2016 年 11 月 7，8 日
久留米シティプラザ中会議室

文部科学省科学研究費補助金 基盤研究 A（15H01678）
「大規模複雑データの理論と方法論の総合的研究」
研究代表者: 青嶋 誠（筑波大学）
シンポジウム

複雑な生命現象を読み解くための大規模データ解析とモデリング

開催責任者
植木優夫（久留米大学バイオ統計センター）

内容・目的

大規模データは、医学・生物学・遺伝学・農学等の生命科学分野で利用が進んでおり、新たな知見を次々と生み出しています。今後、さらに多様でかつ網羅的なデータが収集されることが予想されますが、これらの複雑なデータには様々な不確実性がつきまとい、ノイズのなかから有益な情報を抽出していく作業が重要となります。そのためには、統計学、機械学習、人工知能、データマイニング、バイオインフォマティクスなどの複数のデータ科学技術を横断的に利用し、既存モデルと融合させた、先端的なデータ解析が有用であると考えられます。本シンポジウムでは、大規模データを用いて生命現象の解明や新現象の発見に資する統計解析法、新規理論・方法論、モデリング、アルゴリズム、データ解析事例と知識発見、シミュレーション研究などに関わる講演を頂きます。

【プログラム】

11 月 7 日（月）

13:00-13:30

二宮 嘉行（九州大学 マス・フォア・インダストリ研究所）

「傾向スコア解析のための二重頑健基準」

13:30-14:00

大谷 隆浩（統計数理研究所 リスク解析戦略研究センター 独立行政法人科学技術振興機構 CREST 立川）、野間 久史（統計数理研究所データ科学研究系 独立行政法人科学技術振興機構 CREST 立川）、西野 穣（名古屋大学大学院医学系研究科 独立行政法人科学技術振興機構 CREST 名古屋）、松井 茂之（名古屋大学大学院医学系研究科 独立行政法人科学技術振興機構 CREST 名古屋）

「ゲノムワイド関連解析における多重検定補正手法の再評価」

14:10-15:00

特別講演 1:

瀬々 潤（産業技術総合研究所 人工知能研究センター 機械学習研究チーム）

「多重検定補正法の改良による生命科学への寄与と課題」

15:00-15:50

特別講演 2:

大橋 順（東京大学 大学院理学系研究科 生物科学専攻 ヒトゲノム多様性研究室）

「ゲノムワイド関連解析による遺伝子・環境相互作用の検出: 相互作用モデルと検出力」

16:00-16:30

荒木 由布子（静岡大学 情報学部 行動情報学科）、岡田 栄作（ 浜松医科大学 医学部 健康社会医学講座）、尾島 俊之 （浜松医科大学 医学部 健康社会医学講座）、近藤 克則（千葉大学 予防医学センター環境健康学研究部門）

「大規模追跡調査に基づく要介護認定リスクモデルの提案」

16:30-17:00

重水 大智（東京医科歯科大学 難治疾患研究所 ゲノム応用医学研究部門 医科学数理分野）

「疾患原因となる遺伝子変異同定のためのエクソーム解析の現在」

17:00-17:30

小野木 章雄（JST さきがけ・東京大学 大学院農学生命科学）

「データが育種を加速する」

18:00-20:00
情報交換会

11 月 8 日（火）

9:30-10:00
塩濱 敬之（東京理科大学 工学部 情報工学科）
「Non-linear State-Space Modeling for Wind Speed and Direction」

10:00-10:30
阿部 俊弘（南山大学 理工学部 システム数理学科）
「Weibull 分布と sine-skewed von Mises 分布を用いたシリンダー上の分布の生成」

10:30-11:00
松井 秀俊（滋賀大学 データサイエンス教育研究センター）
「経時測定データの判別と変数・決定境界の選択」

11:10-11:40
小森 理（福井大学 学術研究院工学系部門）
「非対称ロジスティックモデルによる水産資源評価」

11:40-12:10
大森 亮介（北海道大学 人獣共通感染症リサーチセンター バイオインフォマティクス部門）
「性感染症流行動態解析の展望」

12:10-12:40
小島 要（東北大学 東北メディカル・メガバンク機構）、河合 洋介（東北大学 東北メディカル・メガバンク機構）、長崎 正朗（東北大学 東北メディカル・メガバンク機構）
「ハイスループットシークエンサによるマイクロサテライトリピート数推定法の高精度化」

傾向スコア解析のための二重頑健基準

九州大学 マス・フォア・インダストリ研究所 二宮 嘉行

1. 序章

周辺構造モデル (Robins 1997, Robins et al. 2000) はいまや因果推論において基本となるモデルである．これは潜在結果変数モデルであってある種のデータが欠測しており，その欠測メカニズムと結果変数に相関があるにもかかわらずナイーブに推定をおこなうと，推定値がバイアスをもつことになる．このバイアスはその相関を正しくモデリングできれば取り除けることもあるが，その困難なモデリングをせずに逆確率重み付け (IPW) 推定 (Robins et al. 1994) や二重頑健 (DR) 推定 (Scharfstein et al. 1999, Bang and Robins 2005) をおこなうセミパラメトリックアプローチがよく用いられる．

例えば，$y_i^{(h)}$ をサンプル i が処置 $x^{(h)}$ を受けたときの潜在結果変数, $t_i^{(h)}$ をそのときは 1 でそれ以外のときは 0 をとる割り当て変数，ε_i を誤差変数とし，$y_i^{(h)} = \sum_{j=0}^{p} b_{j+1} x^{(h)j} + \varepsilon_i$ という最も基本的な周辺構造モデルを考える (Platt et al. 2013, Talbot et al. 2015)．このモデルでは $t_i^{(k)} = 0$ なる $y_i^{(k)}$ が欠測していると捉えられ，$y_i^{(h)}$ と $t_i^{(h)}$ に相関があるにもかかわらず最小二乗法で回帰構造 $\sum_{j=0}^{p} b_{j+1} x^{(h)j}$ を推定すると，当然バイアスが生じることとなる．そこで，$y_i^{(h)}$ と $t_i^{(h)}$ 間の交絡 $\boldsymbol{z}_i$ が観測されているとし，$e_i^{(h)} \equiv \mathrm{P}(t_i^{(h)} = 1 \mid \boldsymbol{z}_i)$ なる傾向スコア (Rosenbaum and Rubin 1983) を用いたセミパラメトリックアプローチがよく用いられる．この設定のもと，興味ある回帰構造に関するモデル選択問題，上の例でいえば多項式の次数 p の選択問題を扱う．

驚くべきことに，この基本的な問題に対処するために古典的な情報量規準を調整したものは一つを除いて存在しない．その価値ある一つが Platt et al. (2013) による QIC_w である．この基準は，QIC (Pan 2001) の第一項目である最大疑似対数尤度を欠測に対処するよう最大重み付き疑似対数尤度に置き換えたものである．一方でこの QIC_w は，第二項目であるペナルティ項として欠測がないときと同じものを用いている．本報告では，情報量規準元来の考え方に基づいてペナルティ項を評価すると，欠測がないときとははるかに異なる量が現れることを確認する．

Platt et al. (2013) に書かれているように，回帰構造に関するモデル選択はほとんど扱われていないが，交絡変数選択については Brookhart and van der Laan (2006) や Vansteelandt et al. (2012) などで議論が進められている．これらの論文では，計算コストの高い交差検証法や，局所モデル誤特定という特殊な仮定に基づく FIC (Claeskens and Hjort 2003) が用いられている．本報告では，これらを本問題適用のために拡張することは考えず，高い計算コストや特殊な仮定に頼らない方法を構築する．

2. 講演内容

H 個の群を想定し, 処置 h $(\in (1, 2, \cdots, H))$ を受けた時の結果変数を $\boldsymbol{y}^{(h)}$ $(\in \mathbb{R}^m)$ とした周辺構造モデル $\boldsymbol{y} = \sum_{h=1}^{H} t^{(h)} \boldsymbol{y}^{(h)}$ を考える. ここで, $t^{(h)}$ を $\boldsymbol{y}^{(h)}$ が観測されると 1, 観測されないと 0 となる指示変数とする $(\sum_{h=1}^{H} t^{(h)} = 1)$. また, 仮定するモデルを $f(\boldsymbol{y}^{(h)} | \boldsymbol{x}^{(h)}; \boldsymbol{\theta})$ とし, $\boldsymbol{x}^{(h)}$ $(\in \mathbb{R}^p)$ をランダムな共変量ベクトル $\boldsymbol{z} (\in \mathbb{R}^q)$ の一部を含んでもよい独立変数ベクトルとする. さらに, 一般化傾向スコアを $e^{(h)}(\boldsymbol{\alpha}) = P(t^{(h)} = 1 | \boldsymbol{z}; \boldsymbol{\alpha})$ (Imbens 2000) とし, $\boldsymbol{\alpha}$ $(\in \mathbb{R}^s)$ を一般化傾向スコアのパラメータとする. 本報告を通して弱く無視できる割り当て条件 $\boldsymbol{y}^{(h)} \perp\!\!\!\perp \boldsymbol{t}^{(h)} \mid \boldsymbol{z}$ を仮定する. このモデルに対して標本数 N の独立なサンプルが与えられているとし, i 番目のサンプルには $\tilde{}$ と i を付ける.

DR 推定量 $\hat{\boldsymbol{\theta}}^{\mathrm{DR}}$ を用いた場合を考える. 弱く無視できる割り当て条件の下で推定量の漸近的性質を用いることにより, 次の結果を得る. ただし, $\boldsymbol{\beta}$ $(\in \mathrm{R}^r)$ は結果変数をモデリングするときに用いられるパラメータである.

定理 傾向スコアか結果変数のどちらか一方が正しく特定された場合の DR 推定量に対する AIC として次式が導かれる.

$$
\begin{aligned}
&-2\sum_{i=1}^{N}\sum_{h=1}^{H}\frac{\tilde{t}_i^{(h)}}{e_i^{(h)}(\boldsymbol{\alpha})}\log f(\tilde{\boldsymbol{y}}_i^{(h)}|\tilde{\boldsymbol{x}}_i^{(h)};\hat{\boldsymbol{\theta}}^{\mathrm{DR}})+2\Bigg(\mathrm{tr}\{\boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\}\\
&+\sum_{h,h'=1}^{H}\mathrm{E}\left[\boldsymbol{s}^{(h)}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta})'\boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{s}^{(h')}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta})\right]-\sum_{h=1}^{H}\mathrm{E}\Big[\frac{1}{e^{(h)}(\boldsymbol{\alpha})}\boldsymbol{s}^{(h)}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta})'\boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{s}^{(h)}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta})\Big]\\
&+\sum_{h=1}^{H}\mathrm{tr}\Big\{\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})\mathrm{E}\Big[\frac{1}{e^{(h)}(\boldsymbol{\alpha})}\frac{\partial e^{(h)}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\boldsymbol{s}^{(h)}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta})'\Big]+\boldsymbol{D}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})\mathrm{E}\Big[\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{s}^{(h)}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta})'\Big]\Big\}\Bigg)
\end{aligned}
$$

ここで, $\boldsymbol{s}^{(h)}(\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\beta})=\mathrm{E}[\partial\log f(\boldsymbol{y}^{(h)}\mid\boldsymbol{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}\mid\boldsymbol{z};\boldsymbol{\beta}]$, $\boldsymbol{A}(\boldsymbol{\theta})=\mathrm{E}[-\sum_{h=1}^{H}\partial^2\log f(\boldsymbol{y}^{(h)}\mid\boldsymbol{x}^{(h)};\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']$, $\boldsymbol{B}(\boldsymbol{\theta})=\mathrm{E}[\sum_{h=1}^{H}\{\partial\log f(\boldsymbol{y}^{(h)}\mid\boldsymbol{x}^{(h)};\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}\{\partial\log f(\boldsymbol{y}^{(h)}\mid\boldsymbol{x}^{(h)};\boldsymbol{\theta})/\partial\boldsymbol{\theta}'\}/e^{(h)}(\boldsymbol{\alpha})]$, $\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})$ と $\boldsymbol{D}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})$ は $N(\hat{\boldsymbol{\theta}}^{\mathrm{DR}}-\boldsymbol{\theta})$ をスコア関数の線形和で漸近的に表現したときにそれぞれ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ のスコア関数に付く係数行列である.

参考文献

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models, *Biometrics*, **61**, 962–972.

Brookhart, M. A. and van der Laan, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model, *Comput. Statist. Data Anal.*, **50**, 475–498.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion, *J. Amer. Statist. Assoc.*, **98**, 900–945, With discussions and a rejoinder by the authors.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions, *Biometrika*, **87**, 706–710.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations, *Biometrics*, **57**, 120–125.

Platt, R. W., Brookhart, M. A., Cole, S. R., Westreich, D., and Schisterman, E. F. (2013). An information criterion for marginal structural models, *Stat. Med.*, **32**, 1383–1393.

Robins, J. M. (1997). Marginal structural models, *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, 1–10.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American statistical Association*, **89**, 846–866.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology, *Epidemiology*, **11**, 550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association*, **94**, 1096–1120.

Talbot, D., Atherton, J., Rossi, A. M., Bacon, S. L., and Lefebvre, G. (2015). A cautionary note concerning the use of stabilized weights in marginal structural models, *Statistics in Medicine*, **34**, 812–823.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference, *Statistical Methods in Medical Research*, **21**, 7–30.

ゲノムワイド関連解析における多重検定補正手法の再評価

大谷 隆浩[1,2]，野間 久史[2,3]，西野 穣[2,4]，松井 茂之[2,4]
[1]統計数理研究所 リスク解析戦略研究センター，[2]独立行政法人科学技術振興機構 CREST,
[3]統計数理研究所 データ科学研究系，[4]名古屋大学大学院医学系研究科

１．はじめに

ゲノムワイド関連解析 (genome-wide association study; GWAS) は，ゲノム全体をカバーする数十万～数百万の一塩基多型 (single nucleotide polymorphism; SNP) の遺伝子型を決定し，遺伝子型と疾患・量的形質との関連を網羅的に評価する研究手法である. 過去十数年の間に世界中で行われた GWAS の成果により，多くの疾患関連遺伝子が同定されている．この手法では数百万程度の対象となる SNP ごとに，検定によって単変量の関連解析を行うことが一般的である．このため，数百万規模の多重検定による多重性を適切に考慮した上で，疾患関連遺伝子のスクリーニングを行う必要がある．

近年の研究により，GWAS によって同定された疾患関連 SNP を集めても，家系分析から推定される遺伝率のごく一部しか説明できないことが示され，「失われた遺伝率 (missing heritability)」の問題として知られている．この問題はさまざまな要因が複合して引き起こされていることが考えられるが，統計学的な問題としては，関連の有無を判定するために，$P < 5 \times 10^{-8}$を閾値とする「Genome-wide significance 基準[1]」がこれまでのほとんどの GWAS で用いられてきたことが 1 つの原因として考えられる．この基準は，有意水準 5%の検定を 100 万回行う際に，Family-wise error rate (FWER) が 5%未満になるように，Bonferroni 補正法によって多重性の補正を行うことに対応する．これは直感的にも非常に保守的な基準であり，偽陽性は低い頻度に抑えられているとしても，疾患に関連する多くの SNP が見逃されている可能性がある．そこで本研究では，実データを使用した解析例とシミュレーション実験を通して，GWAS における多重検定手法のストラテジーを再評価する．特に，現在，ほとんどの GWAS で採用されている Genome-wide significance 基準の妥当性・効率性についての批判的な再評価を行い，False discovery rate (FDR)[2] をはじめとするその他の方法論との比較・再評価を行う．

２．GWAS データセットに基づく平均検出力の推定

まず，既存の GWAS における多重検定手法の効率性を評価するために，リウマチ性関節炎[3]および統合失調症[4]を対象とした GWAS のデータセットを用いて，疾患関連 SNP の効果サイズ (オッズ比) の分布を考慮した上で，特定の有意水準のもとでの「平均検出力」を推定した．具体的には，全ての疾患関連 SNP を対象とした検出力の平均値 (平均検出力) と，指定した効果サイズ以上の SNP を対象とした検出力の平均値 (部分検出力; partial power[5]) を評価基準として用いた．指定した有意水準のもとでの平均検出力と部分検出力，FDR の推定には，階層混合モデルに基づく経験ベイズ法[6]を用いた．

経験ベイズ法により，Genome-wide significance 基準を採用した場合の FDR はリウマチ性関節炎のデータセットでは 0.0061，統合失調症のデータセットにおいては 0.0011 と推定された．この場合の平均検出力はどちらのデータセットにおいても極めて低く，リウマチ性関節炎においては 0.09%，統合失調症においては 0.04%となった．この結果は，Genome-wide significance 基準の過度な保守性を示唆している．有意水準を緩めることで平均検出力は多少改善でき，例えば FDR=5%の基準を採用した場合の平均検出力は，リウマチ性関節炎において 0.4%，統合失調症において 0.5%と推定された．しかしながら，この場合にお

いても十分な平均検出力を達成できているとはいえず, 既存の GWAS の中で最大規模の研究であっても, ごく少数の疾患関連 SNP しか検出できていないと考えられる．効果サイズのごく小さなものを含めた, すべての疾患関連 SNP を網羅的に検出するのは, 現実的なサンプルサイズの GWAS ではほとんど不可能である可能性がある．

一方で，部分検出力については比較的高く推定されており，効果サイズの大きい SNP については適当な有意水準のもとで，その多くを効率的に検出できることが示唆された．例えば，FDR=5%の基準のもとで，オッズ比が 1.1 以上の SNP を対象とした部分検出力は，リウマチ性関節炎のデータセットにおいては 20%，統合失調症のデータセットにおいては 75%と推定された．一定以上の効果サイズを持つ疾患関連 SNP をターゲットとした場合，このように，現実的な規模の GWAS で相応の割合での検出が可能である．これらの結果は，研究のデザインや全般的なストラテジーの策定において有益な知見となり得るであろう．

３．大規模 GWAS における効率性の評価

前節の結果より， Genome-wide significance 基準では既存の最大規模の GWAS においても，疾患関連 SNP 全体のごく一部しか検出できないことが示唆された．次に，同じ解析のストラテジーを用いて，同様の GWAS を継続して実施したときに，将来，どの程度の疾患関連 SNP が同定できる見込みがあるのかをシミュレーションによって評価した．このために，多因子性の疾患に対する GWAS を模したシミュレーション実験を行った．既存の統合失調症に対する GWAS の結果に基づいて疾患関連 SNP の効果サイズの分布を仮定し，最大で 10 万サンプルのケース・コントロール GWAS を行なった場合の模擬データを作成し，各多重検定手法の平均検出力と部分検出力を評価した．その結果，Genome-wide significance 基準を採用した場合，十分な平均検出力を達成するためには膨大な数のサンプルが必要であり，10 万のサンプルを集めたとしても平均検出力は 17%程度しか達成できないと推定された．一方，FDR の制御に基づく手法ではこれを改善することができ，FDR=5%の基準のもとでは 50%以上の平均検出力を達成できた．さらに，部分検出力については非常に高い水準を達成でき，オッズ比 1.05 以上の SNP に関しては 80%以上となった．これらの結果から，Genome-wide significance 基準に基づく解析のストラテジーでは，継続的に GWAS を実施したとしても，すべての疾患関連 SNP を網羅的に同定することは現実的には困難であると考えられる．このため，FDR の制御に基づく多重検定手法を採用することや，効果サイズの大きい SNP に対象を絞ることなどが，今後の研究において有効なストラテジーとなる可能性がある．

参考文献

[1] Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol. 32, 227–234.

[2] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B, Methodol. *57*, 289–300.

[3] Okada, Y., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381.

[4] Ripke, S., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. 43, 969–976.

[5] Matsui, S., and Noma, H. (2011). Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. Biometrics 67, 1225–1235.

[6] Noma, H., and Matsui, S. (2015). Univariate analysis for gene screening: Beyond the multiple testing. In Design and Analysis of Clinical Trials for Predictive Medicine, S. Matsui, M. Buyse, and R. Simon, eds. (CRC Press), pp. 227–251.

多重検定法の改良による生命科学への寄与と課題

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。更に、現状の問題点と今後の課題を示すことで、皆様と議論を行いたい。

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.

ゲノムワイド関連解析による遺伝子・環境相互作用の検出
-相互作用モデルと検出力-

東京大学大学院理学系研究科生物科学専攻
大橋　順

１．背景

ゲノムワイド関連解析（GWAS）により、様々な疾患の感受性多型（主に単塩基多型；SNP）が検出されてきた。その大部分は、症例と対照の対立遺伝子頻度の比較から検出されたものであり、環境因子（例えば生活習慣）との相互作用が疾患感受性に影響を与える多型についてはほとんどわかっていない。個人の遺伝情報をもとに病気の発症を未然に防ぐ、個別化予防の実現を目指すには、環境改善によってリスクを減らすことができる感受性多型を数多く検出することが望ましい。本稿では、強い相互作用を仮定した遺伝子・環境相互作用モデルの下で、ロジスティック回帰分析により環境と強い相互作用を示す疾患感受性変異を検出する研究のデザインについて考える。

２．遺伝子・環境相互作用モデル

疾患感受性遺伝子座には、感受性アリルAと非感受性アリルaの2つの対立遺伝子が存在すると仮定する。AとaのAの対立遺伝子頻度を、それぞれ f と $1-f$ とする。環境因子Eの保有者頻度を K_1 とする。Eをもたないか、遺伝子型が aa である個体の疾患発症確率を K_2 とする。そのような個体に対し、環境因子を保有し、遺伝子型が Aa だと r 倍、AA だと r^2 倍疾患リスクが上昇すると仮定する。遺伝因子と環境因子は独立であるとすると、遺伝子型と環境因子の双方を考慮した疾患発症確率は右図で与えられる。

	E+	E-
AA	r^2K_2	K_2
Aa	rK_2	K_2
aa	K_2	K_2

３．集団罹患率

疾患感受性遺伝子座はハーディ・ワインバーグ平衡にあると仮定すると、集団罹患率 K は

$$K = K_1K_2\left[f^2r^r + 2f\left(1-f\right)r + \left(1-f\right)^2\right] + \left(1-K_1\right)K_2 = K_1K_2\left(fr+1-f\right)^2 + \left(1-K_1\right)K_2 \quad \text{式(1)}$$

で与えられる。研究デザインを考える際に、対象疾患の集団罹患率は既知であることが多いので、K、K_1、f、r を与え、式(1)より K_2 を決めるのが望ましい。例えば、K=0.01、K_1=0.1、f=0.05、r=3 とすると、K_2=0.0098 となる。

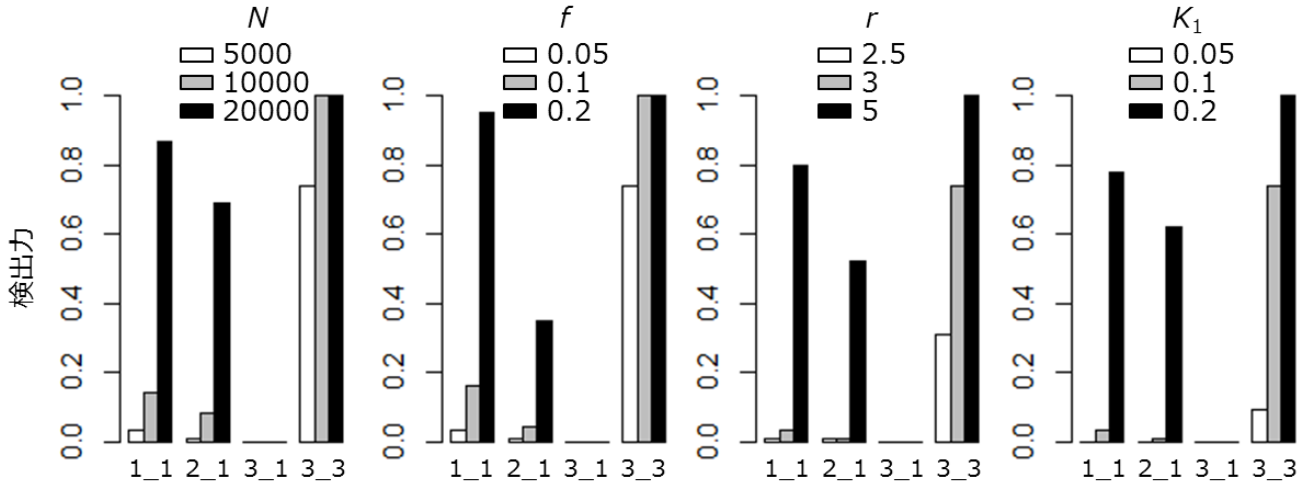
４．ロジスティック回帰モデル

本稿では、次のような３つのロジスティック回帰モデルを考える。ここで、x_1 は個体が保有するAの個数（x_1 = 0, 1, 2）、x_2 は環境因子の有無（x_2 = 0, 1）である。

$$logit(p) = \beta_0 + \beta_1x_1 \qquad \text{モデル１}$$
$$logit(p) = \beta_0 + \beta_1x_1 + \beta_2x_2 \qquad \text{モデル２}$$
$$logit(p) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 \qquad \text{モデル３}$$

５．検出力

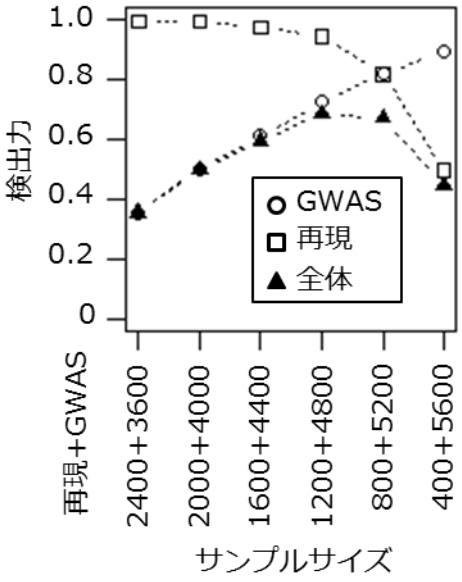
N=5,000（症例と対照のそれぞれ）、K=0.01、K_1=0.1、f=0.05、r=3 を基本とし、着目する１つのパラメタのみを変更してシミュレーションを行い、各パラメタセットに対して 100 個のシミュレーションデータを作製した。各シミュレーションデータに対し、有意水準を$\alpha=5\times10^{-8}$ として、x_1 もしくは x_1x_2（モデル３の交互作用項）の偏回帰係数の検定（Wald 検定）を行い、有意差が検出された割合を各偏回帰係数の検定に対する検出力とした。$\alpha=5\times10^{-8}$ としたのは、GWAS で主に採用されてきた有意水準（いわゆるゲノムワイド有意水準）だからである。なお、シミュレーションを含め、全ての計算は統計分析フリーソフト「R」を用いて行った。結果を次ページに示す。



上図の１－１、２－１、３－１、３－３は、それぞれロジスティック回帰モデル１のβ_1、ロジスティック回帰モデル２のβ_1、ロジスティック回帰モデル３のβ_1、ロジスティック回帰モデル３のβ_3 に対して行った検定の検出力である。ここで試したパラメタでは、ロジスティック回帰モデル３のβ_3 に対する検定が常に最も高い検出力を示した。ロジスティック回帰モデル１のβ_1 に対する検定の検出力は、ロジスティック回帰モデル２のβ_1 に対する検定よりも常に検出力が高いことから、強い相互作用が存在する場合には、環境因子を調整しない方が感受性 SNP を検出しやすいといえる。なお、ロジスティック回帰モデル３のβ_1 に対する検定の検出力は極めて低かった。

６．再現研究まで含めたデザイン

GWAS では、全ゲノムレベルでの解析により有意な関連（P 値 ＜ 5×10^{-8}）を示す SNP を検出した後、独立なサンプルセットでそれらの関連が再現されるかを確認すること（再現研究）が要求される。再現研究では、各 SNP について、有意水準 0.05 で片側検定が行われることが多い。したがって、GWAS 研究のデザインを考える際は、最も検出力が高くなる（GWAS でも再現研究でも関連が有意）ように、GWAS 用と再現研究用にサンプル数を振り分ける必要がある。右図は、N=6,000（症例と対照のそれぞれ）、K=0.01、K_1=0.1、f=0.05、r=3、K_2=0.0098 を仮定した場合の、振り分け方と検出力との関係を示したものである（右図）。このパラメタセットに対しては、GWAS 用に 4,800 検体（症例と対照のそれぞれ）、再現研究用に 1,200 検体（症例と対照のそれぞれ）を振り分ける場合に、全体として最も高い検出力（GWAS の検出力×再現研究の検出力）が達成される。



７．まとめ

本稿で仮定したような強い相互作用が存在し、環境因子の保有者頻度が低い場合には、ロジスティック回帰モデル３の交互作用項の偏回帰係数に対する検定の検出力は高いが、モデル１やモデル２の検出力は低いことが確認された。したがって、環境因子を考慮しないで行われたこれまでの GWAS では、環境因子と強い相互作用を示す感受性多型が見過ごされている可能性がある。K=0.01、K_1=0.1、f=0.05、r=3、K_2=0.0098 の場合に、環境因子を考慮することなく、遺伝因子のみから計算したアリル OR は 1.22 である。（糖尿病や高血圧などの）ありふれた疾患に対する GWAS により、これまでに検出されてきた感受性変異のアリル OR は 1.1～1.3 程度である。このような変異の中には、環境因子と強い相互作用を示すものが含まれている可能性もある。今後、SNP だけではなく、詳細な臨床情報も解析に利用した GWAS が行われ、環境と相互作用を示す感受性多型が数多く同定されることに期待したい。

大規模追跡調査に基づく要介護認定リスクモデルの提案

荒木 由布子[1]，岡田 栄作[2]，近藤 克則[3]，尾島 俊之[2]
[1]静岡大学情報学部行動情報学科，[2]浜松医科大学医学部健康社会医学講座，
[3]千葉大学予防医学センター環境健康学研究部門

１．はじめに

厚生労働省の調査では，2015 年 3 月時点で要支援・要介護の認定を受けた人の数は今年度になり初めて 600 万人を突破し，これは国民の約 20 人に 1 人が要支援・要介護認定者であることを示している．認定者は増加の一途で，2025 年には団塊の世代が 75 歳以上になるため増加ペースはさらに上がることが予想される．このような状況の中，2000 年より要介護認定に至るリスクを解明すべく JAGES（Japan Gerontological Evaluation Study，日本老年学的評価研究）プロジェクトが立ち上がった．本プロジェクトは健康長寿社会のため予防政策の科学的な基盤づくりを目的とした研究プロジェクトである．本研究では，観測された数百の調査項目の中から要介護認定の予後因子を特定し，その機序を解明するための方法論の探求を行った．2003-2013 コホートでは，2003 年 11 月 1 日より 2013 年 3 月 28 日の間の 3436 日における要介護認定を受けていない高齢者の 10 年間追跡データである．

本研究の特徴の 1 つとして，要介護認定の予後因子の候補となる説明変数を予測モデルに用いる際に，探索的なリスク因子の同定を可能とする手法を用いた点である．従来の JAGES データ分析の様にすでに先行研究から既知である変数を投入して検定手法で確証的な分析を行うのではなく，説明変数の候補として考え得る全ての調査項目を選択される変数の候補として用いた．さらに生成したモデルと既存の理論との整合性を検証し，先行研究による知見とデータからの探索的な新たな知見の両面からのモデル構築を行った．また 10 年間という長期追跡調査に基づき要介護認定リスク因子を探索的に 300 以上の項目の中から精度よく抽出し各項目に適切な得点を与える事で，汎化能力の高い予後因子スコアの構築を可能とした．

2. 方法

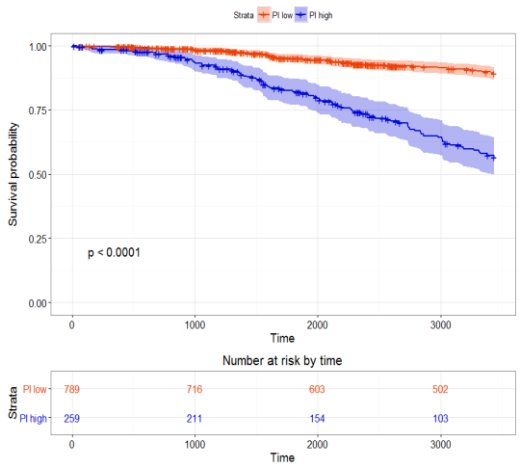
対象とするデータは変数が多くかつ変数によっては欠測の割合も大きい．そこで欠測に MAR (Missing at random)が仮定できると判断し，かつデータの分布に大きな偏りのない 198 個の変数に対して，多重補間法の中でも特に連鎖方程式による方法により欠測値を補完した．また，性別によりリスク因子が異なる事が先行研究より明らかなため男女別の解析を行った．

変数は調査項目に種類があるため,それぞれの種別で対応分析や主成分分析を用いて代表的な変数を抽出した. さらに, 要介護認定２のレベルまで健康状態が悪化する予後因子の探索を行うため, スパース性を仮定した Cox 比例ハザード回帰モデル(Elastic Net による罰則)を用いた. 研究結果である予後モデルの信頼性が重要となるが, 交差検証法により内的妥当性を考慮し, C 統計量を用いて予後モデルの予測の精確性を評価した.

3. 結果

最終的に分析対象となった 14171 人のうち, 10 年間の追跡期間中の要介護認定 2 以上発生は 2596 人であった. A 連合地域の男性については, 個人的属性では 75 歳以上である事が要介護認定に有意に高いハザード比 2.6(95%IC 1.8-3.7)を, 今の市町村の滞在期間が長い事が有意に低いハザード比 0.9(95%IC 0.8-0.99)を示した. 同地域男性の身体的・心理的・生活習慣・社会的特性では, 高血圧あり, 友人の家を訪ねる事がない, 昼寝の週間あり, 家事をしていない, 過去 1 年間に転んだ経験あり, 趣味がない, スポーツ的活動頻度, ドライブ・旅行などの頻度が低い, 今の市町村の在住期間が短いことが要介護 2 以上認定の高いリスクと有意に関連していた.その他の 6 地域の男女別リスク要因も同定した. また, 各自治体で男女別に推定された Cox 回帰モデルの回帰係数の推定値と標準誤差を用いて予後指標を作成した. 指標の大小によりリスクの高い群と低い群に分け, 生存率関数をプロットした（図１）. 図からもわかるように, 予後予測可能な予後指標を作成する事が出来た.

図１. A 連合地域の男性の予後指標別生存率関数のプロット



4. 考察

本研究では, 10 年間の追跡調査に基づいて予後スコアを作成した. この結果, 厚生労働省が介護予防の重点としたリスク要因は要介護リスク要因としてある程度の妥当性があり, さらに身体・心理・社会的要因も要介護認定のリスク要因であると結論付けた先行研究をほぼ裏付ける結論となった.今回は東海地域に限った分析であったが, 全国で地域ごとの特徴を捉えるべく, 今後の課題として日本全国で観測された JAGES データを用いて, 空間統計学の要素を加えて要介護認定のリスク要因を丁寧に見ていく必要がある.

演題：疾患原因となる遺伝子変異同定のためのエクソーム解析の現在

所属：東京医科歯科大学 難治疾患研究所 医科学数理分野 講師
氏名：重水 大智

要旨：次世代シークエンス技術の目覚ましい発展により、全ゲノムシークエンスに比べ低コストで、より深く遺伝子コード領域（エクソーム）をシークエンスする全エクソームシークエンス解析が、疾患原因変異を同定する解析の応用研究として盛んに行われている(図 1)。

一方で、次世代シークエンサーで読まれる配列は、比較的短いため(100-150bp)、そのショートリードを標準ゲノム配列に正確にマッピングし、高精度にバリアントを同定することが難しく、洗練された手法の開発が課題のひとつとしてあげられている。

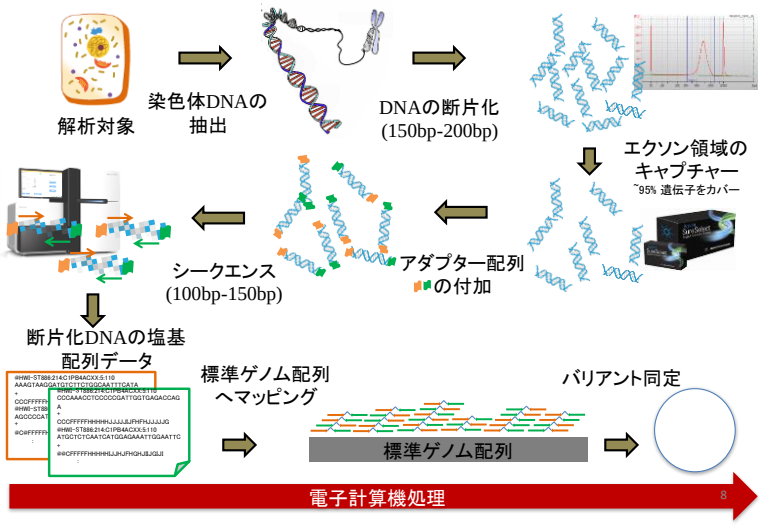


図1. Whole Exome Sequencing(WES)解析の流れ

本セミナーでは、①この次世代シークエンサーで読まれたショートリードのデータから高精度に一塩基置換（Single Nucleotide Variant: SNV）、短い挿入・欠失（Insertion and Deletion: InDel)の同定に成功した我々が開発した手法の紹介(SNV 一致率 99.9%、INDEL 一致率 97.3%） 1)-2)、②家系情報をもとに家系内の数検体をシークエンスし、遺伝形式を想定することにより疾患原因変異を絞り込んでいく「家系解析」によって疾患原因変異を同定した研究成果の紹介 3)-4)、③その他のこれまでの我々の研究成果の紹介 5)-9)、④これからのエクソーム解析の展望(図 2)、この四つに焦点を当てて述べたい。

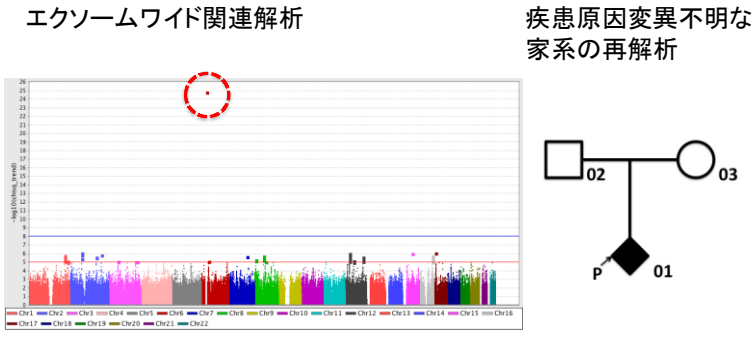


図2. 今後のエクソーム解析

【参考文献】

1. **Shigemizu, D.,** Fujimoto, A., Akiyama, S., Abe, T., Nakano, K., Boroevich, KA, Yamamoto, Y., Furuta, M., Kubo, M., Nakagawa, H., and Tsunoda, T. (2013) A practical method to detect SNVs and indels from whole genome and exome sequencing data. ***Sci Rep*** 3:2161.
2. Fujimoto, A., Furuta, M., Shiraishi, Y., Gotoh, K., Kawakami, Y., Arihiko, K., Nakamura, T., Ueno, M., Ariizumi, S., Nguyen, H., **Shigemizu, D.,** Abe, T., Boroevich, K., Nakano, K., Sasaki, A., Kitada, R., Maejima, K., Yamamoto, Y., Tanaka, H., Shibuya, T., Shibata, T., Ojima, H., Shimada, K., Hayami, S., Shigekawa, Y., Aikata, H., Ohdan, H., Marubashi, S., Yamada, T., Kubo, M., Hirano, S., Ishikawa, O., Yamamoto, M., Yamaue, H., Yamaue, H., Chayama, K., Miyano, S., Tsunoda, T., and Nakagawa, H. (2015) Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity. ***Nat Commun*** Jan 30;6:6120**.**
3. **Shigemizu, D.,** Aiba, T., Nakagawa, H., Ozaki, K., Miya, F., Satake, W., Toda, T., Miyamoto, Y., Fujimoto, A., Suzuki, Y., Kubo, M., Tsunoda, T., Shimizu, W., and Tanaka, T. (2015) Exome analyses of long QT syndrome reveal candidate pathogenic mutations incalmodulin-interacting gene. ***PLoS One*** 10(7), e0130329.
4. Aiba, T., Ishibashi, K., Wada, M., Nakajima, I., Miyamoto, K., Okamura, H., Noda, T., **Shigemizu, D.,** Satake, W., Toda, T., Kusano, K., Kamakura, S., Yasuda, S., Sekine, A., Miyamoto, Y., Tanaka, T., Ogawa, S., and Shimizu, W. (2014) Clinical Significance of Whole Exome Analysis Using Next Generation Sequencing in the Genotype-negative Long-QT Syndrome. ***Circulation*** 130(Suppl 2) A13817.
5. Ichikawa, M., Aiba, T., Ohno, S., **Shigemizu, D.**, Ozawa, J., Sonoda, K., Fukuyama, M., Itoh, H., Miyamoto, Y., Kubo, M., Tsunoda, T., Makiyama, T., Tanaka, T., Shimizu, W., and Horie, M. (2016) Phenotypic Variability of *ANK2* Mutations in Patients with Inherited Primary Arrhythmia Syndromes. ***Circ J*** Oct. 25.
6. Yagihara, N., Watanabe, H., Barnett, P., Duboscq-Bidot, L., Thomas, AC., Yang, P., Ohno, S., Hasegawa, K., Kuwano, R., Chate, S., Redon, R., Schott, JJ., Probst, V., Koopmann, TT., Bezzina, CR., Wilde, AAM., Nakano, Y., Aiba, T., Miyamoto, Y., Kamakura, S., Darbar, D., Donahue, BS., **Shigemizu, D.**, Tanaka, T., Tsunoda, T., Suda, M., Sato, A., Minamino, T., Endo, N., Shimizu, W., Horie, M., Roden, DM., and Makita, N. (2016) Variants in the *SCN5A* promoter associated with various arrhythmia phenotypes. ***J Am Heart Assoc*** Sep 13;5(9).
7. **Shigemizu, D.,** Momozawa, Y., Abe, T., Morizono, T, Boroevich, K., Takata, S., Ashikawa, K., Kubo, M., and Tsunoda, T. (2015) Performance of comparison of four commercial human whole-exome capture platforms. ***Sci Rep*** 5:12742.
8. Miya, F., Kato, M., Shiohama, T., Okamoto, N., Saitoh, S., Yamasaki, M., **Shigemizu, D.,** Abe, T., Morizono, T., Boroevich, K., Kosaki, K., Kanemura, Y., and Tsunoda, T. (2015) A combination of targeted enrichment methodologies for whole-exome sequencing reveals novel pathogenic mutations. ***Sci Rep*** 5:9331**.**
9. Makita, N., Yagihara, N., Crotti, L., Johnson, CN., Beckmann,BM., Roh, MS., **Shigemizu, D.,** Lichtner, P., Ishikawa, T., Aiba, T., Homfray, T., Behr, ER., Klug, D., Denjoy, I., Mastantuono, E., Theisen, D., Tsunoda, T., Satake, W., Toda, T., Nakagawa, H., Tsuji, Y., Tsuchiya, T., Yamamoto, H., Miyamoto, Y., Endo, N., Kimura, A., Ozaki, K., Motomura,H., Suda, K., Tanaka, T., Schwartz, PJ., Meitinger, T., Kääb, S., Guicheney, P., Shimizu, W., Bhuiyan, ZA., Watanabe, H., Chazin, WJ., and George, AL. (2014) Novel Calmodulin (CALM2) Mutations Associated with Congenital Arrhythmia Susceptibility. ***Circ Cardiovasc Genet*** Aug 7(4):466-474.

データが育種を加速する

小野木　章雄
JST さきがけ・東京大学大学院農学生命科学研究科

育種とは人間にとって有用な動植物、例えば乳牛やブタ、コムギやイネを、より人間にとって好ましいように遺伝的に改良することを指す。例えば乳牛であればより多くの乳を生産するように、またコムギであればより高い収量が得られるように改良する。どのように育種するかは繁殖方法の違いに従い生物種によって多様であるが、共通しているのは優れた親同士を交配し、得られた子の中からまた優れたものを選び交配するということを繰り返す点である。この「優れたものを人間が選ぶ」という点を、「環境に適応したものを自然が選ぶ」と置き替えると、育種の仕組みは自然選択による生物進化と基本的に同じであることがわかる。私達が普段食するほとんどの穀物や野菜、肉類は育種により改良されてきた。

育種の効率を左右する重要な点はいかに正確に優れたものを選ぶかにある。実際に飼育あるいは栽培すればどの個体が優れているか分かるように思えるが、それほど単純ではない。例えば肉質のよいウシを選ぶことを考えると、屠畜して肉質を評価した時点でその個体は子孫を残すことはできない。収量の高いコムギを選ぶことを考えると、ある年ある場所で収量が高くても、次の年あるいは別の場所では全く駄目かもしれない。また何千もの候補から選ぶ場合には、その何千を全て栽培あるいは飼育し評価しなければならない。

こうした問題の中で優れたものを選ぶためには、肉質や収量などを実際に評価する前に予測する、あるいは環境による変動を除いて遺伝的な優劣を推定する、という操作が有効であり、そのための統計学的手法が研究されてきた。先のウシの例で言えばこれまでに蓄積された肉質のデータから、選びたいウシの肉質を予測するのである。これであれば屠畜することなく選ぶことができそのウシは子孫を残すことができる。動物育種において予測・推定ために発展してきた統計学的な手法に混合方程式モデル及びその解である Best linear unbiased prediction（BLUP）がある。従来の動物育種では個体間の類似度を家系で定義して、新たな個体に対する予測を可能としてきた。

近年になり様々な生物種において大規模データ、特にゲノムに関するデータが得られるようになってきた。すなわちゲノムワイドな DNA マーカー遺伝子型や、全塩基配列が比較的安価に決定できるようなっている。そして現在の育種はこの大規模データに基づく予測を利用した方法に移行しつつある。ゲノミックセレクション（GS）と呼ばれる育種手法では、個体間の類似度を DNA マーカー対立遺伝子の共有割合により定義することで予測を行う。GS はまず乳用牛に取り入れられ、育種速度の劇的な向上に寄与し（1)、肉用牛やブタなど他の動物種でも検討・導入が進んでいる（2、3)。

GS は植物育種においても急速に利用が進んでいる。全ゲノム情報を用いた予測はイネ（4)、コムギやトウモロコシ（5)、ダイズ（6）などの様々な主要穀物、ナシ（7）やリンゴ（8）などの果樹、マツ（9）などの樹木など、多様な植物種で有用であることが確認さ

れてきた。全ゲノム情報という大規模データにより今後の育種が加速されていくのは明らかであろう。

一方で今日、大規模データはゲノム以外にも得られている。例としてトランスクリプトームやメタボロームなど他のオミックスデータが挙げられる。例えば葉のメタボロームからのトウモロコシ収量の予測や（10)、血液中のメタボロームからのブタ肉量の予測などが報告されている（11)。また大規模データは動植物の育つフィールドからも得られる。例えば乳用牛であれば搾乳ロボットにより搾乳回数やその時間、給餌量などが各個体について得られる。植物であれば気温や日射量などの気候情報や、土壌水分センサーなどのデータロガーによる環境情報、さらに定点カメラやドローンによる画像などが日々蓄積されている。現在のところこういった大規模かつ多様なデータを統合して育種に利用する方法は考え出されていないが、このようなデータがゲノムデータとともに予測能力の向上を始めとする育種の効率化に寄与できる可能性は十分にある。今後育種はこれまで以上にデータ駆動型の応用科学分野となっていくと予想され、データとそれを利用するための統計学・機械学習的手法の重要性が増していくと考えられる。

参考文献

1. Garcia-Ruiz A et al. (2016) Proc Natl Acad Sci U S A. 113: E3995-4004.
2. Onogi A et al. (2014) J Anim Sci 92: 1931-1938.
3. Esfandyari H et al. (2016) Genet Sel Evol 48:40.
4. Onogi A et al. (2015) Theor Appl Genet 128:41-53.
5. Crossa J et al. (2010) Genetics 186:713-724.
6. Jarquin D et al. (2016) G3 (Bethesda) 6:2329-2341.
7. Iwata H et al. (2013) Breed Sci 63:125-140.
8. Kumar S et al. (2012) PLoS One 7:e36674.
9. Resende MFJ et al. (2012) Genetics 190:1503-1510.
10. Riedelsheimer C et al. (2012) Nat Genet 44:217-220.
11. Rohart F et al. (2012) J Anim Sci 90:4729-4740.

Non-linear State-Space Modeling for Wind Speed and Direction

Takayuki Shiohama*
Department of Information and Computer Technology
Faculty of Engineering, Tokyo University of Science
6-3-1 Niijuku, Katsushika, Tokyo 125-8585, Japan
shiohama@rs.tus.ac.jp

Abstract

The state space form is a useful framework for estimating unobserved state variables from some given observations. The applications can be found in diverse areas of natural science and engineering such as ecology, epidemiology, meteorology and economics and finance. The wind speed and directions have complex time series probability structures involving highly non-Gaussian and nonlinear transition. In this study, we consider a simulation-based inference using the sequential Monte Carlo methods for computing the posterior distributions for the state variables given all available observations. We propose here an alternative approach that allows us to extend the methods of importance sampling distributions incorporating with the class of circular Markov transition densities. The resulting methods are compared with various resampling schemes with real data applications.
Keywords: Circular data, state-space modeling, particle filter

1. Introduction

Circular (or directional) data refer to data recorded as points for which directions are measured, typically in the fields of biology, geography, medicine, and astronomy. For such data, which are usually expressed in terms of compass angles or pairs of sine and cosine variables, the beginning and end of the scale in the domain coincide. Owing to this periodicity, analyzing circular data is challenging because traditional statistics are not meaningful, and may even be misleading when the particular definition of the domain is ignored. Although most circular data are in the form of time series, little research has been carried out in the field of circular time series analysis compared with the number of circular time series modeling approaches.

In general, three main approaches are used to model circular time series. The first method is used to obtain circular-valued random variables by wrapping; one example is the wrapped autoregressive process of Breckling (1989). The second approach is based on a link function that maps a line onto a circular domain, called a linked autoregressive moving average process. This model was proposed by Fisher and Lee (1994). The last

*Research reported here was supported by JSPS KAKENHI Grant Number 26380401

approach specifies the density of the conditional distribution, including the Markov process of Wehrly and Johnson (1980), Möbius transformation of Kato (2010), and hidden Markov models of Holzmann et al. (2006). Abe et al. (2016a) studied the circular Markov process of Wehrly and Johnson (1980) and obtained theoretical circular autocorrelation structures under simple model assumptions. According to their results, circular autocorrelations are determined by the mean resultant length of the underlying circular density of the process. Abe et al. (2016b) considered the circular Markov processes whose concentration parameter could be time-varying.

Sequential Monte Carlo (SMC) methods are the set of simulation-based methods which provide a convenient and attractive approach to computing posterior distributions. Over the last few years, there has been a proliferation of scientific papers on SMC methods and their applications. Several closely related algorithms, under the names of bootstrap filters, condensation, particle filters, Monte Carlo filters, interacting particle approximations and survival of the fittest, have appeared in several research fields.

Many data in directional time series applications display nonlinear features such as heteroskedasticity and a nonlinear relationship between wind direction and speed. These features become more and more relevant as the length of the observed time series increases and as the series itself is subject to changes in the dynamic structure. In this study, we address the circular process of Wehrly and Johnson (1980), which allows time-varying concentration parameters, along with the stochastic volatility process of wind speed. The proposed model can incorporate the time-varying autocorrelations of the observed circular time series, and be estimated by a sequential Monte Carlo methods.

References

Abe, T., Ogata, H., Shiohama, T., and Taniai, H. (2016a). A circular autocorrelation of stationary circular Markov processes. *Revision submitted*.

Abe, T., Ogata, H., Shiohama, T., and Taniai, H. (2016b). Modeling circular Markov processes with time-varying concentration. *Submitted*.

Breckling, J. (1989). *The analysis of directional time series: application to wind speed and direction*. Lecture Notes in Statistics, **61**, Springer-Verlag, Berlin.

Fisher, N. I. and Lee, A. J. (1994). Time series analysis of circular data. *Journal of the Royal Statistic Society, Series B*, **70**, 327 – 332.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental Ecological Statistics*, **13**, 325 – 347.

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

Wehrly, T. E. and Johnson, R. A. (1980). Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, **66**, 255–256.

Weibull分布とsine-skewed von Mises分布を用いたシリンダー上の分布の生成

阿部 俊弘
南山大学 理工学部 システム数理学科
E-mail: abetosh@ss.nanzan-u.ac.jp

シリンダー上の分布で, どういった性質を持つのが “良い” と言えるのであろうか？良いモデルの条件として, 多様な形状を表現可能であり, 扱いやすい確率密度関数 (これは性質を調べたり, パラメータ推定をする際に重要) であることが挙げられる. さらに, 周辺分布や条件付き分布が知られている既存の分布であり, 長さと角度の従属構造が妥当な結合分布であるとさらに扱いやすくなる. 一方で, (後で例として用いる)blue periwinkle データのように, シリンダーデータのいくつかを眺めていると, 線形成分が増加すると, 角度部分の集中の度合いも増えている傾向がある. これらの条件を満たすモデルとして, Abe & Ley (2016) は確率密度関数が

$$(\theta, x) \mapsto \frac{\alpha\beta^\alpha}{2\pi\cosh(\kappa)} \left\{1 + \lambda\sin(\theta-\mu)\right\} x^{\alpha-1} \exp\left[-(\beta x)^\alpha \left\{1 - \tanh(\kappa)\cos(\theta-\mu)\right\}\right] \quad (1)$$

で与えられるシリンダー上のモデルを提案している. ここで, $(x, \theta) \in [0, \infty) \times [-\pi, \pi)$, $\alpha > 0$ は (長さの分布の) 形状パラメータ, $\beta > 0$ は (長さの分布の) 尺度パラメータ, $-\pi \leq \mu < \pi$ は (円周分布の) 位置パラメータ, $\kappa \geq 0$ は (円周分布の) 集中パラメータであり, $-1 \leq \lambda \leq 1$ は (円周分布の) 歪みを表すパラメータである.

我々はシリンダー上の分布 (1) を WeiSSVM と呼ぶことにする: “Wei” は $\mathbb{R}^+$ 上の (線形の)Weibull 分布 $x \mapsto \alpha\beta^\alpha x^{\alpha-1}\exp\left\{-(\beta x)^\alpha\right\}$ を表し, “SSVM” は sine skewed von Mises 分布 (Abe & Pewsey, 2011) の省略である.

$\lambda = 0$ のとき, Weibull 分布と von Mises 分布を組み合わせたシリンダー上の分布を得る. 我々の知る限り, この特殊な場合も新しい分布となっている. もし $\alpha = 1$(指数分布) ならば, (1) は

$$\frac{\beta}{2\pi\cosh(\kappa)} \exp\left[-\beta x \left\{1 - \tanh(\kappa)\cos(\theta-\mu)\right\}\right]$$

となり, Johnson-Wehrly-1 モデルと同値であり, 指数分布と von Mises 分布の組み合わせである. これを ExpVM と書くことにする. この分布は, Johnson-Wehrly-1 モデルと同値であるが, パラメータ β と κ の間に大小関係の制約がつかない.

集中パラメータが $\kappa = 0$ となるとき, WeiSSVM の確率密度関数は

$$\frac{1}{2\pi}\left\{1 + \lambda\sin(\theta-\mu)\right\} \times \alpha\beta^\alpha x^{\alpha-1}\exp\left\{-(\beta x)^\alpha\right\}$$

となり, これは Weibull 分布と cardioid 分布の積となる.

Johnson & Wehrly (1978) は彼らの最初の二つの分布に対して, “A major limitation of the two previous densities is that if X and Θ are independent, then Θ is forced to be uniformly distributed on the circle” と述べているが, WeiSSVM は sine skewed circular 分布の構造から, このような欠点に悩まされることはない.

尤度比検定を用いて, WeiSSVM の部分モデルの検定が容易にできる. 各パラメータ $\eta \in \{\alpha, \beta, \mu, \kappa, \lambda\}$ に対して, $\hat{\eta}$ を制約なしの最尤推定値とし, $\hat{\eta}_0$ を各帰無仮説の下での最尤推定値とする. WeiSSVM の部分モデルである, Johnson-Wehrly-1, または, ExpVM の検定は有意水準を α とすると, 帰無仮説 $\mathcal{H}_0 : \{\alpha = 1\} \cap \{\lambda = 0\}$ の下で, 検定統計量

$$T_{\mathrm{JW1}} := -2(\log \ell(1, \hat{\beta}_0, \hat{\mu}_0, \hat{\kappa}_0, 0) - \log \ell(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\kappa}, \hat{\lambda}))$$

が $\chi^2_{2;1-\alpha}$ (自由度 2 の χ^2 分布の上側 α 点) を超えたら棄却すればよい.

他にも, 検定統計量

$$T_{\mathrm{Indep}} := -2(\log \ell(\hat{\alpha}_0, \hat{\beta}_0, \hat{\mu}_0, 0, \hat{\lambda}_0) - \log \ell(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\kappa}, \hat{\lambda}))$$

を $\chi^2_{1;1-\alpha}$ と比較してシリンダー分布の変量間の独立性の検定もできる.

参考文献

[1] Abe, T. & Ley, C. (2016). A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econometrics and Statistics*, in press.

[2] Abe, T. & Pewsey, A. (2011). Sine-skewed circular distributions. *Statist. Pap.*, **52**, 683–707.

[3] Johnson, R.A. & Wehrly, T.E. (1978). Some angular-linear distributions and related regression models. *J. Amer. Statist. Assoc.*, **73**, 602–606.

経時測定データの判別と変数・決定境界の選択

滋賀大学データサイエンス教育研究センター　松井秀俊

1　概要

複数の個体に対して経時的に観測，測定されたデータを関数化処理し，得られた関数化データ集合に基づく解析を行う方法は関数データ解析（Ramsay and Silverman, 2005）とよばれ，生命科学や気象学など幅広い分野において研究が行われている．本研究では，関数データに基づくロジスティック回帰モデルにスパース正則化を適用する方法について検討する．

回帰モデルに含まれるパラメータ推定において，L_1 ノルムを含む制約を課したスパース正則化法は，推定量に安定性を与えると同時に変数選択の役割も担う方法として注目を集め，理論，応用の両側面から幅広い適用例が報告されている（例えば，Hastie et al., 2015）．スパース正則化を関数ロジスティック回帰モデルの推定に適用することで，パラメータの推定と同時に，判別に寄与している関数データの変数を選択することができる (Kayano et al., 2016)．本研究では，sparse group lasso (Simon et al., 2013) 型に基づく制約を用いることで，関数データとして与えられた変数と決定境界の選択を同時に行う方法について議論する．モデルの推定については，正則化最尤法の枠組みで推定法を紹介する．以上の手法を，イースト遺伝子発現データ解析へ適用した結果について報告する．

2　関数ロジスティック回帰モデル

いま，関数データとして与えられた n 個の p 変量説明変数 $x_{ij}(t)$ $(i = 1, \ldots, n,\ j = 1, \ldots, p)$ と，L (≥ 3) 群のラベルを表す二値変数 y_{il} $(l = 1, \ldots, L-1)$ が得られたとする．このとき，関数ロジスティック回帰モデルは，データが観測された下での l 群への判別確率 $\pi_l(\boldsymbol{x}_i; \boldsymbol{b})$ を用いて次で与えられる．

$$\log\left\{\frac{\pi_l(\boldsymbol{x}_i;\boldsymbol{b})}{\pi_L(\boldsymbol{x}_i;\boldsymbol{b})}\right\} = \beta_{0l} + \sum_{j=1}^{p}\int x_{ij}(t)\beta_{jl}(t)dt. \quad (l = 1, \ldots, L-1) \tag{1}$$

ここで β_{0l} は定数項，$\beta_{jl}(t)$ は係数関数で，$\boldsymbol{b}$ はモデルに含まれる未知パラメータベクトルとする．説明変数および係数関数が基底関数展開によって表現できるという仮定を置くことで，モデル (1) は次式のように，ベクトルと行列に基づく回帰モデルの形で表現できる．

$$\log\left\{\frac{\pi_l(\boldsymbol{x}_i;\boldsymbol{b})}{\pi_L(\boldsymbol{x}_i;\boldsymbol{b})}\right\} = \sum_{j=1}^{p} Z_j \boldsymbol{b}_{jl}. \tag{2}$$

ただし Z_j は説明変数から構成される既知の行列，$\boldsymbol{b}_{jl}$ はパラメータベクトルである．

3　モデル推定

関数ロジスティック回帰モデル (2) の係数パラメータ $\boldsymbol{b}$ を正則化最尤法，すなわち次で定義される正則化対数尤度関数の最大化により推定する．

$$\ell_\lambda(\boldsymbol{b}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{b}) - nP_{\lambda,\alpha}(\boldsymbol{b}).$$

ただし $P_{\lambda,\alpha}(\boldsymbol{b})$ はパラメータに対する制約関数で，制約そのものの強さを規定する正則化パラメータ $\lambda > 0$ と，追加の調整パラメータ $\alpha \in [0,1]$ である．これに L_1 ノルムを含む関数を仮定することで，いくつかのパラメータを 0 と推定できる．本研究では，次の sparse group lasso 型制約を用いる．

$$P_{\lambda,\alpha}(\boldsymbol{b}) = n(1-\alpha)\sum_{j=1}^{p}\lambda_j\left\{\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2^2\right\}^{1/2} + n\alpha\sum_{j=1}^{p}\lambda_j\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2.$$

ただし $\lambda_j = \sqrt{M_j}\lambda$，$M_j$ はパラメータベクトル $\boldsymbol{b}_{jl}$ の次元とする．右辺第一項は変数選択の役割を持ち，第二項が決定境界の選択の役割を持つ．この制約に基づく推定量を解析的に導出することは困難であるため，反復的に更新値を得るためのアルゴリズムを適用する．

4　適用例

提案した手法を，経時測定されたイースト遺伝子の発現データへ適用した．ここで扱うデータは，6 種類の実験によって得られたイースト遺伝子の発現量と，各遺伝子の 5 種類の細胞周期からなる．本研究では，各実験が細胞周期の判別に実際に寄与しているかを調べるために上記提案手法を適用した．

参考文献

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalization*, Boca Raton: Chapman & Hall/CRC.

Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S., and Miyano, S. (2016), Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to timedependent biomarker detection, *Biostatistics*, 17, 235–248.

Ramsay, J. and Silverman, B. (2005), *Functional data analysis 2nd ed.*, New York: Springer.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), A sparse-group lasso, *J. Comput. Graph. Statist.,* 22, 231–245.

Asymmetric logistic regression model for fishery assessment

Osamu Komori[1], Shinto Eguchi[2], Shiro Ikeda[2],
Hiroshi Okamura[3], Momoko Ichinokawa[3],
& Shinichiro Nakayama[3].

[1]Department of Electrical, Electronic and Computer Engineering,
University of Fukui
[2]The Institute of Statistical Mathematics
Midori-cho 10-3, Tachikawa Tokyo 190-8562, Japan.
[3]National Research Institute of Fisheries Science, Fisheries Research Agency
2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan

Abstract

The long time-series data such as the RMA legacy data are essential for single and multispecies stock assessment because the biomass in that data reflects the stock status properly. However, the available assessment data usually have limited sample sizes and the ratio of collapsed stocks to non-collapsed stocks is highly imbalanced (very few collapse stocks in comparison with the large number of non-collapsed stocks). Moreover the stock status (collapse or non-collapse) has (often large) uncertainty because it is usually estimated by a population dynamics model. To allow for the imbalancedness and uncertainty involved in the fishery data, we propose a new binary regression model with mixed effects for estimation of stock status by employing an asymmetric model. In the proposed model, we assume that the small part of observations of the collapsed stocks are distributed in the same way as those of the non-collapsed stocks, resulting in a mixture model of conditional probability of collapsed status given explanatory variables in fishery-related data. In the estimation equation, we observe

that the weights for the non-collapse stocks are relatively reduced, which in turn puts more importance on the small numbers of observations of collapse stocks. When we applied our approach to the RAM legacy data, the estimated collapse probabilities were much improved with a little degeneration of the estimated probabilities of non-collapsed stocks. This improvement of the probability to identify the collapsed stocks in the RAM legacy data suggested that unassessed stocks in the FAO data are in the risk of being collapsed. We clarify the characteristics such as ISSCAAP groups of species in the risky stocks in FAO data to contribute to the global fishery sustainable management. We demonstrated that the proposed method was promising by several simulation studies.

Key words: ecological binary data, mixed effect logistic regression,model fitting

Acknowledgements

This research was supported by Japan Science and Technology Agency (JST), Core Research for Evolutionary Science and Technology (CREST).

性感染症流行動態解析の展望

北海道大学 人獣共通感染症リサーチセンター バイオインフォマティクス部門　大森亮介

感染症流行を制御する為にはそのダイナミクスを定量的に理解する必要がある。そのために、感染症流行モデルとして最も基本的なモデルである SIR モデルが広く使われてきた。SIR モデルとは、宿主集団を感染状態で分類し、それぞれの感染状態の個体数の時系列変化を表現したモデルである。最も基本的な SIR モデルでは、宿主集団を *S*:未感染, *I*:感染, *R*:回復で分類した以下のようなモデルである。

$$S' = -\beta SI,$$

$$I' = \beta SI - \gamma I,$$

$$R' = \gamma I.$$

新規感染は、未感染者と感染者の接触で起きるために、新規感染者数の期待値は未感染者数 S と感染者数 I の両方に比例する非線形現象となり、感染症流行ダイナミクスの解析の困難さの原因となっている。この数理モデルを感染者数の時系列データにあてはめ、モデルパラメータを推定することが感染症疫学解析の第一歩となる。推定されたモデルパラメータより、多くの先行研究は基本増殖率(R_0)と呼ばれる統計量を推定している。R_0 とは、宿主集団がごく少数の感染者を除き全員が未感染者の状態の時の、流行が起きる為の条件(R_0 が 1 以上でなければ流行は起きない)であり、生物学的な解釈としては一人の感染者が引き起こす二次感染者数の期待値である。また、R_0 から流行規模が推定でき、流行を未然に防ぐ為のワクチン接種率の推定などにも有用であり、感染症疫学において R_0 は非常に重要な統計量となっている。

感染症疫学において、性感染症は解析が困難である感染症の一つである。インフルエンザウイルスや麻疹のような飛沫感染により伝搬が起きる感染症と違い、性感染症は性的接触を行わない限り通常は伝搬が起きず、感染経路が限定されている。最も基本的な SIR モデルではモデルの対象である宿主集団ではすべての宿主が等しく接触するという仮定を置いているが、宿主個人間の接触の不均一性をモデリングする必要がある。そこで、May and Lloyd 2001 は、パートナー数は宿主個人間で異なり、感染確率はパートナー数に比例すると仮定し、R_0 と宿主集団レベルでのパートナー数の分布の関係性を以下のように導出した。

$$R_0 = \rho_0(1 + C_v^2).$$

ここで ρ_0 は一人の感染者が引き起こす二次感染者数の平均、C_vはパートナー数の分布の変動係数である。パートナー数の分布は冪乗則に従う様な分布と似ており、もしパートナー数の分布がスケールフリーネットワークであれば(パートナー数の分散が無限大であれば)、上式の R_0 は無限大となる。つまり、性感染症の侵入は絶対に起きることを意味する。もちろん実際のパートナー数の分散は無限大ではないが、分散は十分に大きいために非常に大きい R_0 となっており侵入を防ぐことは困難であると予想される。この解析は、モデルの仮定が非常に大胆なものではあるものの（感染確率はパートナー数に比例するという仮定）、性感染症の流行動態は、誰と誰が性的接触を行うパートナーシップを結んでいるかを記述した性的接触ネッ

トワーク、少なくともそのネットワークの次数分布を知る事が必須であることを明らかにした。

このような背景から、現実の性的接触ネットワークを調査する性行動調査が様々な場所で行われた。しかしながら、性行動を調べる方法は基本的には質問表を用いた自己申告であり、かつ、性行動という調査対象の特性から、調査結果が非常に大きなバイアスを含む事は明白である。我々は USAID が行ったサブサハラアフリカ２５カ国における最大規模の性行動調査から次数分布を推定し(Omori et al. 2015)、HIV 有病率(感染者数の割合)との関連性を検討したが有意な関連性は検出されなかった(Omori et al. 2016)。性行動調査データは性感染症の流行動態を明らかにするにはその正確性が不足しており、他のデータからの性感染症の流行動態を推定する必要がある。

そこで我々は、性感染症の流行規模から他の性感染症の流行動態を推定する方法を開発している。この手法は以前から提唱されており、特に、HSV-2 の流行規模から HIV の流行リスクを推定する方法が注目されてきた。先行研究では、HIV と HSV-2 の流行を記述した SIR モデルを改良した数理モデルを用いて、HSV-2 の流行規模が大きければ大きいほど HIV の流行リスクは高いという正の相関を予想している(Abu-Raddad et al.)。しかしながら、このモデルは May and Lloyd 2001 と同様に、陽に性的接触ネットワークを記述しておらず、性的接触ネットワークの構造との関連性が明らかでない。よって本研究は個体ベースモデル（エージェントベースモデル）を用いて時々刻々と変化する性的接触ネットワークを記述し、そのネットワーク上での HIV と HSV-2 の流行の関連性を調査した。性的接触ネットワークを記述するモデルとして、上記の性行動データから得られた、パートナーを新たに獲得する確率とパートナーシップを解消する確率の分布の推定量をベースラインとした、時間変動する重み付けランダムグラフを構築した。新規パートナーシップ形成の際に rewiring を行う事でクラスター係数と次数相関の取り得る値を純粋な重み付けランダムグラフから拡張させた。結果として、HIV と HSV-2 では、ネットワーク統計量との関連性に大きな違いがある事が明らかになった。HIV の有病率は次数相関と弱い負の相関、クラスター係数と弱い正の相関を示したが、HSV-2 の有病率は次数相関と弱い正の相関、クラスター係数と負の相関を示した。また、HIV 流行ポテンシャル(平衡状態での有病率)の推定は HSV-2 の有病率のみからでは精度が悪く、次数相関やクラスター係数といったネットワークの統計量と合わせて推定する必要があることが判明した。

Reference

1. May, R. M., & Lloyd, A. L. (2001). Infection dynamics on scale-free networks. *Physical Review E*, *64*(6), 066112.

2. Omori, R., Chemaitelly, H., & Abu-Raddad, L. J. (2015). Dynamics of non-cohabiting sex partnering in sub-Saharan Africa: a modelling study with implications for HIV transmission. *Sexually transmitted infections*, sextrans-2014.

3. Omori, R., & Abu-Raddad, L. J. (2016). Population sexual behavior and HIV prevalence in Sub-Saharan Africa: missing links?. *International Journal of Infectious Diseases*, *44*, 1-3.

ハイスループットシークエンサによる マイクロサテライトリピート数推定法の高精度化

小島 要　　河合洋介　　長崎正朗

東北大学　東北メディカル・メガバンク機構

1　背景

マイクロサテライトにおけるリピート数変異のいくつかは疾患と関連することが分かっており，例えば，ハンチンチン遺伝子における CAG リピートとハンチントン病の発症率の関連が報告されている．ハイスループットシークエンスデータからのマイクロサテライトのリピート数推定法は大まかには二つに分けられる。一つは、マイクロサテライトをまたぐ形でアラインメントされたリードについて単位配列のリピート数を直接数え上げることでリピート数を推定する方法である。リピート数変異そのものの検出とリピート数の推定の双方において精度が高いが，十分な長さのフランキング領域が正確なリードのアラインメントにおいて必要であることから，リード長が対象のマイクロサテライト領域より十分に長い必要がある．もう一つは，対象のマイクロサテライトの両端のフランキング領域にアラインメントされたペアドエンドリードを用いる方法である．通常インサート長はリード長より長いことから，直接リード上で単位配列を数え上げる方法では扱うことができなかった比較的長いマイクロサテライトについても扱うことできる利点がある．しかしながら，特に低深度の HTS データにおいては，直接単位配列を数え上げる方法に比べて精度が低い問題があった．この問題を解決すべく，ペアドエンドリードの距離をもとにした各個人のリード生成モデルを遺伝子系図によりつなぎ合わせ，多検体についてリピート数を同時推定する coalescentSTR と呼ばれる手法について提案を行っており，本発表で紹介を行う．

2　モデルについて

2.1　ペアドエンドリード距離からのリピート数の推定

一検体についてペアドエンドリードの距離からマイクロサテライトにおけるリピート数を推定する統計モデルについて述べる．以下ではこのモデルを Basic モデルと呼ぶ．d 番目のアラインメントされたリードペアの正鎖リードの開始位置を $s^{(d)}$ とする．また，d 番目のアラインメントされたリードペアの逆鎖リードの終了位置を $e^{(d)}$ とする．d 番目のリードペアのインサートサイズは $e^{(d)} - s^{(d)}$ で与えられ，$e^{(d)} - s^{(d)}$ を $l^{(d)}$ で表すとする．ある一検体のゲノム配列において $s^{(d)}$ と $e^{(d)}$ の間に x bp の挿入がある場合，$l^{(d)}$ は実際のインサートサイズより x bp だけ短くなる．一方，$s^{(d)}$ と $e^{(d)}$ の間に x bp の欠失がある場合，$l^{(d)}$ は実際のインサートサイズより x bp だけ長くなる．これを利用して，$l^{(d)}$ と実際のインサートサイズの違いを検出することで Basic モデルによりリピート数を推定する．u，n_r，n_1 及び n_2 をそれぞれリピートする単位配列の長さ，参照ゲノムでのリピート数，ハプロタイプ 1 のリピート数，そしてハプロタイプ 2 のリピート数とする．d 番目のリードペアがハプロタイプ 1 からのマイクロサテライトをまたぐ DNA フラグメントをもとに生成された場合，実際のインサートサイズは $l^{(d)} + u \cdot (n_1 - n_r)$ で与えられることから，シークエンスデータのインサートサイズの分布を表す $\mathcal{F}$ により $l^{(d)}$ の発生確率は $\mathcal{F}(l^{(d)} + u \cdot (n_1 - n_r))$ で与えられる．但し，DNA フラグメントの開始または終了位置がマイクロサテライト内にある場合，該当の DNA フラグメントはリピート数の推定に利用することはできないことから，e_m を参照ゲノムにおけるマイクロサテライトの終了位置とする時，$l^{(d)}$ は $e_m - s^{(d)}$ より長い必要がある．また，DNA フラグメントが十分大きな長さ K より大きい場合についても除外する，すなわち $l > K$ の時 $\mathcal{F}(l)$ を 0 に設定する．$l^{(d)}$ の発生確率は $\mathcal{F}$ を正規化することで次の式で与えられる:

$$P(l^{(d)} \mid n) = \begin{cases} \frac{\mathcal{F}(l^{(d)} + u \cdot (n - n_r))}{N(s^{(d)}, n)} & \text{if } l^{(d)} > e_m - s^{(d)} \ \&\ l^{(d)} \leq K - u \cdot (n - n_r) \\ 0 & \text{otherwise} \end{cases} .$$

ここで，$N(s,n)$ は次で与えられる正規化係数とする:

$$N(s,n) = \sum_{l=e_m-s+1}^{K-u\cdot(n-n_r)} \mathcal{F}(l+u\cdot(n-n_r)).$$

各リードペアが 2 つのハプロタイプから同じ確率で生成されるものとし，$l^{(d)}$ の尤度は次で与えられる:

$$\prod_{d=1}^{D} P(l^{(d)} \mid n_1, n_2) = \prod_{d=1}^{D} \frac{1}{2}\left(P(l^{(d)} \mid n_1) + P(l^{(d)} \mid n_2)\right). \tag{1}$$

ここで，D はリードペアの総数とする．最大と最小のリピート数を考え，それぞれ $n_{\max}$，$n_{\min}$ とし，式 (1) を最大化する n_{i1} と n_{i2} のペアを $\{n_{\min},\ldots,n_{\max}\}\times\{n_{\min},\ldots,n_{\max}\}$ から探索する．

2.2 多検体の遺伝子系図を考慮したリピート数の推定

coalescent 木についてマイクロサテライト周辺のフェーズ済み遺伝子型情報から推定し，多検体の Basic モデルについて結合しリピート数の同時推定を行う coalescentSTR について以下で述べる．リピート数の変化は推定された coalescent 木に従うものとし，対象のマイクロサテライト周辺のフェーズ済みの遺伝子型を V とする時，リピート数の事前分布を V から推定された coalescent 木で考慮することで，ペアドエンドリードから推測されるインサートサイズの発生確率が表される．まず，$l_i^{(d)}$ を i 番目の検体の d 番目のリードペアのインサートサイズとし，n_{i_1} と n_{i_2} を i 番目の検体のハプロタイプ 1 とハプロタイプ 2 のリピート数とする．この時，インサートサイズ $l_i^{(d)}$ の尤度は下記で表される:

$$P(L,N \mid V) = \prod_{i=1}^{I}\prod_{d=1}^{D_i} P(l_i^{(d)} \mid n_{i_1}, n_{i_2}) \sum_{g} P(N \mid g) P(g \mid V). \tag{2}$$

ここで，I は検体数を，D_i は i 番目の検体のリードペア数を表す．また，L は $l_i^{(d)}$ の集合を，N は n_{i_1} と n_{i_2} の集合を表し，g は coalescent 木を表す．式 (2) の右辺の第一項は Basic モデルの尤度関数を表す．第二項では，coalescent 木 g を介してリピート数が次のように結合される:

$$P(N \mid g) = \sum_{n_{c_1}=n_{\min}}^{n_{\max}} \cdots \sum_{n_{c_{I-1}}=n_{\min}}^{n_{\max}} \prod_{v\in C_g}\prod_{u\in o_{v|g}} P(n_u \mid n_v, t_{v,u|g}; \mu_s). \tag{3}$$

ここで，C_g は g における内部ノード $c_1,\ldots,c_{I-1}$ の集合を，$o_{v|g}$ は g におけるノード v の子ノードの集合を表す．また，$t_{v,u|g}$ は，g において v から u への coalescent 時間を表し，n_v はノード v のリピート数を表す．$P(n_u \mid n_v, t_{v,u|g}; \mu_s)$ は親ノード v から子ノード u への時間 $t_{v,u|g}$ と変異率 μ_s によるリピート数の変化を表す条件付確率である．$P(g \mid V)$ は周辺のフェーズ済み遺伝子型 V が与えられたもとでの coalescent 木 g の発生確率を表す．式 (2) にて全ての coalesncet 木 g について総和計算をすることは困難であるため，フェーズ済み遺伝子型 V から MCMC でサンプリングした coalescent 木に限定して総和計算する．リピート数は式 (2) を最大化する N を求めることで推定される．

3 性能評価

既存手法との比較による性能評価について口頭発表にて報告するが，詳細は Kojima *et al.* (2016) [1] を参照されたい．

References

[1] K Kojima, Y Kawai, N Nariai, T Mimori, T Hasegawa, and M Nagasaki. Short tandem repeat number estimation from paired-end reads for multiple individuals by considering coalescent tree. *BMC Genomics*, 17(Suppl 5)(494), 2016.