文部科学省科学技術試験研究委託事業

数学・数理科学と諸科学・産業との協働による
イノベーション創出のための研究促進プログラム

「統計科学の新展開と産業界・社会への応用」
講演要旨集

主催：金沢大学人間社会研究域　経済学経営学系
共催：文部科学省（統計数理研究所）・統計関連学会連合

2016年9月5-6日

金沢大学角間キャンパス
総合教育講義棟A1教室

目次　ワークショップ プログラム

9月5日（月）

13:00～　数学協働プログラムの活動紹介　伊藤 聡（統計数理研究所）

(pp.1-5) 13:00～15:00 欠測データの解析：基礎理論と実践的な方法論の発展

オーガナイザー：星野崇宏（慶應義塾大学）・野間久史（統計数理研究所）
座長：星野崇宏（慶應義塾大学）・野間久史（統計数理研究所）

①諸分野での欠測データ解析の動向と研究テーマの潮流のレビュー　星野 崇宏（慶應義塾大学）
②多重代入法におけるロバストな推測方法　野間 久史（統計数理研究所）
③重み付き推定方程式と二重ロバスト推定法　逸見 昌之（統計数理研究所）
④臨床試験における欠測データの解析, MMRM とその最新の研究　五所 正彦（筑波大学）

(pp.6-11) 15:30～17:30 超高速グラフ列挙法と統計学への応用

オーガナイザー：水田正弘（北海道大学）・湊真一（北海道大学）・栗原考次（岡山大学）
座長：栗原考次（岡山大学）

①データ解析における超高速グラフ列挙技法および連結成分列挙技法の活用について　水田 正弘（北海道大学）
②離散構造処理系プロジェクトと超高速グラフ列挙技法　湊 真一（北海道大学）
③公的統計の地域別集計分析への利用可能性　谷道 正太郎（統計センター）
④一票の格差が小さな選挙区割の列挙
　川原 純（奈良先端科学技術大学院大学）・堀山 貴史（埼玉大）・
　堀田 敬介（文教大）・湊 真一（北海道大学）
⑤空間データに対するホットスポット検出手法の性質評価について
　石岡 文生（岡山大学）・栗原 考次（岡山大学）・水田 正弘（北海道大学）

9月6日（火）

(pp.12-17) 10:00～12:00 スポーツアナリティクスの広がり

オーガナイザー：酒折文武（中央大学）
座長：酒折文武（中央大学）

①野球選手における脊椎・体幹部障害のマネジメント - 競技復帰時期予測の苦労 -　加藤 欽志（福島県立医科大学）
②成績・試合情報をもとに先発投手を予想する方法
　大川 恭平（データスタジアム株式会社）・宮崎 誠也（東京工業大学）・
　金沢 慧（データスタジアム株式会社）・上原 早霧（データスタジアム株式会社）
③MLBトラッキングデータを用いた捕手のフレーミング評価法について
　永田 大貴（慶応義塾大学）・南 美穂子（慶応義塾大学）
④サッカートラッキングデータから守備戦術技能を測る
　松岡 弘樹（筑波大学）・猶本 光（筑波大学）・田原 康寛（筑波大学）・
　見汐 翔太（筑波大学）・安藤 梢（筑波大学）・西嶋 尚彦（筑波大学）
⑤サッカートラッキングデータに関する統計的モデリング　酒折 文武（中央大学）

(pp.18-23) 13:00～15:00 ライフイノベーションを推進するバイオメディカルビッグデータ解析の新潮流

オーガナイザー：島村徹平（名古屋大学）・新井田厚司（東京大学）・白石友一（東京大学）
座長：島村徹平（名古屋大学）・新井田厚司（東京大学）・白石友一（東京大学）

①局所距離相関に基づくモジュレーター因子の網羅的探索法
　島村 徹平（名古屋大学）・松井 佑介（名古屋大学）・宮野 悟（東京大学）
②大規模がんゲノム変異データマイニングのための統計学的手法　白石 友一（東京大学）
③がんの進化シミュレーションによる腫瘍内不均一性生成原理の探索　新井田 厚司（東京大学）
④多重検定補正法の生命系大規模データへの応用　瀬々 潤（産業技術総合研究所）
⑤共発現解析による軽度認知障害の血漿 microRNA マーカーの検出
　茅野 光範（帯広畜産大学）・檜垣 小百合（国立長寿医療研究センター）・
　佐藤 準一（明治薬科大学）・松本 健治（国立成育医療研究センター）・
　滝川 修（国立長寿医療研究センター）・新飯田 俊平（国立長寿医療研究センター）

(pp.24-29) 15:30～17:30 ヒトゲノムデータの遺伝統計解析

オーガナイザー：鎌谷洋一郎（理化学研究所）
座長：鎌谷洋一郎（理化学研究所）

①Chromatin configuration QTL mapping using ATAC-seq
　熊坂 夏彦（英国サンガー研究所）・Andrew Knights（英国サンガー研究所）・Daniel Gaffney（英国サンガー研究所）
②全ゲノムシークエンスによる肝癌の変異の包括的解析
　藤本 明洋（京都大学）・古田 繭子（理化学研究所）・十時 泰（国立がん研究センター）・角田 達彦（理化学研究所）・
　加藤 護（国立がん研究センター）・柴田 龍弘（国立がん研究センター）・中川 英刀（理化学研究所）
③遺伝統計解析で迫る疾患病態の解明とゲノム創薬　岡田 随象（大阪大学）
④ゲノムワイド関連解析による高血圧遺伝子の解明　竹内 史比古（国立国際医療研究センター）
⑤統計遺伝学モデルを用いた多因子疾患の発症リスク予測法　八谷 剛史（岩手医科大学）

欠測データの解析：基礎理論と実践的な方法論の発展

諸分野での欠測データ解析の動向と研究テーマの潮流のレビュー

慶應義塾大学経済学部　　星野　崇宏

医学・疫学や経済学・社会学・心理学・政治学をはじめとする社会科学、さらには企業のマーケティングなどヒトにかかわるデータを扱う様々な分野では部分的あるいは特定のユニット・測定単位全体にわたるデータの欠測が生じるが、各分野の応用研究では長い間あまり注目を浴びておらず、欠測を無視した解析によって生じる様々な問題点が過小評価されてきたと言ってよい。しかしこの十年ほど各分野で「データの欠測を無視あるいは軽視して単純な解析を行うことで大きなバイアスが生じる可能性がある」といった問題意識が共有され、より適切な解析を行わないと一定レベル以上の学術誌の査読ではリジェクトされる、全米学術研究会議（National Research Council,2010）によって欠測データ解析に関する報告書が発行されるなど、応用研究でも欠測データの適切な扱い方の知識に対するニーズが高まってきた。また、企業の解析実務などでも欠測値に適切な値を代入することが求められることが増えてきている。

欠測データに関連する統計学での理論的研究については、Rubin(1976)の欠測メカニズムの議論をエポックメイキングとして盛んにおこなわれており、特にこの 20 年で統計学の非常に重要な研究分野として認知されてきている。

本企画セッションではこのような欠測データ解析の応用研究での重要性の認知と統計学の理論研究の深化を踏まえ、主に理論的な観点から 4 つの講演を行う。星野を除く 3 名の応用研究は医学薬学分野であり、星野の応用分野は主に社会科学である。医学分野と社会科学での欠測データ解析はその方法論は共通しながらも独自に発展した経緯があるが、近年では両者の交流が盛んとなってきている。具体的には Rubin(1987)に端を発する多重代入法アプローチはもともと社会科学での公的統計の個票公開と利用において代入者と解析者が異なりえることを考慮した手法であるが、近年では医学分野で非常によく応用研究で利用されてきた。一方医学分野では推定方程式アプローチを元に様々な手法が発展し、社会科学においても利用されるようになってきている。このような現状を踏まえて、本講演では特に多重代入法として医学研究でも盛んに利用されている連鎖式による多重代入法(Multiple Imputation by Chained Equation; MICE, van Buren,2012)の問題点と関連する研究を中心に、医学統計や計量経済学などでの欠測データ解析に関する理論研究と応用研究のレビューを紹介し、続く 3 つの講演への橋渡しを行うものである。

引用文献

Liu, J.et al. (2014). On the stationary distribution of iterative imputations. *Biometrika*, 101(1), 155-173.

National Research Council. (2010) *The Prevention and Treatment of Missing Data in Clinical Trials,* Washington DC: National Academic Press.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63(3), 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York, NY: Wiley.

高井啓二・星野崇宏・野間久史(2016)『欠測データの統計科学―医学と社会科学への応用』岩波書店

Van Buuren, S. (2012). *Flexible Imputation of Missing Data,* Chapman & Hall/CRC, Boca Raton, FL.

多重代入法におけるロバストな推測方法

統計数理研究所　野間 久史

多重代入法（multiple imputation; Rubin, 1987）は，医学研究における不完全データの解析において，現在，最も広く利用されている方法のひとつであり，その原理の明快さと理論上の有効性から，今後，ますます実践での普及も進むものと予想される．多重代入法の基本的な原理は，適当なベイズモデルのもとで，補完値を欠測データの事後予測分布から生成する Rubin (1987) による正則（proper）な補完法によるものであるが，一般的な統計ソフトウェアでは，マルコフ連鎖モンテカルロ法を利用した正則な補完法は，限定的なモデルのもとでしか利用可能ではなく（例えば，SAS ver 9.4 では，proc mi が対応しているのは，多変量正規分布の構造が仮定できる場合のみである），実践的にはむしろ回帰モデルに基づく近似法や予測平均マッチングのような非正則（improper）な補完法のほうが広く利用されている．

実践上の悩ましい問題のひとつとして，この補完値の生成モデルの妥当性の問題がある．欠測データの分布そのものは，基本的には，関心のある治療・曝露効果の推測においては局外要因であり，強い理論的制約を置くことは望ましくないが，実際には，補完値の生成のために少なくとも便宜上付加的なパラメトリックな仮定を置く必要がある．その仮定に誤りがあると，Rubin (1987) による標準的な推測の枠組みでは，当然ながら関心のある治療・曝露効果の推測の妥当性も失われる．1990 年代後半以降，いくつかの研究により，このような非正則な補完法を含めた一般的な枠組みのもとでの推定方程式の理論に基づく頻度論的な推測理論が構築されており（例えば，Robins and Wang (2000)），Rubin (1987) による理論的枠組みとは異なる，モデル誤特定に対してロバストな推測を行うための方法論が飛躍的に発展している．本講演では，これらの非正則な補完法を含めた一般的な枠組みのもとでの多重代入法のロバストな推測理論について，演者らの近年の研究成果も踏まえた平易な総説・解説を行い，特に実践的な方法論に関しての最新の知見を共有することを目的とする．

参考文献

Robins, J. M., and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113-124.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

重み付き推定方程式と二重ロバスト推定法

統計数理研究所　逸見　昌之

欠測を含むデータの統計的な解析法には様々なものがあるが、中でも重み付け推定法 (Inverse Probability Weighting) と二重ロバスト推定法 (Doubly Robust estimation) は、1990 年代後半から主に医学統計学の分野で研究され、発展してきたセミパラメトリックな手法である。通常、欠測のメカニズムが MAR (Missing At Random)、すなわちデータの欠測確率が観測データのみに依存するという仮定の下で、前者の重み付け推定法では、欠測が全くない場合の推定方程式 (完全データ推定方程式) に対して、データの観測確率の逆数で重み付けられた推定方程式 (観測データ推定方程式) を用いる。例えば、i 番目の個体に対するデータを $\boldsymbol{Z}_i = (\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T)^T$ $(i = 1, \ldots, n)$ とし、$\boldsymbol{X}_i$ は常に観測され、$\boldsymbol{Y}_i$ (の成分の一部) は欠測し得るとして、R_i を $\boldsymbol{Y}_i$ の観測の指示変数とする (すなわち、$R_i = 1$ ならば $\boldsymbol{Y}_i$ の各成分は全て観測され、$R_i = 0$ ならば欠測し得るとする)。そして、欠測メカニズムについて $P(R = 1|\boldsymbol{Z}_i) = P(R = 1|\boldsymbol{X}_i)$ (MAR) が成り立つことを仮定し、ある興味あるパラメータ $\boldsymbol{\theta}$ に対する不偏な完全データ推定関数 $\boldsymbol{u}(\boldsymbol{z}, \boldsymbol{\theta})$ が与えられたとすると、$\boldsymbol{\theta}$ に対する重み付き推定方程式は

$$\sum_{i=1}^{n} \frac{R_i}{w(\boldsymbol{X}; \hat{\boldsymbol{\alpha}})} \boldsymbol{u}(\boldsymbol{Z}_i, \boldsymbol{\theta}) = \boldsymbol{0}$$

となる。但し、$w(\boldsymbol{X}; \boldsymbol{\alpha})$ は $\boldsymbol{Y}_i$ の (条件付き) 完全観測確率 $P(R = 1|\boldsymbol{X}_i)$ に対するパラメトリックモデル (通常はロジスティック回帰モデル) で、$\hat{\boldsymbol{\alpha}}$ は $\boldsymbol{\alpha}$ の最尤推定量である。この推定方程式の解として得られる $\boldsymbol{\theta}$ の推定量 (IPW 推定量) は、もしこのモデルが正しければ一致性および漸近正規性を持つが、誤特定されていればその性質は崩れる。また、正しく特定されていたとしても、$\boldsymbol{Y}_i$ の各成分が全て観測されているデータしか用いていないので、推定の (漸近) 効率はあまり良くない。二重ロバスト推定法は、上記の重み付き推定方程式に、ある特別な項を付加することによってその欠点を改善するもので、以下のような推定方程式を用いる。

$$\sum_{i=1}^{n} \left[\frac{R_i}{w(\boldsymbol{X}; \hat{\boldsymbol{\alpha}})} \boldsymbol{u}(\boldsymbol{Z}_i, \boldsymbol{\theta}) + \left\{ 1 - \frac{R_i}{w(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}})} \right\} \boldsymbol{m}(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}) \right] = \boldsymbol{0}$$

但し $\boldsymbol{m}(\boldsymbol{X}_i; \boldsymbol{\beta})$ は、条件付き期待値 $E\{\boldsymbol{u}(\boldsymbol{Z}_i, \boldsymbol{\theta})|\boldsymbol{X}_i\}$ を、条件付き分布 $p(\boldsymbol{y}|\boldsymbol{x})$ に対するパラメトリックモデル $p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\beta})$ によって計算したもので、$\hat{\boldsymbol{\beta}}$ は $\boldsymbol{\beta}$ の最尤推定量である。この推定方程式の解として得られる推定量は、(i) ２つのパラメトリックモデル $w(\boldsymbol{X}; \boldsymbol{\alpha})$ と $p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\beta})$ のうち、少なくともどちらか一方が正しく特定されていれば、一致性および漸近正規性を持つ、(ii) ２つのモデルがどちらも正しく特定されていれば、IPW 推定量よりも漸近分散は小さくなる、という特徴を有し、(i) の性質から、二重ロバスト推定量と呼ばれている。

本発表では、まず単純な平均の推定を行う場合において、重み付け推定法と二重ロバスト推定法の仕組みとこれまでの発展を概観し、また、回帰分析における欠測や経時測定データ解析における脱落の問題について議論する。

臨床試験における欠測データの解析，MMRM とその最新の研究

筑波大学 五所正彦

臨床試験において，データ欠測は避けては通れない問題である．近年，この問題は多くの学術論文で取り上げられ，2010 年に発表された全米学術研究会議の報告書は，データ欠測の防止策やその取扱い方法を詳しく論じている．この報告書は，臨床試験データの主解析として，長きにわたり頻用されてきた Last Observation Carried Forward 法の利用を否定した．その影響もあり，最近では主解析の方法として，Mixed Models for Repeated Measures (以下，MMRM) 法を採用することが多い．MMRM 法は普遍的に使用できるものではないが，Missing at Random の下で妥当であること，その実行が比較的容易なこと等から，実務統計家にとって有用なツールといえる．本発表では，MMRM 法に関連する 2 つの最新の研究を紹介する．

第 1 の研究では，連続変数を経時的に測定する臨床試験での小標本問題を扱う．MMRM 法を適用する際, 反復測定されるデータの共分散構造として unstructured 構造を指定することが定石である．しかし，小標本の場合，最適化問題により収束解としてパラメータ推定値が得られないことがある．代替法として，より単純な共分散構造を指定し，ロバスト分散を利用することが考えられるが，ロバスト分散推定量には小標本バイアスの問題がある．本研究では，Kauermann & Carroll (2001) および Mancl & DeRouen (2001) によって提案された小標本バイアスを補正したロバスト分散推定量を MMRM 法に適用した．

第 2 の研究では，合成変数が経時的に測定される臨床試験データを扱う．このデータを解析する際には，2 種類の欠測，すなわち，被験者レベルの欠測（被験者脱落等）と項目レベルの欠測（合成変数を構成する一部の項目が未測定等）を考える必要がある．このときの主解析として標準的に利用されるアプローチは，i) 項目レベルに欠測がある被験者を除外した解析，もしくは，ii) 単一値代入法により項目レベルの欠測を補完した上で計算した合成変数の解析，であろう．本研究では，2 種類の欠測を同時に扱えるように MMRM 法を拡張し，その有用性を検討した．

参考文献

[1] Mancl LA and DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; 57: 126–134.

[2] Kauermann G and Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; 96: 1387–1396.

超高速グラフ列挙法と統計学への応用

データ解析における超高速グラフ列挙技法および連結成分列挙技法の活用について

北海道大学情報基盤センター　水田正弘

１．はじめに

統計学の歴史において、多くの他分野との出会いが学問の発展に寄与している。確率論との出会い、コンピュータとの出会い、ビッグデータとの出会いが新たな分野を切り開いていった。本企画セッション「超高速グラフ列挙法と統計学への応用」も小さな出会いから始まった。最近、同じ専攻に所属するようになった湊教授は、離散構造に関するアルゴリズムについて革新的な技法を提案している。また、長年、共同研究などでお付き合いさせていただいている栗原教授は、空間データ、特に Hot Spot について国際的な業績を出されている。この 2 つの融合をきっかけとして、新たな分野への展開を目指す。

２．超高速グラフ列挙技法および連結成分列挙技法

統計学で利用される数学としては解析学を主体とした確率論が大きな地位を占めている。それに対して、ノンパラメトリックを中心とした分野などでは、離散的な組合せ数学が広く使われている。しかし、いくつかの離散的な問題は、サイズに関する強い制限が存在する。それに対して、フロンティア法と呼ばれる方法により、指定した条件を満たす無向グラフや部分集合を列挙する問題において計算可能な問題の範囲が拡大された。

３．データ解析への活用例

空間データにおける Hot Spot の検討において、連結部分集合をすべて列挙できれば、新たな検出法の提案、および従来の scan 法の評価が可能になる。現時点で、地点数が 100、辺の数が 246 である SIDS データ（乳幼児突然死データ）について、すべての連結部分集合の列挙が実行できる。さらに、北海道の市町村に対応した 177 点、辺の数 452 でも、連結成分数が 11 以下であれば列挙できている。これは、地方自治レベルにおける市町村の活性化に寄与できるパフォーマンスであると言える。

その他、樹木構造接近法、クラスタ分析、シンボリックデータ解析法などに本アプローチが利用できる可能性があると思われる。また、[1]では、ビッグデータに対するミニデータを報告するが、グラフ自体をミニデータとして扱うことも可能である。

詳細な理論や興味深い適用例について､本企画セッションの他の講演で紹介される予定である。

参考文献

[1] 水田正弘 (2016) ビッグデータを扱うためのミニデータアプローチについて、統計関連学会連合大会

離散構造処理系プロジェクトと超高速グラフ列挙技法

北海道大学　　湊 真一

1　離散構造処理系と列挙索引化

論理関数や組合せ集合などの離散構造を表す大規模データを計算機上にコンパクトに表現し演算処理を効率よく行う技法は，計算機科学の様々な応用分野に共通する基盤技術として非常に重要であり，現代社会に対する大きな波及効果を持つ．昨年度までの約 6 年間に渡って展開された「JST ERATO 湊離散構造処理系プロジェクト」での研究活動により，様々な離散構造を統合的に演算処理する技法の体系化と，分野横断的かつ大規模な実問題への応用が進められてきた．その中でも，種々の制約条件を満たすグラフ構造を ZDD（ゼロサプレス型二分決定グラフ）[4] を用いて全列挙して索引化する技法（通称フロンティア法）は，Knuth が近年示したパス列挙アルゴリズム Simpath[3] をさらに一般化したものであり，多くの実用的な問題で従来より桁違いに優れた性能を示すことから，今後の発展が大いに期待されている．ERATO の研究活動を引き継ぎ，離散構造処理系のコアとなる部分に研究者が集まる「場」を継続的に提供し，情報科学の様々な応用分野の競争力の源泉となるアイデアを醸成し続けることを目的として，科研費・基盤研究 (S)「離散構造処理系の基盤アルゴリズムの研究」が採択され，昨年度より 5 年間の研究活動が続けられている．

2　超高速グラフ列挙索引化の技法

ZDD は，複数の因子の組合せの集合（一般には組合せ爆発を起こす）を，コンパクトに圧縮して網羅的に列挙・索引化し，しかも圧縮したままの状態で高速に演算処理を行うアルゴリズム技術である．本プロジェクトにおける最近の主要な成果として，以下が挙げられる．

(1) 実用規模の配電網構成（10^{63} 通りの正解数）の網羅的な列挙・索引化と損失最小化に世界で初めて成功
(2) 超高速な数え上げアルゴリズムに基づく正確で実用的な統計多重検定法 LAMP の開発
(3) 日本科学未来館での研究成果展示「フカシギの数え方」（展示作品が YouTube で 180 万ビュー超）

基盤的な研究成果である組合せ集合の演算処理技術を，社会的・産業的な応用につなげるための解りやすい切り口の 1 つとして，本プロジェクトでは，ZDD を用いて種々のグラフ列挙問題を高速に解くソフトウェアツール「Graphillion」[2] を開発し，オープンソフトウェアとして公開している．そして，その Graphillion の簡単な使い方から，実問題への応用例，処理系内部のアルゴリズム技法の解説までをカバーした技術専門書 [1] を出版し，研究成果の普及に努めている．

3　統計分野への応用と今後の展望

列挙・索引化の技法は，確率解析や検定など統計分野と密接な関わりを持つ．本プロジェクトでは以前より，選挙学会の専門家と連携し，選挙区割りの列挙・索引化と 1 票の格差の分析を行う新手法を提案している．ZDD 技法により，これまで不可能と思われていた網羅的解析が実用規模の問題で可能になってきている．さらに最近では，全国の都道府県の隣接ブロックの総数（約 953 億通り）を ZDD を用いて網羅的に列挙・索引化することに成功しているが，これは廃藩置県が行われた明治以来，初めて可能になったことである．疫学調査におけるいわゆるホットスポットの発見や有意性検定への応用が期待されている．また都道府県に限らず，国内外の州や大都市の行政区・市町村などへの適用も可能と見られている．詳細については本企画セッションで紹介される予定である．

アルゴリズム技術の研究においては,「最適化」と「列挙」は車の両輪と言える．最適化の技法は世界的に競争が激しいが，列挙系の技法はまだ未開拓の領域が多く，伝統的に日本が強い分野でもある．本研究プロジェクトでは，離散構造処理系による列挙と圧縮索引化の技法で世界をリードし，さらに発展させることを目指している．数学的理論と工学的応用の中間に位置する「Art 層」の研究者コミュニティの維持発展を図るとともに，周辺分野の大型研究プロジェクトとも連携しながら研究活動を進めている．今回の企画セッションをきっかけに統計分野との連携が活発化すれば幸いである．

参考文献

[1] 湊真一（編）ERATO 湊離散構造処理系プロジェクト（著）. 超高速グラフ列挙アルゴリズム－〈フカシギの数え方〉が拓く，組合せ問題への新アプローチ－. 森北出版, 2015.

[2] Takeru Inoue and et al. Graphillion. http://graphillion.org/, 2013.

[3] D. E. Knuth. *The Art of Computer Programming: Bitwise Tricks & Techniques; Binary Decision Diagrams*, Vol. 4, fascicle 1. Addison-Wesley, 2009.

[4] Shin-ichi Minato. Zero-suppressed BDDs for set manipulation in combinatorial problems. In *Proc. of 30th ACM/IEEE Design Automation Conference (DAC'93)*, pp. 272–277, 1993.

公的統計の地域別集計分析への利用可能性

独立行政法人統計センター　谷道　正太郎

１．はじめに

地域の実情を把握・分析し、各種計画の企画立案をはじめ様々な活動を行うための基盤的情報として政府統計データが収集・整備されており、これらのデータは政府統計データのポータルサイト e-Stat（イースタット）(http:www.e-stat.go.jp) を通じて提供されている。

また、政府におけるオープンデータ化が進展している中で、統計分野は政府全体の取組の牽引役として、データ利用環境の高度化が進められている。

本報告では、地域分析のための基礎的情報として公開されている公的統計データについて紹介するとともに、超高速グラフ列挙法の公的統計の地域別集計分析への利用可能性について報告する。

２．公的統計の地域別集計分析について

公的統計においては、様々なデータを都道府県別、市区町村別や地方別といった形で提供・分析している（例えば、「社会・人口統計体系」として、人口・世帯、自然環境、経済基盤、行政基盤、教育、労働、居住、健康・医療、福祉・社会保障など、国民生活全般の実態を示す様々な地域別統計データを収集・加工しており、これを、都道府県別に約 440 種類、市区町村別に約 100 種類の統計データに編成している）。本報告では、公的統計データに関し、隣接情報を用いた地域性の集計分析など、新たな形での公的統計データの応用の可能性について報告する。

一票の格差が小さな選挙区割の列挙

奈良先端科学技術大学院大学　川原 純
埼玉大学　堀山 貴史
文教大学　堀田 敬介
北海道大学　湊 真一

1. はじめに

日本の選挙制度において選挙区を策定する際には、2016 年現在では原則として市区群を構成単位とし、与えられた数の地区に分割することになっている。選挙区の最大人口と最小人口の比を一票の格差と呼び、なるべく小さくすることが求められている。市区群を頂点とし、市区群同士が境界を共有する場合に辺を張ったグラフを考えると、選挙区割を求める問題は、グラフの連結成分分割を求める問題と考えることができ、一票の格差を最小化するグラフ最適化問題として定式化できる。本研究では、一票の格差が最小の選挙区割を一つ求めるだけではなく、指定した格差より小さな選挙区をすべて列挙することを考える。

2. 提案手法

連結成分分割の列挙には、与えられたグラフに対して部分グラフを列挙するための手法であるフロンティア法[1]を用いる。フロンティア法では、例えばグラフ上の 2 点間の経路をすべて列挙することが可能である。経路だけでなく、全域木やマッチングなど、様々なグラフ構造を列挙できる。列挙した部分グラフはゼロサプレス型二分決定グラフ（ZDD）[2]と呼ばれるデータ構造により、圧縮して保持される。このデータ構造は部分グラフを単に圧縮するだけではなく、和集合や共通部分の計算、解の数え上げ、指定した解のフィルタリング、一様ランダムサンプリングなどの演算が高速に行えるという特徴をもつ。本研究ではフロンティア法を連結成分分割に適用するための手法の開発を行った。以下に列挙を行った結果を示す。大阪府の結果を見ると、10^{29} を超える連結成分分割を、通常の計算機のメモリ量で扱うことができる。

府県名	頂点数	辺数	区割数	区割解の個数	計算時間
茨城県	41	87	7	11,893,998,242,846	0.22 秒
神奈川県	50	114	18	2,356,100,754,933,627,279,208	0.48 秒
大阪府	69	161	19	172,119,047,292,061,592,625,618,553,239	12.20 秒

一票の格差を指定した値以下に限定することも可能である。茨城県の場合の結果を示す。

指定格差	解の個数	時間	指定格差	解の個数	時間
1.1	16,252	7.13 秒	1.3	5,574,807	615.54 秒
1.2	515,982	112.50 秒	1.4	25,730,669	1925.21 秒

参考文献

[1] ERATO 湊離散構造処理系プロジェクト著、湊　真一編集。超高速グラフ列挙アルゴリズム―〈フカシギの数え方〉が拓く，組合せ問題への新アプローチ―、森北出版、2015。

[2] Shin-ichi Minato. Zero-suppressed BDDs for set manipulation in combinatorial problems. In Proceedings of the 30th ACM/IEEE Design Automation Conference, pages 272–277, 1993.

空間データに対するホットスポット検出手法の性質評価について

岡山大学　石岡文生
岡山大学　栗原考次
北海道大学　水田正弘

1. はじめに

ある地方における感染症の発生状況や，自然災害におけるハザードマップ等に対し，「どの地域で問題が生じているのか」を特定することは，環境保全や安全管理のための対策を講じたり，その原因解明の重要な手がかりとなる。そんな中，Kulldorff（1997）が提唱した空間スキャン統計量は，解析を行う対象地域全体において，ある特定の地域に有意な集積性（hotspot cluster；ホットスポット）が存在しているか否かを検定するための方法として広く利用されている。

2. 空間スキャン統計量

ホットスポット候補となる「一つ以上のある連結した地域」をウィンドウZと考え，Z 内の観測値を$o(Z)$，その期待値を$E(Z)$ と表すと，帰無仮説：$o(Z) = E(Z)$，対立仮説：$o(Z) > E(Z)$ の下で，ポアソン分布に基づく尤度比統計量は，

$$\lambda(Z) = \left(\frac{o(Z)}{E(Z)}\right)^{o(Z)} \left(\frac{o(Z^c)}{E(Z^c)}\right)^{o(Z^c)} \cdot I(o(Z) > E(Z))$$

となる。ホットスポットの同定には，すべてのウィンドウの中から最大尤度比となるウィンドウをみつける必要があるが，解析を行う対象地域の地域数が多くなると，「すべてのウィンドウ」は膨大な数になるため，計算コスト等の面からも現実に求めるのは困難である。そのため，高い尤度比となるウィンドウを，効率よく見つけるための様々な工夫がなされた手法が，これまでに数多く提案されている。

3. ホットスポット検出手法の性質評価

本報告では，実際の空間データに対して ZDD を用いた列挙技法（湊, 2011）を利用することにより，解析対象領域のすべてのウィンドウを取得し，そこから最大尤度比となるホットスポット候補の同定を試みる。加えて，これまでに提案されている既存手法との比較・検討を行い，その性質の違いや現状の課題等について考察する。

参考文献

Kulldorff M. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26, 1481-1496, 1997.

湊真一. BDD/ZDD を基盤とする離散構造と演算処理系の最近の展開. 電子情報通信学会 基礎・境界ソサエティ Fundamental Review, 4(3), 224-230, 2011.

スポーツアナリティクスの広がり

野球選手における脊椎・体幹部障害のマネジメント
- 競技復帰時期予測の苦労 -

福島県立医科大学　整形外科　加藤欽志

【はじめに】

当科では、2012 年 10 月以降、38 名(投手 15 名、野手 23 名、平均年齢 24.3 歳)のプロ野球選手の脊椎・体幹部障害に対する診療を経験している。プロ野球選手に対する診療においては、正確な診断と適切な治療のみならず、初診時に正確な「復帰時期の予測」を求められる。「復帰時期の予測」は、チーム戦略にも影響を及ぼす重要な情報であるが、過去のデータに基づいた復帰時期予測の方策は、これまで確立されていない。初診の担当医が、医学情報を含めた様々な情報を総合して、経験則に基づいて予測しているのが現実であり、様々な苦労がある。

【復帰時期予測の苦労】

実際の復帰時期は、治療の経過だけではなく、ポジション、一軍登録抹消が行われるか否か(一旦登録が抹消されると 10 日間は再登録できない)、選手の立場、選手の取り組み、および選手の心理状態などにも左右される。また、チーム事情から、医学的に適正な復帰時期よりも、早めの競技復帰を求められる場合があり、完治前の選手のパフォーマンス評価、再発リスク評価などが課題となる。

疾患別における特徴としては、体幹部外傷(肋骨骨折など)の選手は、疲労骨折や腹斜筋損傷などの場合を除き、ほとんど選手が、医学的な完治時期(骨癒合)の前から競技復帰をしている。例えば、肋骨骨折後の選手が、骨癒合に至る前に復帰する場合には、骨折の癒合状況、折れている方向、位置(負荷のかかる動作が異なる)、ポジション、投・打側などから、防具の使用やプレー動作における注意点の指導など、きめ細かい対応が求められる。一方で、神経症状を伴う脊椎疾患(腰椎椎間板ヘルニアや胸椎黄色靭帯骨化症など)は、診断や治療の経過が理想通りに運ばない場合もあり、正確な競技復帰時期の予測は困難な場合が多い。

【まとめ・展望】

本講演では、プロ野球選手の脊椎・体幹部障害に対する診療経験から、現場における問題点について考察し、データ活用の可能性について考察する。また、現在行っている選手の脊椎疾患のリスク評価等についても紹介する。

【参考文献】

1. 加藤欽志ほか：プロ野球選手の腰下肢痛に対する診断と治療. Locomotive Pain Frontier 3:92-99,2014.
2. 加藤欽志ほか：腰部障害 - 腰椎分離症と腰椎椎間板ヘルニア - 臨床スポーツ医学 32. 臨時増刊号「野球の医学」 213-219 2015

成績・試合情報をもとに先発投手を予想する方法

データスタジアム株式会社ベースボール事業部　大川 恭平
東京工業大学　宮﨑 誠也
データスタジアム株式会社ベースボール事業部　金澤 慧
データスタジアム株式会社ベースボール事業部　上原 早霧

１．はじめに

野球において投手は試合の行方を左右する．試合に先発する投手の情報は，チームにとっても，試合を見守るファンにとっても重要である．日本プロ野球（以下 NPB）では予告先発制度が導入されて以降，試合の前日にどの投手が先発するか知ることが出来るようになった．しかし，翌々日以降の先発投手に関しては公式な情報はなく，事前に先発投手を知ることはできない．そこで，本研究では NPB12 球団を対象に，投手の登板記録やチームの試合スケジュールなどを基にモデルを作成し，直近 7 日分の先発投手を予測することを試みた．

２．予測モデル

本研究では先発投手の決定に，どの要素がより影響を与えているか明らかにするため，ランダムフォレストを用いて予測モデルを作成した．2014 年と 2015 年の NPB 公式戦のデータを用い，2014 年データを学習データ，2015 年データを検証用データとした．予告先発が発表された時点で，投手に関するデータやチームに関するデータを基に翌日から 7 日間の先発投手を予測する．なお，本研究では教師データに対して回帰を行い，求めた予測値を [0, 1] に変換して登板確率として算出している．

３．結果

構築したモデルを用いて 2015 年に行われた NPB 公式戦全試合の先発投手の予測をおこなった．結果を表 1 に示す．各試合，最も登板確率が高い値の投手を予測投手とし，実際の先発投手と比較することで正答率を求めている．

表1 2015年NPB先発投手正答率

予測対象日	正答率
1日後	100%
2日後	73%
3日後	61%
4日後	59%
5日後	58%
6日後	60%
7日後	55%

また，2014 年と 2015 年を学習用データとして再度モデルを構築し，2016 年のデータに対しても予測した結果は当日に示す予定である．

参考文献

[1] 平井 有三（2012）．はじめてのパターン認識．森北出版株式会社．
[2] 宮﨑誠也．“予告先発より先を予想する －機械学習の手法を用いて－”．Baseball LAB．2016．http://www．baseball-lab．jp/column/entry/294/，（2016-6-26）．

MLB トッラキングデータを用いた捕手のフレーミング評価法について

慶應義塾大学大学院　永田 大貴
慶應義塾大学　南 美穂子

1　はじめに

近年アメリカにおいて，PITCHf/x などのトラッキングシステムにより得られるデータの解析が活発に行われてきている．本研究では捕手のフレーミングというテクニックに着目し，PITCHf/x により得られる座標データを用いて評価をおこなった (Pavilidis and Brooks(2014))．フレーミングとは，投球に対してその投球をよりストライクに見せる捕球技術であり，日本においてはキャッチングなどとも表現される．

2　一般化加法混合モデル (GAMM) を用いた解析

目的変数 $Y_i \sim Bernoulli(p_i)$ を、見逃しストライクかボールかをとる二値変数としたロジスティック回帰モデルを考える．審判のストライク判定に関しては，投球のコースや高さによって決まる．トラッキングデータにより取得できるホームプレート到達点のデータを用いて回帰分析を行いたい．プレート到達点に対してスプライン関数を適用した一般化加法モデル (Woods(2006)) を用いる．プレート到達時における x 座標，z 座標をそれぞれ $plate.x, plate.z$ とする．

$$\log \frac{p_i}{1-p_i} = \alpha + f(plate.x_i, plate.z_i)$$

各投球に対するコールストライク確率を推定し，実際の判定との差を考えることによりフレーミングの貢献度を算出し，Run Value を用いて得点に換算する．Run Value はあるプレーが得点に対してどれほどの価値があるかを平均化した指標である．
次にモデルの拡張を考える．ストライクの判定に対して関係している要因は様々考えられる．ここでは，審判・投手・打者の影響を変量効果として捉え，線形和に組み込むことにより拡張を行った混合効果モデルによる推定を行う．各変量効果の推定を行った上で，フレーミングの貢献度を測ることを目指す．

$$\log \frac{p_i}{1-p_i} = \alpha + f(plate.x_i, plate.z_i) + \gamma_i^u + \gamma_i^p + \gamma_i^b$$

$$\gamma_i^u \sim N(0, \sigma_u^2)\ , \gamma_i^b \sim N(0, \sigma_b^2)\ , \gamma_i^p \sim N(0, \sigma_p^2)$$

モデルの推定結果や，算出したチームごとの Run Value については当日示すことにする．

参考文献

[1] Wood,S.N.(2006). Generalized Additive Models : an introduction with R. Chapman and Hall/CRC. New York.
[2] Pavilidis,H.,and Brooks,D.(2014). https://www.baseballprospectus.com/article.php?articleid=22934.

サッカートラッキングデータから守備戦術技能を測る

筑波大学大学院　松岡弘樹，猶本光，田原康寛，見汐翔太，安藤梢，
筑波大学　西嶋尚彦

１．はじめに

近年，サッカーゲームのトラッキングデータが使用されてきた．ゲームスピード，攻撃・守備組織の centroid position，縦幅，横幅を用いた分析（Frenchen et al., 2011）が可能となり，サッカーの守備戦術技能の計量へ一歩進展がみられた．しかし，守備戦術技能を計量する測定項目，守備結果への関連性は十分に明らかにされていない．本研究の目的は，サッカーゲームのトラッキングデータから構成した守備戦術技能の測定項目の妥当性と守備結果の達成基準を分析することであった．

２．方法

データスタジアム株式会社より提供された 2016 年明治安田生命 J 1 リーグ 1st ステージ第 11 節浦和レッドダイヤモンズ vs 大宮アルディージャの試合におけるボールタッチデータとトラッキングデータを用いた．分析対象は守備成功プレー76，守備失敗プレー13，合計 89 プレーであった．守備結果（成功=0，失敗=1）を従属変数とする 2 項ロジスティック回帰分析を用いて，項目の関連性を分析した．CRT を用いた決定木分析を適用して，守備結果に対する戦術項目の達成基準を分析した．IBM SPSS ver. 23 を用い，有意水準は α= 5%と設定した．

３．結果

サッカー守備戦術技能を測定する 38 項目について，守備結果に対する基準関連妥当性が確認された. 守備終了時のボールの位置が自軍ゴールから 22.3m 以内では，守備成功率は 50.0%であった．第 1 基準の守備プレー中の第 3 DF からボール保持者の平均距離が 14.7m 以下では，守備成功率は 72.2%に向上した．第 2 基準の守備終了時のボール保持者の移動速度が秒速 5.19 m/s 以下では，守備成功率は 92.9%に向上した. この結果は守備戦術として，①相手を自軍ゴール近くでプレーさせない，②守備選手間の距離を小さくし，ボール保持者を自由にプレーさせない，③攻撃選手をスピードに乗った状態でプレーさせないことを示すと推察された．

４．結論

サッカートラッキングデータから構成された守備戦術技能項目は，妥当であり，守備結果を達成するための基準が明らかとなった．

５．文献

Frencken, W., Lemmink, K., Delleman, N., & Visscher, C. (2011). Oscillations of centroid position and surface area of soccer teams in small-sided games. European Journal of Sport Science, 11(4), 215-223.

サッカーのトラッキングデータに関する統計的モデリング

中央大学理工学部　　酒折 文武

プロスポーツやオリンピック競技を中心としたスポーツの現場において, 選手の行動履歴, ボールや選手の軌跡を表すトラッキングデータ, 動画データなど様々なデータが収集され, 選手のパフォーマンスの向上や戦略の決定, チームマネジメントなどに活用されるようになってきた. とくに, IoT やデータの計測技術の飛躍的な発展によりデータの収集が身近となり, それらに基づいた統計分析の必要性がますます広がってきている.

サッカーにおいては, 軍事技術として使われている自動追尾（トラッキング）システムを応用しスタジアムに設置した複数台の専用カメラでピッチ全体を撮影することにより, 選手やボール, 審判の動きの軌跡を測定した, トラッキングデータの収集と活用が行われるようになってきた. 日本においても欧米から遅れはとったものの, 2015 年より J1 リーグでトラッキングシステムが導入され, 各試合ごとの選手の走行距離やスプリント回数などを公開している.

トラッキングデータのさらなる活用のため, さまざまな統計分析もなされつつある. 酒折 (2015) では, 個々のプレイヤーに関する基本的な特徴量について調べ, 混合分布モデルによる選手の走行速度のあてはめと分類, スプリントの速度推移と間欠性回復（スプリントの時間間隔）に関する傾向についてを明らかにした. 成塚他 (2016) ではさらに, 複数選手間の相互作用や集団としてのダイナミクスを表現しうる特徴量として, 慣性半径の差と重心中点の関係, 秩序変数と速さ平均の関係, 相手選手との角度や距離を考えた. また, 神谷他 (2016) では戦況を表すと考えられる, ボール位置, 前線位置, コンパクトネス, 守備脆弱度, 攻撃率といった複数の変数に関する VAR モデルを考え, パラメータのオンライン学習に変化点検出法を用いることにより戦況の変化を検出する方法を提案した.

本研究ではそれらの結果を踏まえつつ, 選手間の位置や角度など複数選手間の関係を表す特徴量や, 試合状況に関する統計的モデリングについて検討する. 具体的な分析例や課題については当日詳しく報告する.

謝辞

本研究はデータスタジアム株式会社の協力を受けて行っている. ここに感謝申し上げる.

参考文献

[1] 神谷啓太, 中西航, 泉裕一朗 (2016). 「トラッキングデータを用いたサッカーの試合における戦況変化の抽出」『統計数理研究所共同研究リポート 363』pp.77-82.

[2] 酒折文武 (2015) . 「サッカーのトラッキングデータに関する統計的分析の可能性について」『2015 年統計関連学会連合大会講演報告集』.

[3] 成塚拓真, 卯田純平, 山崎義弘 (2016). 「サッカーの対戦的特徴に現れる普遍的な統計性の探求」『統計数理研究所共同研究リポート 363』pp.83-90.

ライフイノベーションを推進するバイオメディカルビッグデータ解析の新潮流

局所距離相関に基づくモジュレーター因子の網羅的探索法

島村徹平[1]、松井佑介[1]、宮野　悟[2]
[1]名古屋大学大学院医学系研究科、[2]東京大学医科学研究所

1.　はじめに

近年開始されたがんゲノムアトラス（The Cancer Genome Atlas: TCGA）や国際がんゲノムコンソーシアム（International Cancer Genome Consortium: ICGC）といった大型がんゲノムプロジェクトの進展により、多くのがんにおけるゲノム、エピゲノム異常の全体像が明らかにされつつある。その一方で、同定された異常がどのように下流の遺伝子発現制御機構に影響を与え、最終的にがんの表現型に寄与するかは未だ解明されていない点が多い。本報告では、ゲノム、エピゲノム、トランスクリプトームといった多階層オミックス情報に基づき、遺伝子発現制御に影響を及ぼすモジュレーター因子（ゲノム変異、コピー数異常、DNA メチル化など）を網羅的に探索する手法を提案する。

2.　局所距離相関に基づくモジュレーター因子の網羅的探索

距離相関は、Szekely ら（2007）によって提案された任意の次元の確率ベクトル$X \in R^p$、$Y \in R^q$間の依存性尺度である。有限の一次積率をもつすべての分布に対して、距離相関$\mathcal{R}$は、$0 \le \mathcal{R} \le 1$を満たし、XとYが独立である場合のみ、$\mathcal{R} = 0$である。本報告では、距離相関が他の変数ベクトル$Z \in R^r$に依存すると仮定し、$Z = \boldsymbol{z}_\alpha$周辺での局所距離相関を考える。ここで、ある転写因子とそのターゲット遺伝子の発現量をX、Y、モジュレーター因子のゲノム変異、コピー数、DNA メチル化量をZとし、n組のデータ$\{(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{z}_k): k = 1, \dots, n\}$が与えられたとき、$Z = \boldsymbol{z}_\alpha$周辺での経験局所距離共分散を以下で定義する。

$$\mathcal{V}_n^2(X, Y|Z = z_\alpha) = S_1(X, Y|Z = z_\alpha) + S_2(X, Y|Z = z_\alpha) - 2S_3(X, Y|Z = z_\alpha).$$

ここで

$$\begin{aligned}
\mathrm{S}_1(X, Y|Z = z_\alpha) &= \textstyle\sum_{k,l=1}^n w_{k\alpha} w_{l\alpha} |x_k - x_l||y_k - y_l|, \\
\mathrm{S}_2(X, Y|Z = z_\alpha) &= \textstyle\sum_{k,l=1}^n w_{k\alpha} w_{l\alpha} |x_k - x_l| \sum_{k,l=1}^n w_{k\alpha} w_{l\alpha} |y_k - y_l|, \\
\mathrm{S}_3(X, Y|Z = z_\alpha) &= \textstyle\sum_{k=1}^n w_{k\alpha} \sum_{l,m=1}^n w_{l\alpha} w_{m\alpha} |x_k - x_l||y_k - y_m|,
\end{aligned}$$

である。ただし、$w_{k\alpha}$は$Z = \boldsymbol{z}_k$と$Z = \boldsymbol{z}_\alpha$の距離に依存する重みパラメータである。このとき、$Z = \boldsymbol{z}_\alpha$周辺での経験局所距離相関は以下で与えられる。

$$\mathcal{R}_n(X, Y|Z = z_\alpha) = \frac{\mathcal{V}_n^2(X, Y|Z = z_\alpha)}{\sqrt{\mathcal{V}_n^2(X, X|Z = z_\alpha)\mathcal{V}_n^2(Y, Y|Z = z_\alpha)}}.$$

実際のモジュレーター因子の網羅的探索では、

$$H_0 : \mathcal{R}_n(X, Y|Z) = c \leftrightarrow H_1 : \mathcal{R}_n(X, Y|Z)) \neq c$$

で与えられる仮説検定を考え、局所距離相関がZに依存するかどうかを判定する。

参考文献

Szkely, G. J., *et al.*, Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6), 2769-2974, 2007.

Shimamura, T., *et al.*, Genome-wide identification of biological modulators using local energy statistics, Submitted.

大規模がんゲノム変異データマイニングのための統計学的手法

東京大学医科学研究所　白石友一

がんの変異には，がんの種類に応じて明らかな傾向があることが知られていた．例として，喫煙歴のある肺がんについては，タバコに含まれる化学物質によりもたらされる C>A の変異が多く観察されること，また皮膚がんにおいては，紫外線による C>T, CC>TT の変異が多く観察されることが知られていた．近年のシークエンス技術の発展により，新規がん原因遺伝子の同定だけではなく，個々のがんゲノムにおける変異のプロファイルの違いをこれまでにない精度で検出することが可能になった（Nik-Zainal et al., Cell, 2012, Alexandrov et al. Nature, 2013）．今後新たな変異パターンの発見，またそれに付随する発がん物質を同定することにより，新規発がん物質の発見や評価につながり，がんの予防に繋がることが多いに期待されている．一方で， がんゲノムシークエンス解析による大量の変異データから，特徴的なパターンを抽出するために，新たな情報学的・統計学的手法の開発が求められるようになった．現在，支配的となっている NMF（nonnegative matrix factorization）による方法論には，「組み合わせ爆発の問題から，考慮する因子を増やそうとすると，パラメータ数が指数的に増大し，推定が著しく不安定になってしまう」ということなどの問題点があった．

上記の問題点を解決するために，新しい統計的手法，probabilistic mutation signature (pmsignature, Shiraishi et al., PLoS Genetics, 2015)を開発した．提案手法は，

- 条件付き独立性を仮定したモデリングにより，変異パターンの因子数を増やしても，首尾よく推定が可能である
- 提機械学習分野で文書分類に利用されるトピックモデル（Blei et al. JMLR, 2003）と類似したモデルとなっており，過去にこれらの分野で蓄積されてきた膨大な知見を利用することが出来る

などの特徴を備えている．本統計手法を実装したソフトウェアは https://github.com/friend1ws/pmsignature に公開されている．またウェブアプリケーション https://friend1ws.shinyapps.io/pmsignature_shiny/ から利用することが可能である．

[参考文献]

Nik-Zainal et al., “Mutational Processes Molding the Genomes of 21 Breast Cancers”, Cell, 2012.

Alexandrov et al., “Signatures of mutational processes in human cancer”., Nature, 2013.

Shiraishi et al., “A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures”, PLoS Geneitcs, 2015.

Blei et al., “Latent Dirichlet Allocation”, Jornal of Machine Learning Research, 2013.

がんの進化シミュレーションによる腫瘍内不均一性生成原理の探索

東京大学医科学研究所ヘルスインテリジェンスセンター 助教　新井田厚司

がんは細胞のゲノムに変異が蓄積し増殖能力が高いものが進化的に選択された結果生じる。この進化の過程で様々なクローンが生み出され一つの腫瘍内においてゲノムレベルの不均一性を生み出していると考えられている。演者は九大別府病院との共同研究で一人の患者からの大腸がんの複数の部位から得たDNA をシークエンスすることにより大腸がんに広汎な腫瘍内不均一性が存在するのを見出した。また他の癌腫についても同様の解析により腫瘍内不均一性の報告がなされているが、それを生み出す原理の探求についての試みはほとんどなされていない。この目的のために演者は腫瘍内不均一性を再現する、がんの進化シミュレーションモデル、Branching evolutionary Process (BEP)モデルを構築した。また本研究ではスーパーコンピュータ「京」を利用して様々な組み合わせのパラメーターセットでBEPモデルによるがんの進化シミュレーションを行い、Approximate Bayesian Computation 法を用いて実験データを再現する条件の探索を試みた。その結果、高い遺伝子変異率を仮定すると高い腫瘍内不均一性が再現できることを見出した。更にシミュレーション結果から細胞の増殖に寄与するドライバー遺伝子は進化の初期に獲得され全てのがん細胞に共有されている一方で、不均一性を生み出している変異の大部分は細胞の増殖速度に影響を与えない中立変異であることが示唆された。以上、本研究により腫瘍内不均一性を生み出している進化原理の一端ががんの進化シミュレーションにより明らかにされた。

多重検定補正法の生命系大規模データへの応用

産業技術総合研究所 人工知能研究センター　瀬々　潤

ゲノム情報を始めとする、いわゆる生命情報データの解析には、多重検定補正が欠かせない。観測技術やデータベースの充実に従って、遺伝子数分の検定、着目する転写因子や変異数分の検定、データベース中の項目数分の検定が発生するためである。更に、生命の本質は要素の組合せである。我々の体は 38 兆個もの細胞が組み合わさったものであるし、細胞は様々なたんぱく質が組み合わさり骨格が形成されている。しかしながら、変数の組合せを考えた上で妥当な検定を行おうとすると、一見有意に見えるものでも、多重検定補正の結果、有意なものが現れないことが起こる。広く使われている多重検定補正法は、変数の数に比例して補正を厳しくするが、その結果、保守的になることが問題であると考えられ、その改善を行った。

本講演では、説明変数が 0-1 の 2 値の場合に絞り、かつ、多数の説明変数が存在する場合を考える。一般には、各説明変数と目的変数の間で検定を行い、有意な説明変数を探すが、ここでは、各説明変数に加えて、説明変数の積で表されるものも全て説明変数と考える。その上で、多重検定補正をしても有意となる説明変数（の集合）を発見する問題を考える。Bonferroni 補正を考えると、4 変数の場合は 2^4-1=15 通りの変数を考え、有意水準 α に対し p 値が α/15 未満であれば有意となる。変数の数の増加に従って、指数級数的に補正項が増加し、補正後の有意水準は小さくなる。

組合せを考えた場合、2 つの問題点が存在する。ひとつが過剰な補正を避けること。Bonferroni 補正は、和集合の上界を利用しているため、 特に変数の数が増えると過剰に補正が行われている可能性がある。もうひとつが、計算時間の問題。変数が増えると、それらの組合せと統計量を計算するだけで膨大な時間を要する。これら 2 つの問題を同時に解決するため、前者には Tarone の補正を、後者にはデータマイニング分野で長年研究されているアイテム集合マイニング法を活用した。これらを融合することで、理論的に全ての組合せを考えた場合と同等の回答と、Bonferroni に比べてより厳しい補正後の有意水準を計算することが可能になった。この手法を Limitless Arity Multiple-testing Procedure (LAMP; 無限次数多重検定法)と名付けた。

LAMP を組合せで働く転写因子の発見問題（説明変数 400 程度）や、ゲノムワイド関連解析（説明変数 25 万程度）に適用することで、今までの多重検定補正の手順では有意に差があることが見逃されていた例を挙げることに成功しているので、紹介する。

参考文献

1) Terada A, Okada-Hatakeyama M, Tsuda K, Sese J. (2013) Statistical significance of combinatorial regulations.110(32), 12996-13001.

2) 瀬々 潤, 浜田 道昭. (2015) 生命情報科学における機械学習: 多重検定と推定量設計. 講談社サイエンティフィック.

共発現解析による軽度認知障害の血漿 microRNA マーカーの検出

帯広畜産大学／国立長寿医療研究センター　茅野光範
国立長寿医療研究センター　檜垣小百合
明治薬科大学　佐藤準一
国立成育医療研究センター　松本健治
国立長寿医療研究センター　滝川修
国立長寿医療研究センター　新飯田俊平

1. はじめに

認知症の早期発見は疾病の二次予防（早期発見・早期対応）の点で非常に重要である。軽度認知障害は健康な状態と認知症（アルツハイマー病を含む）の中間に位置し、平均して軽度認知障害の 50%以上が 5 年以内に認知症へと進行する。しかし、軽度認知障害の中には、病状が進行しない場合も健康な状態に戻る場合もある。軽度認知障害を血液検査によって発見できれば、早期介入による対応が可能で、発病を遅らせることが出来るかもしれない。

2. 方法

多層的疾患オミックスプロジェクト[1] に提供された、健常者 30 人、軽度認知障害 23 人の血液サンプルを用いて 745 個の microRNA(miRNA) の発現量を計測した。全体の下位 20%までの量しか発現していないデータを除き、かつ、各群で 80%以上の個体で発現していた 85 個の miRNA を解析対象とした。共発現解析により軽度認知障害マーカーの検出を試みた[注]。

3. 結果

共発現解析により 20 個の miRNA からなる 20 組が軽度認知障害のマーカーとして検出された。上位 5 組の miRNA を用いて高い精度で軽度認知障害を判別可能であった。20 個中 2, 3 個の miRNA がハブとなって、軽度認知障害では、それらの miRNA を中心とした新しい相関が現れる場合と消失している場合があった[注]。

注：詳細は当日報告する。

参考文献：

[1] 多層的疾患オミックスプロジェクト：　http://gemdbj.ncc.go.jp/omics/
[2] Kayano M., Higaki S., Sato J., Matsumoto K., Takikawa O., Niida S., Plasma microRNA biomarkers for mild cognitive impairment using differential correlation analysis (submitted for publication)

ヒトゲノムデータの遺伝統計解析

Chromatin configuration QTL mapping using ATAC-seq

Natsuhiko Kumasaka, Andrew Knights, Daniel Gaffney

Wellcome Trust Sanger Institute, UK

ATAC-seq (assay for transposes accessible chromatin followed by sequencing) is now becoming a common biological assay to characterize chromatin accessible regions in vivo. Those regions are known to be associated with gene regulation by mapping genetic variants genome-wide. However, the mechanisms by which those variants affect gene regulation are not fully elucidated at the molecular level. Therefore, this study aims to map a novel, more complex, class of cellular quantitative trait loci (QTL) in the chromatin accessible regions, which alter not only chromatin accessibility but also transcription factor (TF) binding and cause nucleosome remodeling between individuals according to the underlying DNA sequence. We develop a hidden Markov model that utilizes the abundant information of paired-end sequencing technology to infer an ensemble of nucleosomes, linkers and various TFs hidden behind the V-plot (plot of fragment length against chromosomal location). The model also combines the information from individual diploid DNA sequences with putative TF binding affinity (position weight matrix) to address which TFs are bound in the region. We apply this model to the ATAC-seq data of 100 British samples (GBR in the 1000 Genomes Project) to uncover nucleosome and TF binding change between individuals governed by the chromatin configuration QTLs.

全ゲノムシークエンスによる肝癌の変異の包括的解析

藤本明洋[1,2]、古田繭子[1]、十時泰[3]、角田達彦[1]、加藤護[3]、柴田龍弘[3]、中川英刀[1]
1; 理化学研究所　統合生命医科学研究センター，2; 京都大学　医学研究科, 3; 国立がん研究センター がんゲノミクス研究分野

がんでは、ゲノムに突然変異が生じており、正常な分子経路が破綻して無秩序な細胞増殖をきたすことが分かっている。近年、次世代シークエンサーを用いてがんの突然変異の包括的カタログを作成するプロジェクト（ICGC; 国際がんゲノムコンシーアムなど）が進行している。現在までに ICGC からは 6,000 症例以上のデータが公開されており、今後数年で益々大量のゲノム情報が公開され、がん化のメカニズムの解明や創薬研究に大きな進歩をもたらすことが期待される。

我々は、ICGC に参加し、次世代シークエンサーのデータを解析した。研究開始当時（2009 年)、シークエンサーが産出する膨大なデータの解析手法は、確立されていなかった。そこで、解析手法開発のために、日本人 1 個体の全ゲノムの解析を行い、データ解析プログラムを開発した (1,2)。さらに、この経験に基づいて、がんゲノム解析パイプラインを構築し、点突然変異、挿入・欠失変異、コピー数変異、構造異常、HBV（B 型肝炎ウイルス）のゲノムへの挿入を検出する方法を開発した(3,4)。これらの方法を用いて、300 症例の肝臓癌の全ゲノムシークエンスを解析した(5)。点突然変異や短い挿入・欠失の数を解析し、有力なドライバー遺伝子候補として 38 遺伝子を同定した。また、ノンコーディング領域においては、ノンコーディング RNA（*NETA1*、*MALAT1*）や複数の遺伝子のプロモーター（*TERT*, *BCL6* など）に統計的に有意に多くの変異が存在した。構造異常を詳細に解析したところ、構造異常の数は DNA 複製タイミングと相関することが明らかになった。*CDKN2A*、*CCND1*、*TERT* などの遺伝子に複数のサンプルで構造異常が存在していた。構造異常と遺伝子発現量の相関を解析したところ、構造異常が遺伝子発現量に影響することが明らかになった。特に、*TERT* 遺伝子では、プロモーター領域に HBV の挿入や構造異常が検出され、それらは *TERT* 遺伝子の高い発現と相関していた。

本講演では、情報解析手法を紹介するとともに、肝臓癌の変異の全体像について述べる。

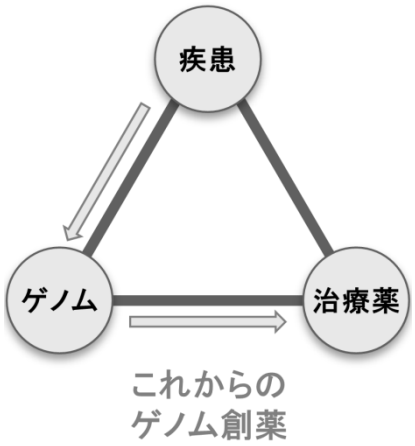
参考文献

1. Fujimoto et al. ***Nat Genet*** (2010), 2. Shigemizu, Fujimoto et al. ***Sci Rep*** (2013), 3. Fujimoto, Totoki et al. ***Nat Genet*** (2010), 4. Fujimoto, Furuta et al. ***Nat Comms*** (2015), 5. Fujimoto, Furuta, Totoki, Tsunoda, Kato et al. ***Nat Genet*** (2016)

遺伝統計解析で迫る疾患病態の解明とゲノム創薬

大阪大学大学院医学系研究科遺伝統計学　岡田随象

遺伝統計学（Statistical Genetics）とは、遺伝情報と形質情報の因果関係を統計学の観点から研究する学問分野である。次世代シークエンサー（next generation sequencer; NGS）やゲノムワイド関連解析（genome-wide association study; GWAS）に代表されるゲノム解析技術の著しい発達により、膨大なゲノム・エピゲノムデータが得られる時代が到来している。一方で、一次的なデータ解析処理を施され、ゲノム配列情報やエピゲノム修飾情報として蓄積された大容量のデータを適切に解釈し、社会還元するためのデータ解析学問へのニーズが高まっている。遺伝統計学は多彩な学問分野におけるビッグデータの分野横断的な統合に適した学問であり、近年その重要性が認識されている。例えば、大規模ヒト疾患ゲノム解析により同定された数多くの疾患感受性遺伝子の情報を、遺伝統計解析を通じて多彩な生物学・医学データベースと分野横断的に統合することにより、新たな疾患病態の解明（関節リウマチの病態における制御性Ｔ細胞の関与）[1]や、疾患バイオマーカーの同定（HLA imputation 法による HLA 遺伝子多型の同定、MIGWAS によるマイクロ RNA の同定）[2,3]、ドラッグ・リポジショニングを通じた新規ゲノム創薬[1,4]、疾患疫学の謎の解明（統合失調症と関節リウマチの低合併率の理由）[5]、個別化医療の推進などに貢献できることが明らかになりつつある。特に、疾患感受性遺伝子情報に基づき直接的に創薬標的を探索する遺伝統計解析は、ゲノム創薬の新たな方向性を示したものとして注目を集めている（右図）[1,4]。本講演では、遺伝統計学における最新ゲノム・エピゲノム解析手法や成果の紹介と共に、重要性に比較して専門家が不足している（と考えられている）遺伝統計学分野における人材育成についても議論したい。



参考文献

[1] Okada Y et al. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. ***Nature*** 506:376-81.

[2] Okada Y et al. (2015) Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. ***Nat Genet*** 47:798-802.

[3] Okada Y et al. (2016) Significant impact of miRNA-target gene networks on genetics of human complex traits. ***Sci Rep*** 6:22223.

[4] Imamura M et al. (2016) Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. ***Nat Commun*** 7:10531.

[5] Lee SH et al. (2015) New data and an old puzzle: the negative association between schizophrenia and rheumatoid arthritis. ***Int J Epidemiol*** 44:1706-1721.

ゲノムワイド関連解析による高血圧遺伝子の解明

国立国際医療研究センター　竹内史比古

高血圧は、食事などの生活習慣、家族歴に現れる遺伝的体質など、複数の要因が作用し合って発症する。約 30 億塩基対からなるヒトゲノムは、わずかに個人差があり（約千塩基対ごとに 1 箇所）、これを DNA 多型とよぶ。DNA 多型により遺伝子の機能が変化して、病気の罹り易さが変わる。約 2 万個ある遺伝子のうち、どれが高血圧への罹り易さを規定しているかが分かれば、病気の仕組みが解明でき、治療薬の開発にもつながる。

「DNA-RNA-タンパク質-細胞-組織-器官-個体」の生体階層構造において、DNA がコードするゲノムと個体の健康状態である疾患は、両端に位置しているにも拘わらず、疾患ゲノム研究は疾患の解明と治療法開発の強力な手段である。それが可能なのは、DNA 多型と疾患の関連が統計的に解析でき（関連解析）、また統計的関連が因果関係を示唆するからである。

ゲノムワイド関連解析（GWAS）では、多数の罹患者と健常者について DNA 多型をゲノム全体に渡って測定し、両グループで有意に頻度が異なる DNA 多型を探索する。疾患と関連する DNA 多型の近傍に位置する遺伝子が疾患原因遺伝子の候補となる。DNA 多型測定技術と遺伝統計学の発展により、これまでに数百の疾患や形質について GWAS が行われ、数千の DNA 多型との関連が同定された[1-3]。

高血圧などの生活習慣と関連する個々の DNA 多型は（本物ではあるものの）関連が極めて弱いことが分かってきた。検出力を上げるために大規模なサンプルが必要であり、複数の GWAS を統合するメタ解析、多人種を統合するメタ解析が行われている。iGEN-BP 研究では、欧米人・東アジア人・南アジア人合計 320,251 名について血圧の GWAS メタ解析を行い、新たに 12 の血圧関連遺伝子座を同定した[4]。これらの GWAS や GWAS メタ解析で用いられている遺伝統計解析の手法を紹介する。

参考文献

[1] Wellcome Trust Case Control Consortium (2007) *Nature* 447:661

[2] M.I. McCarthy, G.R. Abecasis, L.R. Cardon et al. (2008) *Nature Review Genetics* 9:356

[3] http://www.ebi.ac.uk/gwas/

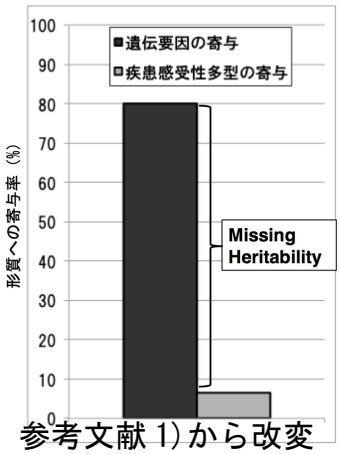
[4] N. Kato, M. Loh, F. Takeuchi et al. (2015) *Nature Genetics* 47:1282

統計遺伝学モデルを用いた多因子疾患の発症リスク予測法

岩手医科大学　八谷 剛史

ある形質について，遺伝要因と環境要因の相対的な寄与の割合を“遺伝率(heritability)”という[1)]．双生児研究や家系研究により，ヒトの様々な多因子形質（疾病発症リスクを含む）について遺伝率が見積もられ，遺伝要因の寄与は決して少なくないことが示されている．このことから，遺伝情報から形質を予測することは，原理的に可能だと考えられる．

一方，ゲノムワイド関連解析（genome-wide association study; GWAS）が進められ，ヒトの多因子形質に影響を与えている感受性多型が数多く同定された[1)]．その結果,「双生児研究等によって見積もられた遺伝要因の寄与率（＝遺伝率)」と「GWAS によって同定された形質関連多型の寄与率」に大きな齟齬が生じることが明らかになった．この齟齬（差分）のことを，“missing heritability”と呼ぶ（右図）．すなわち，GWAS によって同定された形質関連多型を変数に持つ統計モデルの形質予測能力は低く，そのために，遺伝情報に基づく形質予測法はほとんど医療応用されるに至っていない．



参考文献 1)から改変

そこで,（感受性多型として同定されていない）コモンバリアントの寄与を予測に役立てる遺伝統計学的な手法が着目されている[1)]．高密度 SNP アレイでジェノタイピングされるコモンバリアントの個数（≒予測モデルのパラメータ数）は数十万～数百万個あり，一方，予測モデルのパラメータを学習するために利用可能なサンプルサイズは数万～数十万人である．サンプルサイズ（N）よりパラメータ数（p）が多く,「$N \ll p$」となっている．「$N \ll p$」の状況では，個々のパラメータを吟味する（個々のゲノム多型の機能や寄与を明らかにする）よりも，数理モデルを用いて現象を説明・予測するアプローチが有用になる．本発表では，ジェノタイプデータの分散共分散行列をカーネル行列として捉え，線形混合モデルやベイズモデルを用いた形質・疾病発症リスクの予測法について発表する．

参考文献 1）八谷剛史：Missing heritability のルーツと形質・疾病発症リスク予測の可能性．実験医学　東京：羊土社．In press.