シンポジウム「生命・自然科学における複雑現象解明のための統計的アプローチ」

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）
開催責任者：松井 秀俊（滋賀大学）

内容・目的

近年の情報通信機器の発達に伴い、大量かつ複雑な形式を持つデータが多く取得されるようになってきました。これに伴い、様々な分野で、取得されたデータを分析するための高度な方法論に対する需要が高まっています。本シンポジウムでは、統計学、機械学習、バイオインフォマティクスなどに基づく手法による、生命科学をはじめとした多様な分野におけるデータ分析の応用事例に関する講演を広く募集します。講演内容としては、新たな分析手法の提案のみならず、応用分野の側面から見た分析上の問題提起なども歓迎します。 参加者の交流を通じて、知識の共有だけでなく、新たな研究の発展や問題解決に繋げる場にすることを目的としています。

プログラム

2 月 16 日（金）

13:30-13:40　Opening

13:40-14:05　松井佑介（名古屋大学大学院医学系研究科）
「がんの複雑性と進化を読み解くデータ科学駆動型アプローチ」

14:05-14:30　Heewon Park（山口大学国際総合科学部）、井元清哉（東京大学医科学研究所）、宮野悟（東京大学医科学研究所）
「Cancer characteristic-specific analysis via L1-type regularized regression modeling」

14:30-14:55　新村秀一（成蹊大学）
「Cancer Gene Analysis using Small Matryoshka （SM） Found by Matryoshka Feature Selection Method」

15:10-15:35　江田智尊（九州大学大学院数理学府）、恩田義彦（理化学研究所セルロース生産研究チーム）、松井秀俊（滋賀大学データサイエンス学部）、西井龍映（九州大学マス・フォア・インダストリ研究所）、持田恵一（理化学研究所セルロース生産研究チーム）
「統計的遺伝子データ解析と低炭素社会への貢献」

15:35-16:00　田中凌慧（東京大学大学院農学生命科学研究科）
「ベイズ的最適化に基づく遺伝資源の効率的探索」

16:00-16:25　島谷健一郎（統計数理研究所）、荒木希和子（立命館大生命科学部）
「多年生草本の地上部 - 地下部データを用いる動態モデル」

16:40-17:05　Rizky Reza Fauzi（九州大学大学院数理学府）、前園宜彦（九州大学大学院数理学研究院）
「Boundary free estimators of distribution function with transformation」

17:05-17:30　中山優吾（筑波大学大学院数理物質科学研究科）、矢田和善（筑波大学数理物質系）、青嶋誠（筑波大学数理物質系）
「Asymptotic properties of SVM with Gaussian kernel for high-dimensional data」

18:30-　懇親会

2/17（土）

10:00-10:25　草野元紀（東北大学大学院理学研究科数学専攻）
「分布に対する位相的データ解析」

10:25-10:50　今泉允聡（統計数理研究所）
「関数データ回帰の信頼バンド構成法」

10:50-11:15　永井勇（中京大学国際教養学部）
「高次元データにおける精度行列の罰則付推定とその最適化」

11:30-11:55　菅澤翔之助（統計数理研究所）、野間久史（統計数理研究所）
「個別化医療への機械学習的アプローチ」

11:55-12:20　山西芳裕（九州大学生体防御医学研究所）
「AI 創薬：機械学習による様々な疾患に対するデータ駆動型の新薬開発」

12:20-12:45　白井剛（長浜バイオ大学バイオサイエンス学部）
「バイオインフォマティクスによる古代遺伝子の再現と機能解析」

12:45-14:15　昼食

14:15-15:05　特別講演：井元清哉（東京大学医科学研究所）
「統計学とスパコンでがんゲノムにチャレンジ」

15:20-15:45　岩山幸治（滋賀大学データサイエンス教育研究センター）
「トピックモデルによる Shallow RNA-Seq データの補完」

15:45-16:10　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）、新飯田俊平（国立長寿医療研究センター）
「認知症の超早期発見のための血中マイクロ RNA の共発現解析」

16:10-16:35　植木優夫（理化学研究所革新知能統合研究センター）、川崎能典（統計数理研究所 モデリング研究系）、田宮 元（東北大学東北メディカル・メガバンク機構/理化学研究所 革新知能統合研究センター）
「双方向グラフ上の最短経路を利用した遺伝関連解析」

16:50-17:15　大谷隆浩（統計数理研究所）
「精密医療・予防に向けた分子バイオマーカーの探索：階層混合モデルを用いた最適発見手法の応用」

17:15-17:40　野間久史（統計数理研究所）
「多変量メタアナリシスにおける高次漸近理論を用いた推測手法」

17:40-17:50 Closing

がんの複雑性と進化を読み解くデータ科学駆動型アプローチ

松井佑介
名古屋大学大学院医学系研究科・システム生物学分野

サブクローン進化構造の分類手法(phyC)の開発

多数のがんサブクローン進化の推定手法が開発され、データが次々と蓄積されている。しかし一方で、推定した進化構造の解釈については、個々の生命科学者が一つ一つ手作業で生物学的な解釈をしているのが現状で、がんサブクローン進化推定を応用へとつなげる時のボトルネックとなっている。我々は、そのような背景を踏まえて、多数のがんのサブクローン進化構造を統計的に分類する手法を開発した（Matsui et al. 2017）。

がんサブクローン進化構造の特徴を捉える特徴量として、進化系統樹のトポロジー構造と各サブクローンに含まれる変異数の二つを考慮して分類する手法(phyC)を考案した（図 1）。まず、phyC が扱う進化構造のモデルは、各点（ノード）がサブクローンを表し、トポロジーは進化構造を表す（図 1A）。また、サブクローン同士をつなぐ線（エッジ）の長さは、親のサブクローンから子のサブクローンの間に新たに蓄積した変異の数を表す（図 1B）。

木構造の比較では、数学的に取り扱いが容易な二分木を扱うことが多いが、サブクローン進化構造の場合は複雑な木の構造をとるため、患者間での木構造の比較は容易ではない。逆に、複雑な木構造を何らかの形で元の情報を失うことなく二分木に「変換」できれば、患者間でのがんサブクローン進化の比較が可能となる。phyC では、参照木と呼ばれる共通の巨大な二分木を用意して、そこに患者ごとに推定された、がんサブクローン進化構造をマッピングする過程を通じて、「変換」を実現している。マッピングの際には、実際には分岐がない部分のエッジ長は「ゼロ」と考えることで、二分木以外の分岐構造を二分木に埋め込むことが可能である（図 1C; 詳しいアルゴリズムと数理的な性質については Matsui et al. 2017 を参照）。このようにして共通の参照木に埋め込んだ二分木は、枝分かれのない部分のエッジ長はゼロ、枝分かれのある部分については対応するエッジ長のある、単純なエッジ長の組みとして定量的に表すことができる。さらに、患者間で同定される変異は異なるため、サブクローンに含まれる変異数を患者ごとの全変異数で正規化することで、エッジ長の比較を可能としている。これらを元に非類似性（距離）を計算することで、類似したがんサブクローン進化構造をグルーピングしている（図 1D）。

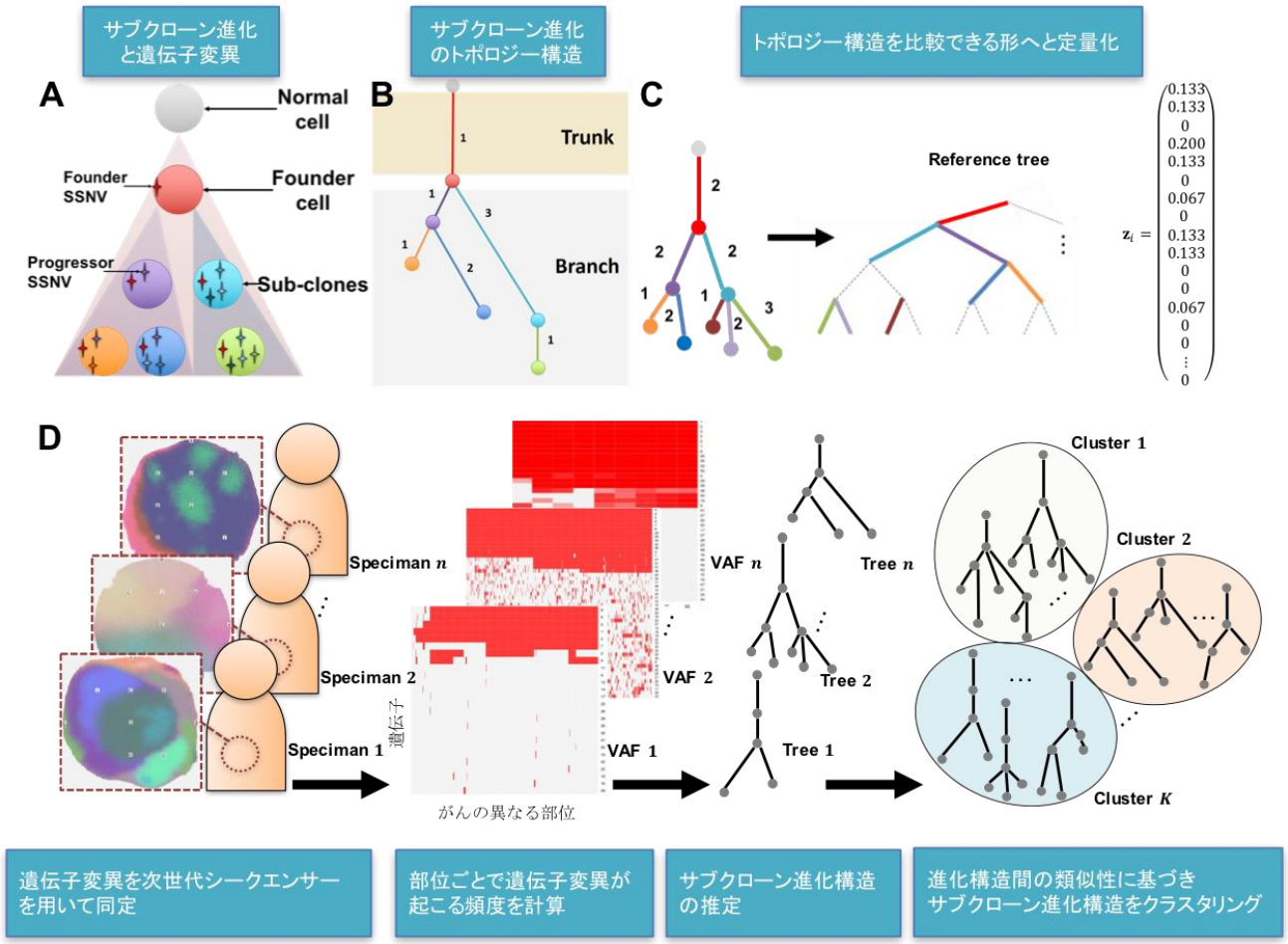


図 1:phyC の概要と解析の流れ

今後の課題と展開

これまでに、がんのサブクローン進化構造を推定するための統計・数理モデルは多数開発されているが、多くの方法では推定したがんのサブクローン進化構造の候補は複数あり、それらの選択規準は特に定まっていない。また手法間でも結果が異なることもあり、推定したサブクローン進化構造の妥当性や結果の解釈については重要な課題となっており、今後それらを評価する手法開発も必要であると思われる。

参考文献：

1. Matsui Y, Niida A, et al: PLoS Comput Biol, 13(5), e1005509, 2017.

Cancer characteristic-specific analysis via L1-type regularized regression modeling

Heewon Park（山口大学国際総合科学部）
井元清哉（東京大学医科学研究所）
宮野悟（東京大学医科学研究所）

Over the last few decades, research on the individual genomic characteristic identification has been received much attention, especially personalize anti-cancer therapy has drawn a large amount of attention in various fields of research (e.g., statistics, bioinformatics, computer engineering, etc.). In order to understand the complexity and heterogeneous system of cancer, various statistical strategies have been proposed, and used to genome-scale information analysis. Although the statistical methodologies have been successfully applied to infer gene regulatory networks in cancer progression, the existing methods, such as L_1-type regularization methods, provide averaged modeling results for all samples. This implies that the existing methods cannot reveal sample (patient)-specific characteristics of cancer.

Shimamura et al. (2011) proposed the statistical method, called as the NetworkProfiler, for sample-specific analysis based on a kernel based L_1-type regularization method. The NetworkProfiler groups samples according to specific genomic characteristic (i.e., similarity of the modulator values) by using the Gaussian kernel function, and performs modeling a target sample based on the grouped neighborhood around the target sample, i.e., the NetworkProfiler constructs gene networks for a target sample based only on samples having similar characteristics with the target sample. By using the NetworkProfiler, Shimamura et al. (2011) revealed the system changes in epithelial mesenchymal transition.

We construct anti-cancer drug sensitivity-specific gene regulatory networks through the data analysis of ``Sanger Genomic of Drug Sensitivity in Cancer dataset from the Cancer Genome Project''. In the dataset, anti-cancer drug sensitivities are given as IC50 values (i.e., half maximal inhibitory drug concentrations) of each drug. In other words, we consider the IC50 value of an anti-cancer drug as a modulator, and then infer the gene regulatory effects related to the drug sensitivity.

The NetworkProfiler, however, cannot perform well sample-specific analysis for a target sample in sparse region (i.e., the value of the modulator is in sparse region of modulator distribution), since the method is based on a constant bandwidth in the Gaussian kernel function. In other words, the NetworkProfiler imposes an extremely small amount of weight to almost samples for modeling a target sample in sparse region. In short, the modeling for

a target sample in dense region is based on many samples, while the modeling for a target sample in sparse region is based only on a few samples. This implies that the NetworkProfiler cannot provide effective results for modeling a target sample in sparse region, because the method may lead to extremely high dimensional data situation. Consequentially, the NetworkProfiler is not suitable to the anti-cancer drug sensitivity-specific gene network construction, because the IC50 values of anti-cancer drug in Sanger Genomic data set, which is used as a modulator in this study, are non-uniformly distributed.

To settle on this issue, we propose a novel kernel based L_1-type regularization method based on an adaptive bandwidth of the Gaussian kernel function. We consider the k-nearest neighbor (KNN)-Gaussian kernel function based on the Euclidean distance to effectively group samples according to the cancer characteristic. The KNN-Gaussian kernel function, however, also cannot perform properly inferring the drug sensitivity-specific gene regulatory networks, since the Euclidean distance between IC50 values is not large enough to properly impose a weight for constructing drug sensitivity-specific gene networks. Thus, we modified the KNN-Gaussian kernel function by imposing an additional hyper parameter incorporating dispersion of a modulator into a bandwidth, and propose a novel NetworkProfiler based on the KNN with range-Gaussian kernel function, called as an adaptive NetworkProfiler. The proposed adaptive NetworkProfiler effectively groups samples for modeling a target sample in not only dense region but also sparse region, because the hyper parameter, i.e., range of modulator, can adjust the width of a kernel function depending on location of the target sample on the modulator distribution. The proposed method can overcome the drawback of the ordinary NetworkProfiler that an extremely small amount of weight is imposed on almost samples.

We demonstrate through Monte Carlo simulations the effectiveness of the proposed adaptive NetworkProfiler (i.e., robust performance against the distribution of a modulator). We also apply the proposed methods to publicly available Sanger dataset from Cancer Genome Project (http://www.cancerrxgene.org/), and construct total 19,800,000 drug sensitivity-specific gene regulatory networks for 99 drugs with 2,000 target genes in each 100 cell lines. This computation may be the largest scale of gene network construction, and thus we perform the huge amount of computation based on super computers.

Cancer Gene Analysis using Small Matryoshka(SM) found by Matryoshka Feature Selection Method

成蹊大学　名誉教授　新村秀一（しんむら　しゅういち）

1. 初めに

大学卒業後、NEC と大阪府立成人病センターの「心電図自動診断システム」のプロジェクトで、判別関数で正常所見と異常所見の診断論理を 4 年間研究した。しかし、医師の「枝分かれ論理」に歯が立たなかった。これが、「Fisher の仮説」に基づく、Fisher の線形判別関数（F-LDF)が多くのデータ特に医学診断に適していないと考えた理由で、研究の動機である。

実証研究を通して 4 つの深刻な問題を発見し、応用問題として、Microarray データを用いた「癌の遺伝子診断」を、僅か 2015 年の 10 月 28 日から 12 月 20 日の 54 日間で Matryoshka Feature Selection Method（新手法 2）で簡単に解決した。成果を 2016 年末に Springer から出版したが、重要な結論は「遺伝子空間は簡単に 10 個から 40 個程度の遺伝子の組で「癌と正常症例」の 50 から 200 程度の小さな部分空間（Small Matryoshka, SM)に分割できる。SM は小標本なので簡単に統計分析できると考えたが、できなかった。しかし、ある工夫をして簡単に癌の悪性度指標で診断できることが分かったので、成果を Amazon の Kindle から 2017 年に出版(＄9.99)した。

2. 判別分析の新理論

IP-OLDF、改定 IP-OLDF、改定 LP-OLDF と改定 IPLP-OLDF を開発した。そしてロジスティック回帰、F-LDF、ハードマージン最大化 SVM(H-SVM)とソフトマージン最大化 SVM（S-SVM)でペナルティｃを 1 にした SVM1 と 10000 にした SVM4 と比較する。また、判別分析は推測統計学でないので「100 重交差検証法(新手法 1)」と癌の遺伝子解析のための「Matryoshka Feature Selection（新手法 2)」と MNM と RatioSV の新統計量を開発した。

さらに IP-OLDF の定式化で**「誤分類数（NM）と判別係数の関係」**を世界で初めて示し、NM の代わりに最小誤分類数（Minimum NM, MNM）を提案した。

3. 判別分析の問題

3.1 Fisher の線形判別関数（F-LDF)

Fisher は 2 群が平均値だけが異なる正規分布に従うという「Fisher の仮説」で、 F-LDF を定式化し判別分析の世界を開いた。分散共分散で簡単に計算でき、相関比最大化基準で定式化される。しかし医学診断データ、各種格付けや、線形分離可能なデータ（LSD)を正しく判別できない(問題 2)。また一般化逆行列の瑕疵(問題 3)、判別分析は推測統計学でない(問題 4)そして Microarray データの「癌の遺伝子解析」に全く役に立たない(問題 5)がある。Fisher の同世代の研究者は、分散共分散が異なる場合 2 次判別関数を薦めた。これは、現実のデータが「Fisher の仮説」を満たさないデータがあることを認識していた証拠である。

3.2 5 つの問題

問題 1 は、NM は問題が多く、特に判別境界上のケースの判定ができないことと、線形分離可能なデータ（LSD)を正しく判別できず誤分類確率が 3 割になるものもあることである。問題 2 は LSD の判別研究がないが、1）MNM の単調減少性（$MNM_k \geqq MNMK_{(k+1)}$）と、2）$MNM_k=0$ であればこの k 変数を含むモデルの MNM は全て 0 になる。これは、LSD はその中に MNM=0 になる小さな部分空間（Small Matryoshka、SM）を含む特殊なデータ構造であることを示す。この他、一般化逆行列の瑕疵（問題 3)と推測統計学でない（問題 4)ことを解決した。

4. 癌の遺伝子解析と新理論 2

4.1 問題 5 の再確認

2015 年 10 月 25 日に富山県民会館で問題 4 の判別係数の 95%CI の発表し、新理論を完成したと考えた。しかし、石井の発表で応用問題として**問題 5** を思い出した。10 月 28 日に Higgins の HP から、1999 年から 2004 年の間に一流紙に掲載の論文で使用した、6 種類の Microarray データを入手し、12 月 20 日までの 54 日で、癌の遺伝子解析を完成した。結論は、1)改定 IP-OLDF（RIP)で 6 種の Microarray データは全て MNM=0 である。2）かつ n 個以下の判別係数が非ゼロで残りは自然に 0 になった（LASSO 入らない）。3)高次元の Microarray データは Matryoshka 構造で、その中に SM1 を含む。4)全体から SM1 を省いて判別すると別の SM2 が得られる。

そこで、Matryoshka Feature Selection Method（新手法 2)と LING Program3 を開発し、6 種類全部の SM と雑音空間を分離し、2016 年末に Springer から「判別分析の新理論と癌の遺伝子解析」を出版した。

4.2 1970 年以来の「癌の遺伝子解析の問題」

Golub 他(1999)は“Although cancer classification has improved over the past 30 years, there has been no general approach for identifying cancer classes (class discovery) or for assigning tumors to known classes (class prediction).” と論文で述べ、問題 5 は 1970 年来の難問である。またこれらのデータをテーマ

にする論文は、3つの言い訳がバズワードである。1)Small n large p データ、2)NP-hard、3)雑音から信号の分離が困難。しかし、RIPは3つの言い訳を簡単に解決した。なぜか？

4.3 分析結果

新手法2をLINGOのProgram3で実現し、全てのSMを探索し、64個から269個のSMを見つけた。高次元の遺伝子空間の信号空間はSMの和集合に分割され、残りは雑音空間である。

4.4 癌の遺伝子解析で分かったこと

分散共分散行列に基づくF-LDFやLASSOはMicroarrayデータがLSDであることを認識できず、NMは大きい。これは相関比最大化基準は、理論的にLSDを正しく判別できないためである。

H-SVMはLSDを正しく判別できるが、なぜか誰も試みていない。しかし、RIPのようにSMを見つけることができない。これは2次計画法のQPは、全遺伝子空間で唯一の極小値(最小値)を求めるためである。

RIPが簡単にSMを見つけることができるのは、IPのアルゴリズムの分枝限定法は、言ってれば「全てのモデルの組み合わせ」を探索し、部分空間の多数の最適解の一つを出力できるからである。

4.5 Alonらの64個のSMsの分析

全てのSMは小標本であり、統計分析が可能。ロジスティック回帰は、全SMのNM=0である。しかし、F-LDFやQDFはNM=0でないものが多い。また、PCAとクラスター分析は線形分離可能な兆候を示さない。

4.6 RatioSVと BGS

Program 3が求めたSMの中に最小次元のSM(BGS) が含まれている。Alon他で求めたが、SMとBGSの評価が必要でRIP の判別スコアの範囲に対するSVの距離の比RatioSVを開発した。Alonの130個の BGSの範囲は[0.00%, 0.9%]。一方、64個の SMs は[2.35%, 26.76%]でしかも63個の RationSVs は5%以上である。さらにShipp他の最大値は約40%で、2群は60%の狭い範囲にある。以上から、BGSは医学診断に役に立たないと判断して、SMを癌の悪性度指標と考えた。

5. RIP判別スコアデータによる革新

SMは小標本で、標準の統計手法で有用な情報を得られると考えたが失敗。これを解決する秘策として、RIPの判別スコアを変数とするデータを作成した。クラスター分析とPCAは2群が完全に分かれた。特に、6データとも2群の症例は、ほぼPCAの第1主成分上に布置した。このことから個々のRIP判別スコアでRatioSVが大きいものは、癌の悪性度指標と考えられる。転置データの分析で、一部のRIP判別スコアは、外れ値になり、癌のサブクラスなどを表すと考えられる。

6. まとめ

1970年以来行われてきた「癌の遺伝子解析」は、質の高い6つの研究データを用いて、SMあるいはBGSと呼ぶ信号空間に簡単に分割できた。日本では血液から49個以下の遺伝子検査を行う検査センターが多い。SMはこれ以下であるので、RatioSVの大きなものから悪性度指標が医学的に検証できれば、社会的に貢献できる。

また、転置データで外れ値になるものは、何らかの癌の特異的な点を示すと考えられる。

データを集めた6研究グループに、Research GateとAmazon の書籍の中で、共同研究を求めたが、反応はない。結局「癌の悪性度指標」と呼ぶ「癌の医学診断」を完成させるには、医学研究者グループとの協力が必要になり、今後の対応を決めかねている。誰かアドバイスをください！

付録1：New Theory of Discriminant Analysis After R. Fisher (Springer)

6種類の実データを用いて、全てのモデルで新手法1を用いて検証標本で平均誤分類確率が最小モデルを比較する8種類のLDFのBestモデルとして選ぶ。そしてRIPのBestモデルがよく、ロジスティック回帰とSVM4、次にSVM1そしてF-LDFが一番悪いことが分かった。2章から7章まではこれらのデータの分析結果であり、各データ固有のテーマを扱っている。2章は Irisデータだけが Fisherの仮説を満たしF-LDFのNMがMNMに収束することを示す。3章は3個の共線性があるCPDデータで、共線性の解消法と変数選択法等を紹介。ロジスティック回帰が問題1に弱いことを指摘。4章は、40人の学生の合否判定を5変数で判別し、超平面上に10人の学生が来るために各LDFの問題点と問題1を説明。5章は18種類の合否判定を得点を用いて判別。F-LDFの誤分類確率が高いことを実証。6章は、スイス銀行1000フラン紙幣の真札と偽札各100枚を6個の計測値でLSD判別。 (X4,X6) の2変数でMNM=0になる。Program3で判別するとMicroarrayデータと同じことが分かり、新手法2の内容を本データで紹介。7章は小型車15車種と普通車29車種を排気量 (X1) や座席数 (X3) の6変数で判別。小型車のX1とX3は、普通車より小さく、この2変数は2個のBGSになる。Program3でSMの一つとしてX3が求まるが、1変数なのでBGSであることを解説。8章は、6種のMicroarrayデータによる「癌の遺伝子解析」。9章は、LINGOのProgram1と2の解説。

付録2：From Cancer Gene Analysis to Cancer Gene Diagnosis (Amazon Kindle1102円、Unlimitedは無償)

6種類の全てのSMとAlon他のBGSの癌の悪性度指標の検討。全てがデータで、2群の症例がPCAの第1主成分上に布置し、第1固有値だけが非常に高い。青嶋・矢田は子次元空間の分布を調べ、2群が完全に分かれていること、PCAの結果第1主成分の固有値が高い結論を得たことと、同じ結果である。さらに、信号空間を複数のSMの和集合に分割した。これらの研究は、将来癌の制圧に役立つと考える。

統計的遺伝子データ解析と低炭素社会への貢献

江田 智尊 (九州大学大学院 数理学府)，恩田 義彦，持田 恵一 (理化学研究所 セルロース生産研究チーム)
松井 秀俊 (滋賀大学データサイエンス学部)，西井 龍映 (九州大学マスフォアインダストリ研究所)

1　はじめに

地球環境に優しいエネルギーの開発は 21 世紀においても喫緊の課題である．その解決策の一つとしてバイオ燃料が挙げられている．実際に現在，各国の政府が主導しバイオ燃料の使用を促進する政策を執っている．例えば米国では 2005 年より，ガソリンへのバイオエタノール混合率を引き上げる政策を施行している．ブラジルでも同様の取り組みを従来から行っており，今後更なる混合率引き上げを検討している．EU においても，2020 年までに全体のエネルギー消費の内 20% を，バイオ燃料を含む再生可能エネルギーによって賄おうとする取り組みを行っている [1]．このような状況下においては当然，バイオ燃料を提供する植物資源の大量生産が必要になってくる．アメリカでは断続的な穀物の価格高騰が問題となっているが，これはバイオ燃料資源として穀物が大量に用いられるようになったことが一つの原因である．従って，バイオ燃料資源としての植物をより安定的に大量生産する技術開発が，需要過多な現状を打破し，ひいては環境政策，低炭素社会へと貢献することが期待される．

この大量生産を実現する一つのキーワードが，人工ヘテローシス技術である．ヘテローシスとは，子の生物品種が親よりも優れた特徴を示す現象である．優れた特徴とは，個体のサイズや生長スピード，環境耐性などである．こういった植物は明らかに植物資源の安定的大量生産の狙いに即している．しかし一方でヘテローシス現象の分子機構は未だに不明な点が多い．そこで我々は遺伝子レベルでヘテローシス現象を分析し，人工的にヘテローシスを付与することのできる技術の開発を志している．将来的に開発した技術で，低炭素社会への貢献を実現したいと考えている．

2　統計的遺伝子データ解析への期待

遺伝子は相互の働きを制御しており，その制御関係は非常に複雑な構造を持っているとされる．そうした複雑な状況下でヘテローシスに深く関連する遺伝子群を，植物から採取した遺伝子データから特定する必要がある．しかし取得できる遺伝子データは一般にノイズが多く，また観測される遺伝子数は数万と大量にある一方でサンプルが非常に少ないため，解析には注意が必要である．そういった背景から，不確実性が評価可能な統計解析のバイオロジー分野への適用可能性は大きい．

その中でも特に遺伝子ネットワーク推定は，遺伝子間の制御関係を表すツールとして用いられ，しばしば生物学的な考察を我々に提供する．遺伝子ネットワークにおいては，有向グラフのノードが遺伝子に対応し，エッジが制御関係を表す．図 1 はその例である．ネットワークにおいてハブとなる遺伝子は多くの遺伝子を制御していることを示し，従って植物の生長過程で重要な役割を果たす遺伝子である可能性が高い．

3　ARX モデルと Group SCAD に基づく遺伝子ネットワーク推定

そこで本研究では RNA-seq に基づいて採取された時系列遺伝子発現量データを用いて遺伝子ネットワーク推定を行った．推定時に用いたモデルは Auto-Regressive eXogenous (ARX) モデルである．例として，ARX(2) モデルに基づく推定式を記述する．記号 X_t^g $(g = 1, \ldots, N, t = 1, \ldots, T)$ を，遺伝子 g の時刻 t

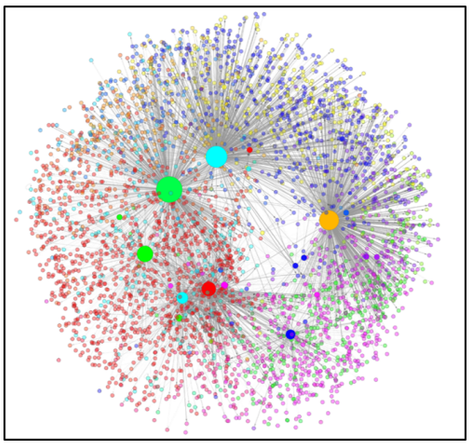


図 1　推定された遺伝子ネットワーク．ノードの大小は繋がっているエッジの多さに比例しており，大きな円で描かれたノードに対応する遺伝子はハブとなっていることを表す．

における発現量とする．ARX(2) モデルは以下のように記述できる．

$$X_t^g = \sum_{i=1}^{N} \left(\beta_1^i X_{t-1}^i + \beta_2^i X_{t-2}^i \right) + \varepsilon_t^i. \tag{1}$$

ここで ε_t^i はガウシアンノイズである．係数 β_1^i, β_2^i が遺伝子 $i \to g$ の関係を定量化する．係数の推定時には Group SCAD 正則化項を加えてパラメータ推定を行った [2]．グループ推定を適用した理由は，同遺伝子に起因する説明変数の係数，ここでは β_1^i と β_2^i をグループ化することで，特定のタイムラグに依らない遺伝子間の従属関係を明らかにする狙いがある．加えて，観測時点数の少なさから生じる解析不安定性を解消するために，ブートストラップを用いて各エッジの信頼度を評価した．

4　成果

図 1 は本研究で実際に推定された遺伝子ネットワークである．推定されたネットワークは，少数のみのハブから構成されるスケールフリー性を有することがわかる．34,000 個以上の遺伝子から本解析を通して 76 個の遺伝子がネットワークのハブ遺伝子として現れ，それらに対してヘテローシス現象解明のための生物学的な検証と追加実験を行った．実際にハブ遺伝子として現れた遺伝子のいくつかが，植物の生育過程において重要な役割を果たす可能性があることが判明した．詳細については文献 [3] を参照されたい．

参考文献

[1] K. Araujo, and D. Mahajan, R. Kerr, and M. Silva, Global biofuels at the crossroads: An overview of technical, policy, and investment complexities in the sustainability of biofuel Development, *Agriculture*, 7(4), 2017.

[2] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 96(456): 1348-1360, 2001.

[3] S. Koda, Y. Onda, H. Matsui, K. Takahagi, Y. Yamaguchi-Uehara, M. Shimizu, K. Inoue, T. Yoshida, T. Sakurai, H. Honda, S. Eguchi, R. Nishii, and K. Mochida. Diurnal Transcriptome and Gene Network Represented Through Sparse Modeling in Brachypodium distachyon. *Frontiers in Plant Science*, 2017. doi:10.3389/fpls.2017.02055

ベイズ的最適化に基づく遺伝資源の効率的探索

田中　凌慧
東京大学大学院 農学生命科学研究科

1. 背景

DNA マーカーの登場によって、長い時間や膨大な思考錯誤、人間の勘と経験に大きく依存してきた植物育種（品種改良）は、データ駆動型の、より合理的な営みに変わりつつある。優れた個体を選んで交配することを繰り返すのが典型的な育種法であるが、優れた個体を判別するには実際に圃場で多数の種子を栽培するしかなく、これには長い時間と労力がかかってしまう。しかし、改良したい形質の値を DNA マーカーに回帰し、DNA マーカーの多型から形質を予測する（ゲノミック予測）モデルを構築すれば、圃場で評価せずとも優れた個体を選抜できる。すなわち、DNA マーカー（入力 **x**）と形質（出力 y）のデータに基づく育種が可能となる。Meuwissen ら[1]が提案したこの育種法はゲノミック選抜と呼ばれ、乳牛ではすでに有効性が実証[2]されるなど、めざましい成果を挙げている。

植物育種をより活性化するためには、遺伝資源と呼ばれる多数（数万〜）の野生種を利用することが重要だとされている[3]。近代品種の多くは、収量性や食味を改良する過程で、病害抵抗性や不良環境への適応性に関与する遺伝子を失っており、これを回復するには野生種のもつ抵抗性・適応性遺伝子を導入する必要があるためである。しかしながら、実際の育種における遺伝資源の利用は極めて限定的なのが実情とされる。野生種のうち優れた遺伝子を持つ系統はごく一部であり、それを探索するために数万もの系統を対象地域の圃場で栽培・評価するのは、金銭面・労力面でコストに見合わないためだ。

そこで、ゲノミック予測を用いて、多数の遺伝資源から優れた適応性をもつ系統を絞り込むことで、圃場評価のコストを削減しようとする試みが始まっている。例えば Yu らは、実際に遺伝資源約 1,300 系統を用いた実験を行い、300 系統ほどの訓練データで、テストデータ 1,000 系統の形質値を予測できることを確認した[4]。このように、遺伝資源の一部を圃場評価して形質値を測定し、残る系統の形質値を予測することを繰り返せば、数万系統の遺伝資源を全て評価せずとも、優れた系統を見つけ出すことができることが期待される。

選抜と交配を繰り返す通常の育種では、予測値の大きな系統から順に選抜することで、生まれてくる子孫の形質値を（期待値の意味で）最大化できる。しかしながら、上述の遺伝資源探索では、そうした単純な戦略は最適ではないことが予想される。本研究では、ベイズ的最適化（Bayesian optimization）の問題設定と遺伝資源探索問題との類似性に着目し、期待改善量（expected improvement）を選抜基準とする効率的探索法を提案した。

---

1 Meuwissen THE et al. (2001) Genetics 157: 1819-1829.
2 Gracia-Ruiz A et al. (2016) Proc Natl Acad Sci U S A. 113: 3995-4004.
3 Tanksley SD et al. (1997) Science 277: 1063-1066.
4 Yu et al. (2016) Nat Genet 2: 16150.

2. 方法・結果

ベイズ的最適化は、評価コストが高く、かつ、関数型が未知の目的関数 y=f(**x**)について、その大域最適解を効率的に求めるためのアルゴリズムといえる[5]。関数型が未知であるから解析的に微分するといった操作はできず、また、関数評価のコストが高いことから、多数の値を代入することも望ましくない。遺伝資源探索について再考すると、DNA マーカーと形質の間の関係性は未知であり、育種家にできることは、圃場試験を実施して、系統 **x** に対応する形質値 y を調べることだけである。また、前述したコストの問題から、この操作を最小限にとどめる必要がある。

ベイズ的最適化は、２つのステップを交互に繰り返すことで、これらの状況において効率の良い最適化を実行する。１つめのステップは、既知のデータ点を用いて、ガウス過程を用いた目的関数の近似を行うことである。これにより、確率的に未知の目的関数を表現する。２つめのステップは、獲得関数（acquisition function）と呼ばれる事前に定めた関数をガウス過程から計算し、その最大値を与える点を次なる入力データとして採用することである。獲得関数は、近似したガウス過程が与えられた下での、未知の入力点の「良さ」を表現するように設計される。代表的な目的関数は期待改善量と呼ばれ、ある入力 **x** を次のデータとして採用した場合における、データの最大値の増加量を計算する。

本研究では、期待改善量により選抜する場合と、通常の予測値の大きいものから順に選抜する場合の探索効率を、幾つかの公開データを用いてシミュレーションにより比較を行った。その結果、期待改善量を用いた選抜戦略により、最も良い場合で約半分ほどの選抜回数で、遺伝資源に含まれる有用系統を発見できることが示唆された[6]。

3. まとめ

植物育種は、まさにデータ駆動型の営みに変わりつつある段階であり、多くの未解決な問題が存在する。植物におけるゲノミック選抜は、『予測ができること』は示されつつあるものの、予測を用いた育種のプロセス全体は、いまだ確立されていない。従来の育種理論は、当然ながらゲノミック予測を想定したものではないため、既存の育種法において、形質値による選抜を、ゲノミック選抜に置き換えただけでは、それは最適ではない可能性が高い。育種プロセスを通して、どのタイミングで、どのようなデータを、どれくらい、どのように取得するかは、非常に重要な研究課題の１つである。農業もまた産業である以上、無制限に金銭と時間を費やすことは許されず、データの収集コストを抑え、予測精度の向上を犠牲とする育種戦略が、経済的な意味で最適である可能性もある。今後、データ解析手法を発展させるだけではなく、データの収集を含めた育種全体の最適化を進めることが望まれる。

---

[5] Mockus J (1994) J Global Optim 4: 347-365.
[6] Tanaka R et al. (2017) Theor Appl Genet 131: 93-105.

多年生草本の地上部-地下部データを用いる動態モデル

島谷健一郎(統計数理研究所)、荒木希和子(立命館大生命科学部)

固着性を有する植物の長期モニタリングは世界各地でさまざまな種について行われ、データが蓄積されている。ところで、植物は根という地下部を持つが、この地下部については驚くほど調査が進んでいない。

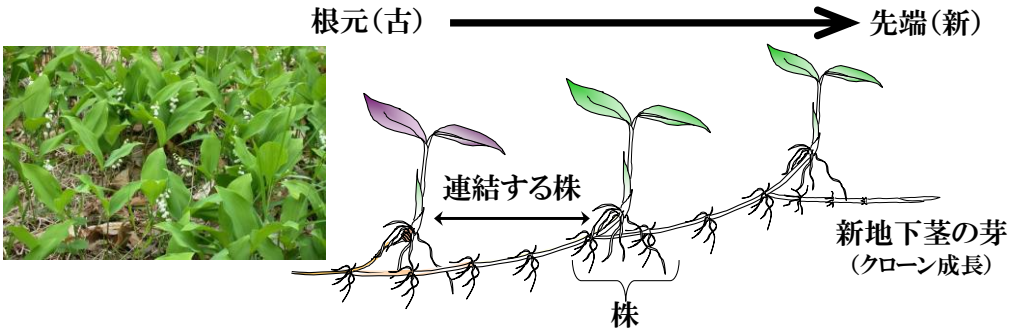
草本には地下茎でクローン繁殖し、地上部の株の分布を拡大する種がある。クローンの広がりは遺伝子型を調べることでだいたい知ることができるが、時間と共にクローンの株が拡大する様子は、そうしたスナップショットデータだけではわからない。

地上部を数年モニタリング調査した後に地下部を掘り起こせば、互いに連関している地上部と地下部のデータが得られる。本研究では、クローナル植物スズランについて、3 年間のモニタリング調査の後に掘り起こし、地上部と地下部の情報を組み合わせてクローン繁殖の分析を試みた。

まず、地下部と地上部の情報を統合し、先端部がさらに先端にクローン繁殖で株を作る場合と、中途から分枝する場合に分けた。さらに、先端部をその春に生まれた株か、それ以前に生まれたが先に延びることなく停滞していた株かに分類し、それぞれについてクローン繁殖率を求めた。そして、クローン繁殖による地上部の株の拡大をマルコフ推移行列でモデル化した。並行して、1 株から順次クローン繁殖で広がるシュミレーションモデルも作った。

クローン繁殖について、地下ゆえ実測の難しい数値を推定し、さらにその実測値をもとにモデル化して長期クローン成長を再現した。

1. クローン繁殖する植物・スズラン



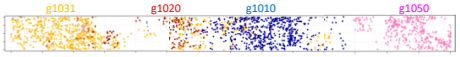
2. 地上部のモニタリングデータと地下部の掘り起し

・ラベルを付ける
・位置(x-y座標)、葉の数 (1-4枚, なし, 地上部なし) 、葉の長さ(最長)、花数、果実数、
・2005-2007年、6ー7月
・2m x 28m
・遺伝子型でクローン識別

ラベル	x	y	遺伝子型	2005 葉の数	2005 葉の長さ(cm)	2005 花数	2005 果実数	2005 備考	2006 葉の数	2006 葉の長さ(cm)	2006 花数	2006 果実数	2006 備考	2007 葉の数	2007 葉の長さ(cm)	2007 花数	2007 果実数	2007 備考
P1	2042	5038	1031	2	16.3				2	17.3				2	19.6			
P2	2077	5017	1031	2	12.5				2	14				2	17.5			
P3	2085	5006	1031	2	13.3				2	18.6				2	17			
P7	2001	5095	1031	2	18.1	6	2		1	14				2	13.6			
P14	2090	5081	1031	2	18.3				2	20.1	7	1		2	20.3			
P4	2082	5036	1031	2	17.2								non	2	17			
P12	2093	5046	1031	2				次無し	2	17				2	13.9			
P13	2098	5045	1031	2	18.9				2	20.8								non
P89	2175	5173	1031	2	14.1								non					non
P8	2025	5085	1031	2	16.1				2	19.2				2	19			
P9	2030	5085	1031	2	16.4				2	14.6				1	17.6			
P10	2056	5073	1031	2	15.9				2	21				2	19.9			
P11	2082	5050	1031	1	17				2	18.9				2	19.2			
P38	2013	5054	1031	2	16.5				2	14.5				2	13.6			

地上部作らなかった
?
死亡と仮定

2. 地上部のモニタリングデータと地下部の掘り起し

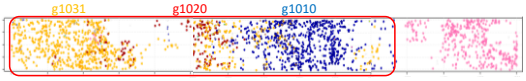


2 m × 28 m

遺伝子型でクローン識別、大きいのは4クローン
各クローンの3年間 (2005-2007) の観察された株数

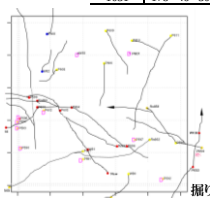
遺伝子型	年	観察数	前年からの生残	地上部新規	地上部出なかった	1年休んで復活	死亡と推定	2007測り忘れ
1031	2005	600						
	2006	701	581	120	19			
	2007	700	603	89	97	8	11	1
1020	2005	108						
	2006	137	104	33	4			
	2007	131	119	11	18	1	3	
1010	2005	477						
	2006	550	451	99	26			
	2007	538	448	74	102	16	10	
1050	2005	236						
	2006	297	227	70	9			
	2007	319	255	61	42	3	6	

2. 地上部のモニタリングデータと地下部の掘り起し



遺伝子型ごとの各断片の株数による分類

Genotype	連結していた株数 1	2	3	4	5	6	7	8	10	12	13	19	22	断片数	掘り返した株数	1断片当たりの平均株数
1010	61	19	11	1	4	4		1		1	1		1	124	235	2.26
1020	28	12		4	3	2		3		1			1	57	136	2.52
1031	178	46	30	6	5	1		1						290	423	1.58

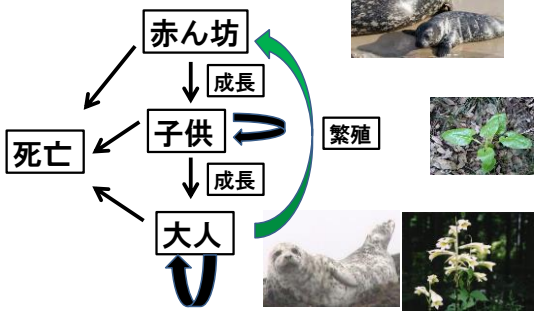


掘り起し地下茎の連結図の例

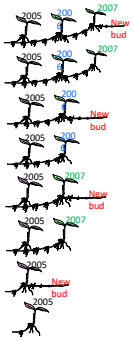
地下茎による連結

地下茎

3. マルコフ推移行列モデル
3.1. 種固有の生活史の定式化



4. クローン繁殖の行列モデル化 行列モデルに必要な確率の計算 (2007年に掘り起こした場合)



2006年	2007年
R_1 (2006→2007)	R_1 (2007→Bud)
R_1	noR_1 (2007→no)
noR_1 (2006→no)	R_2 (2005, 2006→Bud)
noR_1	noR_2 (2005, 2006→no)
R_2 (2005→2007)	R_1
R_2	noR_1
noR_2 (2005→no)	R_2
noR_2	noR_2

5. 結果と考察

観察された割合で確率とした
クローン繁殖のサイズ依存性もあった(ここでは略)

遺伝子型	当年のstage	前年のstage 1年目先端	前年のstage 成長 2年目以降先端	前年のstage 途中	前年のstage 1年目先端	前年のstage 繁殖 2年目以降先端	前年のstage 途中	前年のstage 推移行列 1年目先	前年のstage 推移行列 2年目以降先	前年のstage 推移行列 途中
1010	1年目先端	0.000	0.000	0.000	0.229	0.143	0.021	0.229	0.143	0.021
	2年目以降先端	0.771	0.834	0.000	0.000	0.000	0.000	0.771	0.834	0.000
	途中	0.229	0.143	0.974	0.000	0.000	0.000	0.229	0.143	0.974
1020	1年目先端	0.000	0.000	0.000	0.344	0.080	0.031	0.344	0.080	0.031
	2年目以降先端	0.656	0.890	0.000	0.000	0.000	0.000	0.656	0.890	0.000
	途中	0.344	0.080	0.968	0.000	0.000	0.000	0.344	0.080	0.968
1031	1年目先端	0.000	0.000	0.000	0.348	0.231	0.045	0.348	0.231	0.045
	2年目以降先端	0.652	0.752	0.000	0.000	0.000	0.000	0.652	0.752	0.000
	途中	0.348	0.231	0.978	0.000	0.000	0.000	0.348	0.231	0.978

参考文献

Araki K., Shimatani K. and Ohara M. (2007) Floral Distribution, Clonal Structure, and Their Effects on Pollination Success in a Self-Incompatible *Convallaria keiskei* Population in Northern Japan. Plant Ecology 189: 175-186.

Araki, K., Shimatani, K., and Ohar, M. (2009) Dynamics of distribution and performance of ramets constructing genets: a demographic-genetic study in a clonal plant, *Convallaria keiskei*. Annals of Botany 104: 71–79.

島谷健一郎 (2017) 現場主義統計学のすすめー野外調査のデータ解析. 近代科学社.

Boundary-free Estimators of Distribution Function with Transformation

九州大学大学院数理学府　Rizky Reza Fauzi
九州大学大学院数理学研究院　前園宜彦

$X_1, X_2, ..., X_n$ を互いに独立で同じ分布にしたがう確率変数とし，分布関数を F_X，密度関数を f_X とする．この分布関数 F_X の推定量としては経験分布関数

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), \quad x \in \mathbf{R}, \tag{1}$$

がある．ここで $I(A)$ は定義関数である．経験分布関数は階段関数で連続ではない．連続になるような推定量でよく利用されるのがカーネル型推定量である．

Parzen (1962) と Rosenblatt (1956) によるカーネル型密度関数は

$$\widehat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbf{R}, \tag{2}$$

で与えられる．ここで K はカーネル関数と呼ばれ，$h > 0$ はスムーズさを制御するバンド幅と呼ばれるものである．K は原点について対称で，非負とし $\int_{-\infty}^{\infty} K(v)dv = 1$ の条件を満たし，$n \to \infty$ のとき $h \to 0$ かつ $nh \to \infty$ とする．この密度関数推定量を積分すると，カーネル型分布関数推定量（Nadaraya (1964)）

$$\widehat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \quad x \in \mathbf{R}, \tag{3}$$

が得られる．ただし $W(v) = \int_{-\infty}^{v} K(w)dw$ である．この推定量が一致性や漸近正規性を持つことは容易に示すことができる．$\widehat{F}_X(x)$ の漸近的性質は多くの研究がなされ，経験分布関数 $F_n(x)$ よりも良いことが示されている．適当な条件の下で $n \to \infty$ のとき

$$Bias[\widehat{F}_X(x)] = \frac{h^2}{2} f'_X(x) \int_{-\infty}^{\infty} v^2 K(v)dv + o(h^2), \tag{4}$$

$$Var[\widehat{F}_X(x)] = \frac{1}{n} F_X(x)[1 - F_X(x)] - \frac{2h}{n} r_1 f_X(x) + o\left(\frac{h}{n}\right) \tag{5}$$

となる．ただし $r_1 = \int_{-\infty}^{\infty} vK(v)W(v)dv$ で，多くの場合非負の値をとる．したがって平均積分二乗誤差は

$$\begin{aligned} MISE(\widehat{F}_X) &= \frac{h^4}{4}\left[\int_{-\infty}^{\infty} v^2 K(v)dv\right]^2 \int_{-\infty}^{\infty} [f'_X(x)]^2 dx \\ &\quad + \frac{1}{n}\int_{-\infty}^{\infty} F_X(x)[1 - F_X(x)]dx - \frac{2h}{n} r_1 + o\left(h^4 + \frac{h}{n}\right) \end{aligned}$$

で与えられる．

上記の成果は密度関数のサポートが $\mathbf{R}$ を仮定しており，サポートが有界であるときには境界バイアスの問題が生じる．もしサポートが $\mathbf{R}^+$ や $(0, 1)$ のときには境界バイアスが生じる．$x \in [0, h]$ で K を対称なカーネル関数でサポートを $[-1, 1]$ とすると

$$\begin{aligned} Bias[\widehat{F}_X(x)] &= \left[\int_{-1}^{c} K(v)dv - 1\right] F_X(x) - h f_X(x) \int_{-1}^{c} vK(v)dv \\ &\quad + \frac{h^2}{2} f'_X(x) \int_{-1}^{c} v^2 K(v)dv + o(h^2) \end{aligned}$$

となる．ただし $x \leq h$, $c = \frac{x}{h}$ である．この式から

$$\lim_{n\to\infty} Bias[\widehat{F}_X(x)] = \left[\int_{-1}^{c} K(v)\mathrm{dv} - 1\right] F_X(x) \neq 0$$

となり，この推定量は一致性を持たない．ここではデータの一対一変換を利用してこの境界バイアスの解消を図る．

サポートが $(0, \infty)$ の時に実数 $\mathbf{R}$ への変換としては，対数関数がある．この点に注意すると

$$\tilde{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{\log x - \log X_i}{h}\right), \quad x \in \mathbf{R}^+ \tag{6}$$

が考えられる．ここで $h > 0$ はバンド幅である．この推定量は負の値はは取らない．この推定量のバイアスは h^2 のオーダーで，分散は n^{-1} のオーダーである．適当な条件の下で

$$\begin{aligned} Bias[\tilde{F}_X(x)] &= \frac{h^2}{2}[xf_X(x) + x^2 f'_X(x)]\int_{-\infty}^{\infty} v^2 K(v)dv + o(h^2), & (7)\\ Var[\tilde{F}_X(x)] &= \frac{1}{n}F_X(x)[1 - F_X(x)] - \frac{2h}{n}r_1 x f_X(x) + o\left(\frac{h}{n}\right), & (8) \end{aligned}$$

が成り立つ．ただし $r_1 = \int_{-\infty}^{\infty} vK(v)W(v)dv$ である．

サポートが $(0,\ 1)$ の時の一つの変換としては，標準正規分布の分布関数 Φ を利用した $Y = \Phi^{-1}(X)$ が考えられる．推定量は

$$\tilde{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left[\frac{\Phi^{-1}(x) - \Phi^{-1}(X_i)}{h}\right], \quad x \in (0,\ 1) \tag{9}$$

が考えられる．このときバイアスと分散は

$$\begin{aligned} Bias[\tilde{F}_X(x)] &= \frac{h^2}{2}f'_Y[\Phi^{-1}(x)]\int_{-\infty}^{\infty} v^2 K(v)dv + o(h^2), & (10)\\ Var[\tilde{F}_X(x)] &= \frac{1}{n}F_X(x)[1 - F_X(x)] - \frac{2h}{n}r_1 f_Y[\Phi^{-1}(x)] + o\left(\frac{h}{n}\right), & (11) \end{aligned}$$

で与えられる．ここで ϕ を標準正規分布の密度関数とすると

$$\begin{aligned} f_Y[\Phi^{-1}(x)] &= \phi[\Phi^{-1}(x)]f_X(x),\\ f'_Y[\Phi^{-1}(x)] &= \phi'[\Phi^{-1}(x)]f_X(x) + \phi^2[\Phi^{-1}(x)]f'_X(x), \end{aligned}$$

となる．ただし $x \leq 0$ のとき $\tilde{F}_X(x) = 0$ で $x \geq 1$ の時は $\tilde{F}_X(x) = 1$ とする．

シミュレーションでの検証で，これらの推定量が有効であることが示された．

Asymptotic properties of SVM with Gaussian kernel for high-dimensional data

Yugo Nakayama[1], Kazuyoshi Yata[2] and Makoto Aoshima[2]
[1]Graduate School of Pure and Applied Sciences, University of Tsukuba
[2]Institute of Mathematics, University of Tsukuba

1 Introduction

In this talk, we considered the classification for high-dimensional data. Suppose we have independent and d-variate two populations, $\Pi_i,\ i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. Let $\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$, from each Π_i. We assume $n_i \geq 2,\ i = 1, 2$. Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the two populations. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent.

In the field of machine learning, there are many studies about the classification in the context of supervised learning. A typical method is the support vector machine (SVM). The SVM has versatility and effectiveness both for low-dimensional and high-dimensional data. Nakayama et al. (2017a) investigated several asymptotic properties of the linear SVM in the HDLSS settings. They showed that the misclassification rates tend to 0 as d increases, i.e.,

$$e(i) \to 0 \ \text{ as } d \to \infty \text{ for } i = 1, 2 \tag{1}$$

under the non-sparsity such as $\Delta_\mu \to \infty$ as $d \to \infty$, where $e(i)$ denotes the error rate of misclassifying an individual from Π_i into the other class. Nakayama et al. (2017b) investigated a general framework of the non-linear SVM in HDLSS settings. In this talk, we gave asymptotic properties of SVM with the Gaussian kernel:

$$k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|^2/\gamma),$$

where $\gamma(> 0)$ is a scale parameter.

2 Asymptotic properties and bias-correction of GSVM in HDLSS settings

In this section, we consider asymptotic properties of Gaussian kernel SVM (GSVM). Let $\hat{y}(\boldsymbol{x}_0)$ denote the classifier of GSVM. Let $\beta_i = \exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma\},\ i = 1, 2$ and $\beta_3 = \exp\{-(\mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \Delta_\mu)/\gamma\}$. Let $\Delta = \beta_1 + \beta_2 - 2\beta_3,\ \eta_i = 1 - \beta_i,\ i = 1, 2,\ \Delta_* = \Delta + \sum_{i=1}^2 \eta_i/n_i$ and $\delta = \eta_1/n_1 - \eta_2/n_2$. We note that $\Delta > 0$ when $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ or $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$. Then, we showed the following result.

Lemma 1 (Nakayama et al., 2017b)**.** *Under some regularity conditions, it holds that*

$$\hat{y}(\boldsymbol{x}_0) = \frac{\Delta}{\Delta_*}\left((-1)^i + \frac{\delta}{\Delta} + o_P(1)\right) \quad \textit{as } d \to \infty \textit{ when } \boldsymbol{x}_0 \in \Pi_i \textit{ for } i = 1, 2.$$

Now, we consider the following condition:

(A-i) $\limsup_{d\to\infty} |\delta|/\Delta < 1$.

For the misclassification rates, we gave the following result.

Theorem 1 (Nakayama et al., 2017b)**.** *Under (A-i) and some regularity conditions, GSVM holds (1).*

Without (A-i), we gave the following results.

Corollary 1 (Nakayama et al., 2017b)**.** *Under some regularity conditions, GSVM holds the following properties:*

$$e(1) = 1 + o(1) \quad \textit{and} \quad e(2) = o(1) \quad \textit{as } d \to \infty \textit{ if } \liminf_{d\to\infty} \delta/\Delta > 1; \quad \textit{and}$$
$$e(1) = o(1) \quad \textit{and} \quad e(2) = 1 + o(1) \quad \textit{as } d \to \infty \textit{ if } \limsup_{d\to\infty} \delta/\Delta < -1.$$

We estimate η_i $(i = 1, 2)$ and Δ by

$$\hat{\eta}_i = 1 - \sum_{1\le j<j'\le n_i} 2\frac{k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})}{n_i(n_i-1)} \quad \text{for } i = 1, 2; \text{ and}$$
$$\hat{\Delta} = \sum_{i=1}^{2} \sum_{1\le j<j'\le n_i} 2\frac{k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})}{n_i(n_i-1)} - \sum_{j=1}^{n_1}\sum_{j'=1}^{n_2} 2\frac{k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})}{n_1 n_2}.$$

Let $\hat{\Delta}_* = \hat{\Delta} + \hat{\eta}_1/n_1 + \hat{\eta}_2/n_2$ and $\hat{\delta} = \hat{\eta}_1/n_1 - \hat{\eta}_2/n_2$. We gave a bias-corrected GSVM (BC-GSVM) as

$$\hat{y}_{BC}(\boldsymbol{x}_0) = \hat{y}(\boldsymbol{x}_0) - \hat{\delta}/\hat{\Delta}_*.$$

One classifies $\boldsymbol{x}_0$ into Π_1 if $\hat{y}_{BC}(\boldsymbol{x}_0) < 0$ and into Π_2 otherwise. Then, we gave the following result.

Theorem 2 (Nakayama et al., 2017b)**.** *Under some regularity conditions, BC-GSVM holds (1).*

BC-GSVM has the consistency property without (A-i).

In this talk, we discussed the choice of the scale parameter γ by using the asymptotic properties. Finally, we checked the performance of BC-GSVM and the validity of the tuning parameter by numerical simulations and actual data analysis.

References

[1] Nakayama, Y., Yata, K. and Aoshima, M. (2017a). Support vector machine and its bias correction in high-dimension, low-sample-size settings. *Journal of Statistical Planning and Inference*, 191:88-100.

[2] Nakayama, Y., Yata, K. and Aoshima M. (2017b). Bias-corrected SVM with Gaussian kernel in high-dimension, low-sample-size settings, submitted.

分布に対する位相的データ解析

草野　元紀　(Genki Kusano)*

背景

実験などで得られるデータを解析すると，物質の結晶構造のように，何かしらの形がデータの特徴に関係していることがある．しかし，すべてのデータが結晶のような規則正しい幾何構造を持っているとは限らない．近年のデータ解析において，確率論や統計を用いた従来法では取り出せない幾何的な特徴をトポロジーの視点から記述しようと試みる研究が**位相的データ解析** (*Topological data analysis, TDA*)[Car09] である．ここでの入力データとしては，点集合 $X = \{x_i \in \mathbb{R}^d \mid i = 1, \dots, N\}$ を考えることが多く，材料科学の応用では物質の原子配置を想定している．この時，各点に半径 r の円 $B(x; r) = \{y \in \mathbb{R}^d \mid \|x - y\| \leq r\}$ を設置し，その和集合 $B(X; r) := \bigcup_i B(x_i; r)$ を考える．ある半径 r での $B(X; r)$ のホモロジー群の生成元 (X の穴や空洞の数学的情報) に注目した際，その生成元の発生時間 b と消滅時間 d は形の重要な情報を担う．生存時間 $d - b$ が大きいほど対応する穴は大きいことを意味すると解釈される．これらの発生消滅時刻のペア (b, d) を集めたもの 2 次元ユークリッド空間の多重集合 $D(X) = \{(b_i, d_i) \mid i \in I\}$ とし，これを X の**パーシステント図** (*Persistence diagram*) と呼ぶ．パーシステント図 $D(X)$ はデータ X の形を記述する特徴量であり，SiO_2 の構造解析 [HNH+16] や粉体パッキング [STR+17] など，一見無秩序だが幾何構造が関与しているデータに対しパーシステント図を用いたデータ解析は一定の成果を上げている．

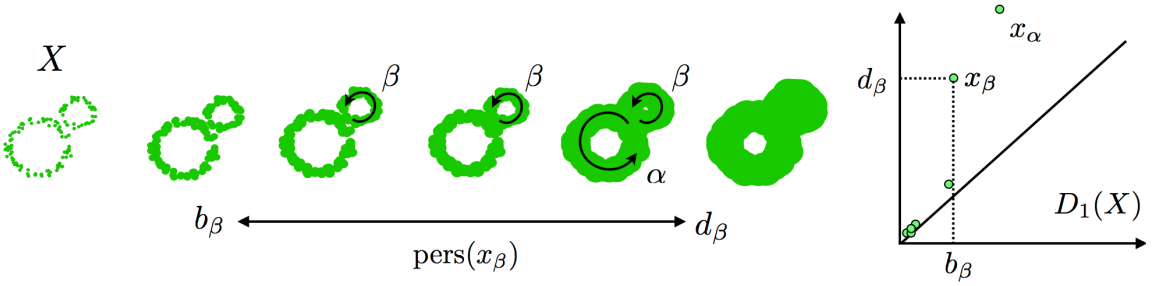


図 1: 点集合 X を中心に半径 r の円の和集合の増大列 (左) と対応するパーシステント図 (右)．穴 β が半径 b_β で発生し，半径 d_β で消滅するとき，その発生消滅の組 $x_\beta := (b_\beta, d_\beta)$ を $\mathbb{R}^2$ に表示する．穴 α は穴 β よりも大きいものであり，対角線からの距離 (生存時間) も大きいものになっている．

研究の動機

様々な応用研究では，一つの点配置からパーシステント図を計算し，それを解析している．しかし，その点配置には観測ノイズや物質に固有の熱揺らぎによる振動などの影響で正確な点配置が得られるとは限らない．ここではむしろ，振動自体もデータに固有の特徴であり，点集合が振動の情報を含んだある分布から生成されている状況を考える．

分布の重要な統計量の一つは期待値である．しかし，パーシステント図は多重集合としてのみ定義され，そこにベクトル空間の構造はないため，直接パーシステント図の “平均” を計算することは困難である．そこで，統計的位相的データ解析の研究ではパーシステント図をベクトルへ変換し，そのベクトル空間で統計解析することが行われている．パーシステント図のベクトル表現として，ここでは *persistence weighted Gaussian kernel* (PWGK [KFH16, KFH18]) を用いる．PWGK によるベクトル表現は正定値カーネル $k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ と重み関数 $w : \mathbb{R} \to [0, \infty)$ を用いて，$V^{k,w}(D) := \sum_{x \in D} w(x) k(\cdot, x) \in \mathcal{H}_k$ として k の再生核ヒルベルト空間 $\mathcal{H}_k$ に値をとるように定義される．適切な状況下では $V^{k,w}$ を $\mathcal{H}_k$ に値をとる確率変数と見なせ，パーシステント図に対する分布を $P_{\mathcal{D}}$ とすると，その期待値 $\mathbb{E}_{D \sim P_{\mathcal{D}}}[V^{k,w}(D)]$ を

本研究は研究は JSPS 科研費 JP17J02401 の助成を受けたものである．

* 〒 980-8578 宮城県仙台市青葉区荒巻字青葉 6-3 東北大学大学院理学研究科数学専攻博士課程 2 年
e-mail: genki.kusano.r5@dc.tohoku.ac.jp

定義することができる.

主結果

パーシステント図の列 $D_1, \ldots, D_n$ がある分布 $P_{\mathcal{D}}$ から独立同分布に生成されているとする. このとき, PWGK ベクトルの列 $V^{k,w}(D_1), \ldots, V^{k,w}(D_n)$ が得られ, 算術平均 $\overline{V}^{k,w}_{(n)} := \frac{1}{n}\sum_{i=1}^{n} V^{k,w}(D_i)$ が定まる. このよう状況の下, 算術平均の真の期待値 $\mathbb{E}_{D\sim P_{\mathcal{D}}}[V^{k,w}(D)]$ への収束を記述する大数の法則と中心極限定理を示す (主結果 1) .

これらの極限定理により, $n \to \infty$ の時の, 算術平均の真の期待値への収束の振る舞いは記述されるが, 現実には n は有限である. そこで, 有限の n から構成した算術平均が真の期待値にどれだけ近いかを確かめるために, $\mathbb{E}_{D\sim P_{\mathcal{D}}}[V^{k,w}(D)]$ の信頼集合を構成する (主結果 2) . この信頼区間の構成には経験分布過程に対するブートストラップ法 [VdV98] の概念を用いている.

パーシステント図に対する真の分布 $P_{\mathcal{D}}$ を求めることは現実的に不可能に近いので, 経験分布などで $P_{\mathcal{D}}$ を近似するような分布 $P'_{\mathcal{D}}$ を使用する. この時, 二つの期待値 $\mathbb{E}_{D\sim P_{\mathcal{D}}}[V^{k,w}(D)]$ と $\mathbb{E}_{D'\sim P'_{\mathcal{D}}}[V^{k,w}(D')]$ が大きく異なることは望まれない. 分布 $P_{\mathcal{D}}$ からその期待値 $\mathbb{E}_{D\sim P_{\mathcal{D}}}[V^{k,w}(D)]$ への対応は距離空間間の写像であるため, その写像の連続性を証明することで, 期待値の変動が分布の変動よりも小さいことを示す (主結果 3) .

分布に対する統計的位相的データ解析の応用として, カーネル法による二標本検定 [GBR+07, GBR+12] の適用とその解釈について紹介する. 対象とするデータとしては, 球面とトーラス, 正方格子を中心に一様分布や正規分布でノイズを加えた実験的なものや, 材料科学への応用に関連のある Matérn hard-core point process[Mat60] に由来するものを用いる. 統計的位相的データ解析でよく用いられる Persistence landscape[Bub15] との比較も行い, PWGK ベクトルの優位性を考察する.

参考文献

[Bub15] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, Vol. 16, No. 1, pp. 77–102, 2015.

[Car09] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, Vol. 46, No. 2, pp. 255–308, 2009.

[GBR+07] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

[GBR+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, Vol. 13, No. Mar, pp. 723–773, 2012.

[HNH+16] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, Vol. 113, No. 26, pp. 7035–7040, 2016.

[KFH16] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Persistence weighted Gaussian kernel for topological data analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2004–2013, 2016.

[KFH18] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Kernel method for persistence diagrams via kernel embedding and weight factor. *accepted, to appear in Journal of Machine Learning Research (arXiv:1706.03472)*, 2018.

[Mat60] Bertil Matérn. Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden Fran Statens Skogsforskningsinstitut*, Vol. 49, No. 5, 1960.

[STR+17] Mohammad Saadatfar, Hiroshi Takeuchi, Vanessa Robins, Nicolas Francois, and Yasuaki Hiraoka. Pore configuration landscape of granular crystallization. *Nature Communications*, Vol. 8, p. 15082 EP, 2017.

[VdV98] Aad Van der Vaart. *Asymptotic statistics*, Vol. 3. Cambridge university press, 1998.

CONFIDENCE BANDS IN FUNCTIONAL LINEAR REGRESSION

MASAAKI IMAIZUMI (THE INSTITUTE OFSTATISTICAL MATHEMATICS)
KENGO KATO (THE UNIVERSITY OF TOKYO)

1. Outline

Data collected on dense grids can be typically regarded as realizations of a random function. Such data are called *functional data*, and statistical methodology dealing with functional data, called *functional data analysis*, has now a wide range of applications including chemometrics, econometrics, and biomedical studies. One of the most basic models in functional data analysis is a functional linear regression model.

This paper develops a simple method to construct confidence bands for the slope function in a functional linear regression model which is applicable to a PCA-based estimator. To be precise, we work with the following setting. Let Y be a scalar response variable and let X be a predictor variable which we assume to be an $L^2(I)$-valued random variable (random function) such that $\int_I \mathrm{E}\{X^2(t)\}dt < \infty$, where I is a compact interval. Consider a functional linear model with a scalar response variable

$$Y = a + \int_I b(t)[X(t) - \mathrm{E}\{X(t)\}]dt + \varepsilon,\ \mathrm{E}(\varepsilon) = 0,\ \mathrm{E}(\varepsilon^2) = \sigma^2 \in (0, \infty), \tag{1}$$

where a is an unknown constant (indeed, $a = \mathrm{E}(Y)$), $b \in L^2(I)$ is an unknown slope function, and X and ε are independent. The error variance σ^2 is also unknown. We are interested in constructing confidence bands for the slope function b centered at a PCA-based estimator. In spite of extensive studies on functional linear regression models, to the best of our knowledge, there is no formal result on confidence bands for the slope function b which is applicable to a PCA-based estimator (see below for the literature review). The purpose of this paper is to fill this important void.

2. Methodology

Let $\{\phi_j\}_{j=1}^{\infty}$ be an orthonormal basis of $L^2(I)$ by the spectral expansion of $K(s,t) = \mathrm{Cov}\{X(s), X(t)\}$, then we have the following expansions in $L^2(I)$:

$$b(t) = \sum_{j=1}^{\infty} b_j \phi_j(t),\ \text{and}\ X(t) = \mathrm{E}\{X(t)\} + \sum_{j=1}^{\infty} \xi_j \phi_j(t),$$

where b_j and ξ_j are defined by $b_j = \int_I b(t)\phi_j(t)dt$ and $\xi_j = \int_I [X(t) - \mathrm{E}\{X(t)\}]\phi_j(t)dt$, respectively. By some calculation, we obtain the following alternative expression of the regression model (1) for each $j = 1, 2, \ldots$, namely,

$$b_j = \mathrm{E}(\xi_j Y)/\kappa_j. \tag{2}$$

Empirical estimation for $\{\xi_j\}_j$ and $\{\kappa_j\}_j$, we obtain an estimator $\widehat{b}_j$ for b_j and consider an estimator for b of the form

$$\widehat{b}(t) = \sum_{j=1}^{m_n} \widehat{b}_j \widehat{\phi}_j(t),$$

where m_n is the cut-off level such that $m_n \to \infty$ as $n \to \infty$. Hall & Horowitz (2007) study the properties of the PCA-based estimator $\widehat{b}$ in detail and provide conditions under which the estimator is rate optimal.

2.1. **Construction of confidence bands.** In the present paper, we aim at constructing a confidence band $\mathcal{C} = \{\mathcal{C}(t) = [\ell(t), u(t)] : t \in I\}$ such that for given $\tau_1, \tau_2 \in (0, 1)$, with probability at least $1 - \tau_1$, the proportion of the set of t at which b is not covered by $\mathcal{C}$ is at most τ_2, i.e.,

$$\mathrm{P}\left\{\lambda\left(\{t \in I : b(t) \notin [\ell(t), u(t)]\}\right) \le \tau_2 \lambda(I)\right\} \ge 1 - \tau_1, \tag{3}$$

where λ denotes the Lebesgue measure. If the band $\mathcal{C}$ satisfies the new requirement (3), then the band $\mathcal{C}$ covers b over more than $100(1 - \tau_2)\%$ of points in I with probability at least $1 - \tau_1$, and so as long as τ_2 is close to 0, the band $\mathcal{C}$ covers b over "most" of points in I with probability at least $1 - \tau_1$.

Under some regularity conditions, it will be shown that

$$n\|\widehat{b} - b\|^2 = \sum_{j=1}^{m_n} \left(n^{-1/2} \sum_{i=1}^{n} \varepsilon_i \widehat{\xi}_{i,j}/\widehat{\kappa}_j \right)^2 + O_{\mathrm{P}}(m_n^{\alpha/2+1} + \sqrt{n} m_n^{-\beta+\alpha/2+1} + n m_n^{-2\beta+1}), \tag{4}$$

where $\varepsilon_i = Y_i - a - \int_I b(t)[X_i(t) - \mathrm{E}\{X(t)\}]dt$ for $i = 1, \ldots, n$, and the last term on the right hand side on (4) is (suitably) negligible relative to the first term (the parameters α and β will be given in the next section). We will approximate the first term of the RHS of (4) and construct the confidence band.

References

Hall, P. & Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70-91.

生命・自然科学における複雑現象解明のための統計的アプローチ

高次元データにおける精度行列の罰則付推定とその最適化

中京大学 国際教養学部 永井 勇

1 導入

本講演では, $n \times p$ の多変量データ $\boldsymbol{Y}$ の分析を考える. ここで, n は個体数, p は次元数である. また本講演では, $E[\boldsymbol{Y}] = \boldsymbol{O}$ ($n \times p$ ゼロ行列), 各個体のデータは互いに無相関であり, 各個体のデータの真の分散共分散行列を $\boldsymbol{\Sigma}$ (未知) とし, $\boldsymbol{\Sigma}$ は $p \times p$ 正定値行列であると仮定する. 一方で, 本講演では特に分布や n や p の大小関係に仮定を置かない. ここで, 多変量データ $\boldsymbol{Y}$ を用いた $\boldsymbol{\Sigma}$ の推定量 $\boldsymbol{S}$ は, 全ての成分が 1 の n 次元ベクトル $\mathbf{1}_n$ を用いて, $\boldsymbol{S} = \boldsymbol{Y}'\{\boldsymbol{I}_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\}\boldsymbol{Y}/(n-1)$ で得られる (Rencher & Christensen (2012) など参照). また, この $\boldsymbol{S}$ は求め方から非負定値行列である.

$\boldsymbol{Y}$ の分析の多くの場面において, $\boldsymbol{S}^{-1}$ (精度行列) が必要となるため, $\boldsymbol{S}$ が正定値行列であること (つまり $\boldsymbol{S}^{-1}$ が存在すること) が仮定されている. しかし, 従来の多変量線形回帰における推定と同様に, $n < p$ の場合などに $\boldsymbol{S}^{-1}$ が存在せず, $n > p$ であっても $\boldsymbol{Y}$ の列に相関の高い変数の組がある場合に $\boldsymbol{S}^{-1}$ が不安定になってしまう. 本講演では, これらの問題を回避する手法について考える.

これらの問題に対する従来の回避法としては, 以下のような手法が提案されている;

1. Moore-Penrose 型一般逆行列を $\boldsymbol{S}^{-1}$ の代わりに用いる手法
   (一般逆行列については, Schott (2017) Section 5 などを参照)
2. $\boldsymbol{S}$ の対角成分だけを用いた行列の逆行列を $\boldsymbol{S}^{-1}$ の代わりに用いる手法
   (詳細は Srivastava, Katayama & Kano (2013) などを参照)
3. $\boldsymbol{\Sigma}$ の Cholesky 分解に着目する手法
   (詳細は Chang & Tsay (2010) などを参照)
4. $\boldsymbol{S}$ にリッジタイプの罰則を付けて逆行列を $\boldsymbol{S}^{-1}$ の代わりに用いる手法
   (詳細は Wang, Pan, Tong, & Zhu (2015) などを参照)

本講演では, 4 番目のリッジタイプの罰則を付ける手法に着目する. この推定法は, 正のパラメータを用いて $\boldsymbol{S}$ に罰則を付けてから逆行列を求め, その後に再調整を行う推定法である. この推定法では, $\boldsymbol{S}$ の全ての固有値に対して一つのパラメータで調整を行っているため, 大きく調整が必要な固有値に対しても調整が不要な固有値に対しても一様に調整する形となっている. また, 二つのパラメータの同時最適化が必要となるという問題がある.

本講演では, リッジタイプの罰則を付ける手法を拡張し, 上述の一様調整・同時最適化の問題点を回避する推定法を提案する. これにより, 調整が必要な部分には適当に調整し, 調整が不要な部分には調整しない形での柔軟な推定が可能となる. さらに, そのときに用いるパラメータの最適な値が陽に求まるため, 最適化のための反復計算が不要となり, 同時最適化が不要な推定量となっている.

2 罰則付推定量を用いる手法

2.1 リッジタイプの罰則を付ける手法とその問題点

上述したように, $\boldsymbol{S}^{-1}$ が不安定になることや存在しないという問題を回避する一つの手法として, リッジタイプの罰則を付ける推定法が提案されている (Wang *et al*, 2015). この手法で用いる推定量は $\boldsymbol{S}^{-1}$ の代わりに, 二つのパラメータ $\lambda > 0$ と $\alpha > 0$ を用いて $\lambda(\boldsymbol{S} + \alpha\boldsymbol{I}_p)^{-1}$ という形の推定量を用いる手法である. ここで, $\lambda = 1$ とすると Chen *et al.* (2011) で用いられている推定量である.

この推定量においては以下のような問題がある;

(I) $\boldsymbol{S}$ の固有値を通して考えると, 0 に近い固有値に対しては大きく調整が必要で, 0 から離れた固有値に対しては調整がほぼ不要だが, それらを α 一つで一様に調整する形となっている

(II) α と λ の同時最適化が必要である

本講演では, これらの問題を回避し, 柔軟な調整が可能かつ同時最適化が不要な推定量を提案する.

2.2 提案する一般化リッジタイプの罰則を付ける手法

リッジタイプの罰則を付ける手法においては, 上述した問題がある. これらの問題を回避する手法として, 次の推定量を用いる手法を提案する;

$$\hat{\boldsymbol{S}}^{-1}(\lambda, \boldsymbol{\theta}) \stackrel{\text{def.}}{=} \lambda(\boldsymbol{S} + \boldsymbol{Q}\mathrm{diag}(\boldsymbol{\theta})\boldsymbol{Q}')^{-1},$$

ここで $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ ($\theta_i \geq 0$, $i = 1, \ldots, p$), さらに $\boldsymbol{Q}$ は $\boldsymbol{S}$ の固有値 $d_1, \ldots, d_p$ を対角に並べた $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_p)$ を用いて $\boldsymbol{Q}'\boldsymbol{S}\boldsymbol{Q} = \boldsymbol{D}$ となる直交行列である. ここで, $\boldsymbol{S}$ の求め方より $d_i \geq 0$ ($i = 1, \ldots, p$) である.

この推定量においては, $\boldsymbol{S}$ の各固有値 d_i に対してパラメータ θ_i がそれぞれ対応している. そのため, d_i が 0 に近い場合は θ_i を大きくし, d_i が 0 から離れている場合は θ_i を小さくすることで柔軟な調整が可能な推定量となっている. また, パラメータ数はリッジタイプの罰則を付ける推定量より多く $(p+1)$ 個となるが, 後述のロス関数を最小にする $\boldsymbol{\theta}$ が陽に求まるため, 同時最適化は不要となる.

2.3 二乗ロス関数

本講演では, λ を固定した下での $\boldsymbol{\theta}$ の最適化をまず考える. そこで, 例えば Wang *et al.* (2015) などで使われている二乗ロス関数を考える. $\hat{\boldsymbol{S}}^{-1}(\lambda, \boldsymbol{\theta})$ の二乗ロスは, 次の形で定義される;

$$L^{[2]}(\hat{\boldsymbol{S}}^{-1}(\lambda, \boldsymbol{\theta})) \stackrel{\text{def.}}{=} \mathrm{tr}\left\{\left(\hat{\boldsymbol{S}}^{-1}(\lambda, \boldsymbol{\theta})\boldsymbol{\Sigma} - \boldsymbol{I}_p\right)^2\right\}.$$

この関数を最小にする $\boldsymbol{\theta}$ は, λ を固定した下で陽に求まることを本講演で示す. さらに本講演では, CV 法を用いたアルゴリズムでの λ の最適化法についても講演する. さらに, James and Stein (1961) などで使われているロス関数やそれぞれのロス関数を最小にする $\boldsymbol{\theta}$ の導出などについても触れる.

数値実験などによる比較など

他のロス関数や各ロス関数を最小にする $\boldsymbol{\theta}$ の導出やその特徴, 数値実験を通じた比較などについては当日の講演で報告する.

引用文献:

[1] Chang, C. and Tsay, R. S. (2010). Estimation of covariance matrix via the sparse Cholesky factor with lasso. *J. Stat. Plann. Infer.*, **140**, 3858–3873.

[2] Chen, L. S., Paul, C., Prentice, R. L., and Wang, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Am. Stat. Assoc.*, **106**, 1345–1360.

[3] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, **1**, 361–379.

[4] Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis* (Third Edition).

[5] Schott, J. R. (2017). *Matrix Analysis for Statistics* (Third Edition).

[6] Srivastava, M. S., Katayama, S., and Kano, Y. (2013). A two sample test in high dimensional data. *J. Multivariate Anal.*, **114**, 349–358.

[7] Wang, C., Pan, G., Tong, T., and Zhu, L. (2015). Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statistica Sinica*, **25**, 993–1008.

個別化医療への機械学習的アプローチ

菅澤翔之助 (統計数理研究所)　　野間久史 (統計数理研究所)

はじめに

2015 年、米国でオバマ前大統領の一般教書演説で示された “Precision Medicine Initiative” のように, 個々人の特性に合わせた最適な医療を行う個別化医療に対する期待はますます高まっている. 特に, 抗がん剤などの薬剤では, 治療効果や副作用の個人差が大きく, それらを正確に予測する分子診断法の開発が実現すれば, 患者個人や社会・医療経済に与える便益は極めて大きい. このような個別化医療に関する研究は, 学術分野における注目度も年々高まっている. 例えば, 米国国立生物科学情報センターが作成しているデータベース PubMed において登録されている個別化医療 (Personalized Medicine) に関連する文献数を調べてみると, 2004 年では 500 件程度だった数が, 2016 年には 7000 件近くまで伸びており, 10 年程度で個別化医療に関連した出版論文数が飛躍的に増加していることが確認できる. 本研究では, 勾配ブースティング木 (Friedman, 2001; 2002) と呼ばれる機械学習手法を導入することで, 大規模な遺伝情報データから, 個別治療効果の正確な推定を可能にする手法を開発し, より精度の高い個別化医療の実現に貢献することを目指す. 個別化医療における統計手法に関しては最新のレビュー論文 (Lopkovich et al., 2017) に包括的にまとめられている.

提案手法

$T \in \{0,1\}$ を処置変数 ($T = 1$ が処置を表す), $Y^{(T)}$ を処置 T のもとでの潜在アウトカム (potential outcome) とする. したがって, 同一個体に対して $Y^{(1)}, Y^{(0)}$ は同時に観測することができず, 代わりに $Y = TY^{(1)} + (1-T)Y^{(0)}$ が観測される. さらに X を p-次元の説明変数ベクトルとする. 多くの場合, X の次元 p は大きい (例えば, X として遺伝子発現量を用いる場合は, p は数万のオーダーになる). 観測データは $\{(Y_i, X_i, T_i),\ i = 1, \ldots, N\}$ で与えられる.

個別治療効果 (ITE; individual treatment effect) は以下のように定義される.

$$\Delta(X) = \mathrm{E}[Y|X, T=1] - \mathrm{E}[Y|X, T=0] \quad \text{(連続値, 二値)}$$
$$\Delta(X) = P(Y \geq t|X, T=1) - P(Y \geq t|X, T=0) \quad \text{(time-to-event)}$$

このような $\Delta(X)$ を推定するために, Tian et al. (2014) では, Y と (X, T) の関係を MC(modified covariate) 法と呼ばれる手法を導入し, Lasso などの正則化法を組み合わせてパラメトリックに推定する方法を提案している. しかし, パラメトリックモデルは X の次元数に応じて必要なパラメータ数が膨大になることやモデルの誤特定の危険性があることから, ノンパラメトリックな方法も提案されている. Lipkovich et al. (2017) のレビュー論文では木回帰による方法が議論されている. また, Foster et al. (2011) では, 木回帰のアンサンブル学習の一種であるランダムフォレスト (Breiman, 2001) を用いて, $\Delta(X)$ を推定する手法を提案している.

本研究では, 木回帰のアンサンブル学習の一種である勾配ブースティング木 (Gradient Boosting Tree; GBT) を用いて $\Delta(X)$ の推定を行うことを提案する. 以下では, アウトカムが time-to-event の場合における提案手法の流れを説明する. Y を time-to-event, C を censoring time, $\delta = I(Y < C)$ を censoring indicator, $\tilde{Y} = \min(Y, C)$ を観測される time-to-event とする. このとき, 各処置群 ($T = 0, 1$) ごとの Cox 回帰を考える.

$$\lambda(s|X, T=t) = \lambda_t(s) \exp\{h_t(X)\}, \quad t = 0, 1$$

損失関数として部分尤度を用いることで, GBT により $h_0(X)$ および $h_1(X)$ を推定することができる. ITE は

$$\begin{aligned}\Delta(X) &= P(Y \geq t_0|X, T=1) - P(Y \geq t_0|X, T=0) \\ &= \exp\left[\exp\{h_1(X)\}\int_0^{t_0} \lambda_1(s)ds\right] - \exp\left[\exp\{h_0(X)\}\int_0^{t_0} \lambda_0(s)ds\right]\end{aligned}$$

と表現することができるので, 推定された Cox 回帰の結果から ITE を推定することができる.

数値実験

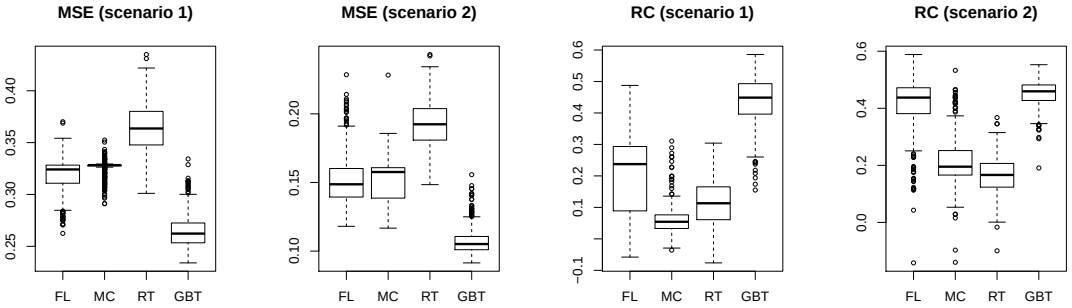
既存手法に対する提案手法のパフォーマンスを確認するため, 数値実験を行う. 以下では, $N = 300$(サンプル数), $p = 1000$(共変量数) とし, サンプルの半分を処置群 ($T = 1$) に割り当てる. さらに, 共変量 $X = (X_1, \ldots, X_p)$ を平均 0, 分散 1, ペアワイズ相関係数 0.3 の多変量正規分布から生成する. Y (time-to-event) を以下のモデルから生成する.

$$Y = \exp\left\{\beta_0 + \sum_{k=1}^{p} \beta_k (X_k + X_k^2) + G\left(\sum_{k=1}^{p} \gamma_k X_k + \sum_{1 \le j < k \le p} \alpha_{jk} X_j X_k\right) + \varepsilon\right\}.$$

ただし, $\beta_0 = 0.1, \beta_1 = \cdots = \beta_6 = 0.2$, $\gamma_1 = \gamma_3 = 0.8, \gamma_2 = \gamma_4 = -0.8$, $\alpha_{15} = \alpha_{16} = \alpha_{17} = \alpha_{18} = 0.6$, $\varepsilon \sim N(0, (1.5)^2)$ であり, それ以外のパラメータの値は全て 0 とした. C を $[0, 40]$ 上の一様分布から生成し, $\tilde{Y} = \min(Y, C)$ を観測値とする.(打ち切り割合はおおよそ 30% であった). 関数 $G(\cdot)$ に対しては, 2 つのシナリオ (1) $G(x) = x$, (2) $G(x) = x + \sin(x)$ を考える. 比較手法として, 正則化 Cox 回帰による方法 (FL), Tian et al. (2014) による方法 (MC), 木回帰による方法 (RT), 提案手法 (GBT) を考える. ITE を

$$\Delta(X) = P(Y \ge t_0 | X, T = 1) - P(Y \ge t_0 | X, T = 0), \quad t_0 = 10.$$

と定義し, 新たに生成した 1000 個のデータから, 対応する $\Delta(X)$(真値) を計算する. 各手法から $\Delta(X)$ の推定値を構成し, 真値との平均 2 乗誤差 (MSE) およびスピアマンのランク相関 (RC) を計算した. 500 回の繰り返しから計算した MSE および RC は以下の図のようになり, 既存手法に対する提案手法の有効性が確認できる.



参考文献

[1] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

[2] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 1189-1232.

[3] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38, 367-378.

[4] Foster, J. C. et al. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30, 2867-2880.

[5] Lipkovich, I. et al. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36, 136-196.

[6] Tian, L. et al. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109, 1517-1532.

AI 創薬：機械学習による様々な疾患に対するデータ駆動型の新薬開発

九州大学生体防御医学研究所・科学技術振興機構さきがけ　山西芳裕
yamanishi@bioreg.kyushu-u.ac.jp

多くの薬物は標的とするタンパク質などの生体分子との相互作用を介して、薬効や副作用という形で人体に影響を及ぼす。そのため、薬物・標的タンパク質間の相互作用の同定は、医薬品の開発過程において最重要課題である。ポストゲノム研究では、ゲノム、トランスクリプトーム、プロテオームなどの遺伝子やタンパク質に関する大量のオミックスデータが得られるようになってきた。同時にハイスループットスクリーニングなどの技術の発展によって、膨大な数の化合物や薬物に関するケミカル情報や生理活性情報も蓄積されている。そのため、これらの様々なオミックスデータを有効活用し、薬物・標的タンパク質間相互作用（もしくは化合物・タンパク質間相互作用）を大規模に予測するための情報技術の開発が期待されている。

薬の分子は目標とした標的タンパク質にのみ結合するとは限らず、本来目標としていない複数のタンパク質に結合し、予想外の薬理作用を起こすことがある。また複数のリガンドを持つタンパク質も多い。つまり、薬物と標的タンパク質の関係は、１対１の関係ではなく多対多の関係となる。これを相互作用のネットワークで表現すると、薬物・標的タンパク質間相互作用ネットワークは、図１のような二部グラフの形で表せる。図では実線が既知の相互作用、点線が潜在的な相互作用を表す。つまり、薬物・標的タンパク質間相互作用予測の問題は、情報科学的には二部グラフの潜在的な辺（エッジ）を予測する問題として捉えることができる。

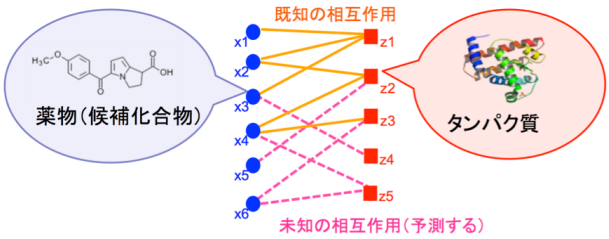


図１ 薬物・標的タンパク質間相互作用ネットワーク

薬物・標的タンパク質間相互作用を情報科学的に予測するための方法論の研究が近年盛んに行われている。これまでの先行研究では、“類似薬物は類似タンパク質と相互作用しやすい”という経験的知見のもと、薬物のケミカル情報を用いる「ケモゲノミクス」、薬物のフェノーム情報を用いる「フェノミクス」、薬物のトランスクリプトーム情報を用いる「トランスクリプトミクス」に大きく分かれる。実際の応用における予測精度は、薬物やタンパク質を表すデータの特性や網羅性に大きく依存する。図２は３つの枠組みを視覚的に示したものである。

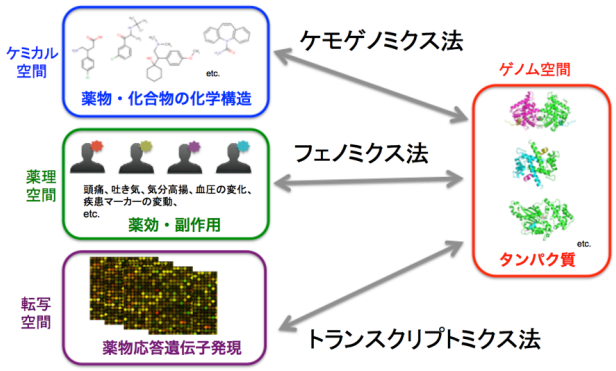


図２薬物・標的タンパク質間相互作用を予測する枠組み：ケミカルゲノミクス、フェノミクス、トランスクリプトミクス

ケモゲノミクスの枠組みでは、化学構造が似ている薬物は配列が似ているタンパク質に相互作用すると予測を行うのが基本的な方針となる。タンパク質の配列情報を使うことで、３次元立体構造が未知のタンパク質にも適用でき、また１つの標的タンパク質だけではなく複数の標的タンパク質を予測モデルに組み込めるので、リガンドが未知のタンパク質にも適用できる点が魅力である。

これまでの先行研究の方法論は、ペアの分類問題として考える枠組みと、次元削減による距離学習として考える枠組みの２つに大きく分けれる[1,2]。ペアの分類問題の枠組みでは、薬物・タンパク質ペアをオブジェクトと見なして、そのペアが相互作用するクラスか相互作用しないクラスかを予測する分類問題として取り扱う [3,4,5,6,7]。分類器としては予測精度の高さからサポートベクターマシンが使われることが多い。この場合、サポートベクターマシンの入力は、ペアのカーネル類似度行列になるので、薬物・タンパク質ペアのカーネル関数を定義する必要がある。

次元削減の方法として我々のグループでは、相互作用予測問題を二部グラフの推定問題と見なし、距離学習に基づく独自の手法を開発した[8,9]。薬物の化学構造類似性を表すケミカル空間とタンパク質のアミノ酸配列類似性を表すゲノム空間があると想定すると、提案手法は以下の３ステップから成る。

1. 既知の薬物・標的タンパク質間相互作用ネッ

トワークを、トポロジーの情報をできるだけ落とさず、有限次元のユークリッド空間に埋め込む。埋め込まれた空間を特徴空間と呼ぶ。ネットワーク上で隣接する薬物と標的タンパク質は、この特徴空間においても近い距離にある。

2. ケミカル空間と特徴空間の間の相関を薬物に関して学習し、相関モデルを構築する。ゲノム空間と特徴空間の間の相関を標的タンパク質に関して学習し、相関モデルを構築する。
3. 標的が未知の薬物（または候補化合物）、リガンドが未知のタンパク質に対して、上のステップで構築した相関モデルを適用し、特徴空間にマッピングする。そこで距離が近い薬物とタンパク質を相互作用するペアとして予測する。

フェノミクスの枠組みでは、薬理作用など人体へのフェノタイプが似ている薬物は同じようなタンパク質に相互作用すると予測を行うのが基本的な方針となる。ケモゲノミクスの手法は薬物の化学構造に依存するが、フェノミクスの手法は薬物の化学構造からは想像がつかないような薬物・標的タンパク質間相互作用を検出できる点が長所である。

先駆的な研究として、医薬品の添付文書に書かれた副作用の類似性を用いて、薬物の標的タンパク質を予測する手法が提案された[12]。我々はそれを発展させる形で、薬物の薬効・副作用による薬理作用類似性とタンパク質のアミノ酸配列類似性の両方を用いた手法を提案した[13]。薬理作用の情報が得られない薬物や化合物に対しても実行可能にするため、薬理作用類似性を化学構造類似性から予測する統計モデルを予測手順を組み込んだアルゴリズムを開発した。

我々のグループでは、フェノミクスの手法の網羅性を上げるため、医薬品添付文書に書かれた副作用の情報だけではなく、市販後に報告された薬物の副作用情報を用いて、薬物・標的タンパク質間相互作用を予測する手法を開発した[14]。米国 FDA（食品医薬局）の AERS（薬害事象報告システム）で公開されている数百万人の患者に対する市販後調査報告書を解析し、副作用の報告頻度のプロファイルから薬物・標的タンパク質間相互作用を予測する手法を提案した。FDA AERS は、製薬会社による医薬品開発時の臨床データには載っていない副作用の情報が多く含まれている点が強みである。他の副作用のデータベースやリソースと比べても、FDA AERS は副作用の情報が得られる薬物の数が一番多いので、より網羅的な解析ができる点が長所である。

本稿では、薬物・標的タンパク質間相互作用（または化合物・タンパク質間相互作用）をゲノムワイドに予測するための様々な手法を紹介した。特に、ケモゲノミクスやフェノミクスの枠組みで近年開発された機械学習の手法の概要を説明した。項数の制限により各手法の詳細部分は省略し、概念的な説明が中心になってしまったので、詳細に興味がある方はオリジナルの論文を参照されたい。

最近、創薬科学における様々な問題に対して開発された機械学習法をまとめた本が出版された[15]。ケモインフォマティクスの機械学習としては世界最初の本である。ケモインフォマティクスや機械学習の最前線で活躍している研究者の方々の貢献によって構成されており、濃い内容となっているので興味のある方は是非参考にして頂ければ幸いである。

参考文献

[1] Yamanishi, Y. and Kashima, H.; Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques (Lodhi, H. and Yamanishi, Y., eds.), pp.304-317 (2010).

[2] Yamanishi, Y.; Data Mining for Systems Biology, Methods in Molecular Biology Series (DeLisi, C., Kanehisa, M. and Mamitsuka, H., eds.), Springer, pp.97-113 (2012).

[3] Nagamine, N. and Sakakibara, Y; Bioinformatics, 23, 2004–2012 (2007).

[4] Faulon, J., Misra, M., Martin, S., Sale, K., Sapra, R.; Bioinformatics 24, 225−233 (2008).

[5] Jacob, L., Vert, J.-P.; Bioinformatics, 24, 2149-2156 (2008).

[6] Bleakley, K. and Yamanishi, Y.; Bioinformatics, Vol.25, pp.2397-2403 (2009).

[7] Yabuuchi, H., Niijima, S., Takematsu, H, Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., Okuno, Y.; Molecular Systems Biology, 7:472 (2011).

[8] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M.; Bioinformatics, Vol.24, pp.i232-i240 (2008).

[9] Yamanishi, Y.; Advances in Neural Information Processing Systems 21 (Koller, D., Schuurmans, D., Bengio, Y. and Bottou, L. eds.), pp.1841-1848, MIT Press, Cambridge, MA (2009).

[10] Mahé, P., Ueda, N., Akutsu, T., Perret, J.L., Vert, J.P.; J Chem Inf Model. 45(4), 939-951 (2005).

[11] Saigo, H., Vert, J.P., Ueda, N., Akutsu, T.; Bioinformatics, 20(11),1682-1689 (2004).

[12] Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., Bork, P., Science, 321, 263-266, 2008 (2008).

[13] Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S.; Bioinformatics, Vol.26, pp.i246-i254 (2010).

[14] Takarabe, M., Kotera, M., Nishimura, Y., Goto, S., and Yamanishi, Y.; Bioinformatics, Vol.28, pp.i611-i618 (2012).

[15] Lodhi, H. and Yamanishi, Y.; Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques, IGI Global (2010).

バイオインフォマティクスによる古代遺伝子の再現と機能解析

白井　剛（長浜バイオ大学）

近年バイオインフォマティクスの応用により、化石や化石 DNA に依存せずに行う古生物学が可能となった。これは分子系統推定技術を利用して、ある遺伝子間の共通祖先遺伝子の配列を推定する方法であり、生命史を配列データに基づき再現し、さらに実験を施す事が可能になる(図 1)。私たちのグループでは、特にタンパク質構造進化の実験による検証に興味を持ち研究を行ってきた。

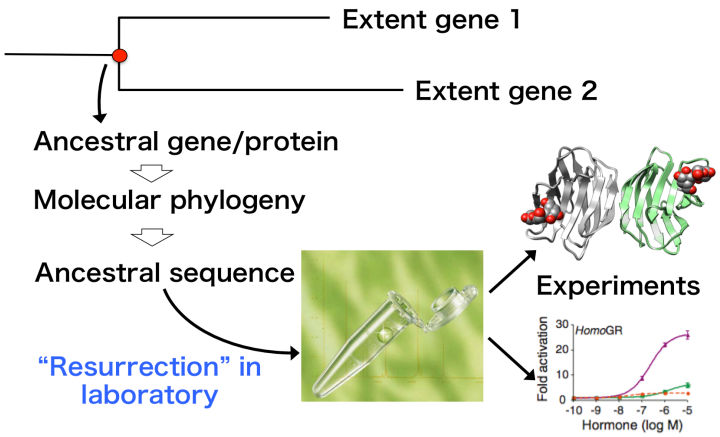


図 1　祖先配列の再現による古生物学

一例として、アポトーシス誘導による生体防御に関連した魚類ガレクチン Congerin の立体構造進化の再現では、サブユニット間の β ストランドの交換(ストランドスワップ)による 2 量体構造の安定化という、比較的大きな構造進化を、アイソフォーム ConI と ConII の共通祖先配列(Con-anc と Con-anc')を経験的ベイズ法で求めてタンパク質分子を再現することで解析した。Con-anc および Con-anc'に対し様々な生化学的実験を行い、さらに X 線結晶解析により、最大 1.4Å 分解能での立体構造解析を行ったところ。Con-anc'の立体構造は、機能的にも構造的にも ConI と ConII の祖型を示した(図 2)。

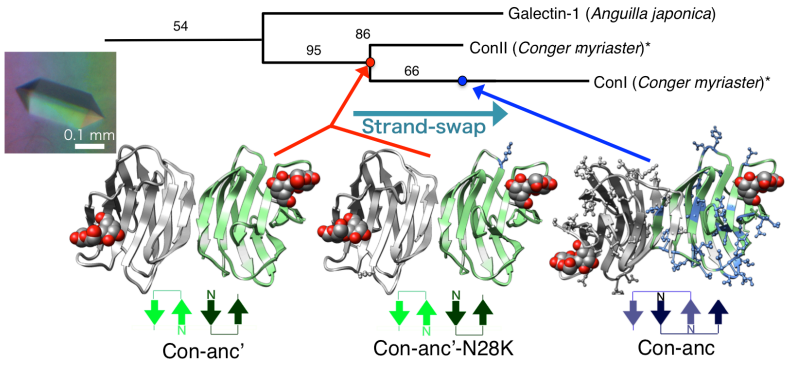


図 2　祖先型コンジェリンの再現

また有酸素運動で酸素の蓄積に関与する myoglobin の潜水適応メカニズムを祖先タンパク質の再現により解析した。クジラなどの潜水動物では

myoglobin 表面の正電荷の増加による静電的反発で、高濃度下での凝集が抑制され溶解度が高くなるとされる。そこで、クジラの類陸上祖先 *Pakicetus* (aMbWp), クジラ類共通祖先 *Basilosaurus* (aMbWb および aMbWb') の myoglobin 配列を計算し再現実験を行った。

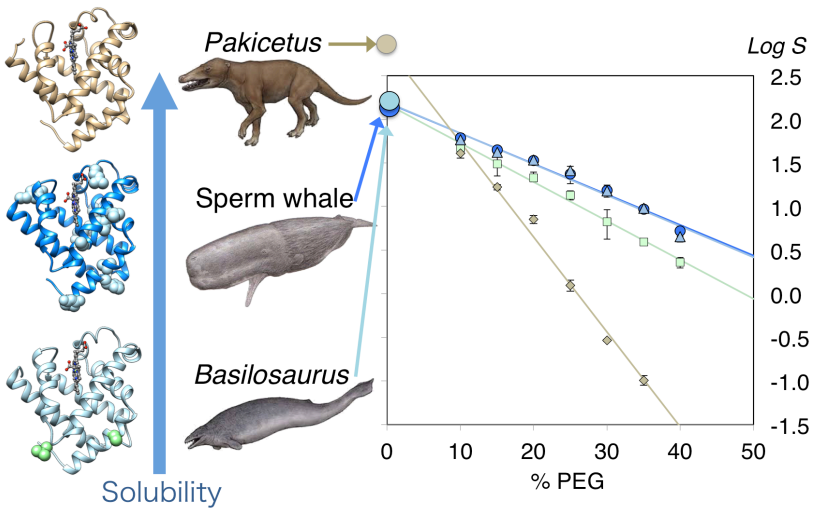


図 3　祖先型 myoglobin の溶解度測定

祖先 myoglobin の溶解度を PEG 沈殿法で求めた結果、aMbWp >> aMbWb' = aMbWb =swMb となり、表面正電荷の増加による適応説から予想される結果とは異なっていることが示された(図 3)。さらに、祖先 myoglobin の立体構造解析と種々の生化学解析を行うことで、沈殿剤耐性と立体構造安定性の増強が潜水適応に重要であることが示唆された。

これらの例から示されるように、祖先配列の再現による古生物学には、実在の分子を対象として実験ができるという大きな魅力がある。このため最近では、数億年〜数十億年前(先カンブリア期)の祖先タンパク質再現の報告があるが、報告された祖先配列の各サイトの平均事後確率は 0.9 を上回っている。しかしながら、平均事後確率が過去に遡っても安定であるのに対して、最低の事後確率は比較的急激に低下する。タンパク質分子はそれ自体が 1 個の複雑系であり、1 個のアミノ酸残基を置換しただけで機能や構造を失う例は多い。つまり全体としての配列の正確さは、大過去の推定においては重大な問題になる。

今後この脆弱性を克服する必要があるが、ひとつには、これが現存遺伝子のデータが多いほど正確性が高くなる方法であることから、現在爆発的に蓄積している配列情報による改善が期待できる。あるいは、現状の方法は遺伝子やタンパク質の配列を単純に文字データと解釈して計算を行うが、これらには分子としての実在性があることを利用して、文字列の自由度を物理化学的に制限する方法を探索するなどの改善が必要になると思われる。

統計学とスパコンでがんゲノムにチャレンジ

井元清哉（東京大学医科学研究所）

次世代シークエンサーの発展に伴い、ヒト全ゲノムのシークエンスは10万円以下となったのは３年程前、いまや 5 万円に近づく勢いである。この技術を用い、世界中至る所でヒトゲノムシークエンスはさまざまな疾患において行われ、膨大なデータが蓄積されている（健常人についてももちろん行われている)。その先端を走っているのががん研究である。がんは、ゲノムに蓄積した多数多様な変異がゲノム上に蓄積され、正常な制御から逸脱した結果生じる疾患であると考えられている。

米国では、The Cancer Genome Atlas（TCGA）において１万以上のさまざまながん患者からマルチオミックスのデータ（ゲノム（公開されているほとんどは全エキソームシークエンス)、トランスクリプトーム、プロテオーム等）がシステマティックに収集されている。また、国際がんゲノムコンソーシアム（International Cancer Genome Consortium（ICGC））では、50のがん種を対象にそれぞれ 500 人の患者をリクルートし、がんゲノム、および対照正常細胞のゲノムを全ゲノムシークエンスし、がんにおけるゲノム異常のカタログを構築している。この２つのビックプロジェクトが共同で行っているのが、ICGC/TCGA PanCancer Analysis of Whole Genome（PCAWG）である。２つのプロジェクトから約 3,000 のがん患者の全ゲノムシークエンスデータ、および RNA シークエンスデータが供出され、10 のデータセンターを用いて virtual machine 上で同じ計算機環境を構築した上で同じ前処理（アライメント、体細胞変異の同定）を行い、その共通データを全てのデータセンターにてミラーし共有している。私の所属する東京大学医科学研究所には、スーパーコンピュータ（SHIROKANE、560 Tera Flops, 30 Peta Byte Lustre Disk Array）が設置されており、10 のデータセンターの一つとして Lustre システムから 1 Peta Byte の領域、および max 4,000 CPU コアを供出している。シカゴ大のスーパーコンピュータセンターやバルセロナスーパーコンピュータセンターなども参画している。

そのデータを世界中から集まった約 600 人のがん研究者・技術者が 16 のワーキンググループに分かれさまざまテーマで解析を行っている。私の研究チームは、PCAWG15 Immunogenomics としてがん免疫に関するデータ解析を PCAWG にお

いて行っている唯一のチームであり、HLA genotyping やneoantigenの予測を行い、我々のチームの研究を進めると共に他のチームにも提供し共同研究としても進めている。この事業を進めるためには、巨大ながんゲノムデータを解析するための計算インフラとしてのスーパーコンピュータと、データを解析するための統計学的な技術が必要不可欠である。本講演では、この ICGC/TCGA PanCancer にて我々、および世界中のがん研究者が行っているチャレンジについて説明したい。

また、このようながんゲノムのビッグプロジェクトの成果やこれまでに出版された論文、薬剤の作用機序の情報（パテント情報)、遺伝子間の制御関係を表したパスウェイの情報などをがんの臨床現場にて活用する「臨床シークエンス」が行われている。今年夏頃からは100程度のがん関連遺伝子をシークエンス（パネルシークエンス）してゲノム変異を同定し、その結果を治療に活用する臨床シークエンスは保険収載され全局の拠点病院においてスタートする。

我々は、2011年から全ゲノムシークエンスデータに基づく臨床シークエンスを実現するための研究をスタートさせ、国際がんゲノムコンソーシアムなど数々のがんゲノム研究で実績のあるデータ解析パイプライン・高セキュリティなデータ解析環境を構築している。この取り組みの中でデータ解析と共に重要となるのが、同定したがん特異的なゲノム変異の臨床翻訳である。基盤となる文献情報は、NIH PubMed には2600万報の論文が登録され、積み上げれば富士山よりも高く、人間が全てを読み理解できるものではない。がん種にもよるが、全ゲノムシークエンスを行うことで数万のゲノム変異が特定され、その臨床解釈は人間の能力を超えたところにある。そこで、我々は、2015 年 7 月より IBM Watson for Genomicsを利用し、このボトルネックを打ち破るための研究を行っている。この研究についても紹介する。

トピックモデルによるShallow RNA-Seqデータの補完

岩山　幸治
滋賀大学データサイエンス教育研究センター

概 要

近年その利用が拡大しているRNA-Seqは，深く読むほどノイズの少ない質の高いデータが得られるが，その分大きなコストを要する．同じコストでより大きなサンプルサイズを扱うために，総リード数の少ない shallow RNA-Seq データから発現量を精度よく推定する手法を提案する．提案手法は，自然言語処理に用いられるトピックモデルを元に，RNA-Seqデータの過分散を説明するために，単語あるいは遺伝子の生成分布に多項分布ではなく負の二項分布を用いる．シミュレーションで生成したRNA-Seqの模擬データに提案手法を適用することで，Shallow RNA-Seqから発現量を精度よく推定できることを示す．

1　はじめに

近年，遺伝子の発現を網羅的に定量するRNA-Seq法 [5] の利用が広がっている．RNA-Seqでは，読んだ配列の数が多いほどノイズの少ない質の高いデータが得られるが，コストが上がるために扱えるサンプルサイズが小さくなってしまう．総リード数の少ない，Shallow RNA-Seqデータから発現量を高精度に推定できれば，より大きなサンプルサイズについて遺伝子発現量を定量することができる．遺伝子間の発現パターンの関係性を適切に抽出できれば，ノイズの大きな定量データからでも他の遺伝子の情報を利用して高精度な発現量の推定が可能になると期待される．

自然言語処理の分野で用いられる，単語の共起に基づいて文書を構成するトピックを推定するトピックモデルの一つである Latent Dirichlet Allocation（LDA） [1] を単一細胞のRNA-Seqデータに適用することで，細胞間の階層的な構造を推定できることが示されている [2]．本研究では，スパース性を導入することで推定の安定性を向上したトピックモデルである Sparse Additive GEnerative model [3] に基づき，RNA-Seqの過分散なデータから発現量を正確に推定するために，単語の生成分布を多項分布から負の二項分布に置き換えたモデルを提案する．

2　提案モデル

以下のような生成モデルを考える．サンプル s におけるトピックの分布は，既存のトピックモデルと同様に Dirichlet 分布から生成され，各サンプルの各遺伝子について，トピックが割り当てられる．各トピックはそのトピックにおける発現量の，平均発現量からの差分 $\boldsymbol{\eta}_k$ で表現される．トピックの各要素をスパースにするため，以下の spike and slab 事前分布を導入する，

$$\eta_{k,i} \sim \pi_{k,i}\mathcal{N}(0, \tau_{k,i}) + (1 - \pi_{k,i})\delta_{\eta_{k,i}}, \tag{1}$$

$$\pi_{k,i} \sim \mathcal{B}eta(1, 1). \tag{2}$$

最後に，各サンプルにおける各遺伝子に対応するリードは割り当てられたトピックに対応する負の二項分布で生成される．

$$P(r_{si}|z_{si}, \boldsymbol{m}, \boldsymbol{\eta}, \boldsymbol{\phi}) \sim \mathcal{NB}(\nu_s \exp(m_i + \eta_{z_{si}i}), \phi_i) \tag{3}$$

ここで，ν_s はサンプル間の総リード数の違いを表すためのパラメータ，ϕ_i は遺伝子 i の dispersion と呼

ばれる過分散の度合いを決めるパラメータである．

推定は，周辺尤度の変分下限を最大化する近似分布を求めることで行う．$\eta_{k,i}$ の事後分布については，Doubly Stochastic Variational Inference [4] により推定する．標準正規分布に従う変数 $\omega_{k,i}$ を用いて，$\eta_{k,i} = \gamma(C_{k,i}\omega_{k,i} + \mu_{k,i})$ と表す．ここで，$\gamma_{k,i} \in \{-1, +1\}$ はベルヌーイ分布に従うとする．変数 $\omega_{k,i}$ を標準正規分布から生成し，そのもとでのパラメータ $C_{k,i}$ 及び $\mu_{k,i}$ に関する勾配を計算し，これらを更新する．潜在変数の事後分布は，式 **??**を $q(z_{si} = k)$ で微分し，2 次までのテイラー展開で近似することで，停留点を求める．発現量は，$\eta_{k,i}$ の事後分布の中央値 $\hat{\eta}_{k,i}$ を用いて，$10^6 \sum_k q(z_{s,i} = k) \exp(m_i + \hat{\eta}_{k,i}$ と推定する．

3　シミュレーション

提案モデルの検証を行うため，統制群と実験群がそれぞれ 20 サンプルある状況を想定して疑似的に作成した RNA-Seq データの解析を行った．遺伝子の総数を 10,000 とし，そのうち 500 遺伝子は実験群で発現量が増加し，500 遺伝子は発現量が減少する．残りの 9,000 遺伝子は二群間で発現量が変わらないとする．カウントデータを負の二項分布で生成した．

推定手法により，真の発現量を推定できることを確認した．また，$\eta_{k,i}$ の事後分布の中央値のトピック間の差を見ることで 3 種類の遺伝子を明確に区別でき，二群間の発現量の変動を精度よく推定できることが示された．

4　おわりに

過分散のある RNA-Seq データを扱うための新しいトピックモデルを提案し，シミュレーションによって推定精度の検証を行った．総リード数が 10^6 程度のデータから推定した発現量は，総リード数 10^8 で定量した発現量よりも真の発現量に近く，1 サンプル当たりの総リード数を増やすよりも，総リード数を減らしてサンプルサイズを大きくしたうえで発現量を推定することの有用性を示している．

さらに，提案モデルはサンプルや遺伝子を特徴づけるうえでも有用である．シミュレーションでは，どのサンプルが統制群でどのサンプルが実験群であるかという情報を全く使わなかったにもかかわらず，サンプルを正しく二群に分割し，発現の変動する遺伝子群を検出できた．今回のシミュレーションの二群比較のように統制のとれた実験のサンプルだけではなく，構成が未知のサンプルから定量した RNA-Seq データに対しても，提案モデルが有効である可能性を示唆している．

参考文献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] D. A. DuVerle, S. Yotsukura, S. Nomura, H. Aburatani, and K. Tsuda. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics*, 17(1):363, 2016.

[3] J. Eisenstein, A. Ahmed, and E. P. E. Xing. Sparse additive generative models of text. *Proc. Int. Conf. Mach. Learn.*, pages 1041–1048, 2011.

[4] M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference, 2014.

[5] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.

認知症の超早期発見のための血中マイクロ RNA の共発現解析

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
国立長寿医療研究センター　新飯田俊平

1. はじめに

健康な状態と認知症(アルツハイマー病を含む)の中間に位置する軽度認知障害患者の 50%以上は 5 年以内に認知症へと進行すると言われている。しかし、軽度認知障害を血液検査などの簡便・低侵襲な検査によって早期に発見し、適切に介入できれば、認知症の発病を遅らせ、健康寿命を延ばすことが出来ると期待されている。

一方、バイオマーカー検出の方法は、t 検定の適用のように分子を個別に扱う方法が伝統的であるが（Gene set enrichment analysis、パスウェイ解析なども個別に検出した分子を後でまとめていることが多い)、実際の分子挙動としては、それぞれの分子は相互に作用して働いていると考えられる。分子の相互作用を考慮したバイオマーカーを検出出来れば、それらは個別の分子に着目して検出したバイオマーカーよりも高精度であることが期待出来る。

具体的なバイオマーカーとして、遺伝子、タンパク質、代謝産物など、どの階層の分子に着目するのがベストか？認知症関連の血中バイオマーカーとしてタンパク質がいくつか報告されている一方で、安定した血中バイオマーカーの候補として、microRNA（miRNA）が注目を集めている。miRNA は、配列長が短く、この RNA 自身はタンパク質を作らないが（non-coding RNA)、他の遺伝子の発現を抑制的に制御する働きを担っている。

2. 方法[1]

多層的疾患オミックスプロジェクト[2]に提供された、健常者 30 人、軽度認知障害 23 人（年齢の幅をマッチさせている)の血漿サンプルを用いて 745 個の miRNA の発現量を計測した。欠測や低発現（全体の 20%未満）が多い miRNA を除き、85 個の miRNA を解析対象に、共発現の差分解析（differential correlation analysis）によりバイオマーカーの検出を試みた。具体的には、85 個の miRNA の全ての組み合せについて、健常群、患者群のそれぞれで Spearman の順位相関係数の値（r1、r2）を計算し、群間の差|r1-r2|が 0.8 を超える miRNA の組をバイオマーカー候補とした。それらの miRNA の各組と、miRNA の複数の組によって軽度認知障害を区別出来るかを判定するために、交互作用項ありのロジスティック回帰を適用した。miRNA の複数の組による判別の場合は、対応する交互作用項を足し合わせた。また、ロジスティック回帰によって推定された判別確率を用いて ROC 解析を行い、AUC 値を求めた。

3. 結果[1]

共発現の差分解析により 20 組の miRNA が軽度認知障害のバイオマーカーとして検出された。特に、２組の miRNA（[miR-191 と miR-101], [miR-103 と miR-222]）を用いて高精度（AUC=0.962）で軽度認知障害を判別可能であった。この他にも、miR-191 と miR-125b を含んだ miRNA の組（２組)も高精度なバイオマーカーとなる可能性があった(AUC≧0.95)。これらの精度は伝統的な t 検定で検出された miRNA より高精度であった。これらの結果等

から、miR191 や miR-125b に関わる相関の崩壊や出現は、軽度認知障害の発症や認知症への進行に重要な役割を担っている可能性が示唆された。

参考文献・資料：

[1] Kayano M., Higaki S., Sato J., Matsumoto K., Matsubara E., Takikawa O., Niida S., Plasma microRNA biomarkers for mild cognitive impairment using differential correlation analysis, Biomarker Research 4:22, 2016（data: GSE90828 in GEO）.
[2] 多層的疾患オミックスプロジェクト：　https://gemdbj.ncc.go.jp/omics/index.html

双方向グラフ上の最短経路を利用した遺伝関連解析

植木 優夫[1]，川崎能典[2]，田宮 元[1,3]
[1] 理化学研究所 革新知能統合研究センター，[2] 統計数理研究所 モデリング研究系，
[3] 東北大学 東北メディカル・メガバンク機構

1. はじめに

ゲノムワイド関連研究 (GWAS) では, 一塩基多型 (SNP) ひとつづつの周辺効果を一変量回帰モデルによって調べる方法が標準的である. これまでの GWAS では，独立な領域に単一の遺伝要因が存在するという仮定の下で検出力を見積もり，研究がデザインされてきた [1, 2]. これは各 SNP を独立に扱うものであるが，実際には SNP 間には連鎖不平衡 (LD) による相関が観察される. もし仮に LD 関係にある複数の遺伝要因が存在すれば，現行の一変量回帰モデルでは検出が困難となる [3]. このような場合，一変量回帰モデルの代わりに重回帰モデルを用いれば，SNP 間の交絡を調整した上で検出を行うことができ，検出力を高められる可能性がある. しかしながら，SNP 数は 100 万個程度もあるため，全 SNP を重回帰モデルに投入することはモデルの高次元化を招き実現不可能である. 一方で, サブセット回帰は候補のサブセットモデルの総数が膨大であるため，やはり計算困難となる. 本研究では,SNP 間に存在する相関構造のために弱い周辺効果しか示さない隠れた遺伝要因を検出するための新たな手法を提案する. 提案法は, 弱い周辺効果を示す SNP の近傍にある SNP によって双方向グラフを作る. 頂点は各 SNP に対応し, 頂点間の隣接性は LD の度合いによって決定する. その後, 各グラフ上で最短経路を集め, 各経路上の SNP を用いて重回帰モデルをあてはめ, 末端の SNP の有意性を検定する. ここで, 経路数とグラフ数の多重性を考慮した保守的な多重検定補正を適用する.

2. 双方向グラフ上の最短経路を利用した遺伝関連解析

y を血圧や血糖値などの連続な応答変数とする. p 個の SNP があるとして, j 番目の SNP を x_j で表す. SNP 間の交絡を調整した上で SNPj の効果を調べるため, 以下の帰無仮説の検定を考える.

$$\forall A \subset \{1, \ldots, p\} \setminus \{j\} : \beta_{j,A} = 0, y = \beta_{0,A} + \beta_{j,A} x_j + \beta_A x_A + \epsilon_A.$$

ここで p は 100 万程度であるため, 検討すべき A の総数は膨大となる. 例えば, $|A| = 2$ であれば, $(p-1)(p-2)/2$ 通り, $|A| = 3$ であれば, $(p-1)(p-2)(p-3)/6$ 通り, などとなる. したがって, まずは計算実行可能性が大きな問題となり, さらには検定の多重性も深刻化するため, 効果の検出にも多大な困難が生じる. 本研究では, 双方向グラフを利用し, 検定する仮説数を大幅に削減可能な新たな検定手法を提案する. 利用するのは, 重回帰係数推定量の以下の表現である. まず, $X_A = (x_k : k \in A)$ と定義する. [4] または [5] によると, SNPj に対応する重回帰係数推定量は以下のように表される.

$$\hat{\beta}_{j,A} = ||Q_{X_A} x_j||^{-2} y^T (Q_{X_A} x_j).$$

ここで $Q_X = I - P_X$, P_X は射影行列である. $y = \mu + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, $\mu = E(y|X)$ としたとき, この期待値と分散はそれぞれ $E(\hat{\beta}_{j,A}|X) = ||Q_{X_A} x_j||^{-2} \mu^T (Q_{X_A} x_j)$ および $\mathrm{var}(\hat{\beta}_{j,A}|X) = \sigma^2 ||Q_{X_A} x_j||^{-2}$ で与えられる. すなわち, $E(\hat{\beta}_{j,A}|X)$ が 0 であるかどうかを評価するためには, $\mu^T (Q_{X_A} x_j)$ に着目すればよい. μ を頂点 0, p 個の SNP を頂点とし, 辺の有無が頂点間の周辺相関係数が非ゼロであるかどうかで定義される双方向グラフを考える [6, 7]. 以下の結果が成り立つ.

命題 1. もし頂点 0 と頂点 j の間に, それ以外の頂点集合のいずれを通る経路もなければ, 頂点 0 または頂点 j でない任意の頂点集合 A について, $\mu^T(Q_{X_A}x_j) = 0$.

これより, μ へ経路のない SNP を無視することができる. それでは, μ への経路がある頂点 j とその中間にある頂点集合 A について, $\mu^T(Q_{X_A}x_j) \neq 0$ となるかというと必ずしもそうではない. 以下の命題が成り立つ.

命題 2. 頂点 0 と頂点 j の間に少なくとも一つの経路があり, そのうち最短のものの中間の頂点が $\{j_1, \ldots, j_l\} \subset \{1, \ldots, p\} \setminus \{j\}$ であれば, $\mu^T(Q_{X_{\{j_1,\ldots,j_l\}}}x_j) \neq 0$.

この命題から, μ への最短経路の末端の SNP は, 非ゼロの回帰係数を与えることがわかる. ただし, 上記の命題はノイズの無い理想的な状況を考えているため，遺伝的効果の有無の評価には仮説検定を用い，多重性を考慮して偽陽性を制御する. 命題 1 と 2 の証明, ならびに GWAS データで動作するアルゴリズムの詳細については，[8] を参照されたい.

参考文献

[1] Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* **273**:1516–7.

[2] de Bakker PI, Yelensky R, Pe ̓er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* **37**:1217–23.

[3] Eberle MA, Ng PC, Kuhn K, Zhou L, Peiffer DA, Galver L, Viaud-Martinez KA, Lawley CT, Gunderson KL, Shen R, Murray SS. 2007. Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet* **3**:e170.

[4] Rao CR, Yanai H. 1979. General definition and decomposition of projectors and some applications to statistical problems. *J Stat Plan Infer* **3**:1–17.

[5] Ueki M, Kawasaki Y. 2013. Multiple choice from competing regression models under multicollinearity based on standardized update. *Comput Stat & Data Anal* **63**:31–41.

[6] Cox DR, Wermuth N. 1996. *Mutivariante Dependencies*. London: Chapman and Hall.

[7] Drton M, Perlman M. 2007. Multiple testing and error control in Gaussian graphical model selection. *Stat Sci* **22**:430–49.

[8] Ueki M, Kawasaki Y, Tamiya G. 2017. Detecting genetic association through shortest paths in a bidirected graph. *Genet Epidemiol* **41**:481–97.

精密医療・予防に向けた分子バイオマーカーの探索：階層層混合モデルを用いた最適発見手法の応用

大谷 隆浩
統計数理研究所

特定の疾患に対する治療の効果や、環境暴露による発症リスクの変化を正確に予測するための分子バイオマーカーの開発は、個々人の特性に合わせた最適な精密医療・予防を実現する上で、最も重要な課題の一つである。この実現に向けて、特定の疾患において治療・環境暴露との交互作用を示す遺伝子を網羅的に探索することを目的とした、ゲノムワイド関連解析 (genome-wide association study; GWAS) が数多く行われている (Thomas, 2010)。しかしながら、現状でのスタンダードとなっている、回帰モデルに基づいて遺伝子と治療・環境暴露の交互作用を検出する検定手法の検出力は極めて低く、これまでに特定されている有用なバイオマーカーは限定的である。この問題を解決するために、Matsui らは多次元セミパラメトリック階層混合モデルを用いた最適発見手法 (optimal discovery procedure; ODP) による効率的な検定手法を開発し、現状の手法を上回る、理論上優れた性能を持つことを示した (Matsui et al., 2017)。この手法では、各遺伝子が疾患に及ぼす効果サイズの分布を、データに基づいて経験ベイズ法により推定する。そして、推定された分布から求められる各遺伝子の疾患関連確率をもとに、確率の高い遺伝子を分子バイオマーカーの候補として検出する。

本研究では、この手法を GWAS のデータ解析に応用し、遺伝型・表現型データベース dbGaP (https://www.ncbi.nlm.nih.gov/gap) で公開されている 2 つの大規模 GWAS を事例として取り挙げ、これらのデータを解析することで、これまでに特定されていない疾患関連 SNP (single nucleotide polymorphism) の探索を試みた。まず、治療の効果を予測するマーカーとなりうる SNP の探索を目的として、脳梗塞を対象とした大規模ランダム化臨床試験「Vitamin Intervention Stroke Prevention (VISP) trial」(dbGaP アクセス番号 : phs000343.v3.p1) における GWAS データの解析を行った。さらに、環境暴露による疾患発症リスクの変化を予測するマーカーの探索を目的として、肺がん発症リスクの遺伝的決定要因を調べることを目的に行われたケース・コントロール GWAS「A Genome Wide Scan of Lung Cancer and Smoking」(dbGaP アクセス番号：phs000093.v2.p2) のデータを解析した。

VISP trial は、マルチビタミン剤（葉酸、ビタミン B6、B12）の高用量日常摂取が、脳梗塞の再発や非致死性心筋梗塞発症の減少につながるかどうか検証することを目的とした、多施設二重盲検ランダム化比較試験である。治療群ではマルチビタミン剤を高用量日常摂取し、対照群では低用量を日常摂取した。対象者数は 2164 名であり、遺伝型決定は Illumina HumanOmni1-Quad_v1-0_B アレイを用いて行われた。品質管理を行った後の対象者数は 1533 名 (治療群：760 名、対照群：773 名)、SNP 数は 774,670 である。本研究では、Wakefield らによって行われた解析 (Wakefield et al., 2014) と同様に、ベー

スライン時と初回観察時での血中ホモシステイン濃度の差をアウトカムとした解析を行った (血中ホモシステイン濃度は循環器疾患と関連する)。

最適発見手法を用いた解析の結果、Wakefield らの解析によって既に検出されていた 2 つの血中ホモシステイン濃度に関連する SNP に加え、3 つの関連 SNP を新たに検出した (FDR<5%)。そのうちの 2 つは、血中ホモシステイン濃度に対する効果サイズが、治療群と対照群で大きく異なることが示唆されるものであった。これらの新たに検出された SNP について、外部データベースを用いた遺伝子機能解析を行ったところ、強い連鎖不平衡にある SNP において、ホモシステインが代謝される肝臓での遺伝子発現や、末梢血単球での遺伝子発現との関連が見受けられた。さらなる検証が必要なものの、これらの結果は検出された SNP の生物学的な妥当性を示唆している。

次に、肺がんを対象とした GWAS データを解析することで、喫煙習慣による肺がん発症リスクの変化を予測するマーカーの探索を試みた。この GWAS における対象者数は 5588 名であり、遺伝型決定は Illumina HumanHap550v3_B アレイを用いて行われた。品質管理を行った後の対象者数は 5527 名 (ケース：2716 名、コントロール：2811 名)、SNP 数は 519,431 である。各対象者について喫煙経験の情報が付与されており、本研究では 6 ヶ月以上の定期的な喫煙を行なった経験がある場合は喫煙者、経験がない場合は非喫煙者とした。非喫煙者は 931 名 (ケース：220 名、コントロール：711 名)、喫煙者は 4596 名 (ケース：2496 名、コントロール：2100 名) である。

解析の結果、連鎖不平衡にない 6 つの独立な疾患関連 SNP を検出し (FDR<5%)、そのうちの 1 つは、喫煙習慣の有無によって肺がんの発症リスクに大きな差があることが示唆されるものであった。遺伝子機能解析によると、これとは異なる 1 つの SNP については肺がん発症リスク・喫煙習慣との関連が示されているものの、他の SNP には生物学的な強いエビデンスがなく、さらなる検証研究が必要である。

謝辞

本研究の遂行にあたり、JST CREST (課題番号：JPMJCR1412)、および、国立研究開発法人日本医療研究開発機構 (AMED) の革新的がん医療実用化研究事業 (課題番号：17ck0106266) の支援を受けました。

参考文献

Matsui, S., Noma, H., Qu, P., Sakai, Y., Matsui, K., Heuck, C., et al. (2017) Multi-subgroup gene screening using semi-parametric hierarchical mixture models and the optimal discovery procedure: application to a randomized clinical trial in multiple myeloma. *Biometrics*, DOI: 10.1111/biom.12716.

Thomas, D. (2010) Gene-environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, **11**, 259–72.

Wakefield, J., Skrivankova, V., Hsu, F.-C., Sale, M. and Heagerty, P. (2014) Detecting signals in pharmacogenomic genome-wide association studies. *The Pharmacogenomics Journal*, **14**, 309–15.

多変量メタアナリシスにおける高次漸近理論を用いた推測手法

統計数理研究所　野間 久史

メタアナリシスとは，過去に行われた複数の臨床試験のエビデンスを統合し，総合的な治療効果の評価を行うための方法論であり，Evidence-Based Medicine における重要な研究手法として広く普及している．近年の医療統計学の方法論の進展により、多変量アウトカムを扱った多変量メタアナリシス（multivariate meta-analysis）の研究が飛躍的に発展し、医学研究の実践にも大きな影響を与えている。

その代表的な手法のひとつが、ネットワークメタアナリシス（network meta-analysis）である。従来のメタアナリシスの方法は，原則として，対象となる治療の対比較（1 対 1 の比較）の結果を統合するという単純化されたものであり，「有効性についてのエビデンスが確立された複数の治療の選択肢の中から，どの治療を行うのが最善なのか？（有効性，安全性は？費用対効果は？）」などの問いに必ずしも本質的な答えを与えてくれるものではなかった．近年，先進諸国における（超）高齢化社会の到来により，医療費・医療資源の節減および効率的な配分のために，比較効用研究（comparative effectiveness research）が世界的に大きな関心を集めており，このような複数の治療を対象とした有効性・有用性を総合的に評価するための方法論に対する要請が飛躍的に高まっている．ネットワークメタアナリシス（network meta-analysis）は，このような要請に応えるために開発された研究手法であり，従来のメタアナリシスの方法を一般化し，複数の治療の有効性・有用性を比較・評価することを可能にした方法論である．ネットワークメタアナリシスでは，複数の対象となる治療を含む臨床試験の結果を系統的に集め統合し，治療間の間接比較の情報も併せたエビデンスの統合が行われる．これにより，直接比較の行われていない治療間の比較も含めて，対象となった治療法すべての有効性・有用性を比較することが可能となる．

しかしながら、多変量メタアナリシスは、この数年で急速に普及した新規な方法論でもあり、現在、実践で普及しているスタンダードな方法でも、さまざまな統計学的問題が存在する可能性がある。その問題のひとつとして、Brockwell and Gordon (2001), Noma (2011) などで議論されている推測手法の妥当性の問題がある。メタアナリシスでは、異なる情報源から得られるエビデンスの異質性を考慮した上で情報の統合を行うために、変量効果モデル（random effects model）を用いた解析が行われるのが一般的である。多変量メタアナリシスでは、以下の多変量変量効果モデルが広く用いられている。

$$\boldsymbol{Y}_i \sim MVN(\boldsymbol{\theta}_i, \boldsymbol{S}_i), \boldsymbol{\theta}_i \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad (*)$$

$$\boldsymbol{S}_i = \begin{pmatrix} s_{i1}^2 & \rho_{i12}s_{i1}s_{i2} & \cdots & \rho_{i1p}s_{i1}s_{ip} \\ \rho_{i12}s_{i1}s_{i2} & s_{i2}^2 & \cdots & \rho_{i2p}s_{i2}s_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{i1p}s_{i1}s_{ip} & \rho_{i2p}s_{i2}s_{ip} & \cdots & s_{ip}^2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \tau_1^2 & \kappa_{12}\tau_1\tau_2 & \cdots & \kappa_{1p}\tau_1\tau_p \\ \kappa_{12}\tau_1\tau_2 & \tau_2^2 & \cdots & \kappa_{2p}\tau_2\tau_p \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{1p}\tau_1\tau_p & \kappa_{2p}\tau_2\tau_p & \cdots & \tau_p^2 \end{pmatrix}$$

$\boldsymbol{Y}_i$は、i 番目の試験における、特定の参照レベル（プラセボなど）に対しての対象となる治療の治療効果の指標（ハザード比，オッズ比など）の推定量からなる確率ベクトルであり、$\boldsymbol{\theta}_i$はその真値を表すベクトル、$\boldsymbol{S}_i$は試験内共分散行列（$\boldsymbol{Y}_i$の共分散行列）、$\boldsymbol{\Sigma}$ は試験間共分散行列を表す。パラメータの推定は、制限付き最尤法（restricted maximum likelihood; REML）によって行われるのが一般的である。通常、試験間の異質性を表す $\boldsymbol{\Sigma}$ に関する情報は、$\boldsymbol{Y}_i$ のデータ数に比例することとなり、試験数が十分に大きいもとで、大標本理論による正当化が可能となる。しかしながら、多くのメタアナリシスでは、統合の対象となる試験数が十分に大きくはなく、古典的な 2 つの治療法の対比較のメタアナリシスでは、20 にも満たない試験での統合が行われるのが一般的である。この場合、大標本近似の崩れから、信頼区間の被覆率が名目水準を大きく下回るなどの問題があることが知られている（Noma, 2011）。

本講演では、シミュレーション実験を通して、まず、現状のスタンダードな推測手法である REML 法に、上記のような条件下で、推測の妥当性が成り立たない状況が多く存在することを示す。そして、演者らが近年開発した、多変量メタアナリシスの解析モデルに汎用的に用いることができる高次漸近理論に基づく有効な推測手法（Noma *et al*., 2017）の紹介を行う。また、シミュレーション実験および統合失調症などのネットワークメタアナリシスの事例解析を通して、その実践的有用性について示す。

文献

Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med.* 2001;20(6):825-840.

Noma H. Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Stat Med.* 2011;30(28):3304-3312.

Noma H, Nagashima, K, Maruo K, Gosho, M., Furukawa TA. Bartlett-type corrections and bootstrap adjustments of likelihood-based inference methods for network meta-analysis. *Stat Med* 2018; doi: 10.1002/sim.7578.