 1: 解パスの比較 (左: 著者らによる提案手法によるもの, 右: Tono et al.（2017）による既存手法によるもの)

参 考 文 献

Tanaka, M. and Takeda, A. (2018). Efficient iterative algorithm for constrained nonconvex sparse optimization, submitted.

Tono, K., Takeda, A. and Gotoh, J. (2017). Efficient DC algorithm for constrained sparse optimization, arXiv:1701.08498v1.

関数推定の理論に基づく深層学習の原理解析

Analysis for Deep Learning by Function Estimation Theory

数理・推論研究系 今泉 允聡（Masaaki Imaizumi）

1. はじめに

本稿では、深層ニューラルネットワーク（DNN）が他手法より良い性能を発揮する原理を、統計理論を用いて解析した。DNN は既存手法よりも高い性能を発揮することが経験的に知られているが、なぜその性能が発揮されるのかという原理は充分には解明されていない。既存の統計理論では、データが滑らかな関数から生成されている場合、多くの既存の統計・機械学習の手法が理論上の最適精度を達成することが示されており、DNN の相対的優位を説明することは難しい。本稿はその困難さを解決するため、データが非滑らかな関数から生成されている状況で各手法の汎化誤差評価を行った。具体的には、DNN による推定量の汎化誤差の収束レートを導出し、そのレートがミニマックスの意味での最適性を満たすことを示した。加えて、いくつかの既存手法がその収束レートを達成しないことを示し、DNN が他手法に理論的な優越する状況を明らかにした。

2. 問題設定

非滑らかな関数による回帰問題を考える。$I = [0,1]$ とし、独立同一分布より生成された観測値の集合 $\{(X_i, Y_i) \in I^D \times \mathbb{R}\}_{i\in[n]}$ が与えられ、またそれらのデータ生成過程は以下の関係を満たしているとする：

$$Y_i = f^*(X_i) + \xi_i.$$

ここで、$f^* : I^D \to \mathbb{R}$ はデータ生成過程を特徴付ける真の関数（未知）であり、また ξ_i は平均 0 で分散 $\sigma^2 > 0$ のガウスノイズであるとする。また、f^* は区分上でのみ滑らかな関数であるとする。即ち、f^* の定義域 I^D が α-Smooth な境界を持つ複数の区分に分割され、その区分の内部で f^* は β-Smooth であるとする。区分の境界線上では、f^* は非連続になりうる。

観測の集合 $\mathcal{D}_n := \{(X_i, Y_i)\}_{i\in[n]}$ による f^* の推定量を考える。DNN によるモデル Ξ_{NN} を用いて、経験リスクを最小化する最小二乗推定量を

$$\widehat{f} \in \underset{f\in\Xi_{NN}}{\operatorname{argmin}} \frac{1}{n} \sum_{i\in[n]} (Y_i - f(X_i))^2,$$

と定義し、この関数 $\widehat{f}$ を f^* の推定量として用いる。

3. 結果

3.1 DNN による汎化誤差の評価

$\widehat{f}$ による汎化誤差は以下のように評価される。

Theorem 1. （$\widehat{f}$ による汎化誤差）
ある定数 $c_1, C_L > 0$ と DNN のあるネットワークのもとでの推定量 $\widehat{f}$ が

$$\|\widehat{f} - f^*\|^2_{L^2(P_X)} \leqslant C_L \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}(\log n)^2,$$

を確率 $1 - c_1 n^{-2}$ 以上で満たす。

収束レートのうち、一つ目の項 $n^{-2\beta/(2\beta+D)}$ は、各区分内部の f^* の滑らかさな部分を推定する影響、二つ目の項 $n^{-\alpha/(\alpha+D-1)}$ は各区分そのものを推定する影響を表現している。

3.2 DNN の最適性

定理 1 で得られた結果の最適性を議論するため、区分上でのみ滑らかな関数 f^* を推定する際のミニマックスな収束レートを導出する。

Theorem 2. （区分上でのみ滑らかな関数推定のミニマックスレート）
$\bar{f}$ を $\mathcal{D}_n$ に依存する任意の推定量とする。この時、ある定数 $C_{mm} > 0$ のもとで以下が成立：

$$\inf_{\bar{f}} \sup_{f^*} \mathbb{E}\left[\|\bar{f} - f^*\|^2_{L^2(P_X)}\right] \geqslant C_{mm} \max\left\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\right\}.$$

定理 2 の結果より、定理 1 で得られた汎化誤差の収束レートは、ミニマックスな汎化誤差の収束レートに対数項の影響を除いて一致している。すなわち、区分上でのみ滑らかな関数の推定問題において、DNN による推定量は理論的な最適性を達成していると言える。

3.3 DNN と他手法の比較

区分上でのみ滑らかな関数を推定する際の、他手法の非最適性について議論する。本稿では、以下の形式で書かれる**線形推定量**と呼ばれる推定量のクラスを考える：

$$\widehat{f}^{\mathrm{lin}}(x) = \sum_{i\in[n]} \Upsilon_i(x; X_1, ..., X_n) Y_i. \tag{3.1}$$

なお、Υ_i は $X_1, ..., X_n$ に依存する任意の可測関数である。この推定量のクラスは、カーネル法、フーリエ法、スプライン法、ガウス過程法などの多くの推定量を含んでいる。

非滑らか関数を推定する問題について、過去の研究が線形推定量が最適性を達成しないことを示している。それを用いることで、以下の結果を得ることが出来る。

Corollary 1. （DNN の理論的優位性）
$\alpha D/(2\alpha + 2D - 2) \leqslant \beta$ が成立するとする。この時、ある f^* が存在し、そのもとで DNN による推定量 $\widehat{f}$ と任意の線形推定量 $\widehat{f}^{\mathrm{lin}}$ に関して、十分大きな n のもと以下が成立する：

$$\mathbb{E}_{f^*}\left[\|\widehat{f} - f^*\|^2_{L^2(P_X)}\right] < \mathbb{E}_{f^*}\left[\|\widehat{f}^{\mathrm{lin}} - f^*\|^2_{L^2(P_X)}\right].$$

この結果により、線形推定量のクラスに分類される推定量は最適性を達成しないため、最適性を持つ DNN による推定量を優越できないことが理論的に示されている。

参 考 文 献

Imaizumi, M., & Fukumizu, K. (2018). Deep Neural Networks Learn Non-Smooth Functions Effectively. arXiv preprint arXiv:1802.04474.