日本統計学会創立75周年記念出版

21世紀の統計科学

国友直人・山本拓 監修

< Vol. II >

自然・生物・健康の統計科学

小西貞則[1]・国友直人[2] 編集

2008年8月 (東京大学出版会)
2012年1月 (増補HP版)

[1]中央大学理工学部教授 (元九州大学数理学府教授)
[2]東京大学経済学部教授

増補HP版・はしがき

シリーズ「21 世紀の統計科学」Vol.I, Vol.II, Vol.III は 2008 年に東京大学出版会より商業出版された。その後、統計学・統計科学に関係するこの種の書籍としては順調に販売が伸び 2011 年半ばにいたり在庫部数が少なくなってきた。

この書籍は 2008 年版の前書き・後書きに説明があるように通常の商業出版物とは異なり、日本統計学会の創立７５周年を契機に、できるだけ多くの人々に統計学・統計科学の最近の動向を紹介することにある。そこでこれを機会に各原稿を可能な範囲で改訂し、更に 2012 年増補版として学会ＨＰより無償でダウンロードする形で広く利用して頂くことにした。

もとより本書・2012 年ＨＰ増補版の各著者は原稿料は要求せず無償で原稿を提供しているわけである。そこで本書の編者・監修者としては各読者にはなるべく本書及び本書の論文を引用等で正確に引用して頂くことを期待したい。

2012 年 1 月

編者・監修者

はしがき

21 世紀を迎えて既に 8 年目を迎えている今日、社会における統計学・統計科学の役割は以前にも増して重要になっている。現代では、社会・経済・経営などの社会科学はもちろん、工学（品質管理、環境、情報、計算機など)、理学（数学・地球物理・天文・化学など)、生命科学（生物・医学・薬学・農学など)、人文科学（教育、心理、言語、文学など）をはじめ、あらゆる学問領域において必要不可欠な基礎的手段として、統計学・統計科学が鍵となる役割を演じている。また、個人・企業・政府、などが直面する様々な問題に適切に対処し、科学的に意志決定を行う多くの状況では、統計データや統計的方法による科学的分析が本質的な役割を演じている。本シリーズは「21 世紀の統計科学」と題して、統計学・統計科学における最新の動向を踏まえつつ、現代の日本社会における重要な幾つかの側面にしぼって統計学・統計科学の現状を鳥瞰するとともに、統計学・統計科学を専攻する第一線の研究者が挑戦している課題について成果をまとめている。全 3 巻の中で、この第 2 巻では特に自然・工学・生物・健康などに関連する諸問題に対する統計学・統計科学のアプローチによる科学的分析をまとめている。

ここで本書の内容を簡単に説明しておこう。現代では自然科学や工学の対象であるわれわれを取り巻く自然現象や日常生活で対処すべき問題では様々な不確実性に直面している。例えば地震・災害など地球環境をめぐる問題をとっても多くの国民の将来の生活に直結しているが、いつ、何処で発生するかを事前に知ることは容易ではない。統計学・統計科学による分析は日常生活では目に見えにくいこともあり、必ずしも多くの人がその発展や成果を理解しているわけではないかもしれない。しかしながら、最近経験した地震・災害などを例に挙げるだけでも、われわれの生活基盤となる問題として、避けて通ることができない重要な問題であり、実はこうした国民生活の「安全や安心」に関わる諸分野における研究活動は統計学・統計科学が真に支えているのである。第 1 章の尾形論文は地震活動の統計解析を長年行っている研究の内容と成果を説明している。とかく話題になりやすい地震の予測をめぐる科学的議論の参考になると思われる。尾形論文に続く第 2 章の北川論文では地震の性質を利用した工学的解析を通じて、統計学的方法による「予測と発見」を目指している、わが国が誇るべき統計数理研究所 (http://www.ism.ac.jp/index.html) における研究活動と成果、さらに今後の方向を紹介している。　北川氏が所長を務めている同研究所では今では New York の金融街 (Wall Street) でもっとも著名な日本人である Ito（確率）解析の伊藤清氏、赤池情報量基準 (AIC) の考案者として知られている赤池弘次氏などをはじめ、これまで多くの世界的な逸材を世に送り出している。本シリーズに収録されている諸論文の中でも伊藤氏の業績は確率過程の統計的解析やファイナンスへの応用（例えば第　巻に収録の林・吉田論文や第　巻の内田論文など）において、また赤池氏の業績は情報量統計学に関わる幾つかの応用的研究（例えば本巻に収録の尾形論文・北川論文・岸野論文など）の中に脈々と受け継がれ、発展を続けていることを多くの読者が理解されるであろう。　次に工学における統計的方法としては、生産現場における品質管理 (quality control) や信頼性理論を挙げることができよう。第 3 章の鎌倉論文はこうした分析には不可欠な統計的解析方法である生存時間分析 (survival analysis) の導入と応用を扱っている。地震・河川の氾濫・金融疲労、といった問題をたどっていくと標準的な初等統計学で説明されていない、「希にしか起きない」事象の科学的分析の必要性にたどり着く。第 4 章の渋谷・高橋論文は統計的極値理論 (extreme value theory) とその応用に関する優れた展望

論文である。多次元極値理論の開拓者の解説により、このテーマも今や一見すると全く別の分野と思われがちな経済・ファイナンスにおける金融リスク管理問題の展開にさえ統計科学の視点から新しい視点が開かれることが期待される（なお鎌倉論文が扱っているテーマは近年ではファイナンス（金融）における企業倒産と信用リスク分析にも有用である)。　統計学の初等的教科書をひもとくと生物学や薬効をめぐる話題が少なくない。実際、統計学の基礎的な概念である t-分布の発見や回帰 (regression) と言った統計的方法は２０世紀初頭の英国における実際的な問題を解決する途中で開発されたことはよく知られている。統計学の成立の時代から今日に至るまで、統計学・統計科学を巡る展開では常に中心的役割を果たしてるこの分野は今日では生物統計・医学統計、あるいは健康科学などとも呼ばれている。生物・医学・薬学などに関連する様々な問題に対して、現代では多くの統計学的・統計科学的発展が見られるが、本巻第　部では３つの大きな問題に絞って取り上げた。第５章の岸野論文は近年の分子進化論の展開を踏まえ、統計的方法がなぜ DNA 解析といった生物をめぐる最先端の研究において有効なのかを説明している。第６章の井元論文は生物が関わる問題についてより新しい統計的方法であるベイズ的分析法の有効性を議論している。また、第７章の吉村論文ではわれわれが日常的に恩恵（時には害も）を受けている薬効をめぐる近年の展開や議論を背景に、統計科学と健康科学の発展の相互依存関係の重要性を説明している。　自然・工学・生物をめぐる様々な問題を解析している中からは、この間、多様な新しい統計学・統計科学の理論が生まれ、進化を続けているばかりでなく、環境から画像処理にいたるまで新しい工学的な応用上の問題を解決する為の統計学的研究が展開している。第８章の矢島論文では近年になり盛んに応用されるようになっている空間統計学 (spatial statistics)・時空間統計学の理論と応用を説明している。近年では工学系の統計科学として機械学習理論 (learning theory)、データマイニング (data-maining)、と呼ばれている新しい分野がめざましく展開している。第９章の福水論文と第１０章の鷲尾論文はこの展開を踏まえ、それぞれ学習理論とデータマイニング法についての最新の議論を説明している。さらに近年では DNA 解析をはじめとする生物学・医学分野における発展を受けて、新たな統計学的問題が発生し、その研究が発展している。第１１章の栗木論文ではこうした新しい生物統計学の理論的深化を議論している。

本書の各章では現代における「自然・工学・生物・健康の統計学・統計科学」をめぐる最新の議論がなるべく多くの人々に理解されるよう、学術的な研究論文というよりもかなり分かりやすい形で統計学・統計科学における最新の動向を理解できるように説明している。さらに本巻に収録している諸論文は本シリーズの他の巻である「社会・経済の統計科学」（第　巻）および「数理・計算の統計科学」（第　巻）が扱っている諸論文とも密接に関連する論考も少なくない。本書の内容が、統計学・統計科学の理論と応用、特に自然科学・工学・農学・生物・薬学・医学などでの統計データの分析に関心がある多くの学生、大学院生、研究者、官庁関係者、実業界の関係者にとって、問題把握や問題の解決へのヒントを与える材料になれば幸いである。

2008 年 5 月

編者

あとがき

本書は 2006 年 5 月と同１ 2 月に日本統計学会創立 75 周年を記念して開催された研究集会 (東京大学と中央大学にてそれぞれ開催) における招待講演・報告の扱いをめぐる議論を一つのきっかけとして企画された。全体としてバランス良い 75 周年記念の書籍としてより意味を持たせるため、編集委員より学会において評価の高い研究者に原稿の書き下ろしをも依頼し、全体を編集しこのたび出版する運びとなった。各章の担当者は元原稿に対するコメントを参考にしてさらに加筆、修正を加えて最終原稿を作成した。

2006 年度は日本における統計学・統計科学では最古で最大の学術団体である「日本統計学会 (Japan Statistical Society,http://www.jss.gr.jp/ja/)」の創立 75 周年に当たっている。草創期における日本での統計学・統計科学の展開や学会周辺の事情については、例えば『日本の統計学 50 年』(1982 年、日本統計学会編、東京大学出版会) により少し知ることができる。創立 75 周年という節目を迎え、日本統計学会の当時の会長（山本拓・一橋大学経済学部教授)、理事長（竹村彰通・東京大学工学部教授）を中心に 75 周年記念事業委員会（委員長：杉山高一・中央大学理工学部教授）が組織され、様々な記念事業が企画され、実施された。この記念事業では数多くの招待講演、研究報告が行われ、「一般社会からはかなり専門的」と評されるかもしれないが、日本統計学会自らが発行している英文・和文の専門学術誌上での学術的資料にとどまるにはあまりにも重要な内容が多かった。そこで 特別に７５周年記念事業委員会の中に編集委員会が組織され、シリーズ「21 世紀の統計科学」の出版が企画されることになった。本書はそうした日本統計学会の７５周年記念出版事業での成果刊行物である。むろん、専門的な研究報告や統計学会・会員向けの展望報告などは日本統計学会・和文誌上に７５周年特集 $I \cdot II \cdot III$ として 2007 年度後半より順次、掲載され始めている。

なお、本企画は意識的に網羅的であることを避け、今回集まった編集委員の判断で内容を厳選し、原稿作成の期限を設けたことをお断りしておく。日本統計学会には今回の全３巻では取り上げることができなかった統計学・統計科学の研究分野や研究・教育などで活躍していると多忙で優れた研究者も少なくない。今回取り上げることの出来なかった多くの研究分野や重要な課題における統計学・統計科学からの貢献・発展・今後の展開についての議論は、別の機会に譲ることとしたい。

ここで第 II 巻「自然・生物・健康の統計科学」と第 III 巻「数理・計算の統計科学」の編集にあたっては、寄稿された諸論文の編集に際して評価者・助言者として、（敬称略）大森裕浩、大屋幸輔、川崎能典、間瀬茂、下平英寿、宮田敏、小林景、田中研太郎、二宮嘉行、清水邦夫、清智也、南美穂子、汪金芳、樋口知之 の各先生方のご協力を得た。これら 諸先生方のご協力に特に感謝したい。また、本シリーズ「21 世紀の統計科学」の出版にあたって、東京大学出版会の黒田拓也氏のご協力にも感謝したい。この第 II 巻が第 I 巻と第 III 巻とともに多くの方々にとっての座右の書となることを期待したい。

2008 年 5 月

監修者

目次

第 1 章
地震活動の統計科学

尾形良彦[1] (情報・システム研究機構 統計数理研究所 教授)

要約

統計モデルによってデータから本質を露出する。これは，望遠鏡や顕微鏡のように, 辛うじて見えるものや見えないものを，見えるようにする。筆者は，地震統計による経験法則や地震学の仮説を，統計的点過程モデルで表現して, 統計的方法の有用性を示すように心がけてきた。地震活動研究と密接に関係する統計地震学の最近の展開と, それらの地震予測における意義をお伝えできればと思う。

目次

[1] ogata@ism.ac.jp

1 はじめに

体に感じないものを入れれば, 地震は日常不断に起きており, 膨大なデータ量となって日々蓄積されている。地殻や地球内部は直接見ることができないので, これらは, 固体地球科学の研究や大地震を予測する重要な手がかりである。データが豊富なところには, 目的に即した統計モデルが必要である。統計モデルには, 実効性のある将来の予測を実現することや, 科学的新知見を導くような役割を求められている。

データが膨大であればあるほど, その隠れた情報を十分汲み取るために，非定常または非一様な複雑モデルを考慮する必要があり, そのために大規模な統計モデルが避けられない様になってきた。逆問題や時空間モデルなどの大量のパラメタを必要とする大規模モデルの開発は, ベイズ法の助けを必要とし, 最適化法などの計算技術の開発に関わり, 推定結果を表示する多次元の動画像解析法などを駆使し, 情報学や数値解析などとの境界分野に及ぶ研究の比重も高くなってきた。

本稿では, 日本などの地震国で, 一世紀近くの長きにわたって蓄積されている, 地震カタログのデータを考える。地震カタログは発生時刻 (破壊の始まった時刻), 震源座標 (破壊の始まった位置), 大きさ (マグニチュード), 発震メカニズム (断層面の向きとすべりの方向) などを編集した膨大なデータである。地震活動は, このデータに基づいて研究されている。しかし, 長きにわたる地震カタログは, データとして均質でなく, 観測の制約に基づく様々な弱点を抱えている。データに含まれている自然界の本質的な情報を抽出するために, 様々な最尤法のモデルとベイズモデルやそれらの解析法を紹介する。これらは, 官庁統計，統計調査，疫学, リスク解析や信頼性理論をはじめとする, 他分野での統計的解析法やモデリングの展開にも通じるところがあると信じている。

2 地震の震源の位置補正 – バイアスの推定

震源の位置は, 地球内部を伝播する地震波の速度のモデルを仮定して, 決められている。地震波は, 震源から地震計まで辿り着く道のりを, 同じ速

度で進んでいない。地球内部の物性は不均質で, 波の伝播速度に関する充分正確な地球の内部モデルが必要である (たとえば宇津, 2001, 4 章参照) 。その上で, 震源の位置と発生時刻は, 地震波の各地の地震計への到着時刻の, 非線形最小二乗法で決められている (たとえば宇津, 2001, 6 章参照) 。たとえば, 各地点 i で観測された地震波の到達時刻が t_i であるとき, 震源位置と発震時刻は, P 波などの地下伝播速度構造のもと, 最短の道のりで予測される到達時刻と実際の到達時刻の残差の二乗和

$$\sum_{i=1}^{N}\{t_i - t_0 - T(\Delta_i, h)\}^2 \tag{2.1}$$

を最小にする様に決められる。ここで, i は地震計を置く観測点で, その数 N は, 解を求めるためには 4 点以上必要であるが, 一般に多いのが望ましく, 地震が大きければ地震波は多くの観測点で捉えられる。さらに, t_0 は断層がすべり始めた発震時刻, Δ_i は震央 (震源を地表に投影した地点) までの距離, h は震源の深さである。関数 $T(\Delta_i, h)$ は予め与えられた，地下モデルによって予測される走時 (travel time) である。これは, 波が伝わってくる経路に関しての, 伝播速度の線積分として与えられるものである。地下における地震波の伝播速度は, 水平方向に比べて, 深さの方向の違いによる変化が大きいので, 第一近似として, 深さ h に関してのみ変化している関数で表現される地球内部モデルが考えられる。このモデルは, 地震の破壊開始の位置や時刻を求めるために, 採用されている。

しかし, 深さだけに依存する (2.1) のモデルの精度は充分でないので, 正確な位置の決定には出来るだけ密に地震計を配置する必要がある。これは陸域では十分可能である。しかし海域では, 海底に地震計を置くため, 電源などの困難があり, 観測期間も極めて限られている。そのため海域の地震は, 陸域の地震計からの片側だけからの決定を強いられている。陸域でも，震源の位置のうちで, 深さの推定の精度が低い。これは, 地震計の配置を 3 次元的に見た時, ほぼ平面 (地表) にあることが原因である。しかし, 不確定性には, 偏差 (バイアス) と不偏な誤差 (ばらつき) がある。特にバイアスは, 推定に使ったモデルの不完全性にあるわけだ

から, これを除去できれば, 震源の位置の補正がある程度可能になる。

長野県にある, 気象庁の松代精密地震観測室は, 松代群列地震計システム (Matsushiro Seismic Array System,MSAS) と称する, 半径 5 ｋｍ程度の円周上と中心付近に 7 つの地震計を設置している。これによって世界中の地震の震央 (緯度, 経度) を推定し, 核実験などの速やかな探査に寄与している。 MSAS の震源の決め方は上記のものと違う。まず，各観測点の P 波到着時刻から方位角と MSAS 内での P 波の見かけの速度を最小 2 乗法 ($N = 7$) で求める。これらのデータと松代の S 波と P 波の到達時刻の差 ($S - P$ 時間) を用いて, 震源距離は IASPEI91 (Kennet, 1991) の走時表に基づいて決める。しかし, その推定位置の偏差は多様で, 誤差は小さくない。図 1a で, MSAS だけで決められた震央と米国地質調査所 (United States Geological Survey, USGS) の世界的地震観測ネットワークで決められた震央の違いを示す。 USGS ののものに比べると大分ばらついている。

しかし，驚く事に, それは MSAS における地震波の読み取り誤差によるものより, 地球内部における地震波の伝播経路と速度の違いによる, 系統的な偏差の方が圧倒的に大きいのである。従って, その空間的な偏りが推定できれば, MSAS で決められた位置から偏差を取り除くことによって, 震央の補正が出来る。

問題の性質上, MSAS で決められた地震 $\{i;\ i = 1, 2, \cdots, N\}$ の震央位置について, MSAS を原点とする極座標 (Δ_i, Ψ_i) を考える。そして, その地震の真の震央位置は (Δ_i^0, Ψ_i^0) であるとする。実際には, 真の位置としては USGS が決めたものが比較的正確なので, これを使う。それぞれのバイアスの成分 f と g は MSAS が決めた震央の位置に依存する関数とし, 補正の線形モデル

$$\begin{aligned} \Delta_i^0 &= \Delta_i + f(\Delta_i, \Psi_i) + \varepsilon_i \\ \Psi_i^0 &= \Psi_i + g(\Delta_i, \Psi_i) + \eta_i \end{aligned} \tag{2.2}$$

を考える。ここで, ε_i と η_i は, それぞれの座標成分の不偏 (平均値 0) な誤差で, 独立な正規分布を仮定する。最小二乗法によって, 真の位置

(Δ_i^0, Ψ_i^0) とのバイアスを予測する変換写像 $\varphi = (f, g)$ を求める。そのため, 変換写像の成分の極座標上の関数 f と g をBスプライン基底関数[1] $B_\cdot(\cdot)$ で双一次,

$$f_{\boldsymbol{\vartheta}}(\Delta, \Psi) = \sum_{i,j} \theta_{i,j} B_i(\Delta) B_j(\Psi); \quad g_{\boldsymbol{\vartheta}'}(\Delta, \Psi) = \sum_{i,j} {\theta'}_{i,j} B_i(\Delta) B_j(\Psi)$$

に展開し, その係数行列 $\boldsymbol{\vartheta} = (\theta_{i,j})$ や $\boldsymbol{\vartheta}' = ({\theta'}_{i,j})$ を推定する。

まず, 距離のバイアスに関するBスプライン関数 (曲面) $f_{\boldsymbol{\vartheta}}$ を求める。実用的には, Bスプライン基底 $B_\cdot(\cdot)$ を定義する区間の両端を定める節点を十分細かく取る必要がある。しかし, 上記の場合, 係数の数は節点数の2乗個以上になるので, (2.2) 式の線形モデルの最小2乗法では, 現実的な解が得られない。そこで, Bスプライン関数 (曲面) $f_{\boldsymbol{\vartheta}}$ の微分係数が全体的になるべく小さく (滑らかに) なるように, 次のような曲面の凸凹 (変動量) に関する2種類の制約 (ペナルティ)

$$\begin{aligned}
\Phi_1(f_{\boldsymbol{\vartheta}}) &= \int\int_A \left\{ \left(\frac{\partial f_{\boldsymbol{\vartheta}}}{\partial \Delta} \right)^2 + \left(\frac{\partial f_{\boldsymbol{\vartheta}}}{\partial \Psi} \right)^2 \right\} d\Delta d\Psi, \\
\Phi_2(f_{\boldsymbol{\vartheta}}) &= \int\int_A \left\{ \left(\frac{\partial^2 f_{\boldsymbol{\vartheta}}}{\partial \Delta^2} \right)^2 + 2 \left(\frac{\partial^2 f_{\boldsymbol{\vartheta}}}{\partial \Delta \partial \Psi} \right)^2 + \left(\frac{\partial^2 f_{\boldsymbol{\vartheta}}}{\partial \Delta^2} \right)^2 \right\} d\Delta d\Psi
\end{aligned}$$

を入れる。ここで, Bスプライン基底は3次多項式なので積分は計算され, ペナルティ関数はパラメタに関する2次形式になる。そして, ペナルティ付きの最小2乗法

$$(2.3)\ Q(\boldsymbol{\vartheta}; w_1, w_2) = \sum_{p=1}^{N} \left\{ \Delta_p^0 - f_{\boldsymbol{\vartheta}}(\Delta_p, \Psi_p) \right\}^2 + w_1 \Phi_1(f_{\boldsymbol{\vartheta}}) + w_2 \Phi_2(f_{\boldsymbol{\vartheta}})$$

によって, 例えば修正コレスキー法 (森, 1987, 参照) によって, 最小解を求めることができる。

しかし, 制約の強さの塩梅をコントロールする重み w_1 と w_2 をどう決めるかという重要な問題が残る。そのためには, 先ず, 最小2乗法が多変

[1] 節点で等間隔に分割された区間上で定義された, 特定の 4 組の 3 次多項式。例えば Ogata & Katsura, 1993, 付録

量の正規分布モデルの最尤法に他ならないということに注目する。(2.3) 式の第一項のデータを含む二乗和は，多変量正規分布の尤度関数 (以下で $L(\boldsymbol{\vartheta})$ と置く) の負の対数である。ペナルティの項はBスプライン関数の係数に関する多変量正規分布の負の対数で，これが事前分布に対応する。ペナルティ付き二乗和全体の (2.3) 式が事後分布の負の対数である。このようにベイズの公式に対応している。事後分布を確率分布とするために割り算する正規化因子

$$\Lambda(w_1, w_2, \theta_M) = \int_{R^{M-1}} L(\boldsymbol{\vartheta}) \cdot prior(\boldsymbol{\vartheta}^r \mid w_1, w_2, \theta_M)\, \boldsymbol{d\theta}^r \tag{2.4}$$

はベイズモデルの尤度と呼ばれ，重み w_1, w_2 と θ_M の関数である。$\boldsymbol{\rho} = (w_1, w_2, \theta_M)$ はパラメタ (スプライン関数の係数) 間の制約を調節するパラメタの役割を果たすので，**超パラメタ**とよばれる。ただし, $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}^r, \theta_M)$ で θ_M は末尾のBスプライン係数であり, θ_M を超パラメタとすることで, 残りのパラメタ $\boldsymbol{\vartheta}^r$ の事前分布を確率分布 (proper prior) として規格化できることが大事である。以下に述べるように, 異なるベイズモデルの比較をするために必要なことである。さもなければ (2.3) 式のペナルティの項に対応する分散・共分散行列は特異となり，多変量正規分布は $\boldsymbol{\vartheta}$ に関しての確率分布にならない (improper prior) 。

ベイズモデルの尤度の値が大きければ大きいほど予測力の優れた適合度の良いモデルと言える (以下参照) 。すなわち, この値を最大化する超パラメタ $\boldsymbol{\rho} = (w_1, w_2, \theta_M)$ が最適のものである。超パラメタに関するベイズモデルの尤度の最大化は, たとえば, シンプレックス法 (Kowalik & Osbone, 1968, 参照) で検索する。

Bスプライン節点を細かくとることで, 係数の数が2乗のオーダーで増えるが, ベイズモデルの尤度もある所まで増えていく。今回のデータでは Δ と Ψ の, それぞれの定義域 $[0°, 180°] \times [0°, 360°]$ 上の節点で分けた区間による矩形領域の数を 6×6 から始めて 2 倍 2 倍と増やして計算していくと, 96×96 分割でその尤度が頭打ちになったので, そこで止めたものである。

双一次Bスプライン関数以外の 2 次元関数表示 (たとえば本稿 14 節

のデロネ関数参照) とも適合性を比較することが可能である。赤池は制約の強さなどを示す重み (超パラメタ $\boldsymbol{\rho}$) の数も考慮した, ベイズモデル選択のための赤池ベイズ情報量基準 (ABIC)

$$ABIC = (-2)\max_{\boldsymbol{\rho}}\{\log \Lambda(\boldsymbol{\rho})\} + 2 \times \dim(\boldsymbol{\rho}) \tag{2.5}$$

を提案している。ベイズ尤度も, この規準も, 統計的予測の向上を目指すエントロピー最大化原理から導かれ, 事後分布の良さ悪さを予測の観点から比較するものである。詳しくは, 赤池弘次論文選集 (Parzen et al. ed., 1998) を参照されたい。また, この節の詳細は (Ogata et al., 1998) を参照されたい。

同様にして, 角度のバイアス関数 $g_{\boldsymbol{\vartheta}'}$ の推定もできる。求められた関数の曲面が図 1c と d で表示されている。これらによって, 新たな MSAS 震央データのバイアスを補正したものが図 1b に示されており，実際と遜色がない。この補正関数は, 松代精密地震観測室の従来の補正法に替えて, 実際の業務に採用されている。

これまで述べたモデルや方法を使って, 次のような提案が可能である。日本は東北沖や南海・東南海沖などのように, 沖合のプレート沈み込み帯での地震活動が非常に高い。しかし, 海底に地震計を常置することが困難であるため, 陸域のみの地震計で決められた沖合の震源 (特に深さ) の精度は低い。これは震源カタログの弱点であり, 地震活動研究の差し支えとなっている。原理的な対策としては, 精密な地震波伝播速度の 3 次元的速度構造モデルを求めることであるが, 全ての海域地震帯で，それを逆問題として求めるためのデータが少ない。そこで, 海底地震計の導入で得られた精度の高い震源データを拠りどころに, 陸上のみで決められた震源のバイアスを 3 次元的に補正するために, (2.2) 式と同様な 3 次元変換モデルを構築することが考えられる。

3 地震の大きさの分布の地域性 – 非定常・非一様性のモデリング

地震の大きさは, 通常, マグニチュードで与えられている。これは原理的

には, 各地震計で計測された地震波の振幅によって, 震源からの距離を勘案して推定される。それは, 星や電球の光度を, 見かけの明るさと光源までの距離で決めているのと同様である。マグニチュードは地震エネルギーの対数と線形関係にある。発生する地震の数 N はマグニチュード M が小さくなるとともに指数的に増えることが知られおり, これは Gutenberg-Richter の法則[2]と呼ばれ,

$$\log_{10} N = a - bM$$

と表される。この式は, マグニチュード M_0 以上の地震のみを考えた時, マグニチュードの確率密度分布が指数分布

$$f(M) = \beta \exp\{-\beta(M - M_0)\}, \quad M \geq M_0, \quad \beta = b \ln 10 \tag{3.6}$$

に従うことを示している。ここで, 指数の係数は b 値と呼ばれ, 地域的に異なった値を取り, 物理量として研究されている。岩石破壊の研究によって, これが地殻内部の地質, 温度やストレスの高さなどに関係しているからである。定められた地域内での平均的な b 値は, 観測されたマグニチュードのデータ $\{M_1, \cdots, M_N\}$ から, 最尤推定値 $\hat{\beta}$ で求められる。実際, 上記の指数分布 (3.6) から, $\hat{\beta}$ はマグニチュード平均値の逆数になる。

大抵の地震カタログでは, 地震の震源要素のなかでマグニチュードだけが, 「M6.8 中越沖地震」などと発表されるように, 2 桁の有効数字 (少数点 1 桁) でしか与えられていない。これは, マグニチュード値の推定誤差がその程度に大きいためであるが, マグニチュードをめぐる統計的解析にとっては不幸なことである。かつて, JMA の旧カタログでは, 深さを精度上の理由から 10km 単位 (時期によっては 20km) の有効数字で与えられたことがあるが, このために, 沈み込む太平洋プレートに沿った地震の 2 重深発面の発見が遅れたと考えられている。

たとえば, 使用している地震計や推定方式の変遷のため, 時間が経つと, マグニチュードの推定バイアス (マグニチュード・シフト) が生じる

[2]例えば, 宇津, 1999, 11 章；宇津, 2001, 5.4 節を参照

ことがある。これを無視すれば地震活動の解析にとって致命的である。同じ地震に関する, 他のカタログのマグニチュードとの差を, 本稿 2 節で述べたようにモデル化して, バイアスの時間変化などを議論することが可能である (尾形, 1998) 。しかし, このような統計的解析のためには, 大きな誤差分布を承知の上で, マグニチュードの推定値として, もう一桁以上まで求めたい。事実, 地震モーメント (宇津, 2001, 5.2.G 節参照) は有効数字 3 桁まで与えてあるので, これを変換すれば 3 桁までのマグニチュードが得られる。有効数字 2 桁のマグニチュードでは, 例えば, b 値の最尤推定値によって偏り無く推定するためにも, 一定の補正を考慮しなければならない (宇津, 2001, 5.4.B 節参照) 。

さて, マグニチュードに関して均質なデータを考えよう。地域的空間変化や時間変化を考えるためには, ベイズモデルが有効である。たとえば, Gutenberg-Richter 法則の係数 b 値に対応する指数分布 (3.6) の係数が, $\beta = \beta(x, y, z)$ の様に, 3 次元関数であると考え, その対数関数 $\phi(x, y, z) = \log \beta(x, y, z)$ を B スプライン基底によって 3 次元展開することとし,

ペナルティ付き対数尤度 = 対数尤度関数 − ペナルティ関数

を最大化することを考えるのである。ここで対数尤度は, モデルが正規分布であれば, 2 節 (2.3) 式で見たように二乗和であるが, 今回は指数分布の対数の和である。

前節のように, モデルの当てはまり具合を測り, ペナルティ関数はパラメタ間の制約の強さ (滑らかさ) を測る。ここで, 滑らかさに関する制約 (ペナルティ) 関数は

$$\begin{aligned}
\Phi_1(\phi|w_1, w_2) &= \int\int\int_A w_1 \left\{ \left(\frac{\partial\phi}{\partial x}\right)^2 + \left(\frac{\partial\phi}{\partial y}\right)^2 \right\} + w_2 \left(\frac{\partial\phi}{\partial z}\right)^2 dxdydz \\
\Phi_2(\phi|w_3, w_4, w_5) &= \int\int\int_A w_3 \left\{ \left(\frac{\partial^2\phi}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2\phi}{\partial x\partial y}\right)^2 + \left(\frac{\partial^2\phi}{\partial y^2}\right)^2 \right\} \\
&\quad + w_4 \left\{ \left(\frac{\partial^2\phi}{\partial x\partial z}\right)^2 + \left(\frac{\partial^2\phi}{\partial y\partial z}\right)^2 \right\} + w_5 \left(\frac{\partial^2\phi}{\partial z^2}\right)^2 dxdydx
\end{aligned}$$

である。そして, (3.6) 式より, ペナルティ付き対数尤度

$$Q(\theta|w_1,\cdots,w_5) = \sum_{i=1}^{n}\log\left\{\beta(x_i,y_i,z_i)e^{-\beta(x_i,y_i,z_i)(M_i-M_0)}\right\} - \{\Phi_1(\phi|w_1,w_2)+\Phi_2(\phi|w_3,w_4,w_5)\}$$

を考え，これを最大にするようなBスプラインの係数を求める。数値的には Davidon-Fletcher-Powell 法 (Kowalik & Osbone, 1968, 参照) に修正コレスキー法 (森, 1987 参照) を組み合わせたものによって, 次元がいくら大きくても, 効率的に最大化することができる。

このとき, 制約に関する重み $w_1,\cdots,w_5$ を, ABIC を最小化するように, 自動的に定めることができる。つまり, 制約関数がパラメタに関して二次形式になっていることから, 十分な重みのもと, 最大値を与えるパラメタの近傍で Taylor 展開すると, ペナルティ付き対数尤度が二次形式で良く近似できる。言い換えると, 事前分布が多変量正規分布であることから, 事後分布を多変量正規分布で近似 (Laplace 近似) することによって, (2.4) 式と同様のベイズ尤度が計算される。かくして, 例えばシンプレックス法 (Kowalik& Osbone, 1968, 参照) で, 制約関数の重み (ベイズモデルの超パラメタ; $w_1,\cdots,w_5$) を自動的に探索することが出来る。

ここで, 地殻上部マントルの深さ方向と水平面方向では, 同じ強さの制約で良いか否かを判定したい。これは, 制約条件の空間的等方性と非等方性に関する, ベイズモデルの事前分布の選択問題である。後者に対しては w_1, w_2, w_3, w_4, w_5 の 5 組の独立な重みに関してベイズ尤度の最大値を求め, 前者は制約つきで 2 組の重み $w_1 = w_2$ かつ $w_3 = w_4 = w_5$ に関し最大値をもとめ, 超パラメタの数が二つ違うベイズモデルの ABIC 値の比較を行なう (前述 2.5 式参照) 。

防災科学技術研究所による関東地方直下 100km 深までの微小地震データに, このベイズモデルを当てはめたところ, 深さ方向に変化が顕著な, 非等方的な制約が選ばれ, 関東直下のプレートに直交方向に大きく変化する b 値の空間変化が求められた (Ogata, Imoto & Katsura, 1991) 。

大規模なベイズモデルのパラメタ間の制約は平滑化のために正規事

前分布を想定することが多いが, 上記のように, 時空間の異なる座標軸の重み付けの判断が必要である。また, 余震のマグニチュード系列のように，不等間隔な時刻で採取された系列データでも, データの豊富な区間と疎らな区間では平滑化の重みが違ってくる。例えば，本震の経過からの時間の対数を取って，できるだけ一様な採取 (変換) 時間のもとで平滑化関数を考え，それをもとの時間の関数に戻した場合がはるかに良い ABIC 値となる (Ogata, 1989) 。

このように, 比較されるべきベイズモデルの多様性は多く, それらは, 尤度関数や事前分布に反映し, その適合度を比べ, 選択する必要がある。こういったベイズモデルの良さを比較するのに ABIC が有要である。

4 地震の検出率の時・空間の変化 – データの欠測のモデリング

発生した地震が全て検知されてはいない。その検知率は, 地震そのものの大きさだけでなく, 地震計の感度や震源からの距離にも関係している。したがって, 地震カタログに掲載されている地震の数やマグニチュードの頻度分布は年代によって変遷し, 地域によっても違うように, データとして不均質なものである。データの有効利用には, それがどの様に不均質なのか, つぶさに定量化して表現することが求められる。これに十分応えられるのはベイズモデルである。その様なモデル化の考え方は, 無回答や欠測などの避けられない, 生物医学統計や各種統計調査などのデータ解析にも必要である (たとえば Ogata, Katsura, Keiding, Holst & Green, 2000) 。

気象庁編集による震源カタログ (1923~ 現在) の内容を見るとすぐ分かるのは, 年代が進むに連れて, 収録されている地震の数が急速に多くなっていることである。これは, 地震計の性能が上がり, 観測点の数が増えたために, 小さい地震が検出されて増えたからである。また, 地域的にも, 地震の多い所とそうでない所がある。これも地震活動の高さ低さのためとは限らない。地震計のネットワークが密で良く整備されている地域 (たとえば内陸部) とそうでない所 (たとえば沖合の地域) では, 地震の

検出率が大きく異なるからである。しかし, 以下の統計モデルを考えると, ベイズ法に基づき, 各時期各地域の地震検出率と真の地震活動 (b 値など) が一挙に推定できる。

先ず, 観測できていないものも含め, 実際に発生している地震のマグニチュード分布が, 図 2a および 2a' で示されているような, 指数分布であると考える (Gutenberg-Richter の法則) 。その中で, 検出される地震の割合はマグニチュード M によって高くも低くもなる。このような検出率関数として正規分布の累積分布関数 (図 2b)

$$q(M|\mu,\sigma) = \int_{-\infty}^{M} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{(2\sigma^2)}\}dx$$

を考える。ここで, 正規分布の平均値 μ は 50% の検出率が見込まれる地震のマグニチュードで, 標準偏差 σ は部分的に検出されているマグニチュードの範囲の狭さを示す。そうすると, 検出された地震の確率密度関数は

$$f(M|b,\mu,\sigma) = \frac{10^{-bM}q(M|\mu,\sigma)}{\int_{-\infty}^{\infty} 10^{-bM}q(M|\mu,\sigma)dM}$$

で与えられる。このようにして得られた分布 (図 2c および c') は, 日本の気象庁カタログや世界の Harvard 大学 CMT カタログ, 国際地震センター (ISC) や USGS のカタログなど, 多くの地震カタログ (宇津, 1999, 5.1 節参照) で経験的に得られるものと大変良く似ている。

マグニチュードの系列データ $\{M_1, \cdots, M_N\}$ が与えられているとする。仮にパラメタ b, μ, σ が定数であるなら対数尤度関数は

$$\log L(b,\mu,\sigma) = \sum_{i=1}^{N} \log f(M_i|b,\mu,\sigma) \tag{4.7}$$

となり, これを最大化することで, 最尤推定値を求めることができる。対数尤度関数の最大化には, たとえば Davidon-Fletcher-Powell 法 (Kowalik & Osbone, 1968, 参照) を採用する。さらに, 推定誤差は最尤推定値での対数尤度の 2 階微分行列 (Hessian matrix) から得られる

しかし, 地震カタログの長期間にわたるデータを見れば, これらのパラメタが定数であり得えないことは明白である。先ず, これらのパラメタが時間 t について変化していることを考える。つまり, 関数 $b_{\theta_1}(t)$, $\mu_{\theta_2}(t)$ および $\sigma_{\theta_3}(t)$ がB スプライン基底によって, 線形に

$$\begin{aligned}
\varphi_1(t) = \log\beta(t;\theta_1) &= \sum_{k=1}^{K} a_k B_k(t) \\
\varphi_2(t) = \quad \mu(t;\theta_2) &= \sum_{k=1}^{K} b_k B_k(t) \\
\varphi_3(t) = \log\sigma(t;\theta_3) &= \sum_{k=1}^{K} c_k B_k(t).
\end{aligned}$$

と展開されている。ここで, Bスプライン関数の係数

$$\theta = (\theta_1;\theta_2;\theta_3) = (a_1, ..., a_K; b_1, ..., b_K; c_1, ..., c_K).$$

の対数尤度は

$$\log L(\theta) = \sum_{i=1}^{N} \log f(M_i | b_{\theta_1}(t_i), \mu_{\theta_2}(t_i), \sigma_{\theta_3}(t_i))$$

と書ける。関数 $b_{\theta_1}(t), \mu_{\theta_2}(t)$ および $\sigma_{\theta_3}(t)$ に関して各々滑らかさの制約 (ペナルティ)

$$\Phi_j(\phi | w_{2j-1}, w_{2j}) = \int w_{2j-1} \left(\frac{d\phi_{2j-1}}{dt}\right)^2 + w_{2j} \left(\frac{d^2\phi_{2j}}{dt^2}\right)^2 dt, \quad j = 1, 2, 3,$$

を付けることによって, ペナルティ付き対数尤度を考えて, 6 つの重み $(w_1, \cdots, w_6)$ を (2.4) 式と同様なベイズモデルの尤度で決め, 最適な解を得ることができる (Ogata & Katsura, 1993) 。数値的なアルゴリズムは前節と同様である。

検出率関数のパラメタ $\mu(t)$ と $\sigma(t)$ が求められると, $p\%$ の検出率を期待できるマグニチュードの関数は $d_p(t) = \mu(t) + \gamma_p\sigma(t)$ で与えられる。ここに γ_p は, 標準正規分布の, いわゆる有意水準 p パーセントの点

(percentile) である。さらに, b 値の変動 $b(t)$ が求まれば, 任意の時間区間とマグニチュード帯 $(t, t+\Delta t) \times (M, M+\Delta M)$ での検出できなかったものを含めた, 真の地震発生率の推定値を与えることが可能である。たとえば, 本稿 7 節では, 余震の確率予測にたいする応用が述べられる。

同様な事が空間についても考えることができる。たとえば, 2 節や 3 節で考えたような B スプライン関数に対するペナルティ関数を考えれば良い (Ogata & Katsura, 1993) 。ここで, 図 3 では, 2001 年 10 月における地震活動の b 値分布と気象庁カタログの 50% 検出率のマグニチュードを, デロネ 3 角分割 (Delaunay triangulation; 14 節の図 7b 参照) の上の三角面で張られた部分的線形関数 (デロネ三角面関数) で表したものを使った。これには一回微分 (三角面の傾き) のみのペナルティ制約 (14 節参照) を課した。デロネ分割の関数表現は, 広域の割に標本点が非一様な場合のデータに効果的な推定が可能である。地震分布の密な所は詳しい変動で, 疎なところは大雑把にという具合である。図 3 から, 前述したように, 内陸部から沖合にいくに従って, 検知率が悪くなっている状況が見える。

時間と供に変化する検出率関数が分かれば, 各マグニチュードの検出率を推定でき, 同時に, 下限マグニチュードに拘ることなく, b 値の時間変動も推定できる。さらに, 検出された全ての地震データを使って, 欠測地震を含む全ての地震の地震発生率の推定値を, 任意のマグニチュード M に対して与えることが可能である。これは, 不均質なデータから地震活動をみるためのデータの有効利用である。たとえば, 13 節と 14 節で述べる時空間 ETAS モデルを不均質なデータに当てはめる場合にモデル化することができる。

5 余震発生頻度の減衰

大きい地震が起きたとき, その後には余震が続く。それがすぐ終息してくれれば良いのであるが, 実際は減衰がゆっくりで延々と長引く。人々は, これが更なる大きな地震の前触れでないかと, 不安に駆られる。その可能性を報道関係者や住民に問われたときに, 地震関係の識者が「余震

活動が順調に推移しているので・・・」と言ったコメントをよく聞く。では一体, 余震活動の順調な推移とは, どのように定義されるものであろうか。

日本では大地震が頻発するため, 余震のデータは古くから豊富である。 19 世紀の大森房吉の研究以来の余震活動の経験法則によると, 余震の単位時間当りの発生頻度は, 本震直後からの推移時間 t について,

$$\nu(t|\,\theta) = \frac{K}{(t+c)^p}, \quad \theta = (K, c, p) \tag{5.8}$$

の様に逆べキに従って減衰する (例えば Utsu et al., 1995, 宇津, 2001, 7.1.B 節参照) 。これは宇津徳治の研究で大きく発展したもので, 宇津自身は改良大森公式と呼んだが, 我々は**大森・宇津の公式**と呼んでいる。減衰の定数 p は通常 1.0 以上の様々な値を取り, 地殻やその深部の地質や温度などの特性を反映していると考えられている。しばらく前まで, 定数 p や c は, 宇津が 1960 年代に始めたように, 余震の単位時間当たりの発生個数 $\nu_\theta(t)$ と本震後の経過時間 t の関係の両対数グラフ上のプロットの漸近直線の傾きなどから推定されてきた。いまでこそフラクタル次元の推定などで両対数プロットは良く使われているが、当時は全く創意的な方法であった。これがなくては余震活動の詳細な研究は進まなかっただろう。

6 非定常ポアソン過程

これに対して, 地震の頻度を数えるのでなく, 発生時刻そのままをデータとして使う点過程の最尤法が，地震活動の研究に，新しい展開をもたらした。余震活動を非定常ポアソン過程と考え, 最尤法によって少数のデータでも推定ができ, 推定誤差も与える。また, データの数に見合って, 2 次余震などの詳しい活動の有無が, 赤池情報量規準 AIC(10 節参照) によって, 議論できるようになった。

本震後の時間区間 (S, T) で発生した余震の時刻データ $\{t_i;\ i = 1, 2, \cdots, N\}$ が, (5.8) 式の大森・宇津の公式の強度関数 $\nu(t)$ に従う非定常ポアソン過程に従っていると仮定する。ここで, 本震直後の区間

$(0, S)$ の余震のデータを当てはめない理由は, 本震直後の余震の欠測問題[3]にある。この場合の対数尤度関数は

$$\log L(\theta) = \sum_{i=1}^{N} \log \nu(t_i|\,\theta) - \int_S^T \nu(t|\,\theta)dt, \quad \theta = (K, c, p) \tag{6.9}$$

といった形で与えられる (Daley & Vere-Jones, 2003, 7 節参照) 。最尤法は, パラメタ $\theta = (K, c, p)$ に関して, これを数値的に最大化するのである。最大値を与えるパラメタ値を最尤推定値と言う。対数尤度関数の最大化には, たとえば Davidon-Fletcher-Powell 法を採用する (Kowalik & Osbone, 1968, 参照) 。さらに, 推定誤差は最尤推定値での対数尤度の 2 階微分行列 (Hessian matrix) または Fisher の情報行列から得られる (Ogata, 1999, 参照) 。

7 余震の確率予報

本震発生後しばらくは, 余震が頻発し大きな余震が多く, その予測情報は 2 次災害を防ぐために緊急性が高い。大森・宇津の減衰公式 (5.8) と地震の大きさの発生頻度分布 (3.6 式: Gutenberg-Richter の法則) を掛け合わせた, マグニチュードと発生時刻の強度関数

$$\lambda(t, M) = m(M)n(t) = \frac{10^{a+b(M_0-M)}}{(t+c)^p} \quad (a, b, c, p; \text{パラメタ}) \tag{7.10}$$

を求めて, 余震の数や大きな余震の確率予測ができる。これは California (Reasenberg & Jones, 1989) で始まったが, 日本でも地震調査委員会 (1998) の推定方法の報告に基づいて, 被害を起こすような地震にたいして, 本震後ほぼ 1 日経ってから, 気象庁によって公表されている。これらのパラメタは，比較的大きな余震 $\{(t_i, M_i);\ M_i \geq M_0\}$ から， $\theta = (a, b, c, p)$ の対数尤度関数

$$\log L(\theta) = \sum_{i=1}^{N} \log \lambda(t_i, M_i|\,\theta) - \int_{M_0}^{\infty}\int_S^T \lambda(t, M|\,\theta)dtdM, \tag{7.11}$$

[3] この理由については 7 節で述べる。

を最大化 (Davidon-Fletcher-Powell 法) して最尤法で求められる[4]。

しかし本震直後, 特に 1 日以内, の余震の予報は次のような困難を抱えている。本震直後には, 余震の発生率は極めて高いが, 検出率は極めて低い。なぜなら, 直後には本震の地震波が激しく, その後も連発する余震の地震波が重なり合っているために, どの波形記録がどの余震のものか識別困難である。このため, 発生時刻などを求められないので欠測となる。しかも, 余震のマグニチュードが小さくなればなるほど, 欠測率が甚だしくなる。このことが本震直後の余震発生確率の算定を困難にしている。

この対策として, (7.10) 式の代わりに 4 節で議論した検出率モデルによって求められたマグニチュード分布を使って

$$\lambda(t, M) = \frac{q\{M|\, \hat{b}(t), \hat{\mu}(t), \hat{\sigma}(t)\}}{(t+c)^p} \quad (a, b, c, p;\ \text{パラメタ}) \tag{7.12}$$

で予測することが考えられる。たとえば, 2003 年宮城県沖の地震 (M7.0) の一年にわたる余震列[5]について, その b 値と検知率関数のパラメタの時間的変化を推定した (図 4) 。推定結果を見ると b 値と σ 値は殆ど変化せず, 50% 検出されているマグニチュード (μ 値) は, 本震後 2 週間から 20 日かけて下がり M0.5 となり, その後は一定に推移している。従って, この余震データは本震後 20 日以降一年にわたり, 各マグニチュードにつきそれぞれ一定の比率で検出され, 図 3c および c' に示す同一分布によったものである事を示している。

そこで, 本震直後の確率予測のために, 次のような少数パラメタのモデルを考える (Ogata & Katsura, 2006) 。すなわち, (7.12) 式において, 図 4 で得られた結果に基づいて, b 値と σ 値を定数として, $\mu(t)$ 値は 4 つのパラメタで特徴付けられた関数

$$\mu(t) = \mu(\tau) = \mu_0 + \mu_1 \exp\{-\alpha\tau^\gamma\} \quad (\mu_0, \mu_1, \alpha, \gamma\ \text{パラメタ})$$

[4]実際は, この対数尤度関数を分解して, (7.10) における関数 $m(M)$ と $n(t)$ を別々独立な対数尤度で求められている。詳しくは地震調査委員会の推定方法の報告参照; http://www.jishin.go.jp/main/yoshin2/yoshin2.htm

[5]ただし, この場合のデータは, 地震カタログの編集が完結した最良版であり, 確率予報の実践についてはデータ編集上の問題が残っている。

(ただし τ は変換時間で $\tau = 3 + \log t$) を考える。これらを対数尤度関数 (7.11) に代入し，$\theta = (b, \mu_0, \mu_1, \alpha, \gamma, \sigma)$ の合計 6 つのパラメタを最尤法で求める。

これで, たとえば, 本震直後 30 分までのデータから推定して, その先 30 分の予測をするというように, 経過時間とともに, 逐次, 推定と予測を繰り返すが, マグニチュードの予測分布は実際の余震頻度と良く合う。

8 ETAS モデル (Epidemic Type Aftershock Sequence model)

余震活動を小さいものまで細かく見ると, 余震の余震 (二次余震) などが顕著に見えはじめ, 非定常ポアソン過程に従っている様な, 単純な余震系列は実際には少ないことが分かる。一般の地震活動になると, 前震, 群発地震などがあり, 特に地震活動が活発な地域では, 本震と思われていたものが, 他の地震に誘発されていたり, 他地域の活動と関係があったりして, 本震と余震の区別をつけることが困難になることが多い。

しかし, これらの複雑な地震活動も, 一つ一つの地震に対する大森宇津関数の重ね合わせによって良く表現されることが分かってきた。点 (地震) の発生履歴や他の情報 (時系列や他の点過程など) に依存して変わるような発生率は**条件付強度関数**と呼ばれる。この確率論的定式については 18 節の付論を参照されたい。 Epidemic Type Aftershock Sequence (ETAS) モデルは, 条件付強度関数が

$$\lambda(t|\,\theta) = \mu + \sum_{\{j;\, t_j<t\}} \nu(t - t_j|\,K, c, p) e^{\alpha(M_j - M_c)}, \quad \theta = (\mu, K, c, \alpha, p) \tag{8.13}$$

によって表現[6]されるものである (Ogata, 1988; 宇津, 1999, 7.4.F 節; 宇津, 2001, 10.5 節参照) 。但し, μ は常時地震活動 (background seismicity) の活動度の高さを示す定数であり, $\nu(\cdot)$ は (5.8) 式の大森宇津関数, 総和 $\sum$ は時刻 t 以前に発生した全ての地震 j について, 本震と余震の区別を問わず, 取るものとする。ここで, t_j と M_j は j 番目の地震の発生時刻とマグ

[6]条件付強度関数 $\lambda(t|\,\theta)$ は, 本稿の付論にあるように, 履歴 H_t の条件が付くが, 簡単のため, 以下これを省略する。

ニチュード, M_c は考慮している地震データの下限のマグニチュード値である。 ETAS モデルは, 各地震の後にそのマグニチュードに見合った数の余震を伴うというモデルである。そして, ある地域全体での地震の発生強度は, 過去に発生した各地震後に減衰する余震の強度を重ね合わせたものに, その地域の常時活動の強度を足し合わせたものとなる。

ETAS モデルは, その領域に適合した地震活動を統計的に予測する。各地の地震活動の特徴は ETAS モデルの 5 個のパラメタ $\theta = (\mu, K, c, \alpha, p)$ によって量的に表現される。とくに, p, K と α が地震活動の特徴を測る量として重要である。 α 値は群れの大きさに対するマグニチュード差の効率性を示している。例えば, 本震と余震の違いがはっきりし余震の余震が顕著でないパタンは大きい α 値をとる。他方, 群発地震は小さい α 値になる。 K 値は 14 節で示すように, 同じ大きさの地震に対する平均的な余震数の違い (余震の生産性) を表す。 ETAS モデルのパラメタは (6.9) 式と形式的に同じ対数尤度で書くことができ, 最尤法で求めることが出来る。対数尤度関数の最大化には, たとえば Davidon-Fletcher-Powell 法 (Kowalik & Osbone, 1968, 参照) を採用する。

ETAS モデルは, 条件付強度関数で一般的に表現された, Self-exciting 点過程 (Hawkes, 1971) を起源とする。その Hawkes の点過程も Epidemic (伝染病) 分枝過程に遡る (Hawkes & Oakes,1974) 。 Hawkes の点過程モデルに起源を持つ同様な地震活動モデルが Lomnitz(1974) や Kagan & Knopoff (1987) によって紹介されているが，これらはモデルの記述やシミュレーションに留まっている。すでに, Hawkes & Adamopoulos (1973) はスペクトル尤度[7]によって地震発生データに対する当てはめや適合度の比較も行っている。 Hawkes の点過程モデル以前に, 別の意味で ETAS モデルに良く似ている, 先駆的な点過程 Trigger モデル (Vere-Jones & Davies) もスペクトル尤度で推定可能である。しかし, ETAS モデルは点過程の尤度法にもとづき, その予測力を発揮するのである。こ ETAS モ

[7]これまで紹介した点過程の尤度関数とは違い, 自己相関関数のフーリエ変換 (スペクトル）の推定量であるピリオドグラムが平均値の異なる指数分布になることに基づいた尤度（正規過程による近似, Whittle, 1962)

デルの起源について詳しくは尾形(1993) を参照されたい。

ETAS モデルの物理的な解釈は, 次の通りである。地震断層のすべりに伴う, 急激で局所的なストレス変化によって, 膨大な数の近傍の中小断層群の破壊やすべりが次々と誘発され, 余震の連鎖性・集中性を生じていると考えられている (Dieterich, 1994) 。この様な余震の発生メカニズムは複雑で観測できないため, 活動予測には統計的経験法則に基づくものが実用的である。つまり, 余震現象を統計的に記述する。 ETAS モデルの役割は, 統計力学において相互作用ポテンシャルで，衝突する運動粒子系の物性を記述するのに似ている (Helmstetter & Sornette, 2002, など) 。

9 点過程モデルによるデータ情報の診断 – 「残差」解析

以下に述べるように, 推定されたETAS モデルを「ものさし」 (基準) にして, モデルから説明できないような, 実際の地震活動の微妙な変化を見いだせる。

通常, モデルやデータの診断解析には, データ値から予測値を差し引いた残差を解析する。しかし, 地震活動や余震活動が予測どおりに推移しているか否かをグラフで見るために, 通常の時系列の残差と違ったものを考える。すなわち, 点過程の「残差」は, 次のような考えに基づき, 非線形的な時間の伸び縮みによって作られる。時刻 t までに発生する地震の期待個数 $\Lambda(t)$ は

$$\Lambda(t) = \int_0^t \lambda(s)ds \tag{9.14}$$

の如く条件付強度関数の積分となり, t に関する単調増加関数である。これを使って地震発生の時刻 $\{t_1, \cdots, t_N\}$ を変換した時刻 $\{\tau_i = \Lambda(t_i);\ i = 1, 2, \cdots, N\}$ を見ると, モデルで予測したものとデータの食い違いが見易くなる。これは次の様に時間が伸縮されたものである。条件付強度関数 (発生強度, 活動度) が高ければ, それに応じて伸び, 低ければ縮む様になっている。そして, もし地震活動が完全にETAS モデルにもとづいて推移しているならば, 通常の時間を伸縮した変換時刻に関する点過程

は定常ポアソン過程 (一様にランダムな系列) になることが証明される (Meyer, 1971; Papangelou, 1972) 。このとき, 変換時刻に関する累積関数は, 傾き 1.0 の直線上に乗る。つまり, $\Lambda(\tau) = \tau$ となる。

実際の地震系列データを, この様にして変換したものを**残差系列** と呼ぶ。これが, モデルが悪くないのに，定常ポアソン過程にならない場合がある。そのようなとき, これが一様なランダム系列と, どの様に違っているのかを詳細にみることによって, その地域の標準的な地震活動からのずれを調べるのである。一般に, データの生成過程を良く理解して, 科学的な意味のある異常変動を示す残差系列を引き出せれば, 当てはめたモデルに含まれて無い新しい情報をつかむことになる。たとえば，後述するように, ETAS モデルによって地震活動を計測し, その変化を検出することによって，広域の地震活動変化を促す未知のストレスの増減変化を見易くするのである。

10 変化点問題と赤池情報量基準

ETAS モデルは余震の発生率とその減衰を組み込みこんだモデルであり, これによる予測発生率と実際の地震の発生頻度の違いをみることによって, 異常現象を際だたせるねらいがあることは前節で述べた。しかし, 実際に, これを測ったり有意性を議論したりするには, まず, 地震活動が或る時点を境に, 前後の期間で変化しているか否かという問題を判断する必要がある。或る時点を境に地震活動に変化がある場合, 前後で異なったパラメタ値をとる 2 組の ETAS モデルが必要となる。他方, 変化が無く順調に推移していれば, 単一の ETAS モデルが良い当てはまりを示す。どちらが良いかの評価には, 後者の AIC の値と前者の AIC 値を比較する。しかし前者は, 異なった 2 つのモデルが必要なため, パラメタ数がその分増える。さらに，変化点自体のパラメタをどのように扱うかという問題がある。

改めて, 赤池の情報量規準 (AIC) は

$$AIC = (-2) \max_{\boldsymbol{\vartheta}} \{\log L(\boldsymbol{\vartheta})\} + 2 \cdot \dim(\boldsymbol{\vartheta})$$

で与えられ, 統計モデルを比べたとき, 小さな AIC 値をとるモデルの方が予測の観点から優れている (Parzen et al. ed., 1998) 。ここで, 上記第 2 項のペナルティ (パラメタ数の 2 倍) は考えているモデルの予測分布 (点過程の場合は条件付強度関数) がパラメタ値の変動に対して滑らかに変化するなどの性質[8]から導かれる。

しかし, 変化点の最尤推定値に対しては, 第 2 項のペナルティを修正した AIC を使う。このペナルティはモデルが正則条件を満たさない上, 解析的な計算の導出は難しく, モンテカルロ・シミュレーションによって求められた。それによると, ペナルティ値は, 4.0 付近から始まり, 地震のデータ数が多くなるに従って単調に増え, 数百個以上で 6 .0 前後の値になる (表や図については, たとえば, Ogata, 1999, 参照) 。従って, このような変化点を挟んだ ETAS モデルの有意性を示すためには, 単一の ETAS モデルより，対数尤度の差が 7 ～ 8 以上大きくなければならず, 変化点のあるモデルが選択されるためのハードルはかなり高い。ただし, 地震データとは別の理由から, 変化点の時刻が予め与えられている場合は, 変化点探索をする必要が無く, 通常の AIC 比較をすることができる。

11 地震活動の相対的静穏化

地震活動パタンの変化が確認された場合, 変化点を境に, 後半区間の地震発生累積数が前半の活動の ETAS モデルによる地震発生予測数 $\Lambda(t)$ に比べて少ない時, これを相対的静穏化現象と呼ぶ。これは予測されている発生率以下の活動が一定期間続いたからである。 (9.14) 式による変換時刻に関しては, 傾き 1 .0 の直線上に乗っている累積関数が, 変化点後, 直線から下方に離れていくことになる。このように, 予測と実際の地震発生頻度の相違を測ることが, 地震活動の変化を検出する物指となる。

たとえば, 2006 年 11 月 15 日の千島列島 (Kuril Islands) の巨大地震 (M8.3) の余震活動に ETAS モデルを当てはめた (図 5 参照；尾形 &遠田, 2007) とき, この余震活動が順調に推移しているか否かはマグニチュード

[8]統計的漸近理論モデルの正則性。たとえば ETAS モデルの場合は Ogata(1978) 参照

対時間図や累積関数図などの目視では明瞭でない。しかし, 前節の解析によって, ETAS モデルによる理論的累積数 (発生率の積分 $\Lambda(t)$) を物差しにして, 余震の実際の累積数との偏りを見ると, 本震後半月頃に有意な変化が認められ, その後, 2007 年 1 月 13 日の地震までの, 1 ヵ月間の余震発生が理論発生率に比べて有意に少ない (図 5a, b 参照) ことが見える。

実は, この余震群は空間的に 2 つの領域に分けられる。一つはユーラシア・プレートと，その下に沈み込む，太平洋プレートとの境界面である。もう一つは，太平洋プレート内部で，沈み込む入り口で折れ曲がっている所 (海溝の東側領域) である。前者は狭義 (普通) の余震群で, 後者は, 本震の断層から少し離れた場所で誘発されたもの (広義余震) である。異常現象を変換された時間で見ると, 静穏化は狭義の余震域には見られず, 広義余震域に見られる (図 5c 参照) 。 2007 年 1 月 13 日に, もう一つの巨大地震 (M8.2) が広義余震の領域で起きた。この理由は 12 節で述べるようなメカニズムで説明できる。

この様にして, 余震活動を調べると, 目だって大きな余震が起きる前や，近隣で大きな地震が起きる前には, 相対的静穏期 (所により活発化) が認められることが多い (Ogata, 2001, 2006; Ogata, Jones & Toda, 2003a) 。中でも, 1995 年兵庫県南部地震や 2004 年福岡県西方沖の地震に関しては, 本震直後からの余震解析によって相対的静穏化現象が認められ報告されたが, この後, 顕著に大きな余震の発生した (松浦, 1995; 尾形, 2005; Ogata, 2006a) 。余震の確率予報 (7 節参照) に基づけば，この程度の大余震の出現確率は, 相当小さかった。今後，静穏化にもとづく大余震の再予測を与える意義は大きい。多くの事例の経験を積む事によって，いづれ実用化されるであろう。

同様な解析で，一般の地震活動の異常な変化を示すことができる。遡及的な解析ではあるが，過去の世界のマグニチュード 9 クラスの超巨大地震や日本のマグニチュード 8 クラスの巨大地震の前には, 広領域で ETAS モデルに対して相対的な静穏化現象が認められた (Ogata, 2001) 。

12 なぜ静穏化するのか

余震活動が静穏化すると, 余震の隣接部に顕著な断層破壊を伴う大きな余震が起きる可能性が高い。静穏化が数ヶ月以上の長期間に及ぶと, 余震域から遠くないところ (例えば 200km 以内) では, たとえば 6 年以内に本震と同規模以上の地震が起きる発生確率が通常より数倍以上高い (Ogata, 2001) 。このような前兆的静穏化現象の仕組みとして, 次の様な仮説が考えられる。すなわち, 来るべき大余震または大地震の, 断層内または隣接部において, 地震計に掛からない様な (非地震性の) ゆっくりとした断層運動 (すべり) があり, それに伴い周辺部でストレスが変化するため, 余震活動が, 減衰法則で予測されるものより低下すると考えられる。実際, 筆者が最近報告しているように, クーロンの破壊応力が低下した領域 (ストレスシャドウ) では, 地震活動が相対的に静穏化することが多い。以下に, このことについて説明する。

地震の誘発を促したり, 活動を抑制したりする現象を説明するためには, クーロンの破壊応力という概念が有効である。地殻や上部マントルの岩石圏の内部は永年一定の方向で増加する応力を受けて歪んでいる。その意味で, 岩石圏といえども, 長期的には弾性体として考えることになる。断層は岩石圏の割れ目で, 大きなものから小なものまで無数にあり, 様々な方向を向いている。ここで一つの断層面を単純化して図 6 で示した。断層面にとって, その向きが大事で, 面に働く応力 (ストレス) は, 断層面を押さえつける成分 (垂直応力) と断層を滑らせる方向に働く成分 (せん断応力) に分解される。断層がすべる臨界状態は, 次の破壊応力 (Coulomb Failure Stress) で決まる。

$CFS =$ **せん断応力 - 断層の摩擦係数** $\times$**(垂直応力 - 断層間隙の流体圧)**

ここで, 最後の項の流体圧は, 通常の地震では 0 の場合を考えるが, 地震断層間隙中の水圧変化 (Hainzl & Ogata, 2005) や火山性群発地震における間隙マグマ流体の圧力変化 (Dieterich, Cayol & Okubo, 2000; Toda, Stein & Sagiya, 2002) などのように, 大きな役割をはたすこともある。このこ

とに関連して, 内陸の浅い地震活動の季節性と降雨量パタンとの関係が考えられる。このような関係をデータに基づいて検証する統計モデルやその応用例については, 筆者の解説論文 (尾形, 1993; Ogata, 1999) を参照されたい。

過去からのストレスの蓄積であるクーロンの破壊応力が一定の閾値を超えると, 断層が滑り地震を引き起こす。破壊応力は一定の率で増加しており, 臨界値に達すると滑る。実際的な時間スケールとして, たとえば内陸直下の断層は, 次に滑るまで, 千年のオーダーの時間を要する。ところが最近, これよりはるかに大きな応力の変化が注目されている。それは, 他の地震による地殻の応力変化 (ΔCFS) である。地震が近くで起ると, その影響で, 断層面の向きによって, 破壊応力が増えたり減ったりする。増えると地震は予定より早く起き, 減ると発生は先延ばしとなる。同じような向きの断層が多数あると, そこでの地震がたくさん起きるか, あるいは静穏化する，つまり, 活発であった活動が不活発になる。

原理的には以上の様な筋書きであるが, 実際の地震活動は相当に複雑であり解析は簡単でない。すなわち, 何処でも，いったん地震が起きると, その断層に隣接する断層系の破壊応力が高まり, 多数の地震が誘発されるからである。これが余震で, 大きい地震には多くの余震が発生し, 小さい地震でもそれなりの余震を誘発する。しかし，地殻中の断層系は見えないし, あまりにも複雑なので, 応力変化の計算は困難である。これが，ゆっくりすべりによる，広域の応力の微妙な変化を見難くして, 地震活動の異常性を見逃しているかもしれない。したがって, 統計モデルが必要となる。余震の経験則から構成した ETAS モデルを最尤法であてはめ, パラメタを決め, 地震の発生率を計算して, 誘発の強さを予測するのである。

大きな地震については, 地震波または GPS (Global Positioning System; 全地球測位システム) 観測値などの逆問題として, その断層の配置やすべりの方向や規模が求められる。これらを入力として ΔCFS の計算を実行するプログラム (Okada, 1992) が普及したので, 近来 ΔCFS と地震の誘発 (トリガリング) に関する研究が盛んになってきた (Harris, 1998, ed.

および Steacy et al., 2005, eds. 参照) 。例えば, 東南海地震と南海地震前後の西南日本における地震活動異常の ΔCFS による解釈もある (Ogata, 2004) 。一方, ゆっくりとした非地震性の断層運動を想定して, 地殻内の微弱なストレス変化と地震活動の静穏化や活発化の関係を論ずることもできる (Ogata, 2005a, b, 2006a, b) 。 ETAS モデルなどを使って, 相対的地震発生率の変化を解析することによって, 非地震性のすべりなどによる広域のストレス変化を検出でき, これが GPS による地殻変動の異常変化と同様の鋭敏なセンサーとなりそうである。

13 時空間 ETAS モデル

時空間 ETAS モデルは ETAS モデルの拡張であり, 空間的には余震域の面積と本震のマグニチュードの相関 (宇津・関のスケーリング則; 宇津, 2001, 7.1.A 節参照) に依拠している。余震活動の強度は距離的に減衰するが, 様々な関数のなかで最も AIC 値の良かったのは, 次のような, 逆冪型減衰関数を持つ, 条件付強度関数であった。時刻 t と位置 (x, y) での地震活動の強度が, 過去の地震発生履歴の関数として,

$$\lambda(t,x,y|\,\theta) \;=\; \mu(x,y) + \sum_{\{j;\, t_j<t\}} \nu(t-t_j|\,K,c,p) \times \left[\frac{(x-x_j, y-y_j) S_j (x-x_j, y-y_j)^t}{e^{\alpha(M_j-M_c)}} + d\right]^{-q} \tag{13.15}$$

で表されるものである (例えば, Ogata, 1999, 参照) 。これを x と y で全平面を積分すると, 通常の ETAS モデル (8.13 式) になる。

ここで, $\mu(x,y)$ は地点 (x,y) における常時地震活動度 (background seismicity), $\nu(t|\,K,c,p)$ は (5.8) 式の大森・宇津の関数, M_c はデータの下限マグニチュード, M_j は j 番目の地震のマグニチュードである。座標 (x_j, y_j) は, 必ずしもデータの震央でなく, 誘発された地震群の重心として計算されたものである。何故なら, 大きな地震の場合, 断層面 (余震域) は中心部から破壊するとは限らないので, 一般に余震活動の中心は破壊開始点である震源の座標とは異なるからである。 S_j は, 余震群の空間の密

度分布の等高線を楕円で近似するような対称行列である。余震が少ない小さな地震などの場合, 余震群の空間分布は等方的 (すなわち S_j は単位行列) で, その中心は震源の位置で十分である。

これらは，AIC 比較によって, 非心性なら座標 (x_j, y_j) を地震群の重心で, 非等方なら行列 S_j を余震座標の分散や相関係数を使って置き換える。その様に編集された時空間領域 $[S,T] \times A$ 内のデータ (t_i, x_i, y_i) に対して, 対数尤度関数

$$\log L(\theta) = \sum_{i=1}^{N} \log \lambda(t_i, x_i, y_i|\, \theta) - \int_S^T \int\int_A \lambda(t, x, y|\, \theta)\, dtdxdy \tag{13.16}$$

(例えば, Daley & Vere-Jones, 2003, 7 章参照) を最大化する最尤法を適用して, パラメタ $\hat{\theta} = (\hat{\mu}, \hat{K}, \hat{c}, \hat{\alpha}, \hat{p}, \hat{d}, \hat{q})$ は推定される。(13.16) 式の第 2 項の空間積分は数値積分をする (詳細は Ogata, 1998, 参照) 。対数尤度関数の最大化には, たとえば, Davidon-Fletcher-Powell 法 (Kowalik & Osbone, 1968, 参照) を採用する。

14 階層的ベイズ型時空間モデル

広域的にみると, 各地の地震活動には個性がある。それをつぶさに捉えるために, 地域的な違いを表現する階層ベイズ型時空間モデルを開発した。これはモデル (13.15) のパラメタ値が場所によって変化するものとし, それらで地震発生様式の地域性を表わし可視化する。すなわち, 常時地震活動度 $\mu(x, y)$ のみならず, 4 組のパラメタ K, α, p, q も位置 (x, y) の関数と考える。たとえば, 図 7a にあるような, 1926~1995 年のマグニチュード 5 以上の地震を解析する時, 地震の位置 (x_i, y_i) を頂点とするデロネ分割 3 角形 (Delaunay triangulation; 図 7b) 内で, 線形的に内挿され, 三角面で張られた部分的線形関数 (デロネ三角面関数) を考える。デロネ分割に基づいて推定する方法の特長は, 空間的標本の集中型または不均質な配置でも, 効率良く関数の変化を表現できるところにある。地震の密な所は詳しく, 疎なところは大雑把にという具合である。空間の次元が高くなっても, デロネ分割だと, データ数に比例した

パラメタ数で表現できる。以後, この (13.15) 式のモデルを階層的時空間 ETAS(HIST-ETAS) モデルと呼ぶ (Ogata, Katsura & Tanemura, 2003b ； 詳細は Ogata, 2004, 参照) 。

それでも, これらのパラメタの関数における, 推定すべき係数の個数はデータの地震数の 5 倍強である。したがって, 安定した解を求めるため, 係数同士の関係に次のような制約を課す。デロネ分割上の各三角面の傾き (一次微分) の 2 乗の積分に対するペナルティ, すなわち, デロネ三角面関数が無意味に凸凹することを抑える。そして, 当てはまりの良さを測る対数尤度 (13.16 式) との釣り合いを, ペナルティ付き対数尤度を通して測るのである。 5 組のパラメタ関数それぞれのペナルティの重みを ABIC によって客観的に決め, ペナルティ付き対数尤度を最大化する係数を解として得る。数値的には Davidon-Fletcher-Powell 法 (Kowalik & Osbone, 1968, 参照) に, 不完全修正コレスキー分解前処理つき共役勾配法 (ICCG 法, たとえば森, 1987, 参照) を組み合わせたものによって, 次元がいくら大きくても, 効率的に最大化することができる。その上で, (2.4) 式と同様なベイズモデルの尤度の最大化については, 例えばシンプレックス法 (Kowalik & Osbone, 1968, 参照) で, 制約関数の重み (ベイズモデルの超パラメタ; $w_1, \cdots, w_5$) を自動的に探索することが出来る。

ここで, モデルのパラメタ関数は, 位置には依存するが, 時間変化 t に関しては変わらず一定とした。これは, 15 節に述べる様に, 異常活動が時間依存するものとし, その診断解析における異常の検出を効果的にするためである。

たとえば, HIST-ETAS モデルを気象庁震源データ (1926 年以降, マグニチュード 5 以上) にあてはめて, 推定されたパラメタ (Ogata, 2004) のうち, 特記すべきものを 2 つ示す。図 7c は余震減衰指数 p 値である。列島の太平洋側は p 値が低く, 日本海側や火山地域では高い。これは茂木 (Mogi, 1967) が日本各地の余震列を一つ一つ解析して求めた p 値のパタンに酷似している。 Mogi (1967) や Kisslinger & Jones (1991) によると, p 値は地殻中の熱流量 (温度) と強い相関がある (例えば Utsu et al, 1995 参照) 。図 7d には, HIST-ETAS モデルの余震活動の強度 K 値を, 東北地

方と周辺部に限って拡大して示した。これは, (5.8) 式の K を (8.13) 式のように，マグニチュードの効果を標準化した余震活動の強さを示し, 余震の生産性 (productivity) と呼ばれる。三陸はるか沖や日本海沿岸では余震活動が強く, 青森県沖や宮城県沖では弱いことが分かる。図中の + 印はアスペリティの中心を示す。アスペリティは断層面内の摩擦強度の強い部分で, これが滑ると大地震になる。地震研究所の山中と菊地 (Yamanaka & Kikuchi, 2004) は大地震の各地での地震波の逆問題を解く事によってこれらを求めた。K 値の高い場所はアスペリティと相補的であり, アスペリティの境界部にあることが分かる。実際, 破壊断層の中で, すべり量の大きい部分と余震の密度の高い部分は相補的である事が知られている。α 値や q 値も日本周辺で独特の空間分布を持っている。

15 どうやって静穏化した地域を見つけるか

各地の実際の地震活動度と時空間モデルで計算された理論的活動度の相対比率から, 相対的静穏化や相対的活発化した地域や期間を見出すのが目的である。以下のように, モデルを基準にした相対的な地震活動変化を可視化することによって, 相対比率 (相対的地震活動度) を算出し, これをもとに地殻の歪変化を観察することを狙った。

$$\eta(t,x,y|\,\phi) = \lambda(t,x,y|\,\hat{\theta}) \cdot R(t,x,y|\,\phi) \tag{15.17}$$

すなわち

実際の地震活動度 = 階層的時空間 ETAS モデル × 相対的地震活動度

である。相対的地震活動度 $R(t,x,y|\,\phi)$ は, 3 次元の時空間のデロネ 4 面体分割 (頂点は地震の時空間座標) 上の部分的線形関数 $\phi(t,x,y)$ で $R(t,x,y|\,\phi) = \exp\{\phi(t,x,y)\}$ として表示される。ここで, 相対的地震活動度モデルのパラメタ関数は, HIST-ETAS モデルと違って, 位置にも時間にも変化して，時間依存する異常活動がを検出するためである。

(15.17) 式のように定められた, 実際の地震活動の強度 $\eta(t,x,y|\,\phi)$ を (13.16) 式の対数尤度に $\lambda(t,x,y|\,\theta)$ の代わりに代入し, デロネ関数

$\phi(t,x,y)$ の傾き (一階微分) の 2 乗の積分にペナルティをかけた制約条件のもとに, ペナルティ付き対数尤度を考え, ベイズ法で最適化して, 最大事後分布のパラメタ解として得られる。

相対的地震活動度を求めることで, 地震の異常活動度を 3 次元的に表現し, AVS などで可視化できる。広域的な歪変化は, 相対的地震活動を考えることで, より鋭敏に検知できる。例えば, 大地震の前に相対的空白域が明瞭に見られる場合がある。図 8 には, 推定値 $R(t,x,y|\,\hat{\phi})$ の 3 次元イメージを，東経 143 度で切取った緯度と時間の平面領域上での, 相対的地震活動度の画像で示した。青森県沖, 三陸沖, 宮城県沖における大地震発生の前に, 相対的な静穏化を示す, 窪み (暗灰色) が明瞭に見られる。これらは元々, HIST-ETAS モデル $\lambda(t,x,y|\,\theta)$ には含まれていない新しい現象である。このようなものが大地震の予測の手掛かりになりうるのか，多くの解析例を積み重ねる価値と必要がある。

16 最後に

地震カタログの不均質性，地震統計の多様な経験法則や仮説を統計的モデルとして表現してきた。これらによって計測し議論できる, 統計地震学 (Statistical Seismology) とも称すべき, 地震活動解析の研究領域が広がり深まったと考えている。ただし, 本稿で紹介したものは, 掲載された学会誌等に基づいているが, その多くが研究途上である。たとえば, 地震活動静穏化ひとつをめぐっても, その意義自体への異議を唱える研究者もあり, 静穏化の解釈にいたっては, 多くの説がある。本稿で紹介した成果は, 近来の豊富な優れた地震カタログデータと使用した統計モデルに帰するところが大きいが, まだ比較的小数例の解析結果でもあり, 定説として受け入れられているわけではない。本稿での解析例は, とりあえず, 統計地震学に対する新しい可能性の事例であり, 主張されているような仮説の真偽が定着するためには, 解析結果の積み重ねが必要である。

本稿で紹介したモデルや方法の一部計算ソフトウェアとマニュア

ル (Utsu & Ogata, 1997)[9] が国際地震学地球内部物理学会 (IASPEI) やアメリカ地震学会 (SSA) から販売されている。最近では, 筆者の研究プロジェクトの WEB[10] または, 統計数理研究所発行の*Computer Science Monograph*[11]から利用可能である。

最後になったが，赤池弘次先生, David Vere-Jones 先生および故宇津徳治先生をはじめ, 共同研究者各位, 諸先輩・同僚には折に触れてご教授・討論を頂き, 著者の研究に多大なる影響を与えてくださった。また, 気象庁をはじめ地震学関連の大学・研究機関の成果である地震カタログを使用させていただいた。匿名のお二人の査読者には本稿改善のための有益なコメントを頂いた。ここに記して感謝する。

文献

ページ数が限られているため, 本文中で引用した論文の多くは，ここに記すことができなかった。ここでは, 総合報告や書籍など手掛かりのために限られたもののみ記した。これらの中の文献表から, 本稿で割愛した関連文献を探すのにも役に立つであろう。関連する先行研究も, それらの文献をご参照頂き, 筆者の WEB ページ[12], または Google Scholar などで, 検索されたい。

D.J. Daley and D. Vere-Jones (2003). *An introduction to the theory of point processes.* **Vol. I***: Elementary theory and methods.* Springer, Berlin.

R.A. Harris, Ed. (1998) *Stress Triggers, Stress Shadows, and Implication for Seismic Hazard; J. Geophys. Res., Special Issue*, American Geopgysical Union, Washington DC.

J. Kowalik and M.R. Osborne (1968) *Methods for unconstrained optimization problems*, American Elsevier, New York, (山本善之・小山健夫

[9] http://www.seismosoc.org/ssa/publications/IASPEI_Software.html
[10] http://www.ism.ac.jp/~ogata/Ssg/ssg.html
[11] http://www.ism.ac.jp/editsec/csm/index_j.html の No.32 および No.33
[12] http://www.ism.ac.jp/~ogata/papers.html

(訳) 1970, 非線形最適化問題, 培風館, 165pp.)

森正武 (1987) **数値計算プログラミング**, 岩波書店, 398pp.

尾形良彦 (1993). **地震学とその周辺の地球科学分野に於ける統計モデルと統計的手法**, **日本統計学会誌 22**, No. 3 (60 周年記念特別号), 413 — 463.

Y. Ogata (1998) Space-time point-process models for earthquake occurrences, *Ann. Inst. Statist. Math.* **50**, 379-402.

Y. Ogata (1999) Seismicity analyses through point-process modelling: A review, in *Seismicity Patterns, Their Statistical Significance and Physical Meaning*, M. Wyss, K. Shimazaki and A. Ito eds., *Pure and Applied-Geophysics* **155**, 471-507.

Y. Ogata (2004). Space-time model for regional seismicity and detection of crustal stress changes, *J.Geophys. Res.*, Vol. 109, No. B3, B03308, doi:10.1029/2003JB002621.

Y. Ogata, A. Kobayashi, N. Mikami, Y. Murata and K. Katsura (1998) Correction of earthquake location estimation in a small-seismic-array system, *BERNOULLI* **4**, 167-184, 1998.

E. Parzen, K. Tanabe and G. Kitagawa (1998) *Selected Papers of Hirotugu Akaike*, Springer-Verlag, New York.

S. Steacy, J. Gomberg and M. Cocco, Eds. (2005) *Stress Transfer, Earthquake Triggering and Time-Dependent Seismic Hazard; J. Geophys. Res., Special Issue*, American Geopgysical Union, Washington DC.

宇津徳治 (1999). **地震活動総説**, 東京大学出版.

宇津徳治 (2001). **地震学**, 第 3 版, 共立出版.

T. Utsu, Y. Ogata and R. S. Matsu'ura (1995) The centenary of the Omori formula for a decay law of aftershock activity, *J.Phys. Earth* **43**, 1-33.

付論: 点過程の条件付き強度関数

$P_\Delta(t|H_t)$ は, 過去の履歴 $H_t = \{(t_i, M_i);\ t_i < t\}$ のもとでの, 微小時間区間 $(t, t+\Delta)$ における事象 (点) 発生の，条件付き確率であるとする。履歴 H_t は, 現在時間 t から遡った, 過去の事象の発生時刻 $\{t_i\}$ のみならず, 関係する他の情報 (時系列や他の事象系列) である。たとえば, 地震の場合, マグニチュード $\{M_i\}$ 系列の履歴は重要な情報である。**条件付き強度関数** (*conditional intensity function*) $\lambda(t|H_t)$ は形式的に

$$\lambda(t|H_t) = \lim_{\Delta \to 0} \frac{P_\Delta(t|H_t)}{\Delta}$$

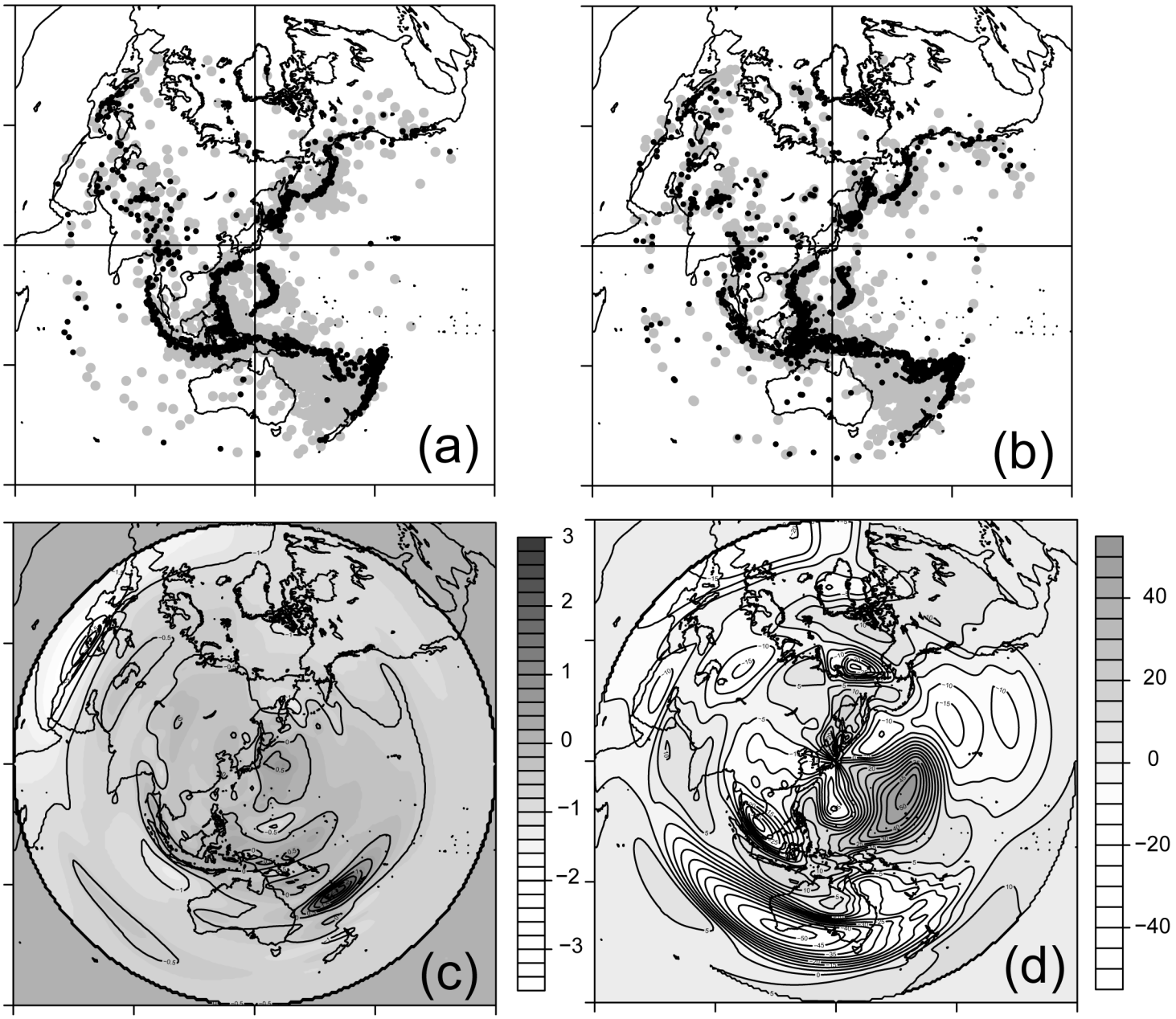
で与えられるものである。条件付き強度関数は点過程を完全に記述する。つまり, 条件付き強度関数と点過程は一対一に対応する (Daley and Vere-Jones, 2003, Chapter 3 参照) 。例えば, 待ち行列, 保険数学や信頼性理論でよく使われる更新過程は, 条件付き強度関数が最近に発生した点にのみ依存した危険率 (Hazard rate) 関数を使って

$$\lambda(t|H_t) = \nu(t - \tau_L) = \frac{f(t-\tau_L)}{1 - F(t-\tau_L)}$$

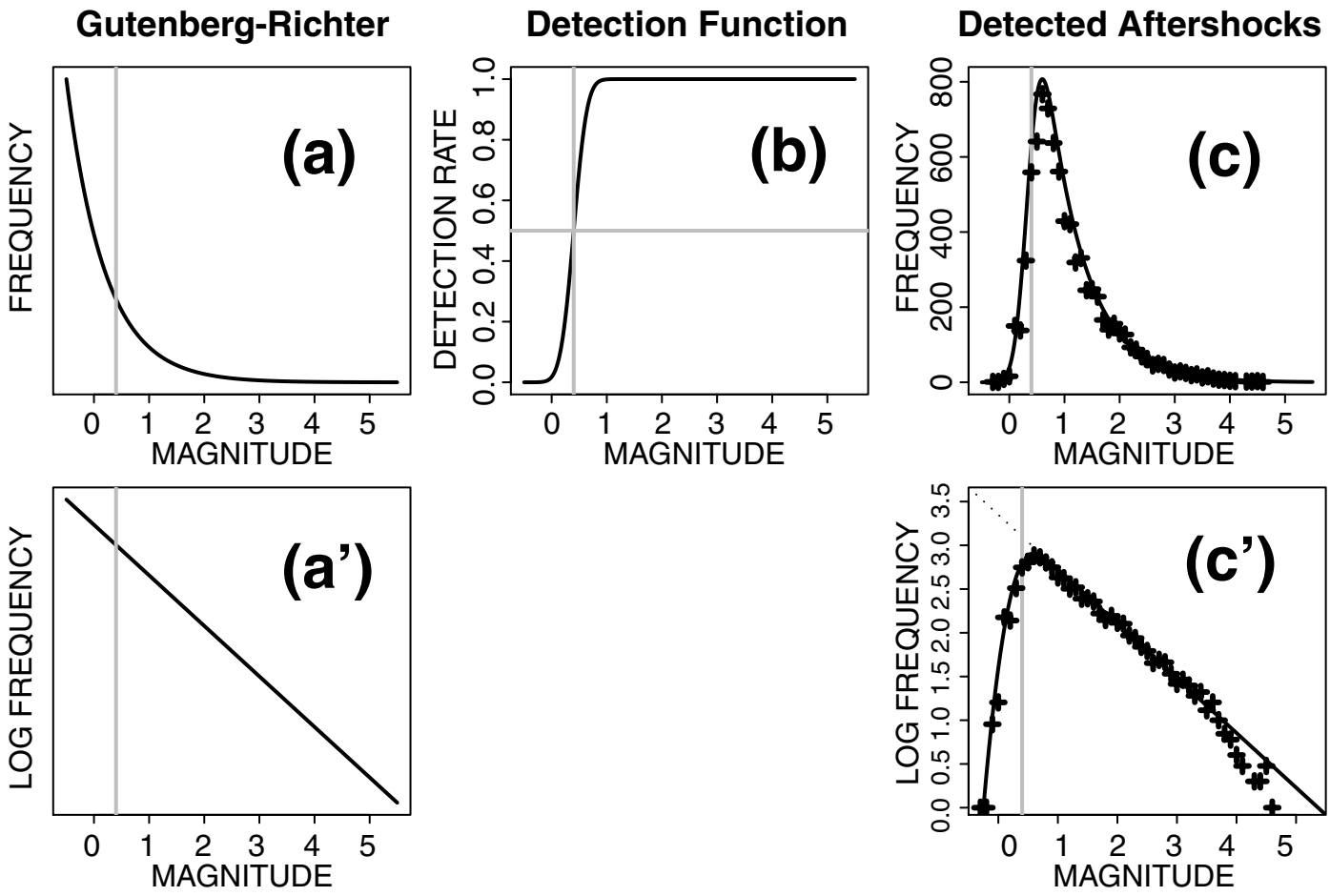
の様に表現される。ここで, $F(\cdot)$ は隣り合う点の間隔の長さの分布関数で, $f(\cdot)$ は, その密度関数である。このタイプの点過程を拡張したものに Wold 過程がある。これは $\lambda(t|H_t) = \nu(t - t_{(1)},\ t - t_{(2)},\ \cdots,\ t - t_{(m)})$ のようになる。ここで, $t_{(i)}$ は, 時刻 t より前にあって, 最後から i 番目に発生した点の時刻である。非定常ポアソン過程は, 条件付き強度関数が $\lambda(t|H_t) = \nu(t)$ のように, 過去の履歴に無関係で, 時刻 t だけの関数である。さらに, $\lambda(t|H_t) = \mu =$ 定数, のとき, 定常ポアソン過程を特徴づける。

条件付き強度関数に基づいた点過程のシミュレーション・アルゴリズムとして, Thinning によるシミュレーション法 (Ogata, 1983; 例えば,

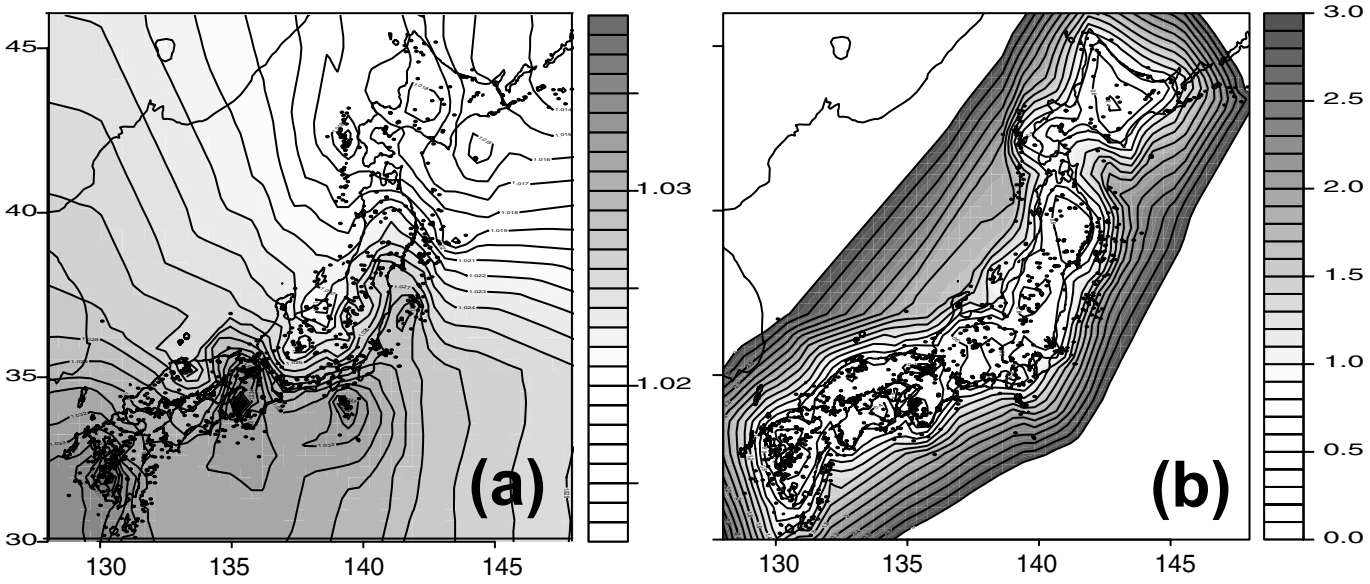
Daley and Vere-Jones, 2003, Chapter 3, 参照) がある。その一般性と速さは画期的なものである。他方，点過程のスペクトル関数や自己共分散関数, 点間の分布関数などは, これらに対応する異なる点過程をいくつも作ることが出来るので, 点過程を同定できないし, 従ってシミュレーションもできない。



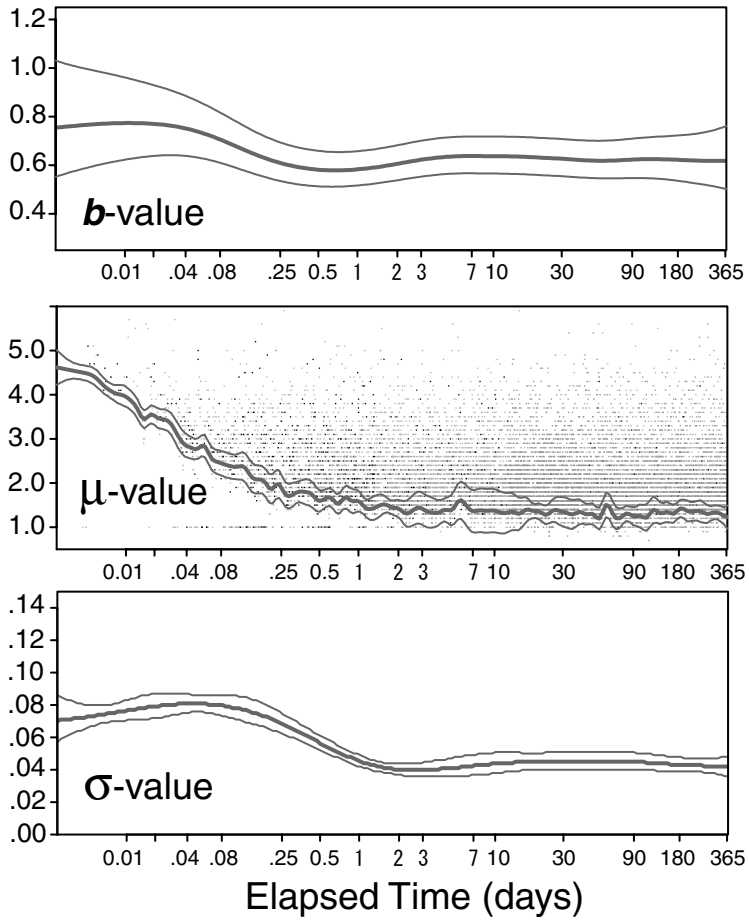
[図1.] 松代群列地震計(MSAS)が決めた世界の地震の位置(aとbの灰色の点)。米地質調査所(USGS)の世界ネットワークが決めた位置(aの黒点)及びMSASを補正した位置(bの黒点)。松代観測室を原点にとって極座標で表した。(a)は1984～88年の地震で(b)は1989～92年の地震。(c)と(d)は1984～88年のMSASとUSGSの震央位置の違いから推定されたバイアスで, これらによって1989～92年のMSAS震央(bの灰色点)が補正された(bの黒点)。(c)は松代観測室からの震源距離の伸・縮(+/-)の程度を示し, その距離の単位は緯度の1度分(= 111km)。(d)は時計回り(+)または反時計回り(-)の角度を示す。



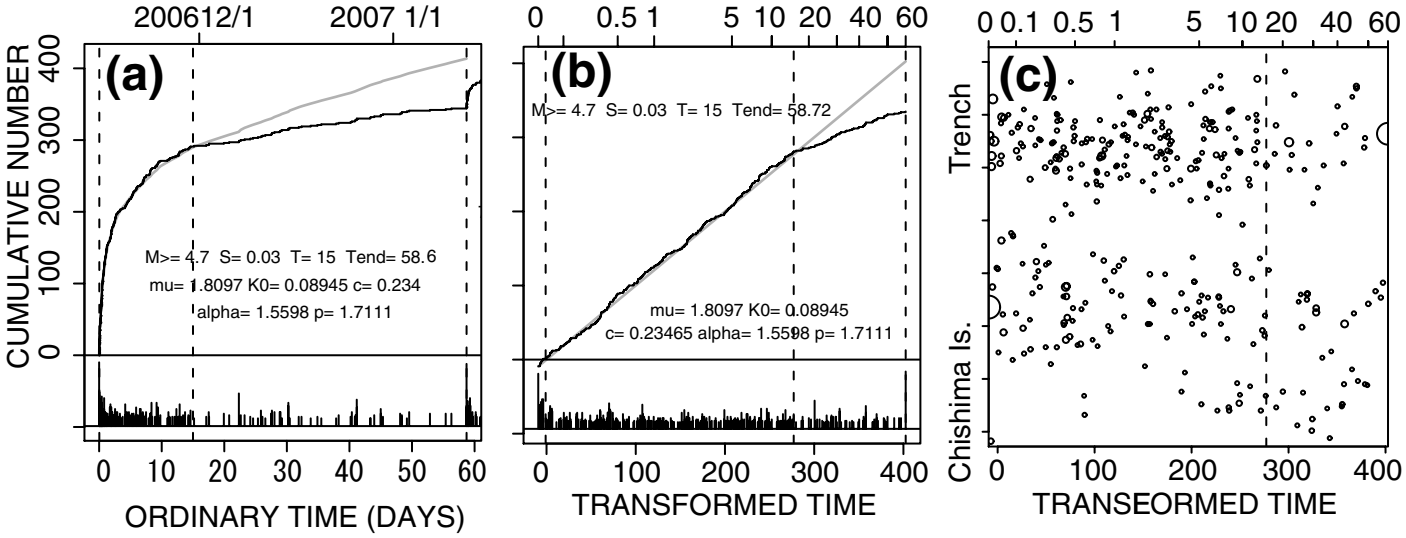
[図2.] 地震検出率とb値を同時に推定するモデルの模式図。+プロットは, 2003年宮城県沖の地震(M7.0)の, 本震後20日から1年にわたる余震のマグニチュードの頻度分布(図4参照)。それぞれのパネルの説明については本文４節参照。



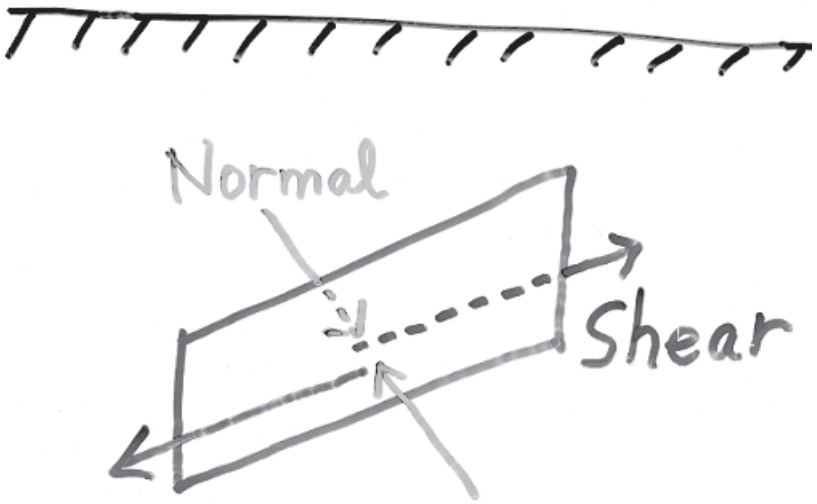
[図3.] 2001年10月中に収録された全ての地震データから求めたb値分布(a)と50%検知率のマグニチュードμ値(b)。



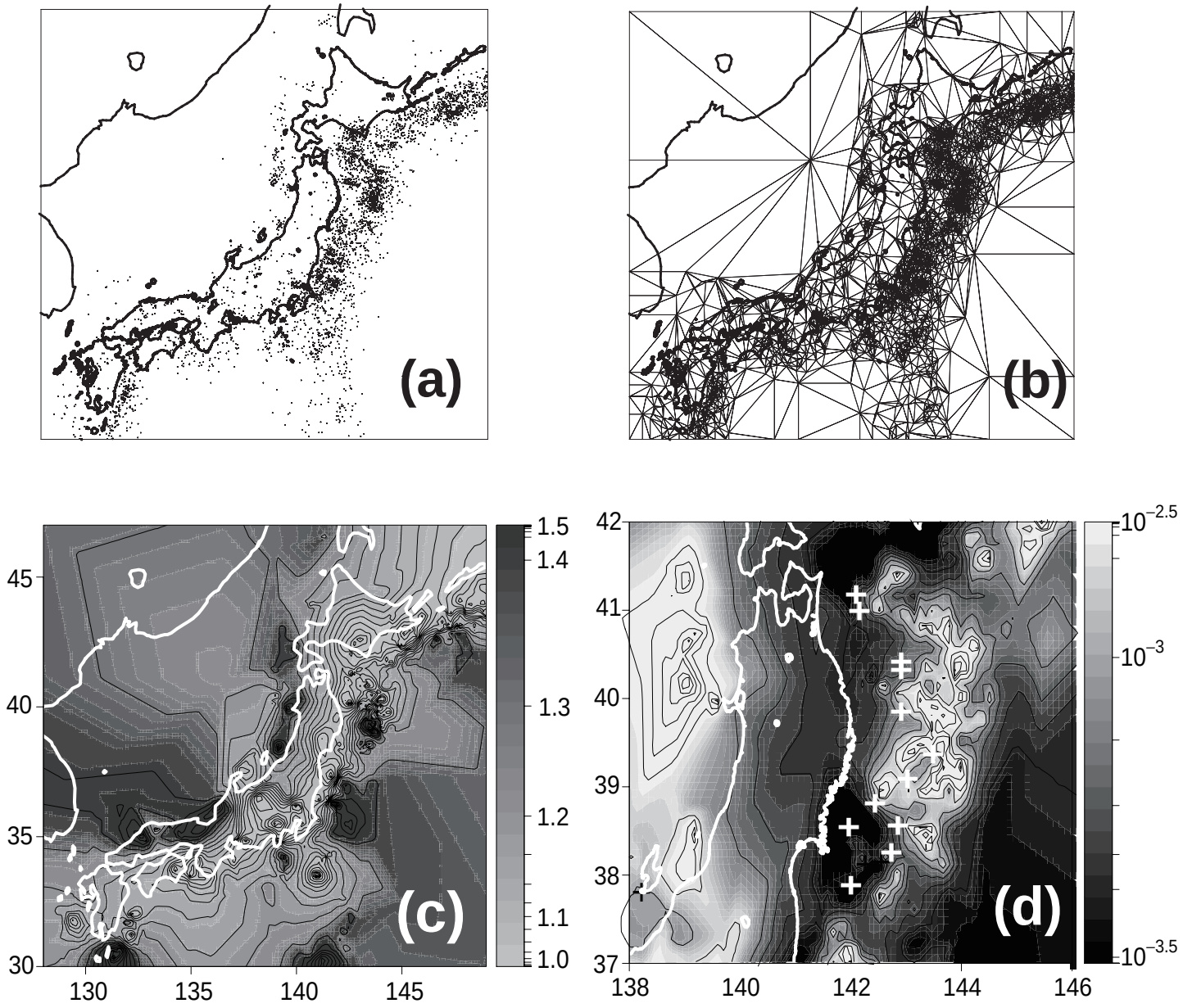
[図4.] 2003年宮城県沖の地震(M7.0)の余震の, b値と検出率(μ値とσ値)の, 本震直後から1年間の推移。太い実線が最大事後分布推定で, 細い実線が95%誤差。ただし, 時間は次の様に変換されたものである。M3.0以上の余震に当てはめた大森・宇津の公式の積分関数が一様になるようにした(９節参照)。中段のパネルの灰色点群プロットは, 検出された各余震の, マグニチュードを示す。



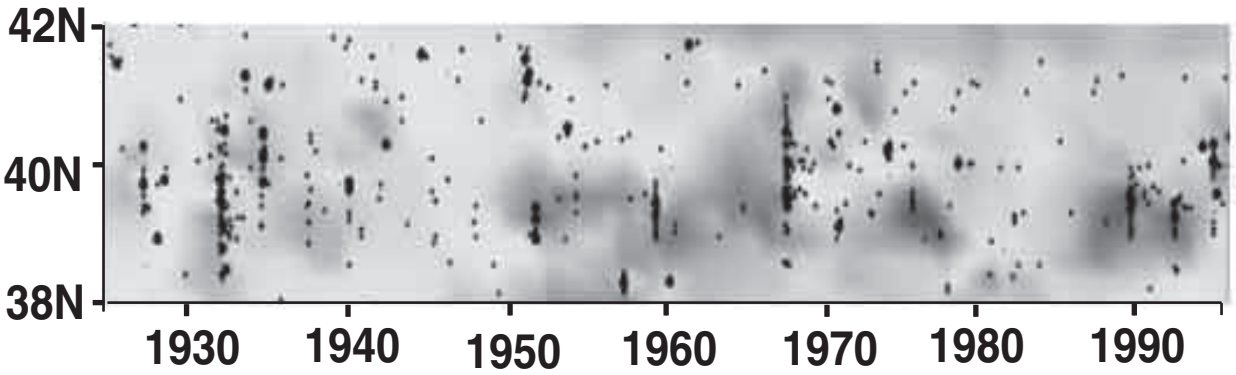
[図5.] 2006年11月15日の千島列島(Kuril Islands)の巨大地震(M8.3)の余震(M≥4.7)。2007年1月13日までの, マグニチュード対時間図と累積関数が示されている。(a)と(b)は, 千島列島に直交する線分に投影した, 震央の時空間プロット(c)。灰色の累積曲線は, 本震直後から半月までの余震データにETASモデルをあてはめ推定し残りの一ヶ月間を予測した, 理論的累積曲線。横軸は (a)が通常の時間推移で, (b)と(c)がETASモデルによる変換時間(９節参照)。



[図6.] 断層に働く応力の成分の模式図。



[図7.] 階層的時空間ETAS (HIST-ETAS)モデルで推定されたパラメタ関数の画像。
(a)は1926〜1995年のマグニチュード5以上の地震の震央図。(b)は震央間を繋いだデロネ三角形分割。そして (c)は余震減衰p値の推定(0.8〜1.5)で, (d)は余震活動の強度K値の推定(単位はevent/day/deg2)で, 東北地方のみを拡大したものである。



[図8.] HIST-ETASモデルの活動度と実際の活動度の相対比率(相対的地震活動)。
理論と実際のものが同じところは1.0で, 相対的地震活動が小さい値から大きな値まで, 暗灰色から明灰色の順で対応している。縦に並んでいる点群は大地震後の余震群(東経143±2度以内のもの)。

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版，2011 年 11 月
第 1 部 自然・工学と統計科学

予測と発見を目指す統計科学

北川　源四郎[1]
(情報・システム研究機構 機構長)

情報社会の到来と情報通信技術の飛躍的発展によって，学術分野でも一般社会でも膨大なデータや情報が蓄積し，広大なサイバー空間が形成されようとしている．データに基づく知識獲得の方法としての統計科学の役割は今後ますます重要になる．しかしながら，同時に研究の目的も真理の探究から，知識の獲得・知識の創造へシフトし，求めるべき知識自体も徐々に変化しつつある．従来のマクロで普遍的な知識に替わって，より個別的な状況に対する知識が要求されるようになっている．これに伴って統計科学が取り組むべき課題も変化しつつある．この個別化に伴う課題への回答のためには，究極の条件付けモデリングが必要であり，実用的な方法の開発は容易ではない．ここに現代の統計科学が取り組むべきグランドチャレンジが存在すると思われる．

本稿の前半では，情報化社会における知識獲得のためには，柔軟なモデリング，能動的なモデリングが必要であり，それがベイズモデルによって実現できることを議論する．後半では，時系列解析の問題に特化して，筆者が取り組んできた一般化状態空間モデリングによる情報統合の問題を考え，海底地震計データの解析例を紹介する．

[1]kitagawa@rois.ac.jp

[1]本稿は日本統計学会誌第 37 巻第 1 号 25-36 (2007) の内容を一部修正したものである．4 節の海底地震計データに関する共同研究者である北海道大学の高波鉄夫氏に感謝する．

1　はじめに

20 世紀には科学・技術の発展に支えられて大量生産・大量消費の物質文明が急速に発展した．しかし，20 世紀後半における社会の情報化および情報通信技術の飛躍的発展によって，情報や知識が物質・エネルギー以上の価値を持つ社会となり，脱工業化の動きは決定的になった．情報は質量を持たないという特性もあって，インターネットの発達により，今や「いつでも，どこでも，誰でも」情報サービスが受けられるユビキタス社会の実現も目前に迫っているように思われる．

このような情報化された 21 世紀社会においては，学術も産業技術もまた社会技術も必然的に大きな変容を遂げつつある．学術の世界においては，20 世紀においてすでに普遍的な「真理の探究」から，生命，地球環境，人間社会など，変化し，進化する対象の理解へとその目的や対象は徐々に変化しつつあったが，情報社会を迎えてその変容は一段と加速している．人間は，短期間に広大なサイバー空間を築き上げ，今では実世界とサイバー世界の両者に跨って活動しているともいえる．しかし，サイバー世界に「普遍の真理」は想定し難い．問題は，如何にこの巨大なサイバー世界を活用し，人間の生活に有用な知識を獲得・創造するかである．

本稿では，この様な認識のもとで統計科学が今後目指すべき方向を検討する．後半では，時系列解析の分野に限定し，能動的なモデリングとその具体例を紹介する．

2　ポスト IT 時代における知識獲得技術としての統計的モデリング

現代社会は，物質・エネルギー以上に情報が重要な意味を持つ情報社会となった．これと相俟って，測定技術の発達によって様々な学術分野や社会において大量のデータが時々刻々集積しつつある．ライフサイエンスにおける DNA データやマイクロアレイデータ，マーケティングにおける POS データ，ファイナンスにおける高頻度データや CRD（Credit Risk Database），環境科学データ，地震学・気象学などの地球科学データ，天文学（全天 CCD 観測）データなどその例は枚挙に暇がない．さらにインターネットとデータベース技術の発展によって，いつでもどこでも情報ネットワークに容易にアクセスして情報サービスを受けることができるユビキタス社会が実現しつつある．

情報の多寡が勝敗を決めた初期の情報社会に対して，万人にとってユビキタスな社会が実現した暁には，大量データからの本質的な知識獲得の技術が

死命を制することになる．本稿ではこのような社会をポストIT社会と呼ぶことにする．ポストIT社会では，大規模データに基づく予測，シミュレーション，情報抽出，知識発見，データマイニングの方法，すなわち予測と発見の方法の研究・開発が必須となる．まさに統計科学の出番である (樋口 2006)．

データが単に巨大化するだけならば，理論的には重大な問題は生じない．実際，求められる情報や知識が従来どおりのものならば，データの巨大化はむしろ歓迎すべきことである．データが増えれば統計的推定の精度は向上する．データの巨大化によって生じる計算上の困難は，計算システムの高速化とアルゴリズムの高度化によってある程度対応できるだろう．

しかし，ポストIT時代の情報処理においては，求められる情報，求められる知識自体が変化していることを忘れてはならない．従来の科学研究においては普遍的真理の探求が主要な目的であったが，情報社会においては知識の創造が重要となる．特に大規模データを背景に，何らかの条件に特化した個別的知識の獲得が求められる．例えば，薬剤疫学や医療においてはゲノム情報を用いて副作用や効果等が異なる個々人に対応した投薬や医療を目指すテーラーメイド創薬 (投薬)，オーダーメイド医療が試みられるようになりつつある．マーケティングにおいては，従来のマスマーケティングに替わって，個人の属性や過去の購買履歴に応じて効果的な対応をするマイクロマーケティングが開始されている．集団の平均的な特性としての知識に替わって，ある時点，ある地点，あるいはある個人に関する個別的な情報が必要になる．今後の「知識」はこのような個別的情報に対応可能なものとなるであろう．

しかし，個別化の問題は究極の条件付けを要求する．カテゴリカルデータの解析法において，Sakamoto and Akaike(1978) は多数の説明変数の中で目的変数の予測に有用な変数を自動的に探索できる方法を提案した．この方法は，最大次元の分割表の作成を回避することによって多数のモデル間の比較・選択を実用化したもので，現在でもデータマイニングの手法として利用されている．しかし，今後マイクロ・マーケティング等において，個人の情報を最大限活用しようとする個別化の要求に答えるためには，最大次元の分割表にも対応可能な方法あるいはそれを経ずとも必要な情報を抽出できる方法の開発が必要である．

同様な問題はマイクロアレイデータ解析にも見られる．数千から数万の特性値に対して被験者数はせいぜい数十程度であることからいわゆる新NP問題（n をデータ数，p を説明変数の数とするとき $n << p$ となる）が発生する (図1参照)．利用できる特性値は今後も増え続けるに違いない．$n \gg p$ を前提とする従来の統計解析の常識を超えた，多数の要因に対応できるモデリングが必要となる．

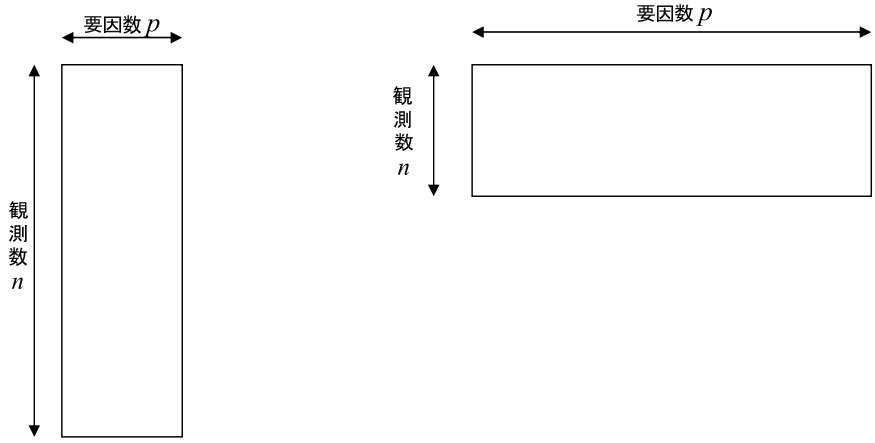


図 1: 従来のデータ行列 (左 : $N > p$) と新 Np 問題におけるデータ行列 (右 : $p \gg N$)

このように，大量データが自在に入手可能な情報社会においても，必要な情報は常に不足し欠落していると考えざるを得ない．データは増えても考えるべきモデルはそれ以上に大規模化し，条件付けに必要な情報はますます疎になるという逆転した状況が発生するからである．この問題の解決には，必要な情報の不足を補う制約充足の方法が不可欠である．

問題は，それを実現するための知識獲得の支援技術の確立である．統計的モデルは確率分布で表されるが，統計的モデリングの真髄は適切な条件付モデルを求めることにある．その具体的方策は，変数選択や次数選択あるいは関数形，変数変換や確率分布の選択の形をとることが多いが，その本質は適切な条件付けにあるといえる．

従来の正統的な数理統計では，データは真の分布 $x \sim f(\cdot|\theta)$ に従うとの前提の下で,「真の」構造の推定を目指してきた．これに対し，赤池氏は「よい」モデルを求めようとする立場から，モデル評価の規準として情報量規準 AIC を導いた (Akaike 1973)．いったん「真の」モデルを離れて「よい」モデルという視点に立てば様々な可能性が見えてくる．不偏性は必須の要件ではなくなる．比較的少数データの環境では，よいモデルを求めるためには多少，不偏性は犠牲にしても簡潔なモデルを選択するべきであることは AIC が明確に示している．

しかし，このモデル選択の方法さえも，有限のデータからよいモデルを求めるための方策のひとつにすぎない．変数選択によって，モデルに含まれる自由度を制限するかわりに，変数間に適切な制約を課すことによって,「良

い」モデルを求めることも可能なのである．Akaike(1980) は，季節調整の問題において，トレンドと季節成分を含む経済時系列 y_n を

$$y_n = T_n + S_n + w_n \tag{2.1}$$

と 3 つの項に分解する方法を提案した．ただし，T_n は経済成長にともなってゆっくり変動するトレンド成分，S_n は毎年ほぼ同じ値を繰り返す季節成分，w_n はその他の不規則成分である．従来のパラメトリックモデルではトレンドを多項式，季節成分を三角関数等の「硬い」モデルで近似するのに対して，赤池氏は T_n および S_n を未知パラメータとみなす大胆なモデルを提案した．いうまでもなく，従来の最小二乗法や最尤法では意味のある推定値は得られない．赤池氏は，T_n や S_n の時間的変動の滑らかさに関する制約を導入し，ベイズモデルの枠組みで問題が解決できることを示した．長年の課題であった，事前分布の設定の問題を解決し，ベイズモデルの実用化に成功したのである．これが先鞭となって，いまや知的情報処理の分野ではベイズモデルが不可欠なツールなりつつある．

後に，赤池氏は「唯一無二の真理」などというものは存在しないとし，統計的推論の効果的な展開のためには，現在の観測データに考察を限ることなく，対象に関する知識や経験も含めた形で考察を進めることが重要であることを指摘している (赤池 1995)．さらに Akaike(2001) では，D_1 を既存の客観的に確立された知識，D_2 を経験的な知識，D_3 を観測データとして，情報データ群 IDS $=(D_1, D_2, D_3)$ を定義し，統計的推論は仮説の提案と検証により情報データ群を適切に構成し発展させる過程を通じて実現される包括的な知的活動であると規定している．

さて，真のモデルの存在を仮定しない立場にたって，大量データが得られる状況を考えて見よう．いったんモデルは真の構造を表現するものであるという立場を離れると，モデルを情報抽出のための道具と考える立場が生じる．すなわち，統計的情報処理においてモデルは，われわれが必要な情報を取り出すために，対象に関する理論，過去の知識，現在のデータ，解析の目的など，あらゆる情報を統合して構築すべきことになる．本稿では，このような立場を能動的モデリングと呼ぶことにする．この場合，データが増加すればするほど，より詳細で複雑な構造を表現できるはずである．統計的モデルはそのような可能性を秘めた柔軟なものである必要がある．

統計科学に限らず，知的情報処理の諸分野では近年，様々な種類の情報統合を実現する枠組みとしてベイズモデルが盛んに用いられている．事前確率に関する議論に加えて，線形・正規モデル以外の場合の適用困難さから従来ベイズモデルの利用は極めて限られていたが，MCMC や逐次フィルタなどの計算統計の方法の発展によって，今やその実用性は飛躍的に発展している．

3　能動的な時系列モデリング

以下では筆者がこれまで研究を行ってきた，時系列解析に限定して状態空間モデルによる能動的なモデリングとその応用について紹介する．

時系列を表現する極めて一般的な枠組みとして，状態空間モデルが知られている．線形性を仮定した場合，ℓ 次元の時系列 y_n に対して，状態空間モデルは以下のように表される．

$$x_n = F_n x_{n-1} + G_n v_n, \quad y_n = H_n x_n + w_n \tag{3.1}$$

ただし，x_n は k 次元の状態ベクトル，F_n, G_n および H_n はそれぞれ，$k \times k$ 行列，$k \times m$ 行列，$\ell \times k$ 行列，また v_n と w_n は平均が 0 ベクトル，分散共分散行列がそれぞれ Q_n，R_n で与えられる多変量正規白色雑音で，システムノイズおよび観測ノイズと呼ばれる (片山 2000)．

状態空間モデルの特長は，未知の k 次元ベクトル x_n をカルマンフィルタによって逐次的に推定できることにあり，その計算効率性についてはよく知られているが，モデリング上の意義については明確に認識されていないように思われる．通常の線形回帰モデル

$$y = Hx + \varepsilon \tag{3.2}$$

が分散を除いて k 次元の未知パラメータ x を持つのに対して，状態空間モデルは時間とともに状態 x が変化するので，n 個のデータ $y_1, ..., y_n$ に対して $k \times n$ 個と極めて多数のパラメータを含んでいる．このモデルでは時間 n が増大するとき，n に比例して未知パラメータの個数も増大するので，通常の統計理論が前提とする (パラメータ数)<<(データ数) がなりたつことはない．実際，状態空間モデルは，回帰係数ベクトル x が時間とともに変化する時変係数の回帰モデルと解釈することができる (北川 2005)．この解釈によれば，状態空間モデルの第 1 式（システムモデル）は，係数ベクトルの時間変化に関する制約モデルであり，ここに様々な形の事前情報や知識を投入し，自由にモデルを構築する可能性がある．

図 2 は状態ベクトルのイメージを示したものである．Akaike(1974) は時系列 y_n に対して，現時点 n までの情報（例えば $y_0, \ldots, y_n$）から構成される空間を $Y_{0:n}$，現時点以降の情報から構成される空間を $Y_{n:\infty}$ として，その共通部分を状態 x_n と説明している．この解釈によれば，状態ベクトル x_n は時系列の将来の予測に必要な過去の情報を集約したものと考えることができる．時系列モデリングにおいては，時系列の特殊な構造を利用して特定の時点に個別化された情報が状態の形で表現できるのである．この状態を利用するこ

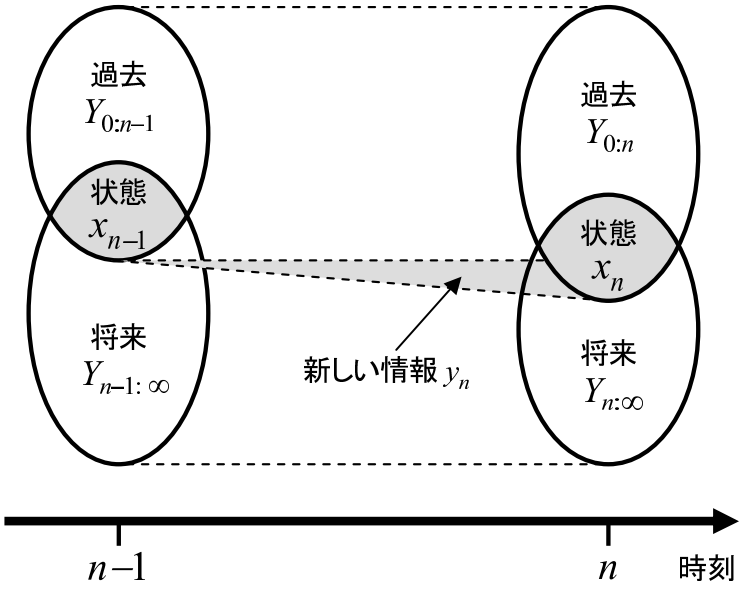


図 2: 状態空間モデルにおける状態の変化

とによって状態空間モデルは巨大なパラメータを持つモデルでありながら，従来のパラメトリックモデルとほぼ同様に取り扱うことが可能である．

状態空間モデルは柔軟な時系列の汎用モデルであるとはいえ，もちろん複雑な現実のシステムをすべて表現できるわけではない．近年は，モデルを非線形とし，またノイズの分布を非ガウス型とした，非線形・非ガウス型状態空間モデル

$$x_n = f(x_{n-1}, v_n), \quad y_n = h(x_n, w_n) \tag{3.3}$$

が利用されている．この一般化された状態空間モデルの状態（パラメータ）推定のために非線形･非ガウス型フィルタが知られているが (Kitagawa 1987)，積分項を含むために高次元の問題への適用は現実的ではない．

そこで，状態ベクトルの条件付分布（予測，フィルタ，平滑化分布）を m 個の粒子によって $\{p_n^{(1)}, ..., p_n^{(m)}\} \sim p(x_n|Y_{n-1})$ $\{f_n^{(1)}, ..., f_n^{(m)}\} \sim p(x_n|Y_n)$ $\{v_n^{(1)}, ..., v_n^{(m)}\} \sim p(v_n)$ と近似することによって，モンテカルロ・フィルタが導出されている (Kitagawa 1996，Doucet et al. 2001，本書 11 章)．モンテカルロ・フィルタ (粒子フィルタ，ブートストラップ・フィルタなどとも呼ばれる) は複雑な非線形高次元システムに適用可能なので，(1) 非ガウス型平滑化 (レベルシフトの検出，非ガウス型季節調整，確率的ボラティリティの推定)，(2) 非線形平滑化 (トラッキング，フェーズアンラッピング)，(3) 信号抽出，

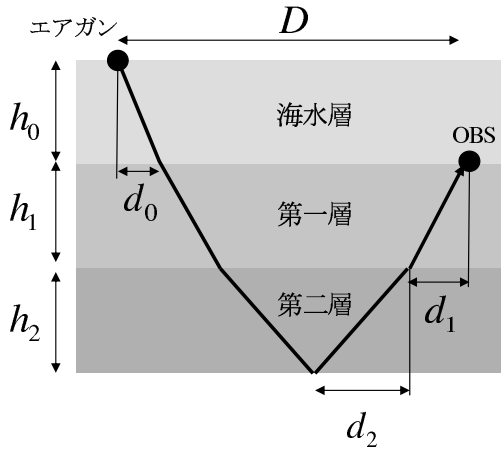


図 3: 地下構造 (2 層の場合) と信号伝播経路

(4) 計数データのモデリング，(5) 自己組織型状態空間モデリング (Kitagawa 1998) など，さまざまな複雑な現実の問題に適用されている (Doucet et al. 2001、本書第 11 章).

今後，さらに大規模な問題への適用のためには，並列計算の利用が不可欠である．その場合，単に計算を並列化するだけでは効率を上げられないことが知られている．並列化モンテカルロ・フィルタのように，並列化に適した計算法の開発も必要である．

4　海底地震計 (OBS アレイ) データの解析

本節では，本稿で紹介した様々な情報を統合して知識を獲得する方法，特に状態空間モデルを利用する情報抽出の一例として，OBS（Ocean Bottom Seismograph, 海底地震計）データの解析例を紹介する．この結果は，北海道大学地震火山研究観測センターの高波鐵夫氏との共同研究であり，Kitagawa et al.(2002) の一部を追加・改良したものである．

実験の目的は，石油等の物理探査を視野に入れた海底の地下構造推定である．そのために，海表面付近で人工的に発生させた振動を，海底に設置した OBS で観測し，その中から地中を通った波動（反射波，屈折波）を検出することによって，構造パラメータ (層厚 h_j，地震波速度 v_j) を推定する (図 3 参照).

実験はノルウェー沖の深さ 1500 − 2000m の海底に 39 個の OBS を設置して，直線上を一定速度で移動する実験船に曳航されたエアガンから 70 秒

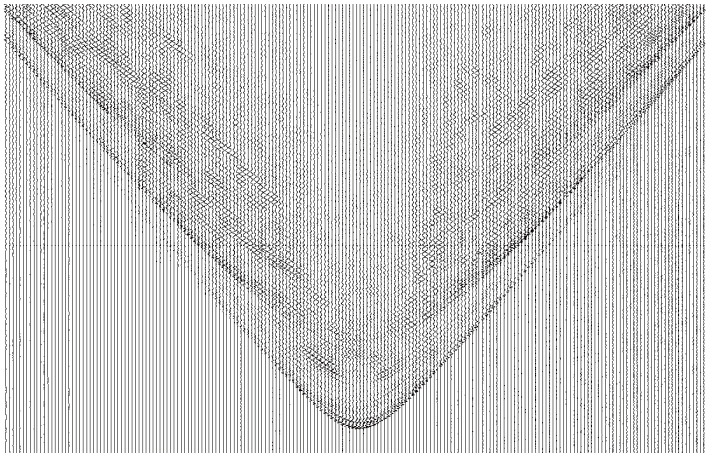


図 4: OBS データ (一部分)

(200m) 間隔で 982 回発振した．このように，人工的な波動を利用することによって，計画的に大量の多変量の時系列を取得することができる．各 OBS では 1/256 秒のサンプリング間隔で 4 チャンネル (東西，南北，上下成分および上下高感度）時系列を観測したので，39 か所の OBS で長さ 17920 の 4 チャンネル時系列が 982 回観測されることになる．982 次元の多変量時系列といっても，同時に観測した時系列ではないことには注意する必要がある．ただし，これによって観測雑音は互いに独立とみなせることから，入力波形を同一にすることによって，ほぼ同一の信号を観測したとみなすことができ，かえって本質的な情報を取り出しやすいという利点もある．

図 4 に，ある OBS で観測された 982 次元 (上下動) 時系列の一部 (200 系列) を示す．時間は下から上方に経過している (約 20 秒)．明瞭な波形が何層かにわたって観測されているが，残念ながらこれらの明瞭な波形のほとんどは，海水中のみを通過した直接波で，地下構造に関する情報は含んでいない．当面の目的からは，これらは単なるノイズといえる．したがって，これらの直接波を分離・除去することによって，海底地中を通過した微小な反射波あるいは屈折波を抽出し，層厚，速度の構造パラメータを推定することが必要になる．ただし，実際の波形は様々なルートを通った波形の重ねあわせとなるので極めて複雑である．また，液体である海水を通過する直接波は P 波 (粗密波) だけであるが，海底に到達すると，S 波への変換も生じ，複数の速度の波が発生する．

まず，エアガンの直下に OBS があり，地下構造が水平層構造をなす最も簡単な場合を想定してみる．海水中，第 1 層，第 2 層を通過する波をそれぞ

表 1: 典型的な経路モデルとエアガン直下における到着時刻

経路モデル	到着時刻
Wave(0^{2k-1})	$(2k-1)h_0/v_0$
Wave($0^{2k-1}1^{\ell}$)	$(2k-1)h_0/v_0 + 2\ell h_1/v_1$
Wave($0^{2k-1}12^m1$)	$(2k-1)h_0/v_0 + 2h_1/v_1 + 2mh_2/v_2$

れ「0」「1」「2」と表すことにすると，直接 OBS に到達した波の経路は 0, 海底と海表面で 1 回ずつ反射して OBS に到達した波は 000 などと，一般に 0^k と表すことができる．k は奇数である．これに対して，第 1 層と第 2 層の境界で反射して OBS に戻ってきた波は 011，第 2 層と第 3 層の境界で反射した波は 01221 などと表現できる．ひとつの層内で複数回反射する波も考えられるので，実際にはさらに複雑である．表 1 は典型的な経路と反射回数 k, 層厚 h_j および速度 v_j を仮定したときの到着時刻を示す．もちろん，このような層構造をなすこと，層内では速度一定であることなどはひとつのモデルである．エアガン直下以外の場所はさらに複雑である (表 2 参照).

以下ではまず，時系列モデルを利用して海水中のみを通った直接波と海底の地中を通過した反射波あるいは屈折波 (以下では簡単のために反射波と総称する）を分離するために，以下のような観測モデルを想定する．

$$y_n = r_n + s_n + w_n \tag{4.1}$$

ただし，r_n は海水中だけを通過した直接波，s_n は地中を通過した反射波，w_n は平均 0，分散 σ^2 に従う観測ノイズである．直接波は水中，反射波は最終的には海底第 1 層を伝播する波で，伝播速度も異なることから各成分は下記のように異なる AR モデルで近似できるものと考えられる．

$$r_n = \sum_{j=1}^{\ell} a_j r_{n-j} + u_n, \quad s_n = \sum_{j=1}^{m} b_j s_{n-j} + v_n \tag{4.2}$$

ただし，u_n，v_n は平均が 0 で分散がそれぞれ τ_{1n}^2，τ_{2n}^2 の正規白色雑音とする．いうまでもなく，直接波や反射波が厳密に AR モデルで表現できると考えているわけではない．異なる AR モデルで表現することによって，変動特性の違いを利用しようとしているのである．

よく知られているように，これらの観測モデルおよび成分モデルは，状態空間モデル

$$x_n = Fx_{n-1} + Gw_n, \quad y_n = Hx_n + \varepsilon_n \tag{4.3}$$

によって表現できる．ただし，

$$x_n = \begin{bmatrix} r_n \\ r_{n-1} \\ \vdots \\ r_{n-m+1} \\ \hline s_n \\ s_{n-1} \\ \vdots \\ s_{n-\ell+1} \end{bmatrix}, \quad F = \left[\begin{array}{cccc|cccc} a_1 & a_2 & \cdots & a_m & & & & \\ 1 & & & & & & & \\ & \ddots & & & & 0 & & \\ & & 1 & 0 & & & & \\ \hline & & & & b_1 & b_2 & \cdots & b_\ell \\ & & & & 1 & & & \\ & & 0 & & & \ddots & & \\ & & & & & & 1 & 0 \end{array}\right], \quad G = \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \hline 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{array}\right]$$

$$H = [\,1 \;\; 0 \;\; 0 \;\; 0 \,|\, 1 \;\; 0 \;\; 0 \;\; 0\,] \tag{4.4}$$

である．したがって，すべての構造パラメータ a_j, b_j, τ_{1n}^2 および τ_{2n}^2 が与えられれば，カルマンフィルタ・平滑化によって，状態ベクトル x_n を推定でき，その成分である r_n および s_n も同時に推定できることになる．

しかしながら, AR 係数 a_j および b_j は全体のデータから前もって推定することが可能であるが，分散 τ_{1n}^2 および τ_{2n}^2 は時々刻々変化する時変パラメータである．そこで，状態ベクトルを新たに $z_n = [x_n, \log\tau_{1n}^2, \log\tau_{2n}^2]^T$ と定義し，分散の対数値の変化はランダムウォーク的に変化すると仮定して

$$\log\tau_{1n}^2 = \log\tau_{1,n-1}^2 + \varepsilon_{1n}, \quad \log\tau_{2n}^2 = \log\tau_{2,n-1}^2 + \varepsilon_{2n} \tag{4.5}$$

というモデルを想定すると，状態ベクトル x_n と分散パラメータの変化を同時に表現する自己組織型の状態空間モデルが得られる (Kitagawa 1998)．このモデルは非線形であるが，モンテカルロフィルタを適用することによって，この拡大された状態 z_n を推定できる．これは，本来の状態ベクトル x_n と時変分散パラメータ τ_{1n}^2 および τ_{2n}^2 を同時に推定できることを意味している．図 7 に，すべての系列に対して，この方法を（独立に）適用して直接波 (左側) と反射波 (右側) を推定し，重ね描きした結果を示す．一見すると，直接波についてはよい推定値が得られているようにみえるが，残念ながら反射波の中には直接波が混入してよい分解が達成できたとは言いがたい．

この失敗の原因は明らかである．これまでの方法は時系列構造だけを用い，空間構造を全く利用しなかったからである．そこで，次に，隣り合った 2 つの時系列間の関係の利用を考えてみる．表 2 にいくつかの代表的な経路モデルについて，その到着時刻を示す．ただし，D はエアガンから OBS までの震央距離，$d_{ij} = v_i h_i / \sqrt{v_j^2 - v_i^2}$，$d_3 = D - d_{03} - 2d_{13} - 2d_{23}$ で各 d_{ij} は j 番目の経路モデルにおける第 i 層内で水平方向に移動する距離を表す．これらは到着時刻の最小化によって決定することができる．

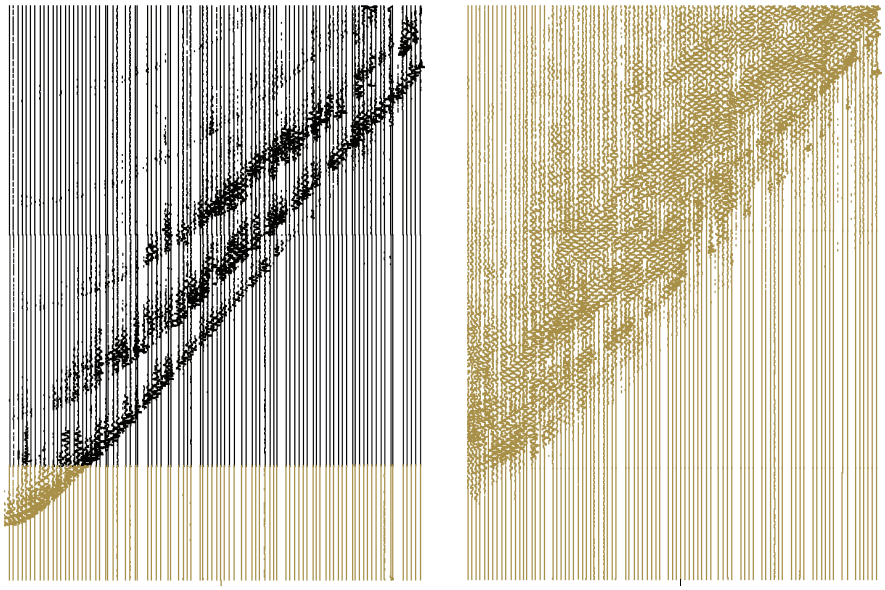


図 5: 時系列モデルにより分離検出された直接波および反射波

表 2: 典型的な経路モデルと到着時刻

経路モデル	到着時刻
Wave(0^k)	$v_0^{-1}\sqrt{k^2h_0^2 + D^2}$
Wave(0^k1)	$kv_0^{-1}\sqrt{h_0^2 + d_{01}^2} + v_1^{-1}(D - kd_{01})$
Wave(0^k121)	$kv_0^{-1}\sqrt{h_0^2 + d_{02}^2} + 2v_1^{-1}\sqrt{h_1^2 + d_{12}^2} + v_2^{-1}(D - kd_{01} - 2d_{12})$
Wave(0^k12321)	$kv_0^{-1}\sqrt{h_0^2 + d_{03}^2} + 2v_2^{-1}\sqrt{h_2^2 + 2d_2^2} + 2v_2^{-1}\sqrt{h_2^2 + d_{23}^2} + v_3^{-1}d_3$

表 3: 代表的な経路モデルと震央距離ごとの到着時点差

	Epicentral Distance (km)				
Path Model	0	5	10	15	20
Wave(0)	1.7	32.2	34.4	34.8	35.0
Wave(000)	0.6	21.6	29.7	32.4	33.5
Wave(00000)	0.3	14.9	24.1	28.8	31.1
Wave(011)	—	10.5	15.4	15.1	15.3
Wave(01221)	—	12.9	14.7	15.1	15.3
Wave(012321)	—	—	10.2	10.2	10.2

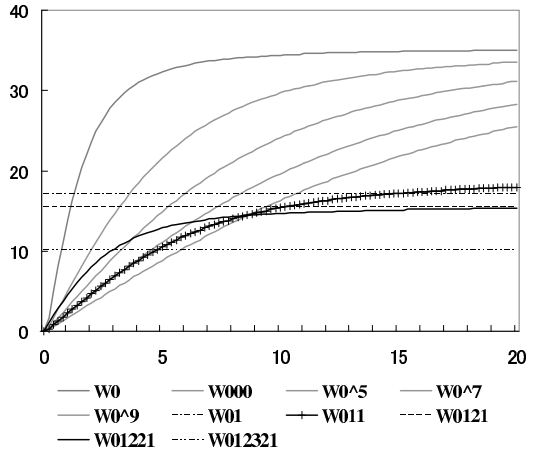


図 6: 代表的な経路モデルの震央からの距離による到着時刻差の変化

この表は地下構造パラメータ (h_j や d_j) が与えられればそれぞれの経路を通った波の到着時刻が計算できることを示しているが，実際にはそれらは未知である．しかしながら，その絶対値は未知でも，その差分で定まる隣接した 2 系列間の到着時刻の差については，観測データの局所的な相互相関関数や，時系列構造だけを用いた方法によって推定されたラフな推定値を利用して求めることができる．

表3および図6はこのようにして求めた経路モデルW の到着時刻差 $\Delta_j(W) = T_j(W) - T_{j-1}(W)$ の変化を示す．ただし，$T_j(W)$ はエアガン直下から j 番目の時系列における到着時刻 (時点) を示す．横軸はエアガン直下からの震央距離 (km)，縦軸は到着時刻差 (時点数，1 秒=256 時点) である．

このとき，j 地点における時刻 n の時系列を $y_{n,j}$ と表すとき下記のような

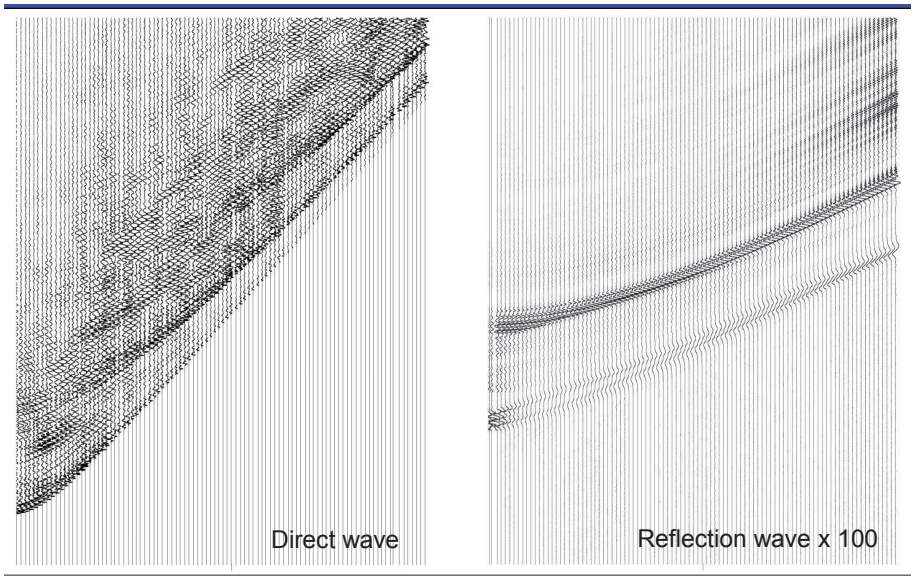


図 7: 分離検出された直接波および反射波

時空間構造モデル

$$y_{n,j} = r_{n,j} + s_{n,j} + w_{n,j} \tag{4.6}$$

を考えることにする．ただし，時系列構造は (6) と同じ AR モデル

$$\begin{aligned} r_{n,j} &= a_1 r_{n-1,j} + \cdots + a_\ell r_{n-\ell} + v^{(r)}_{n,j} \\ s_{n,j} &= b_1 s_{n-1,j} + \cdots + b_m s_{n-m,j} + v^{(s)}_{n,j} \end{aligned} \tag{4.7}$$

を仮定するが，隣接した 2 系列間には経路モデルから得られる以下のような (空間構造) 制約モデルを導入する．

$$\begin{aligned} r_{n,j} &= r_{n-k_j,j-1} + u^{(r)}_{n,j} \\ s_{n,j} &= r_{n-h_j,j-1} + u^{(s)}_{n-j} \end{aligned} \tag{4.8}$$

ただし，W_0 および W_1 はその時空間上に到達する直接波および反射波とするとき，$k_j = \Delta_j(W_0)$，$h_j = \Delta_j(W_1)$ はその到着時刻差を表す．

$z_{n,\,j} = (r_{n,j}, s_{n,j})^T$ とおくとき，下記の条件付分布に関する関係式を利用すると，原理的には時空間フィルタリングおよび平滑化を行うことができる．

$$p(z_{nj}|z_{n-1,j}, z_{n-k,j-1}) = \frac{p(z_{nj}|z_{n-1,j})p(z_{n-k,j}|z_{nj}, z_{n-1,j})}{p(z_{n-k,j-1}|z_{n-1,j})} \tag{4.9}$$

ただし，これを厳密に適用するためには，$(m+\ell)\times p$ (p はチャンネル数で数百) 次元の状態空間モデルの非線形フィルタリングが必要となる．現時点ではこの厳密な計算の代わりに近似式

$$p(z_{n-k,j}|z_{nj}, z_{n-1,j}) = p(z_{n-k,j-1}|z_{nj}) \tag{4.10}$$

を用いた近似フィルタリングを行っている．図 7 にこの方法によって得られた分解を示す．左側が直接波の推定値，右側は検出された反射波で信号が微小なので振幅を 100 倍に拡大して示す．現段階では，期待される多くの反射波は検出されていないが，これまで明らかでなかった Wave(012321) および Wave(00012321) の波形が明瞭に検出されている．

参考文献

[1] Akaike, H.(1973). Information theory and an extension of the maximum likelihood principle. Proc. 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaki eds.) Akademiai Kiado, Budapest, 267–281.

[2] Akaike, H. (1974). Stochastic theory of minimal realization, *IEEE Transactions on Automatic Control*

[3] Akaike, H.(1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain., **AC-19**(6), 667–674.

[4] Akaike, H.(2001)，Golf swing motion analysis: An experiment on the use of verbal analysis in statistical reasoning, Ann. Inst. Statist. Math., **53**, 1–10.

[5] 赤池 弘次 (1995)，時系列解析の心構え，時系列解析の実際 II，朝倉書店，197–203.

[6] Doucet, A., Freitas, F. and Gordon, N. (2001), *Sequential Monte Carlo Mthods in Practice*, Springer, New Nork,

[7] 樋口 知之 (編)(2006)，特集号「予測と発見」，統計数理，第 54 巻，第 2 号.

[8] 片山 徹 (2000)，応用カルマンフィルタ (新版)，朝倉書店，東京.

[9] Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion), *Journal of the American Statistical Association*, **82**, (1987), 1032–1063.

[10] Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space model, *Journal of Computational and Graphical Statistics*, **5**, (1996), 1–25.

[11] Kitagawa, G. (1998). Self-organizing State Space Model, *Journal of the American Statistical Association*, **93**, (1998), 1203–1215.

[12] Kitagawa, G., Takanami, T., Kuwano, A., Murai, Y. and Shimamura, H. (2002). Extraction of signal from high dimensional time series: analysis of ocean bottom seismograph data, *Progress in Discovery Science*, Lecture Notes in Artificial Intelligence, No. 2281, eds. Arikawa, S. and Shinohara, A., Springer-Verlag, Berlin, 449–459.

[13] 北川 源四郎 (2005)，時系列解析入門，岩波書店.

[14] Sakamoto, Y. and Akaike, H. (1978). Analysis of cross-classified data by AIC, Ann. Inst. Statist. Math., **30**, 185-197

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2008 年 5 月
第 III 部 自然・工学と統計科学

第3章 生存時間・再発事象分析：理論と応用

鎌倉稔成[1]

生存時間分析は人間の寿命分析として始まったが、1950 年代から 1070 年代にかけて電子デバイスの信頼性工学で急速に発展した。生命、あるいは機械や電子部品のコンポーネントやアイテムの寿命にいたるまでの長い時間を扱うために、途中までは観測されないという、いわいるデータの不完全性の問題が一つのキーワードになっている。1970 年代以降は医学・薬学系の分野で成熟期を迎え、理論・応用が進んでいるのが現状である。本章では生存時間分析における打切データの扱いを中心としてその基礎を概観し、ワイブルモデル、Cox モデルの扱い、斉らには死亡（故障）の概念を拡張して使われるようになった再発事象分析に至る統計的方法について、応用例を含めて、工学の立場で議論を与える。

[1]中央大学理工学部

生存時間分析と再発事象分析：理論と応用

中央大学・理工学部　鎌倉　稔成

要約：生存時間分析は人間の寿命分析として始まったが，1950 年代から 1970 年代にかけて電子デバイスの信頼性工学の分野で急速に発展した.生命あるいは機会や電子部品のコンポーネントやアイテムの寿命にいたるまでの長い時間を扱うために，途中までしか観測されないという，いわゆるデータの不完全性の問題が 1 つのキーワードとなっている．1970 年代以降は医学・薬学系の分野で成熟期を迎え，理論・応用が進んでいるのが現状である．本稿では，生存時間分析における打切データの扱いを中心としてその基礎を概観し，ワイブルモデル，Cox モデルの扱い，さらには死亡（故障）の概念を拡張して使われるようになった再発事象分析に至るまでの統計的方法について，応用例を含めて，工学上の立場で議論を与えるものである．

1. はじめに

寿命データは信頼性理論における統計解析の中心課題となる対象であるが,動物試験や臨床試験などの生物学，医学・薬学の分野においてもよく見受けられる．また，文字通りの故障，死亡ということばにとらわれずに,ある関心のある事象の生起するまでの時間の分布も寿命データとみることができ,ここで述べる生存時間分析の方法論が適用可能である.基本的な仮定としては,解析の対象となる母集団が何らかの意味で均一であること仮定して,母集団分布に対しての推論をデータに基づいて行う.寿命データが抽出されるところの母集団が何らかの意味で均一であるという条件は重要で,これがくずれてしまうと統計解析で得られた結論がまったく意味のないものになってしまうこともあることに注意しなければならない.

工学の領域で知られている，JIS Z8115 による信頼度の定義では,「アイテムが与えられた条件で既定の期間中，要求された機能を果たすことができる確率」としているが，これも条件をきちんとして，対象を均一なものに制限していることに他ならない．規定外の環境下での機械・装置の使用からは意図した信頼度は期待できないのである．

しかしながら,環境条件のように寿命に大きく影響を及ぼす因子に関して十分な情報が得られ,その因子と寿命分布との関連づけができれば,ある程度の環境の変動に対しても寿命分布が予測可能になってくるはずである．また，新しく製品評価を行うときにも，クリティカルに効いているような環境因子が発見できれば,設計の段階にその情報を活かして寿命の長い製品が開発でき可能性がある．

寿命分布と環境要因とを関連づけたモデルとして，応用上，きわめてよく用いられているモデルに Cox(1972)の比例ハザードモデルがある．比例ハザードモデルは，後で定義するハザードを

対して共変量（説明変数）によって説明しようという，一種の回帰型のモデルであり，モデルの柔軟性から多くの分野で用いられている．

近年は，関心のある事象が繰り返し起こる，再発事象の分析のためのモデリングについても研究が進んでいる（Cowling et al.，2006；Cook and Lawless，2007，Kalbfeish and Prentice，1980）では多重故障(multiple failures)というような名称で再発事象を扱っているが，分析の対象が多岐にわたり，その意味ももはや故障という名称がふさわしくないという意味からも再発事象(recurrent event)という用語が用いられるようになってきた．また，クレジット・デリバティブの分野でも信用事由の到達リスクモデル化のために研究およびその利用が始まっている（Shonbuncher，2003；望月衛訳，2005）．

2. 生存時間分析の基礎事項

ここでは，生存時間分析のための基本的用語等について述べる．

2.1 生存関数（信頼度関数），ハザード関数

分布関数を１から引いた関数を生存関数（信頼度関数）という．場合によっては，生存時間分布関数という．これをR(t)で表す．また，分布関数そのものを不信頼度関数と呼ぶ．つまり，

$$R(t) = 1 - F(t)$$

である．さらに，故障の起こり方を特にその変化に着目してみるときには，分布関数の微分形の１つとしてのハザード関数（故障率関数）が使われる．確率密度関数を$f(t)$（ここでは，確率変数が絶対連続の場合について述べるが，離散型でも同様な扱いが可能である(Cox and Oakes，1984））），ハザード関数をλ(t)とすると，

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0}\frac{P\{t < T \le t+h \mid T > t\}}{h} \\ &= \lim_{h\to 0}\frac{1}{h}\frac{P\{(t < T \le t+h) \cap (T > t)\}}{P\{T > t\}} \\ &= \lim_{h\to 0}\frac{P\{t < T \le t+h\}}{h}\Big/P\{T > t\} \\ &= \frac{f(t)}{R(t)} \end{aligned} \tag{2.1}$$

のように，確率密度関数を信頼度関数で割ったものに等しい．また，上式の最右辺は，

$$\frac{f(t)}{R(t)} = -\frac{d}{dt}\ln R(t) \tag{2.2}$$

のようにも表すことができるので，逆に信頼度関数はハザード関数を用いて表すことができる．

$$R(t) = \exp\left\{\int_0^t \lambda(u)du\right\} \tag{2.3}$$

特に，上の式の指数関数中にある，ハザード関数を 0 から t まで積分した t の関数を累積ハザード関数といい，Λ(t)で表すことにする．

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.4}$$

2.2 パラメトリック分布

これまでいくつかの寿命分布に使われる用語について述べてきたが，ここでは，信頼性工学の分野でもっともよく利用されている，パラメトリックな分布の代表格である指数分布およびワイブル分布について議論する．指数分布，ワイブル分布それぞれの確率密度関数ではそれぞれ次のようになる．

$$f(x) = \lambda e^{-\lambda t} \tag{2.5}$$

$$f(t) = \frac{m}{\eta}\left(\frac{t}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{t}{\eta}\right)^{m}\right\} \tag{2.6}$$

指数分布のパラメータ，λは故障率と呼ばれる．ワイブル分布のパラメータｍは形を表す形状パラメータ，ηは尺度の変更をする，尺度パラメータである．この他，いずれの分布の場合にも，位置パラメータγを入れることもあるがここでは，議論しない．ワイブル分布においてｍ=1 とすれば，λ=1/ηとおいて指数分布となる．つまり，指数分布はワイブル分布の特別な場合ということになるので，以後の議論では，ワイブル分布として説明していく．

３． ワイブル解析

観測された故障寿命のデータにワイブル分布をあてはめてパラメータを求めるにはいくつかの方法がある．ここではその代表的なものについて簡潔に説明する．

3.1 ワイブル確率紙（真壁，1966）

あまり，多くないデータをワイブル分布の適合性を見ながらあてはめるのに適している．ワイブル確率紙の原理は次の通りである．ワイブル分布の信頼度関数 R(t)=1-F(t)の対数をとると，

$$\ln\{1 - F(t)\} = -\left(\frac{t}{\eta}\right)^{m}$$

さらに，両辺にマイナスを付けてもう一度対数をとると，

$$\ln\left[-\ln\{1 - F(t)\}\right] = m\ln t - m\ln\eta \tag{3.1}$$

となる．ここで，左辺を Y，右辺で $\ln t$ を X，$-m\ln\eta$ を C とおけば，

$$Y = mX + C$$

と１次式となることがわかる．つまり，ワイブル分布の形状パラメータは，上の直線の傾きとして，また，尺度パラメータηは，

$$\eta = \exp\left(-\frac{C}{m}\right)$$

として得られることになる．実際には，真の分布関数は未知であるので，何らかの方法で推定しなければならない．もし，故障時間のデータが$t_1, t_2, ..., t_n$のように n 個与えられ，それに対応する分布関数の値が推定できれば，上に述べた直線の傾きを決めることによってワイブル分布のパラメータが推定できることになる．分布関数の推定のし方は後で述べるが，このように直線をあてはめることによってワイブル分布のパラメータを図の上で求められるようにしたのがワイブル確率紙である．

3.2 分布関数のノンパラメトリックな推定法

ワイブル確率紙を利用するには，分布関数を推定しなければならないが，このときには，特定の分布を仮定しないノンパラメトリックな推定値を利用する．この推定法には種々の方法があるが，図の上で，作図によってパラメータを求めるのであるから，あまり神経質になる必要はない．代表的な方法について述べる．ただし，データは

$$(t_i, \delta_i) \quad (i = 1,2,...,n)$$

$$\delta_i = \begin{cases} 1 & (failure) \\ 0 & (censored) \end{cases}$$

であるものとする．ここで δ_i は打切インディケーターと呼ばれ，i 時点で故障が観測されれば１，あるいは，i 時点まで故障が起こらずに生きていたという，故障に関しては不完全な情報であれば０である．　こうした，データに打切を含むという，情報の不完全性は信頼性のデータ解析に特有である．打切の起こり方でデータを分類すると次の図１，２，３のようになる．

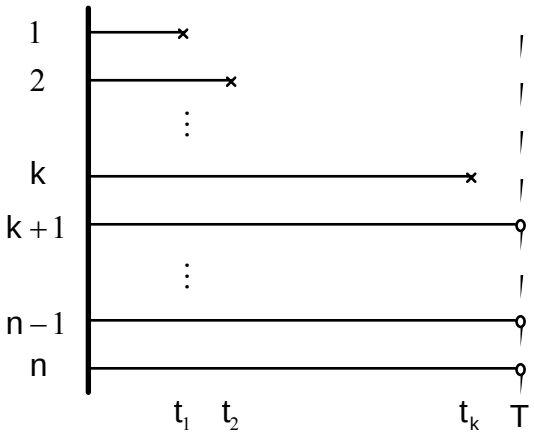


図１．定時打切（時間打切）（Type Ⅰ 打切）

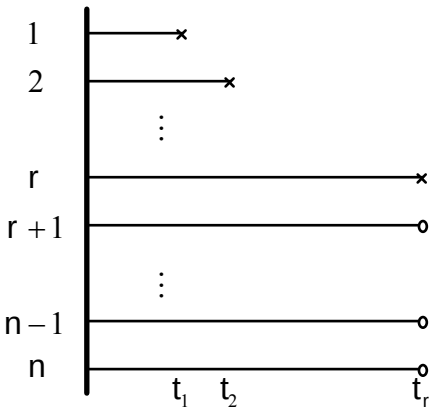


図２．定数打切（故障打切）（Type II 打切）

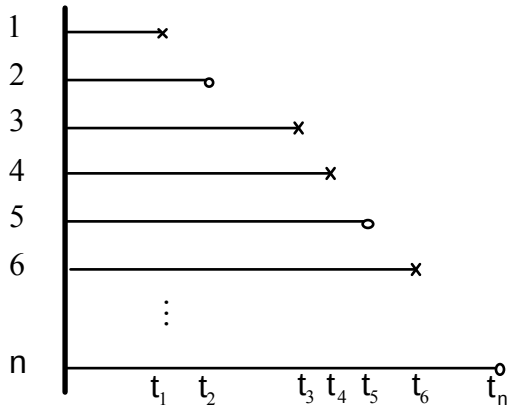


図３．ランダム打切

上の図から明らかなように定時打切では,寿命試験等で時間上の制約から一定時間 T で試験をやめてしまうものであり，定数故障打切では，推定の精度をある一定値以上にするために最初からｒ個の観測が観測されたら試験を打ち切ってしまうというものである．３番目のランダム打切では打切の起こり方がまったくランダムで，寿命試験のデータよりも，フィールドデータとして得られることが多い．医学・薬学の分野では，ランダム打切データは脱落，観測中断，フォロー・アップ不能，他の競合因子での死亡のように種々の打切の概念が混在している．

(1) 経験分布関数

打切がまったくないもの，つまり，完全データであるとき経験分布関数

$$\hat{F}(t_{(i)}) = \frac{i}{n} \tag{3.2}$$

で与える．ただし，$t_{(i)}$は故障寿命データを生起した順番に並べたときのｉ番目の値である．この推定量を使ったときの欠点としてはｎ番目の値がプロットできないということである．

(2) 平均ランク法

$$E\left\{F(T_{(i)})\right\}=\frac{i}{n+1}$$

という一般的な関係式が成り立つのでこれを利用して

$$\hat{F}(t_{(i)})=\frac{i}{n+1} \tag{3.3}$$

で与えようというものである.

(3) メディアンランク法

上の平均ランク法に対して $F(T_{(i)})$のメディアンを利用しようというものであるが, 正確な式は簡単でないので, 表を利用するかあるいは近似式が用いられている. たとえば, 次の近似式が知られている(Nelson, 1982).

$$Med\left\{F(T_{(i)})\right\}\cong\frac{i-0.3}{n+0.4} \tag{3.4}$$

(4) カプラン＝マイアー法 (Kaplan and Meier, 1958)

これまでの方法はいずれも, 完全データ, つまり, 打切がない場合の議論であったが, 打切が入ってくる場合には話はやや複雑になる. 簡単のため, データは次のようにまとめられているものとする.

表 1. 故障データ整理

故障時点	t_1	t_2	$\cdots$	t_j	$\cdots$	t_k
故障数	d_1	d_2	$\cdots$	d_j	$\cdots$	d_k
残存数	n_1	n_1	$\cdots$	n_j	$\cdots$	n_k

ここでいう残存数は, 故障時点の直前で生き残っているアイテムの数である. 信頼度関数 R(t) のカプラン＝マイアー推定量は次のように表される.

$$\hat{R}(t)=\prod_{i:\ t_i\le t}\left(1-\frac{d_i}{n_i}\right) \tag{3.5}$$

故障時点に同順位のものがなく, $(t_i,\delta_i)\ (i=1,...,n)$ の形式でデータがまとめられている場合には, 次のようになる.

$$\hat{R}(t)=\prod_{i:\ t_i\le t}\left(1-\frac{1}{n-i+1}\right)^{\delta_i} \tag{3.6}$$

カプラン＝マイアー推定量はあまり強くない条件で一致性および漸近正規性を持つという優れた性質を持っており, 分布の推測を行うにはきわめて便利な性質であるといえる(Lawless, 1982).

(5) 累積ハザード法

累積ハザード関数を,

$$\hat{H}(t)=\sum_{i:t_i\le t}\frac{d_i}{n_i} \tag{3.7}$$

で推定するものであり,

$$\hat{S}(t)=\exp\left\{-\hat{H}(t)\right\} \tag{3,8}$$

となる.

(6) ジョンソン法

打切データを含む場合に, 観測中断されたアイテムの実際の故障は未知であるが, その故障の起こり方によって, その打ち切られた時点よりもあとに観測された故障寿命の故障順位は変わり得るのでその平均化された値で代用するということが考えられる. たとえば,

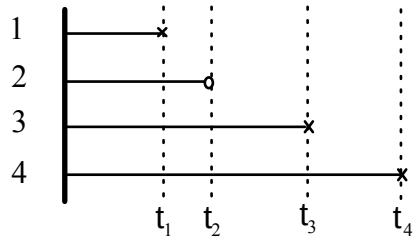


図４. ジョンソン法説明のための図

のように４個のアイテムの故障記録があった場合, ２番目のアイテムについては t_2 まで観測されているが, それから後はいつ故障したのかがわからない. ３番目のアイテムについて考えると, アイテム２が区間$(t_2,t_3]$で故障する場合には故障順位は３, $(t_3,t_4]$で故障する場合には２, (t_4,∞)で故障する場合は２となる. このとき故障順位をその平均の値, (３＋２＋２)／３＝７／３で置き換える. これをメディアン・ランク法のｉに入れて分布関数を推定する. 故障データがたくさんある場合にはこの計算は煩わしいが, 次のようにして再帰的に平均故障順位を求める式が与えられている.

$$J_{i+1}=J_i+\frac{(n+1)-J_i}{\#\left\{k:\,t_i>\textit{previous censored time}\right\}+1}$$

$$J_1=\begin{cases}1 & (\delta_1=1)\\ \dfrac{(n+1)}{\#\left\{k:\,t_i>\textit{previous censored time}\right\}+1}\end{cases} \tag{3.9}$$

3.3 モーメント法

モーメント法では,完全データの場合について１次と２次の標本モーメントと母モーメントを等しくするようにパラメータを決定する方法である．ここでワイブル分布の平均寿命は,

$$E(T)=\eta\Gamma\left(\frac{1}{m}+1\right) \tag{3.10}$$

分散は,

$$Var(T)=\eta^2\left\{\Gamma(1+2/m)-\Gamma^2(1+1/m)\right\} \tag{3.11}$$

であることから，標本平均，標本分散

$$\bar{t}=\frac{1}{n}\sum_{i=1}^{n}t_i$$

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(t_i-\bar{t})^2$$

と等しいとおいて m と η について解けばよいことになるが,このままでは簡単に解けないので次のように変動係数で考えると,

$$CV(T)=\frac{\sqrt{Var(T)}}{E(T)}=\frac{\sqrt{\Gamma(1+2/m)-\Gamma^2(1+1/m)}}{\Gamma(1+1/m)} \tag{3.12}$$

となり，η に関係ない量となる．したがって，標本変動係数と上式の母変動係数を等しくなるようにまず，m を決定し，これを $\hat{m}$ とおいて，次式によって η を推定する．パラメータ m を既知とすれば，t_i^m は尺度パラメータ $\theta=\eta^m$ を持つ指数分布となるので，指数分布の母平均である θ を $\{t_i^m; i=1,...,n\}$ の標本平均と等しいとおいて，η について解くと次式が得られる．

$$\hat{\eta}=\left(\frac{1}{n}\sum_{i=1}^{n}t_i^{\hat{m}}\right)^{1/\hat{m}}$$

m を求める際には，次のようなグラフを利用すると便利である．

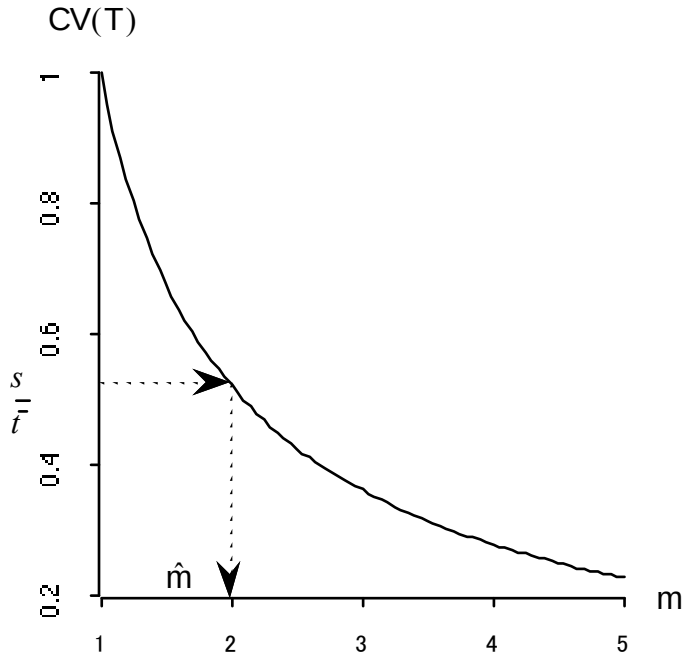


図５．モーメント法の計算ための概要

3.4 最尤法と平均のパラメータ推測

尤度関数

$$L(m,\eta)=\prod_{i=1}^{n}\left\{\frac{m}{\eta}\left(\frac{t_i}{\eta}\right)^{m-1}\right\}^{\delta_i}\exp\left\{-\left(\frac{t_i}{\eta}\right)^{m}\right\} \tag{3.13}$$

を最大にするようにmとηを決めるものである．通常は，これと同値であるが，対数尤度，

$$LL=\log L(m,\eta)=-m\sum_{i=1}^{n}\delta_i\log\eta+\log m\sum_{i=1}^{n}\delta_i$$
$$+(m-1)\sum_{i=1}^{n}\delta_i\log t_i-\left(\frac{t_i}{\eta}\right)^{m}$$

を最大化する．解は最大化の適当なアルゴリズム，たとえば，ニュートン法により計算機を用いて求める．Menon(1963)は 1/m の簡便な一致推定量を提案し，漸近不偏性を示した．Cohen(1965)(Cohen and Whitten, 1988)はワイブル分布の変動係数を利用した実用的な近似推定量を提案している．また，長塚・鎌倉(2004)はW変換を提案し実用的な形状パラメータ推定を与えている．

ワイブル分布における統計的推測に関しての論文はきわめて多岐にわたり，いまなお研究の途

中にある．Kamakura (2004)ではワイブル分布の平均の推定について比較しており，標本平均 $\overline{X}$ の，母平均の最尤推定量 $\tilde{\mu}$ に対する漸近効率を計算し，標本平均の実用性を検討している．漸近効率は，

$$\begin{aligned} ARE(\overline{T}) &= \frac{nA\,\mathrm{var}(\tilde{\mu})}{nA\,\mathrm{var}(\overline{X})} \\ &= \frac{6}{m^2\pi^2}\cdot\frac{1}{CV^2}\left[\frac{\pi^2}{6}+\left\{c-1+\psi(1+1/m)\right\}^2\right] \end{aligned} \tag{3.14}$$

と計算され，ここに，c はオイラー定数，$\psi(\cdot)$ は digamma 関数である．表２のように標本平均の漸近効率が高いことがわかる．特に $m \geq 0.5$ では 90%以上の効率があることが示されている．

表２．標本平均の最尤推定量の漸近効率

m	ARE	m	ARE	m	ARE
0.1	0.0018	1.1	0.9997	2.1	0.9980
0.2	0.1993	1.2	0.9993	2.2	0.9981
0.3	0.5771	1.3	0.9988	2.3	0.9982
0.4	0.8119	1.4	0.9984	2.4	0.9983
0.5	0.9216	1.5	0.9981	2.5	0.9984
0.6	0.9691	1.6	0.9980	2.6	0.9984
0.7	0.9890	1.7	0.9979	2.7	0.9985
0.8	0.9968	1.8	0.9979	2.8	0.9985
0.9	0.9995	1.9	0.9979	2.9	0.9985
1.0	1.0000	2.0	0.9980	3.0	0.9986

表２は形状パラメータの値が小さいときには，漸近効率が非常に悪いことを表しているが，有限の標本サイズについては， MSE で比較すると，シミュレーション結果では図６のように非常におもしろい結果となる．標本サイズが 20 より小さいとき，形状パラメータが１より小さくなるところでは，標本平均が最尤推定量に比べて非常によくなっている．標本サイズが増えると，たとえば，n=50，100 では，m=0.5 以上では MLE の方がいいが，さらに m が小さくなると，逆に標本平均の方が優ってくることがわかる．つまり，有限の標本サイズでは，どんなにサイズを増やしても，十分小さな m に対しては，MSE の評価規準で見て，標本平均が優れていることになる．

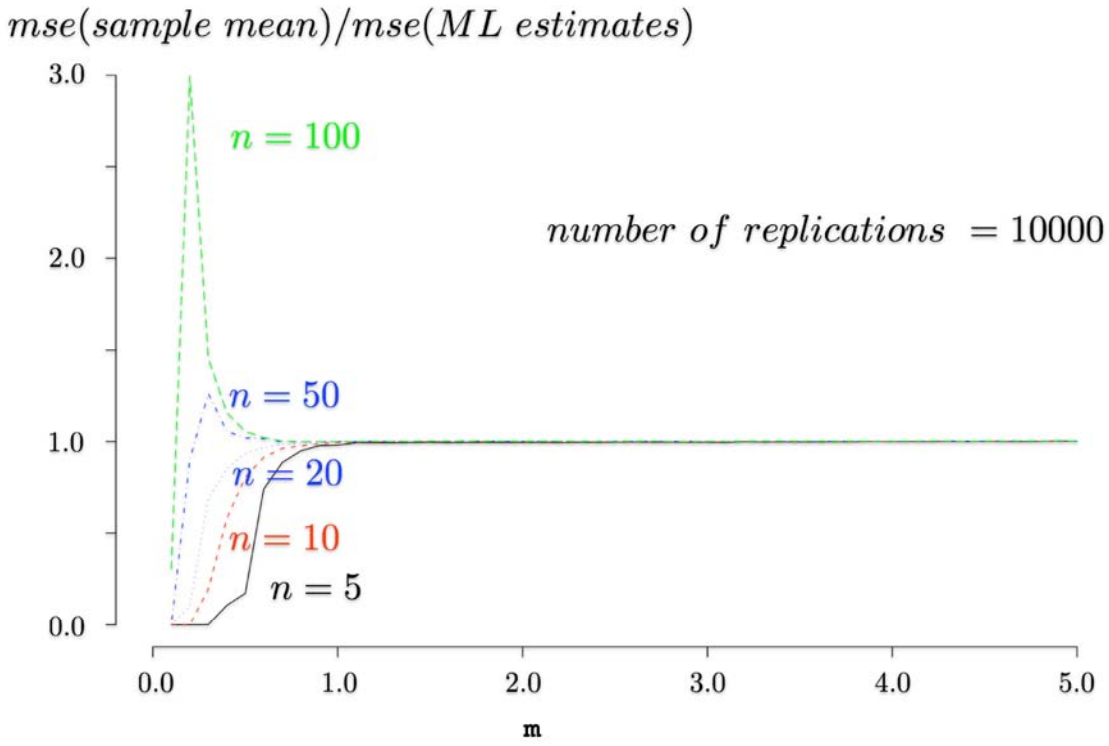


図６．有限標本の MSE の比較（標本平均と最尤推定量）

４．Cox モデル

Cox モデルは比例ハザードモデルとも呼ばれ，Cox(1972)が寿命分布と環境要因などを含む，故障（死亡）時間に付随して得られる共変量$\mathbf{z}$の間に，ハザード

$$\lambda(t;\mathbf{z}) = \lambda_0(t)\exp(\mathbf{z}\beta) \tag{4.1}$$

の関係を与え，その推定法を開発した．ここに，$\lambda_0(t)$は基準ハザード関数(baseline hazard function)と呼ばれ，任意の非負の関数である．共変量には，大きさpのベクトルを考えており，

$$\mathbf{z} = (z_1, z_2, ..., z_p)$$
$$\beta = (\beta_1, \beta_2, ..., \beta_p)'$$

である．回帰係数βの推定法はいろいろ考えられるが，最初に Cox が与えた部分尤度(partial likelihood)による方法，数学的に明快な Kalbfleish and Prentice (1973)の周辺尤度(Marginal likelihood)による方法が代表的である．いずれにせよ，基準ハザード関数に関する推定を行わないでβの情報を得ようというもので，回帰係数の推定と検定は主として寿命に与える要因分析が目標となる．これに対して，予測とういう観点からすれば，基準ハザード関数は必要であり，これに対してのノンパラメトリックな推定が使われるが，必ずしも満足の得られる結果とはならない．

部分尤度は比例ハザードモデルに従う寿命分布の母集団から，

$$\left\{(t_{(i)}, \mathbf{z}_{(i)}) : i = 1,...,k\right\}$$

という形のデータが得られているものとする．ただし，

$$t_{(1)} < t_{(2)} < \ ... < t_{(k)}$$

は死亡時点を小さい順に並べた順序統計量である．$t_{(i)} - 0$ 時点でのリスク集合 $R(t_{(i)})$ とし，$R(t_{(i)})$ が与えられ，$t_{(i)}$ 時点で１つの死亡が起こるとしたときの，(i) が死亡する条件付き確率は，

$$\begin{aligned} & P\left\{(i)\ fails\ at\ t_{(i)}\ |one\ failure\ at\ t_{(i)}\ and\ R(t_{(i)})\right\} \\ &= \frac{\lambda_0(t_{(i)})\exp(\mathbf{z}_{(i)}\beta)}{\sum_{l \in R(t_{(i)})} \lambda_0(t_{(l)})\exp(\mathbf{z}_{(l)}\beta)} \\ &= \frac{\exp(\mathbf{z}_{(i)}\beta)}{\sum_{l \in R(t_{(i)})} \exp(\mathbf{z}_{(l)}\beta)} \end{aligned} \tag{4.2}$$

となる．これを $i = 1, 2, ..., k$ について掛け合わせたものが Cox の部分尤度関数，

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\mathbf{z}_{(i)}\beta)}{\sum_{\ell \in R(t_{(i)})} \exp(\mathbf{z}_{(\ell)}\beta)} \tag{4.3}$$

である．

他方，周辺尤度法は部分尤度の１つ justification として Kalbfleish and Prentice（1973）で扱われている．これは，$u = g^{-1}(t)$（g は狭義単調増加，微分可能）としたとき t がハザード関数 $\lambda_0(t)\exp(\mathbf{z}\beta)$ を持てば，u はハザード関数，$\lambda_0(g(u))g'(u)\exp(\mathbf{z}\beta)$ を持つことになる．つまり，t の単調増加な変換は基準ハザード関数の変化でしかない．よって，β の推定には，t のランクのみが影響するので，データから得られるランクの確率を t に関して周辺をとることによって計算する．

$$\begin{aligned} P\left\{\mathbf{r};\beta\right\} &= P\left\{\mathbf{r} = \left[(1), ..., (k)\right];\beta\right\} \\ &= \int_0^\infty \int_0^\infty \cdots \int_{t_{(k-1)}}^\infty \prod_{i=1}^{k} f(t_{(i)}; \mathbf{z}_{(i)})\, dt_{(k)} \cdots\, dt_{(1)} \\ &= \frac{\exp\left(\sum_{i=1}^{k} \mathbf{z}_{(i)}\beta\right)}{\prod_{i=1}^{k} \sum_{l \in R(t_{(i)})} \exp(\mathbf{z}_{(l)}\beta)} \end{aligned} \tag{4.4}$$

この他，基準ハザード関数をノンパラメトリックなモデルにして，死亡時点間で区分的に定数を仮定すると，Breslow(1974) の方法が得られる．データにタイ（同順位）がない場合は同一の

尤度を与えるが．タイがある場合には，それぞれの方法で尤度が異なり，したがって，推定された回帰係数βの推定量も異なる．推定量の比較に関しては，Kamakura and Yanagimoto (1983) がある．計算量の点から見れば，タイが多い場合には周辺尤度法が最も大変で，実際のデータでは観測精度のうえで，タイが現れることを考えると実用的評価は低い．また，実際のデータ解析の場で使用される例は少ない．部分尤度法もタイが多くなると計算が困難であり，Efron (1977), Farewell and Prentice (1980), Yanagimoto and Kamakura (1984) が考えられている．

回帰係数の推定については，いずれの方法を用いても推定方程式を陽に解くことはできないので，Newton-Raphson 法などの最適化のアルゴリズムを反復的に用いることになる．つまり，一種の最尤推定値を計算することになるのであるが，部分尤度法を用いる限りでは，通常の意味での最尤推定量にはならない．しかしながら，MLE 同様に漸近正規性が成り立つことが証明されている．

比例ハザードモデルでは，任意の２つの共変量$\mathbf{z}_1$と$\mathbf{z}_2$に対して，

$$\lambda(t;\mathbf{z}_1) \propto \lambda(t;\mathbf{z}_2) \quad for\ any\ t \in (0,\infty)$$

が成り立つ．要因によってはこの比例性が成立しないものが考えられる．つまり，本質的に基準ハザードが異なるわけである．これを補正するために，

$$\lambda_j(t;\mathbf{z}) = \lambda_{0j}(t)\exp(\mathbf{z}\beta) \ \ (j = 1,...,q)$$

のように層別してモデルを拡張することが可能である．このとき，部分尤度は，

$$L(\beta) = \prod_{j=1}^{q} L_j(\beta) \tag{4.5}$$

となる．層別すれば，Newton-Raphson 法が収束しやすくなるが，いたずらに層を多くとると，βの推定の立場からは効率が落ち，好ましくない．

５．　再発事象データのモデリング

５.1 確率過程

非修理系の信頼度の評価は通常，故障が起こるまでの時間の分布を解析することによって行なわれる．これに対して修理系では，系を修理することによって繰り返し，使用に供するので故障が複数回ということがあり，これまでの寿命分布の解析方法だけではうまく解析ができない．故障の起こり方が時間とともにどのように変動するか，故障強度にトレンドがあるか等が解析の目的になってくる．

こうした故障の起こり方の時間変動を捉えるために確率過程の概念を導入する．確率過程とは

時間軸上に並んだ，ランダムに変動する，ある変数の集まりのことである．ある変数とは，もし，観測者が故障の起こり方に関心があれば，時刻ｔまでの故障数としたり，ソフトウェアに含まれているバグ数の変動に関心があれば，バグ数ということになる．共変量過程としては，たとえば寿命に影響を与えると考えられる環境要因としての，温度，圧力，湿度等があげられる．この節では，一般的な確率モデルの議論については触れないで，時刻ｔまでの故障数の時間変化に着目した確率モデルについて述べる．

5．２ 故障強度関数

５．２．１ 点過程とポアソン過程

一般に，ある空間上にランダムに配置された点の位置を示す集合を点過程という．ここでは，空間として１次元の時間軸を考え，時間軸上に定義された故障時点のみからなる集合について議論を与える．つまり，０と１の値を持つ確率過程ということになる．

時間軸上の点過程の確率モデルを与えるにあたって，時刻ｔまでの累積故障数 N(t)を考えると便利である．これを計数過程という．点過程と計数過程の関係は次の図の通りである．

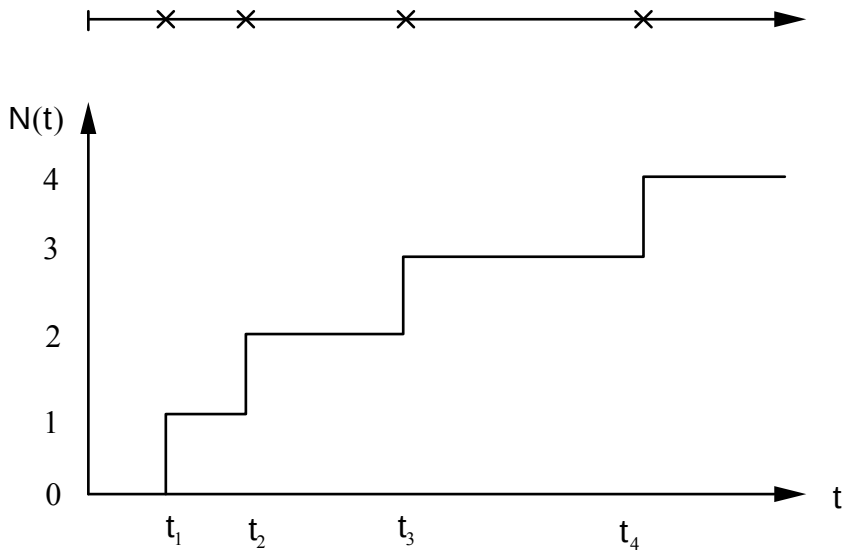


図７．計数過程の例

ここで，N(t)が完全に故障の情報を保持していることに注意する．つまり，N(t)から逆に故障事象の点過程を再現できるのである．このN(t)を用いて修理系の故障強度は，

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{P}\left\{\mathrm{N}(t+\Delta t) - \mathrm{N}(t) \geq 1 \,|\, \mathrm{H}(t)\right\} \qquad (5.1)$$

と定義される． 上の定義の中に現われるH(t)はN(t)のｔまでの履歴である．また，この計数過程のジャンプの幅を１以上としてあるが，以下に述べるポアソン過程を仮定すれば，１に等しいとしてよい．ポアソン過程は計数過程の特別な場合で，あまり無理のない次の３つの条件が必要である．

i) N(0)=0

ii) 任意の２つの交わらない区間について，それぞれの区間の故障総数は互いに独立である．これを独立増分過程 (independent increment)という．

iii) 任意の区間の故障数はポアソン分布に従う．

上の３番目の条件で，ポアソン分布のパラメータについて説明する．任意の区間を[s, t]とし，ポアソン分布のパラメータをΛ(s,t)とすれば，この区間にｎ個の故障が生起する確率は，

$$\mathrm{P}\{\mathrm{N(t)}-\mathrm{N(s)}=n\}=\frac{\Lambda(s,t)^n e^{-\Lambda(s,t)}}{n!} \quad (n=0,1,\cdots) \tag{5.2}$$

で与えられる．パラメータΛ(s,t)は，故障強度関数から計算され，次の関係がある．

$$\Lambda(s,t)=\int_s^t \lambda(t)\,dt \tag{5.3}$$

ここで，簡単のため，Λ(0,t)をΛ(t)で表すことにすれば，N(t)の期待値は

$$\begin{aligned}\mathrm{E}[N(t)] &= \sum_{n=0}^{\infty} n\mathrm{P}\{\mathrm{N}(t)=n\} \\ &= \Lambda(t)\end{aligned} \tag{5.4}$$

と表される．したがって，故障強度関数を N(t)の期待値の微分として次のように表すことができる．

$$\lambda(t)=\frac{d}{dt}\Lambda(t)=\frac{d}{dt}\mathrm{E}[\mathrm{N(t)}] \tag{5.5}$$

ポアソン過程において，λ(t)が時間に依存しないで一定の値，λのとき，定常ポアソン過程といい，時間に依存する場合を非定常ポアソン過程という．

５．２．２　ノンパラメトリックな故障強度関数の推定

故障強度関数の種々なモデルを仮定する前に,ノンパラメトリックな故障強度関数を求めてみることにする．これは，得られたデータを複雑なモデルを仮定しないで，どの辺に故障の現われ方が強く起こっているかを素直に解釈するのに役立つ．まず，データは，$(t_1, t_2, \cdots, t_n)$のように故障時点として与えられているものとして考える．このとき，

$$\lambda(t) = \lambda_i \quad (t_{i-1} < t \le t_i; i = 1,2,\cdots,n) \tag{5.6}$$

と，各故障時点の間で故障強度が一定であるとする．ただし，$t_0 = 0$とする．最尤法によって各パラメータを求めると，

$$\hat{\lambda}_i = \frac{1}{t_i - t_{i-1}} \quad (i = 1,2,\cdots n) \tag{5.7}$$

が得られる．詳しい導出については，Snyder and Miller(1991)を参照されたい．

5.3 信頼度成長のモデル

ソフトウェアのデバッギングを行なって製品として出荷するときには,十分にバグが取れていることが重要である．バグが見つかることと故障とを対応づけると，故障の起こり方が少なくなる傾向，つまり，信頼度成長が十分に確認されることが必要である．また，新製品の開発の段階でも，新たに故障に対処することによって信頼度成長について調査しなければならない．また，逆に長期間にわたって機械あるいはシステムを使用する場合，修理を行なっても，故障間隔が短くなっていく傾向があることが確認されるときには全体を新しいシステムに取り替えるといったことも必要である．この場合，成長と逆で負の成長ということになる．

バグ数のモデルについては,ソフトウェアに含まれるバグ数を有限とした様々なモデルが考えられているが，ここでは，故障といった概念の中で捉えるので故障強度関数によって規定されたモデルのみについて述べる．

5.3.1　確率モデルによる信頼度成長の評価

故障強度関数のモデルとしてよく知られているのは，ワイブル過程モデルである．ワイブル過程モデルでは，故障強度関数として，

$$\lambda(t) = \beta m t^{m-1} \tag{5.8}$$

を仮定している．上の式より，これをｔについて積分したものは，ワイブル分布の累積ハザードと一致する．ただし，ワイブル分布とは別の概念を扱っていることに注意する．しかしながら，１番目だけの故障に着目すれば，ワイブル過程から生成される故障時間はワイブル分布に従う．ここでは，修理を伴う系を扱うため，２番目，３番目とそれ以後の故障についても解析の対象とする．ｍ が１のとき，(5.8)式は，一定の値βをとり，定常ポアソン過程となる．このとき，故障時間間隔の分布は指数分布となる．これに対して，m<1 のときは時間とともに故障強度は減少する．別の言い方をすれば信頼度成長をしているということになる．逆に，m>1 のときは故障強度は増加する.データからｍの値がこの３つのいずれに分類されるかということを統計的に結論付けることに意味がある．つまり，帰無仮説としては，$H_0 : m = 1$，対立仮説として$H_A : m < 1$を考えている．

5.3.2　モデルのパラメータの推定と検定

データとしては,区間$(0, t_0]$に$0 \le t_1 \le \cdots t_n \le t_0$のｎ個の故障データが観測されたものとする．$t_0$は観測中断時間を示す．ｎ個の故障時間データが得られたら観測を中断するという,いわゆる,定数打切（故障打切）データの場合には，若干扱いが異なるので注意を要する．

このときのｍとβそれぞれの最尤法による解は次のように与えられる．

$$\hat{m} = \frac{n}{\sum_{i=1}^{n} \ln(t_0 / t_i)} \tag{5.9}$$

$$\hat{\beta} = \frac{n}{t_0^{\hat{m}}} \tag{5.10}$$

ただし，時間打切の場合には上の式でよいが，故障打切の場合には，

$$\tilde{m} = \frac{n-1}{n} \hat{m} \tag{5.11}$$

がｍの不偏推定量になることが知られているので(5.11)式を利用した方がよい．ｍの検定につい

ては，$2nm/\hat{m}$が時間打切，故障打切でそれぞれ自由度 2n，2(n-1)のχ^2分布に従うことが知られているので，これを利用して行なう．検定統計量として

$$T=\frac{2n}{\hat{m}} \tag{5.12}$$

を計算して，χ^2分布の上側αパーセント点と比較して大きければ有意に信頼度成長が行なわれていると判断する．ただし，時間打切と故障打切では自由度が異なる点に注意する．

5.3.3　他のモデルと例題

他の信頼度成長のモデルで使いやすいのは，故障強度関数が

$$\lambda(t)=e^{\alpha+\beta t} \tag{5.13}$$

である．成長の検定の問題としては，帰無仮説として$H_0:\beta=0$，対立仮説として$H_A:\beta<0$を考える．パラメータの推定は最尤法では，数値計算に頼らねばできない．しかし，故障数ｎが与えられたという条件付きのもとで，検定統計量を作ることが可能である．条件付き尤度に基づくスコア統計量を利用して次の検定統計量が得られる．

$$U=\frac{\sum_{i=1}^{n}t_i-\frac{1}{2}nt_0}{t_0\sqrt{\frac{n}{12}}} \tag{5.14}$$

として，帰無仮説のもとで，近似的に標準正規分布に従う．したがって，上式を計算して標準正規分布の下側αパーセント点と比較して有意性の判定を下すことができる．ここでも，故障打切の場合にはt_0をt_nで置き換え，ｎをn-1とする若干の変更が必要であることに注意する．

例として 100000 時間の観測で次の 10 個の故障時間が得られた場合を考える．

2　5　10　15　22　30　40　55　75　98:（単位 1000 時間）

上式より，ｍの推定値は，

$$\hat{m} = \frac{10}{\ln(100/2) + \ln(100/5) + \cdots + \ln(100/98)}$$
$$\cong 0.6391$$

$$\hat{\beta} = \frac{n}{t_0^{\hat{m}}} = \frac{10}{100000^{0.6391}} \cong 6.3770 \times 10^{-3}$$

となる．検定統計量 T の値は，

$$T = \frac{2n}{\hat{m}} \cong \frac{2 \times 10}{0.6391} \cong 31.30$$

と計算される．これを自由度 20 の χ^2 分布の上側確率のパーセント点と比べる．有意水準を，５パーセントと取ると，31.41 が得られ，５パーセント有意とは言えない．したがって，信頼度成長はこの検定からは言えないことになる．しかしながら，ほとんど棄却値に近いことから，信頼度成長の傾向がうかがえることがわかる．

このデータについてもう１つの別のモデルで検定を行なうと，

$$u = \frac{352000 - (0.5)(10)(100000)}{100000\sqrt{\frac{10}{12}}}$$
$$\cong -1.621$$

が得られ，これを正規分布の下側５パーセントである-1.64 と比べると，前のモデルと同じ結論が得られる．ここでも，統計量の値は棄却値にほぼ近い値であることに注意する．

5.4 コンテンツ評価における再発事象のモデリング

前節までは主として信頼性工学をベースとした工学上の応用に主体をおいてきたが,事象生起を定義することによって様々な問題に適用可能であること示す．Nelson (2003)は点過程のモデリングと類似した平均累積関数(Mean Cumulative Function=MCF)を定義し，これまでの数学的に厳密な確率過程の枠組みだけでなく,事象の生起にともなうコストの変化を取り入れた統一的のモデルを提案し，スプレッドシート型のプログラムまで提供している．

ここでは，映画の評価の問題を点過程のモデリングとして扱うことにする．“いい”映画と批評されるコンテンツは大きな感動を与えるものであり,どのように感動が推移するかが問題であ

る．5 人の被験者から 2007 年邦画部門で一位となった映画「フラガール」の感動場面のタイムスタンプの記録する.その時刻を事象の生起時間と解釈したデータの計数過程をプロットしたものが，図８である（著者の研究室のメンバー5 人を被験者とした実際の観測データに基づく).

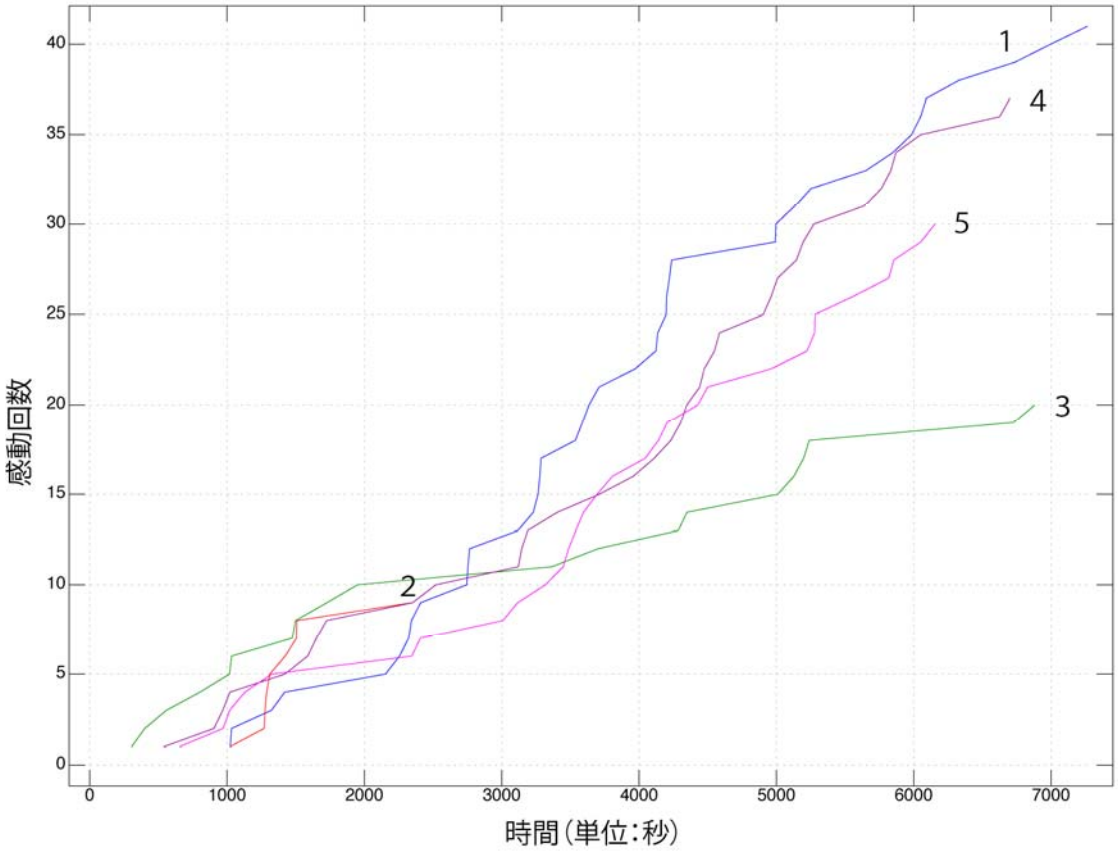


図８．被験者 5 人の計数過程観測値のグラフ

このままでは感動のパターンの変化が見えにくいので微分形である，強度関数の推定を行う．強度関数の推定は，Diggle(1985)により，核関数を利用したノンパラメトリックな強度関数の推定法が有効である．ここでは，その方法を利用して滑らかな強度関数の推定を行う．

$$\hat{\lambda}_t(x) = \left\{ \sum_{i=1}^{n} \delta_t(x - x_i) \right\} \Big/ p_t(x). \tag{5.15}$$

ただし，$\delta_t(x) = t^{-1}\delta(t^{-1}x)$ は原点で対称な適当な確率密度関数であり，時刻 T まで観測がなされているときに，$p_t(x) = \int_0^T \delta_t(x-u)du$ は端点補正の定数である．実際の推定にはガウシアン核関数を利用した．図９は被験者 5 人の結果をまとめて強度関数を推定した結果である．上側と

下側に書かれた曲線は 95%信頼区間を構成したものである． IMSE (Integrated Mean Squared Error)最小化の近似評価を行うために交差検証法によってスムージングパラメータを決定している．

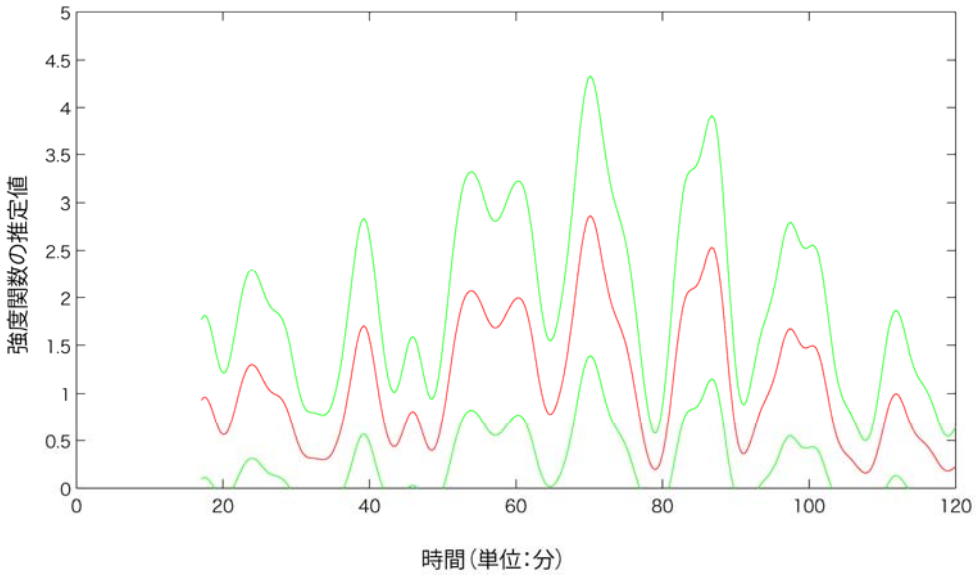


図９．強度関数の核関数を用いたスムーズな推定値

感動のパターンとしては，７つの山があること，最も大きな山は約 70 分にあることがわかる．実際の映像と強度関数との関係を分析するための回帰型の点過程のモデリングについては今後の研究テーマの１つである．

６．議論

生存時間分析の方法論は様々な分野で応用可能である．特に不完全データとして重要な意味を持つ打切データの概念が解析上必要になってくる．打切は本来信頼性試験における定時打切，故障打切から出てきた用語である．したがって，右側打切が標準であるが，この概念は拡張され，任意の区間のどこかに観測値があるという不完全な情報をも取り込んだ形で打切という用語が使われている．JIS では用語の名詞としての利用は“打切”という送り仮名を使用しない定義にしているが，現在は様々拡張がなされ，用語もまた明確になっていない．“打ち切り”ということばも使われいが，学問上の用語としては統一がはかられるべきである．ここでは，“打切”を用いた．

比例はザードモデルのように回帰型のモデルも基本的には生存時間の期待値を共変量で予測しようというものであり，寿命分布が正の分布であり，歪みを持つ分布であることを考えると，

分位点回帰に関するモデルの開発も重要でありこれからの研究が必要な分野である.複数の計数過程におけるパラメータの同時推定の問題における共通パラメータの効率的推定という問題も貴重なデータから有用な情報を効率的に抽出するために重要である(Kamakura, 1996). 共通パラメータ以外の局外母数をベイズモデルによる緩やかにしばる柔軟なモデルの開発も今後の課題である.

謝辞: 査読者の方々には細部にわたり原稿の不備をご指摘いただきました. また大変有益なコメントをいただき感謝申し上げます.

参考文献

Breslow, N.E. (1974). Covariance analysis of censored survival data. Biometrics, 30, 89-99.

Cohen, A. C. (1965). Maximum Likelihood Estimation in the Weibull Distribution Based On Complete and On Censored Samples, Technometrics, 7, 579-588.

Cook, R. J. and Lawless, J. F. (2007). The Statistical Analysis of Recurrent Events, Springer.

Cowling , B. J, Hutton, J. L. . Shaw, and J. E. H. (2006). Joint modelling of event counts and survival times, Appl. Statist., 55, 31-39.

Cox, D. R. (1972). Regression models and life tables (with discussion). J. R. Statist. Soc., B34, 187-220.

Cox, D. R. and Oakes, D. (1984). Analysis of survival data. Chapman and Hall.

Diggle, P. (1985). A Kernel Method for Smoothing Point Process Data, Appl. Statist., 34, 138-147.

Efron, B. (1977). Efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc., 72. 557-565.

Farewell, V. T. and Prentice, R. L. (1980). The approximation of partial likelihood with emphasis on case-control studies. Biometrika, 67, 273-278.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihood based on Cox's regression and life model. Biometrika, 60, 267-278.

Kalbfleisch, J. D. and Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons.

Kamakura, T. and Yanagimoto, T. (1983). Evaluation of the regression parameter estimators in the proportional hazard model. Biometrika, 70, 530-533.

Kamakura, T. (1996). Trend analysis of multiple counting processes. Lifetime Data: Models in Reliability and Survival Analysis Edited by Jewell, N. P., Kimber A. C., Lee, M.-L. T., and Whitmore, G. A., Boston: Kluwer Academic Publishers.

Kamakura, T. (2004). Computational Methods in Survival Analysis, Handbook of Computational Statistics: Concepts and Methods edited by J. E. Gentle, W. Hardle and Y. Mori, , 767-785.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. J. Am. Statist. Assoc. 53, 457-481, 562-563.

Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons.

真壁肇 (1966). ワイブル確率紙の使い方, 日本規格協会.

Menon, M. (1963). Estimation of the shape and scale parameters of the Weibull distributions. Technometrics, 5,175-182.

長塚豪己, 鎌倉稔成(2003). ワイブル分布の形状母数推定の新方法,日本信頼性学会誌, 25, 583-597.

Nelson, W. (1982). Applied Life Data Analysis. New York: John Wiley & Sons. (奥野忠一監訳：柴田義貞, 藤野和建, 鎌倉稔成訳, 寿命データの解析, 日科技連, 1988)

Nelson, W. (2003). Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications, Society for Industrial and Applied Mathematics

Schonbucher, P. J. (2003). Credit Derivatives Pricing Models: Models, Pricing and Implementation, John Wiley & Sons. (望月衛訳, クレジット・デリバティブ, 東洋経済新報社, 2005)

Snyder, D. L. and Miller, M. I. (1991). Random Point Processes in Time and Space, Springer-Verlag.

Yanagimoto, T. and Kamakura, T. (1984). The maximum full and partial likelihood estimators in the proportional hazard model. Ann. Inst. Stat. Math., 36, 363-373.

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2011 年 10 月
第 1 部 自然・工学と統計科学

第4章　極値理論, 信頼性, リスク管理

渋谷政昭[1],　　高橋倫也[2]

(慶應義塾大学理工学部 名誉教授),　　(神戸大学海事科学部 教授)

極値統計理論の歴史的展開を追いながら, 現実世界の要請とそれに対応したモデル, 方法の整備を述べ, 現在の主流研究状況を紹介する.

1 変量で, 区分最大値データが得られる場合, 水準超過データが得られる場合の基本分布とその推測が基本的な統計理論・方法である. 多変量・時系列の場合には複雑となるためにモデルの説明に限り, 超単純多変量極値分布 (§5.3.3) を具体例として導入した. より詳しくは, 参考書・ ソフトウェア (§5.2) を利用されたい.

[1]sibuyam@1986.jukuin.keio.ac.jp
[2]r-taka@maritime.kobe-u.ac.jp

1　1 変量極値理論の要点

1.1　まえがき (極値理論)

21 世紀に入ってから極値理論に関する秀れた参考書が, 改版を含まずに, 数冊出版された. 極値理論に基づく統計的方法が熟成し, 統計パッケージも作られ, ようやく諸分野で活用され始めている.

極値の方法を多用してきたのは, 水文学, 土木工学である. この分野では民主的合意形成のために, 公開されたデータと方式による設計値設定の必要から理論が開発されてきた. 自然災害による建造物の破壊防止もこれに近い. 交通手段の事故, 腐食による機器の破損なども, 実験による設計と, 実世界における使用状況との隔たりにたいして, 極値理論を援用している分野である.

水文学と関連する大きな課題は地球温暖化と異常気象の発生である. 地球規模の平均温度の上昇について論拠が確実になっているが, それと連動する異常気象は, 現象が多様であり, 地域差, 不確定性, 多重関連性のために, 異常の同定と予測が難しい. 水, 熱を含む大気循環の活発化にともなう, 降雨量, 風速, 気温の最大値・最小値の変化が異常気象の指標として調べられている. 測定の低価格化のために測定地点と, 測定頻度が増えているために, 期間は短いが, 時間空間的分解能の高いデータの解析が求められている.

もっともめざましいのが, 金融業における極値理論の利用である. 金融市場の拡大, 流動化のために市場の変動が巨額となり世界の経済活動全般に影響を及ぼすようになった. このような状況の中で始まった, 大銀行を中心とするトップ金融機関の自主的管理活動に対して, ヨーロッパを中心とする統計学者が極値理論を持ち込み, 国際的な基盤構造を作るのに成功した (Embrechts, 2004). その結果, 金融における定量的リスク管理理論 (Qualitative Risk Management) 体系化の中心的な道具の一つとなっている (McNeil et al., 2005).

著者たちはもっぱら工学の分野に関心をもっているために, このように拡大した適用分野の実体を評価することはできず, ここでは基本的な入門と, 重要と思われる新しい理論を紹介する.

1.2　区分最大値 (古典理論)

極値 (extreme value) とは標本の最大値または最小値である. 観測値の符号を変えれば大小が入れ替わるから, 習慣としてもっぱら最大値を扱う.

第 2 次世界大戦中にナチスを逃れてアメリカに移住した Gumbel は極値統計理論の普及に貢献した最初の書物を書いた (Gumbel, 1958) . 彼はこの本で多くの現象に注目しデータを解析した. 典型的なのは水文学, 河川・海岸工学において “安全な堤防の高さを決定する” 課題である. 利用するデータは, その地点における年間最大水位・流量の記録である. 安全の基準は, たとえば “100 年に 1 回の水害に耐える” ものと設定される. これは氾濫危険地域が小中規模ならば, 今日でも合意が得られる規準であろう. 記録データは, 50–200 年, あるいはそれ以下である.

年間最大水位 X が既知の独立同一分布関数 H に従うと, 事象 $X > x$ の待ち時間が期待値 $T = 1/(1 - H(x))$ の幾何分布に従うことから, これを解いた $x_T = H^{-1}(1 - 1/T)$ つまり, 上側確率 $1/T$ の確率点を T **年再現水準** (T-period return level) と呼ぶ. 逆に T は x_T の再現年である. 過去 t 年のデータから T 年 $(T > t)$ の再現水準までも推定するという “統計的補外” の問題である. Gumbel の時代までに Fisher and Tippett, von Mises, Gnedenko, たちにより次の理論が作られていた.

$(Y_n)_{n=1}^{\infty}$ を i.i.d. 確率変数列, その分布関数を $F(y) = P\{Y_n \le y\}$ とする. 標本最大値 $M_n = \max(Y_1, \ldots, Y_n)$ の分布関数 $F^n(y)$ の漸近分布を求める. F に依存する適当な数列 $(a_n > 0)_{n=1}^{\infty}, (b_n)_{n=1}^{\infty}$, と非退化分布関数 $H(x)$ にたいして, $(M_n - b_n)/a_n$ の確率分布が

$$F^n(a_n x + b_n) \xrightarrow{d} H(x), \quad n \to \infty,$$

を満たすとき, H を**極値分布** (extreme value distribution) と呼び, F は H の**最大値吸引領域** (maximum domain of attraction) に属すると言い, $F \in MDA(H)$ と記す. (a_n, b_n) は F の吸引係数と呼ぶ.

定理 1 (極値分布型, Trinity Theorem)**.** 極値分布は（位置, 尺度を別として）次の *3* つの型に限られる.

$$\textit{Gumbel 分布} \qquad \Lambda(x) := \exp(-\exp(-x)), \quad -\infty < x < \infty. \tag{1.1}$$

$$\textit{Fréchet 分布} \qquad \Phi_\alpha(x) := \exp(-x^{-\alpha}), \quad x > 0,\ \alpha > 0. \tag{1.2}$$

$$\text{負の } \textit{Weibull 分布} \qquad \Psi_\alpha(x) := \exp(-(-x)^{\alpha}), \quad x \le 0,\ \alpha > 0. \tag{1.3}$$

これらの極値分布からの標本の最大値の分布は, 以下のように同じ型となる. その性質を最大値安定性と呼ぶ. 逆に最大値安定性は極値分布を特徴づける.

$$\Lambda^n(x) = \Lambda(x - \log n),\ \Phi_\alpha^n(x) = \Phi_\alpha(n^{-1/\alpha}x),\ \Psi_\alpha^n(x) = \Psi_\alpha(n^{1/\alpha}x),$$
$$\forall n \in \mathbb{Z}_{>0},\ \forall x \in \mathbb{R}. \qquad (\mathbb{Z}_{>0} \text{ は, すべての正整数の集合である.})$$

後述のように, 多くの教科書分布が Gumbel 分布の最大値吸引領域に属し, パラメータ α の推定を必要としないため, Gumbel 分布が広く利用された. もっとも簡単には, Gumbel QQ プロットに直線を当てはめ, 直線性を確かめると同時に, 未知の吸引係数を推定し, 再現確率点も読み取る. つまり T 年再現水準を

$$\widehat{x_T} = \hat{b} + \hat{a}\Lambda^{-1}(1 - 1/T)$$

で推定する. $\hat{a}, \hat{b}$ は尺度, 位置の推定量で a_n, b_n の近似である.

用語 “再現水準” が新しい業界用語 Value-at-Risk, VaR, に変わりつつある. しかし, たとえば 1% VaR により時間を無視して上側確率 0.01 の確率点を指している. 地震学者も, 同じことに別の表現を用いている.

1.3 一般極値分布

三つの極値分布をまとめて次のように表わすことができる.

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), & x > -1/\gamma, \quad \gamma > 0, \\ \exp(-e^{-x}), & -\infty < x < \infty, \quad \gamma = 0, \\ \exp(-(1+\gamma x)^{-1/\gamma}), & x < -1/\gamma, \quad \gamma < 0. \end{cases} \tag{1.4}$$

$\gamma > 0$ は Φ_α $(\gamma = 1/\alpha)$, $\gamma = 0$ は Λ, $\gamma < 0$ は Ψ_α $(\gamma = -1/\alpha)$ に対応している.

この分布族は**一般極値分布** (generalized extreme-value distributions), と呼ばれるが, 三つの極値分布の統一型であり, 拡張しているわけではない. パラメータ γ は**裾指数** (tail index), と呼ばれている. γ が増加すると, より裾が重い分布となる.

$\gamma \neq 0$ のときには $G_\gamma(x) = \exp(-(1+\gamma x)_+^{-1/\gamma})$, $(w)_+ = \max(w, 0)$, とまとめて表すことができる. $\gamma = 0$ は $\gamma \to 0$ の場合とみなせる.

一般極値分布導入の意義

形式 (1.4) は統一した型であるというだけでなく, 推測上の意味がある. 第 1 に, どの最大値吸引領域に属するかの知識は不確かである. 第 2 に, 極値分布への収束が必ずしも早くない. たとえば正規分布の Λ への収束は非常に遅い. 下記定理 2 の関数 ϕ を用いると

$$n(1 - F(a_n x + b_n)) \approx (1 + \gamma_n x)^{-1/\gamma_n}, \quad \gamma_n = \phi'(b_n), \quad n \to \infty,$$

(5.3 数学的説明) より

$$F^n(a_n x + b_n) \approx \exp(-(1+\gamma_n x)^{-1/\gamma_n}) = G_{\gamma_n}(x), \quad n \to \infty,$$

となるから, 極限の $G_\gamma(x)$ でなく $G_{\gamma_n}(x)$ による近似の方が望ましいことがある.

第 3 に, 3 種の型の区別を, 裾指数の推定に帰着し, 位置・尺度パラメータと同時に推定することになる. 一般極値分布の分布範囲が裾指数に依存するために, 最尤推定のための正則条件が満たされていないが, Smith (1985) は, Akahira and Takeuchi (1981) の議論を拡張して, $\gamma > -1/2$ ならば最尤推定可能であることを示した. この議論は, 次小節で述べる一般 Pareto 分布でも成り立つ. 一般極値分布の吸引領域に関する次の十分条件は諸分布の判定に便利である.

定理 2 (一般極値分布の吸引領域 von Mises, 1936)**.** F を分布関数, f をその密度関数とし,

$$\phi(x) := \frac{1-F(x)}{f(x)}, \quad \phi'(x) = \frac{d}{dx}\phi(x), \quad x^* = x^*(F) := \sup\{x : F(x) < 1\},$$

とすると,

$$\lim_{x\to x^*} \phi'(x) = \gamma \quad \Longrightarrow \quad F \in MDA(G_\gamma).$$

このとき, 吸引係数 (a_n, b_n) は $n(1-F(b_n)) = 1$, $a_n = \phi(b_n)$ と択べばよい.

逆に, F が 2 階微分可能, $F \in MDA(G_\gamma)$ であれば $\lim_{x\to x^*} \phi'(x) = \gamma$.

□

証明を 5.3 数学的説明で述べる.

次の諸分布は Gumbel 分布 $G_0 = \Lambda$ の最大値吸引領域に属する.

ガンマ, 正規, 対数正規, ロジスティック, 一般双曲, 負の Fréchet.

他の $G_\gamma, \gamma \neq 0$ については, たとえば Beirlant, et al.(2004) 参照.

1.4 水準超過観測値と一般 Pareto 分布

一般に分布の裾は

$$G(x) \approx 1 + \log G(x) =: W(x), \quad (G(x) \to 1) \tag{1.5}$$

と表せる. G を G_γ とするとき, $\log G(x) > -1$ を満たす範囲での変換 $W(x)$ により導かれる分布関数は

$$W_\gamma(x) = \begin{cases} 1-(1+\gamma x)^{-1/\gamma}, & 0<x<\infty, \quad \gamma>0, \\ 1-e^{-x}, & 0<x<\infty, \quad \gamma=0, \\ 1-(1+\gamma x)^{-1/\gamma}, & 0<x<-1/\gamma, \quad \gamma<0, \end{cases}$$
$$= 1-(1+\gamma x)_+^{-1/\gamma}, \quad 0<x<\infty, \quad \gamma\in\mathbb{R}, \tag{1.6}$$

となる. これを **一般 Pareto 分布** (generalized Pareto distributions) と呼ぶ. パラメータ γ は一般極値分布の裾指数に対応し, 同じ名前で呼ばれる. γ が正, 零, 負のとき, 一般 Pareto 分布は, それぞれ Pareto 分布, 指数分布, ベータ分布である.

一般の分布関数 G について

$$P\{u<X\le x \,|\, u<X\} =: G(x\,|\,u) = \frac{G(x)-G(u)}{1-G(u)} = 1-\bar{G}(x)/\bar{G}(u),$$
$$\bar{G}(x) = 1-G(x), \quad x\ge u,$$

となるから, これを (1.5) の関係を満たす W によって近似できる. つまり, $F\in MDA(G_\gamma)$ のとき $F^n(a_nx+b_n)$ を G_γ で近似するかわりに, $F(a_nx+b_n\,|\,b_n)$ をより簡単な W_γ で近似することができる. より正確には次の定理が成り立つ.

定理 3 (Balkema and de Haan, 1974; Pickands, 1975)**.**

$$F\in MDA(G_\gamma) \iff \lim_{\bar{F}(u)\to 0} \frac{\bar{F}(a(u)x+u)}{\bar{F}(u)} = \bar{W}_\gamma(x), \ \forall x\ge 0, \ \bar{F}(a(u)x+u)>0.$$

ただし, $a(\,\cdot\,)$ は吸引係数 a_n を連続化した適当な正の関数である.
□

一般極値分布を適用できるデータは, 堤防の高さの例で述べたように, ある長期間の最大値が十分な個数得られることが前提となっている. これを**区分最大値データ** (block maxima) と呼ぶ. 腐食の測定では, 長方形格子ごとに腐食孔の深さあるいは直径の最大値を求める. 区分最大値の場合では, 区分の中のデータが比較的大きくても小さくても 1 個しか取らない.

それにたいして, 一般 Pareto 分布は, 閾値を大きくしたときにそれを超える観測値の条件付分布である. 分布の裾に関する, 異なるサンプリング法で

ある. このときの観測値を**水準超過データ** (threshold exceedances, or Peaks Over Threshold) と呼ぶ. もしも河川の水位に季節性がないならば, あるいは腐食の測定であれば, このようなサンプリング法の方が効率がよい.

一般 Pareto 分布の形は一般極値分布より簡単であり, 諸計算が楽になる. 一般極値分布の良い特徴はすべて保存されている. ただ一つ, 閾値をどのように決定するかの問題が生じる. 閾値を大きくすれば近似は良くなるが, データ数は減少して推定精度が悪くなることをバランスしなければならない.

1.5 一般 Pareto 分布の性質

一般 Pareto 分布 (1.6) に尺度パラメータを加え, 分布下限を 0 としたものを

$$F(x) = F(x;\gamma,a) = \begin{cases} 1-(1+\gamma x/a)_+^{-1/\gamma}, & \gamma \neq 0, \\ 1-e^{-x/a}, & \gamma = 0, \end{cases} \tag{1.7}$$

とする. その分布範囲は $(0,\infty)$, $\gamma \geq 0$; $(0, a/(-\gamma))$, $\gamma < 0$, である. (1.7) の分布を GPrt (γ, a) で表わす. その期待値, ハザード関数は,

$$E(X) = a/(1-\gamma), \quad \gamma < 1; \quad h(x) = \frac{f(x)}{1-F(x)} = \frac{1}{a+\gamma x}, \quad \forall \gamma \in \mathbb{R},$$

である. $X > x_0$ の条件の下で, $X \sim$ GPrt $(\gamma,\ a)$ の生存関数は

$$P\{X - x_0 > x | X > x_0\} = \frac{1-F(x_0+x)}{1-F(x_0)} = \left(1+\frac{\gamma x}{a+\gamma x_0}\right)^{-1/\gamma}, \ x \geq 0; \quad \gamma \neq 0,$$

ただし $\gamma < 0$ のとき $x < a/(-\gamma) - x_0$. つまり $X - x_0 | X > x_0 \sim$ GPrt $(\gamma,\ a + \gamma x_0)$ であり, 左裾打ち切りが同じ型の分布で, 尺度が x_0 の 1 次式として変わる. $\gamma = 0$ の場合は指数分布が 'memoryless である' と呼ばれる性質である. したがって条件付期待値 (平均余命 mean residual life, **平均超過関数** mean excess function) は $E(X)$ の式の a を $a + \gamma x_0$ に変えた

$$m(x_0) := E(X - x_0 | X > x_0) = \frac{a+\gamma x_0}{1-\gamma}, \quad \gamma < 1, \quad 0 < x_0,$$

である. さらに任意の上側確率の確率点も x_0 の 1 次式である. 期待値は上側確率 $(1-\gamma)^{1/\gamma}$ の確率点である.

1.6 推測

一般 Pareto モデルの最尤推定は裾指数の範囲が限られ ($\gamma > -1/2$), 標本が小さいときの精度が低い. そのために他の推定法が工夫されている.

期待値が有限の範囲では ($\gamma < 1$), 確率点関数が陽に表せることを利用する**確率荷重モーメント推定量** (Probability Weighted Moment est.) が通常のモーメント法より有効である.

$\gamma > 0$ の Pareto 分布関数 $1 - x^{-1/\gamma}, x > 1$, に従う確率変数の対数は指数分布に従い, その順序統計量は良い性質を持っている. この事実を用いた **Hill 推定量** は簡単であるためよく調べられておりその改良版も多数ある. すべての $\gamma \in \mathbb{R}$ で使える Pickands 推定量もある. Hill 推定量については 5.3 数学的説明で述べる.

$F \in MDA(G_\gamma), \gamma \neq 0$ の一つの必要十分条件は, $1 - F$ が正則変動関数になることである. $F \in MDA(G_\gamma)$ からの標本の性質を議論するためには, この条件では不足で, 2 次正則変動関数の理論が必要である (de Haan and Ferreira, 2006).

一般極値分布, 一般 Pareto 分布の Bayes 解析では, 裾指数の (必要ならば位置も) 事前分布として正規分布を, 尺度には対数正規分布あるいはガンマ分布を, そしてこれらが独立であることを前提にする. しかしそれでは $\gamma = 0$ の確率が 0 であり, それが不適切な分野もあるので, $\gamma = 0$ に点確率を与える試みもある.

またモデル構成に専門家の判断を取り入れるアプローチがある. Coles and Tawn (1996) はイギリス南西部の 1 地点での日雨量 54 年データから, 年最大日雨量を Bayes 解析するに当たって専門家の知識を取り入れるため, 次のように事前分布を導入した. まず一般 Pareto 分布の 3 パラメータを, 上側確率 $p_1 > p_2 > p_3$ (再現年 $1/p_1 < 1/p_2 < 1/p_3$) の確率点 (雨量) q_1, q_2, q_3 に変換する. この雨量 (パラメータ) を, 確率変数とみなし, その 50%, 90% 確率点を専門家に予想してもらう. $\tilde{q}_1 = q_1, \tilde{q}_2 = q_2 - q_1, \tilde{q}_3 = q_3 - q_2$ が独立な 2 パラメータガンマ分布に従うと仮定し, q_1, q_2, q_3 の周辺分布の確率点が専門家の予想に合うように, パラメータを決定する.

この専門家はイギリス南西部の気象を熟知しており, その地点の地理も知っているが, その地点の気象統計は, 予想に利用しなかった. $(p_1, p_2, p_3) = (.1, .01, .001)$ とその 1/3 を試み, 結果に差はなかった.

1.7 応用例

1.7.1 Delta 計画

1953 年 1 月 31 日北海の嵐による堤防の破壊でオランダ南部 Zeeland 州など Delta 地帯 1500 km^2 が水没, 1853 人が死亡した. 国家的 Delta Project 1968–1987 が計画実行され, 3.2 km の防波堤防が建設された. de Haan (1990) 参照.

Return Period 10,000 年 を政府が提示した. 別の根拠として, 実際には嵐と最高潮位がずれていたのを, もしも一致していたらという最悪のシナリオから導かれた数値でもあるという.

1.7.2 MV Derbyshire 号遭難事件

1980 年 9 月 9 日, 台風 15 号 (Orchid) の際に, ケベックから川崎に鉄鉱石を輸送していたイギリス 9 万トン貨物船 MV Derbyshire 号が, 連絡のないまま, 乗船していた 44 人とともに日本の南東約 600 km で沈没した. 台風 15 号は暴風雨圏が非常に大きく, 長時間継続したのが特徴であった.

海難審査報告では原因不明となったが遺族の強い要求により再調査が行われた. 沈没船のデジタル写真, ビデオ撮影により空気取入口からの浸水, バラストのアンバランス, から最終的に船倉のハッチカバーの破壊という原因が判明した. 問題は同型船の設計基準が諸運行条件で安全に航海できるものか, の検討となった. 船型, 積載量, 波にたいする進行方向, 台風の特徴, 波浪が主要因で, これらの種々の条件の下での水槽模型船実験が行われ, 実験データの統計的解析により, 問題部位への最大荷重の確率分布が推定された. その結果, 設計基準の改定勧告がなされ, 海洋波浪の衛星による推定が計画された.

Heffernan and Tawn (2001) 参照. 事件の詳細は www.mv-derbyshire.org.uk/

1.7.3 Venezuela 大洪水

1991 年 12 月 Venezuela で大雨・洪水が起こり, 50,000 人が死亡し, 日本の援助で建てられたばかりの病院が崩壊した. 災害からの救出, 医療のために国際的な支援が行なわれた. Maiquetia 国際飛行場の降雨データによると, そのときの日降雨量 410.4 mm は, 過去 50 年間の最大日雨量約 150 mm を遥かに超えるものであった. Coles (2004) は過去 40 年間の Maiquetia 国際

飛行場日雨量データを解析した. 1 年を 3 時期に分け, 各時期における一般極値分布パラメータが, 独立に変動するというベイズ模型を考えて MCMC を用い, その事後分布において降雨量 410.4 mm は不可能な値ではなく, 上側確率 0.67% の確率点であることを示している. 極値理論を適用するための条件が満たされていない可能性がある場合にモデルの不確実性を考慮する手段としてベイズが使われている.

1.7.4 金属疲労と Wicksell's corpuscle problem

金属疲労は多くの事故の原因である. 金属に荷重を加え, 元に戻すと, 弾性変形の範囲内では元の形に戻る. しかしこれを数百万回反復すると, 表面・内部に傷が生じ亀裂が成長して破壊に到る. 村上敬宜は高品質鋼の疲労寿命が, 応力集中部分にある非金属介在物の最大寸法で定まることを理論的・実験的に示した. 介在物はほとんど球状で, 鋼内部にランダムに点在している. 鋼試験片を切断し切断面の円形介在物の寸法データから空間の球の寸法分布を推定しなければならない.

この問題は空間統計学の古典的な問題で **Wicksell's corpuscle problem** と呼ばれている, 逆問題である (Takahashi and Sibuya, 2002). 介在物を球と考えると, 球の中心が強度 λ_V のポアソン過程に従い, 大円面積 S_V がこれと独立で確率密度関数 $f_V(s)$ をもち, $E(\sqrt{S_V}) = m$ とすると, 切断面円形の中心は強度 $\lambda_A = 2m\lambda_V/\sqrt{\pi}$ のポアソン過程に従い, 切断面の面積 S_A の確率密度関数は

$$f_A(t) = \frac{1}{2m}\int_t^\infty \frac{1}{\sqrt{s-t}} f_V(s)ds, \quad 0 < t < \infty,$$

となる. 逆変換, 生存関数についても同様の式が得られ, これらを総称して "Wicksell 変換" という.

f_V, f_A の分布関数 F_V, F_A の一方が一般極値分布の吸引領域 $MDA(G_\gamma)$ に入ると, 他も吸引領域に入るが, 裾指数が次のように変わる. 問題 は F_A の標本に基づく F_V の最大値の予測である.

$$F_V \in MDA(G_\gamma) \Longleftrightarrow F_A \in MDA(G_{2\gamma/(2-\gamma)}).$$
$$F_V \in MDA(G_{2\gamma^*/(2+\gamma^*)}) \Longleftrightarrow F_A \in MDA(G_{\gamma^*}).$$

1.7.5 超高齢者の寿命と表データ

日本人の平均寿命は男女とも世界最高であり, 半世紀以上も平均寿命が確実に増加している. しかし, 最近の超高齢の寿命分布を調べると, 分布そのものはそれほど変化しておらず, 高齢者数の急速な増加が平均寿命を延ばしている原因となっている. 利用可能なデータとして, 人口動態統計データ (死亡表) と全国高齢者調査 (生存統計) が利用できる.

人口動態統計における 84 歳以上, 高齢者調査の 100 歳以上のデータについて, 標本平均余命を計算すると, ほとんど直線的に減少し一般 Pareto 分布 ($\gamma < 0$) が当てはまると予想される. 種々の解析の結果では, 大体当てはまりがよく, 寿命分布には有界な上限があると考える方がよい. つまりハザード関数が分布上限で発散することになるが, 人口学ではハザード関数が指数的に増加する Gompertz 曲線が用いられているが, これは最小値の極値分布である負の Gumbel 分布に他ならず, 極値理論と整合しない.

調査統計は通常, 表の形で公表されている. 死亡統計はある年度に各年齢で死亡した人の数の表である. 生存統計は毎年のある日に生存している各年齢の人の数である. このように表に区分されたデータに対し, 一般 Pareto 分布をあてはめるための理論は十分に研究されていない (渋谷, 華山, 2004).

2 多変量極値分布

2.1 まえがき

1 変量確率標本 (独立同一分布) に関する極値理論は美しく汎用的であるが, 多変量ではかなり複雑となる. 2 変量の場合には視覚化可能であるため扱いやすいが, 3 変量以上となると低次元でも種々の困難が生じる. そのために多変量極値理論は未完成で多くの挑戦が続いている.

$(\mathbf{Y}_n = (Y_{1n}, \ldots, Y_{dn}))_{n=1}^{\infty}$ を分布関数 F をもつ d 変量確率変数の i.i.d. 列とし, 成分ごとの最大値を

$$\mathbf{M}_n := \left(\max_{1\le i\le n} Y_{1i}, \ldots, \max_{1\le i\le n} Y_{di}\right)$$

とする. その同時分布関数は $P\{\mathbf{M}_n \le \mathbf{x}\} = F^n(\mathbf{x})$ である.

適当なベクトル列 $(\mathbf{a}_n > \mathbf{0})_{n=1}^{\infty}$, $(\mathbf{b}_n)_{n=1}^{\infty}$, により各成分を基準化したとき, 非退化分布 H の連続点 $\mathbf{x}$ において

$$\lim_{n\to\infty} F^n(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n) = H(\mathbf{x})$$

のとき, H を**多変量極値分布**と呼び, F は H の**最大値吸引領域に属する**と言い, $F \in MDA(H)$ と表す. 演算は成分ごととする.

F の周辺分布関数 $F_J(\mathbf{x}_J)$, $\mathbf{x}_J = \{x_i : i \in J\}$, $J \subset \{1, \ldots, d\}$, は $F_J \in MDA(H_J)$ を満たす. H_J は, H の周辺分布関数で多変量極値分布である. 特別な場合として 1 変量周辺分布 $H_i, 1 \leq i \leq d$, は一般極値分布に従い, F の周辺分布を $F_1, \ldots, F_d$ とすると $F_i \in MDA(H_i)$ である. 吸引係数列もそのまま利用できる. H_i の分布型が既知なので, $H(\mathbf{x})$ の周辺分布によらない従属性が問題である. 1 変量分布は確率変数の単調連続変換 (確率積分変換) により自由に変換できるから, H_i を同じ分布に固定し, それに応じて F_i も同じ分布に変換して, 議論することができる.

2 つの周辺分布が代表的である. 1 つは一様分布, もう 1 つは標準 Fréchet 分布である. これらについて述べるが, 他にも負の Gumbel 分布, Weibull 分布を用いる接近法がある.

2.2 極値接合分布関数

この節の前半で述べる接合分布関数一般についての詳細は塚原 (2007) 参照.

定義 1 (接合分布関数 (copula))**.** *d 変量接合分布関数, あるいは単に接合関数 $C(\mathbf{u}) = C(u_1, \ldots, u_d)$ とは単位立方体 $[0,1]^d$ 上の確率分布関数であって, その 1 変量周辺分布関数がすべて $[0,1]$ 一様分布のものである.*

より具体的に, d 変量連続確率分布関数が与えられると, 次のように接合分布関数を構成することができる.

定義 2 (F の接合分布関数)**.** *確率変数 $\mathbf{X} = (X_1, \ldots, X_d)$ の同時分布関数を F, 周辺分布関数を F_1,...,F_d とするとき確率変数*

$$(F_1(X_1), \ldots, F_d(X_d))$$

の分布関数を F の接合分布関数と呼び, $C(\mathbf{u}; F)$ と表す.

周辺確率点関数 $F_i^-(u) := \inf\{x : F_i(x) \geq u\}$ を用いると, $C(\mathbf{u}; F)$ を陽に表すことができて,

$$C(\mathbf{u}; F) := F(F_1^-(u_1), \ldots, F_d^-(u_d)),$$

となる. つまり $C(\mathbf{u}; F)$ は F の周辺確率表示である.

逆に $C(\mathbf{u})$ が接合分布関数ならば, 任意の分布関数 $F_1, \ldots, F_d$ にたいして, これを周辺分布とする多変量分布関数を $C(F_1(x_1), \ldots, F_d(x_d))$ により生成できる.

定理 4. $\mathbf{X} = (X_1, \ldots, X_d)$ を連続な周辺分布関数, 連続な接合分布関数 C をもつ確率変数とする. $\varphi_1, \ldots, \varphi_d$ を任意の単調増加関数とすると $(\varphi_1(X_1), \ldots, \varphi_d(X_d))$ の接合分布関数は C で変わらない.

任意の独立連続分布関数の接合分布関数は同一で,

$$C_0(\mathbf{u}) := \prod_{i=1}^{d} u_i,$$

となる. それにたいして

$$C_+(\mathbf{u}) := \min(u_1, \ldots, u_d),$$

は $\mathbb{R}^d$ の "上昇曲線": $(f_1(t), \ldots, f_d(t);\ f_i$ 非減少$)$ の上に退化した**完全正従属** (perfectly positively dependent, comonotone) 確率分布の接合分布である.

$$C_-(u_1, u_2) := \max(u_1 + u_2 - 1, 0),$$

は $\mathbb{R}^2$ の "下降曲線" の上に退化した**完全負従属** (perfectly negatively dependent, countermonotone) 確率分布の接合分布である.

すべての d 変量接合分布関数は不等式

$$\max\{\sum_{i=1}^{d} u_i + 1 - d, 0\} \leq C(\mathbf{u}) \leq C_+(\mathbf{u})$$

を満たす. 下限は $d = 2$ の場合にだけ, 接合分布関数 C_- となる. C が大きいほど正方向の従属性が強くなり, 周辺分布が等しい分布間の自然な半順序となる.

パラメトリックな接合分布関数を推定するには, まず周辺分布関数を推定しなければならない. ノンパラメトリックであれば, 各成分を成分ごとの順位統計量に置き換えればよく, この限りでは簡単である.

確率変数 $\mathbf{X} = (X_1, \ldots, X_d)$ の分布関数を F, 周辺分布関数を $F_1, \ldots, F_d$, 生存関数を $\overline{F} = P\{X_1 > x_1, \ldots, X_d > x_d\}$, とするとき確率変数 $-\mathbf{X} = (-X_1, \ldots, -X_d)$ の接合分布関数は

$$\hat{C}(\mathbf{u}; F) := \overline{F}(F_1^-(1 - u_1), \ldots, F_d^-(1 - u_d)).$$

となる. これを $C(\mathbf{u};F)$ に対応する生存接合分布関数 (survival copula) と呼んでいるが, これは生存関数でないのでまぎらわしい.

極値接合分布関数

$H(\mathbf{x})$ が極値分布関数であれば, それからの確率標本の, 成分ごとの最大値は, 基準化すれば, $H(\mathbf{x})$ に従う. この意味で H は "最大値安定" である.

$$H^n(\mathbf{a}_n\mathbf{x}+\mathbf{b}_n)=H(\mathbf{x}),\quad H_i^n(a_ix+b_i)=H_i(x),$$

両辺の接合分布関数は, 周辺分布関数の基準化に依存しないから,

$$C^n(\mathbf{u}^{1/n})=C(\mathbf{u}),\quad \text{or}\quad C^r(\mathbf{u})=C(\mathbf{u}^r),\ \forall r>0.$$

最後の条件を満たす "最大値安定" 接合分布関数を**極値接合分布関数** (extreme value copula) と呼ぶ. 極値接合分布関数は "最大値無限分解可能" である.

極値接合分布関数の例として次のようなものがある.

1. 独立接合分布関数 C_0 , 完全正従属接合分布関数 C_+ は極値接合分布関数である.

2. ロジスティック接合分布関数

$$C_\theta(\mathbf{u})=\exp(-(\sum_i(-\log u_i)^\theta)^{1/\theta}),\quad 1\le\theta<\infty.$$

3. 任意の極値接合分布関数 $C(\mathbf{u})$ にたいして

$$C_\alpha(\mathbf{u})=\mathbf{u}^{\mathbf{1}-\alpha}C(\mathbf{u}^\alpha),\quad \alpha=(\alpha_1,\ldots,\alpha_d),0\le\alpha_i\le 1,$$

   は極値接合分布関数である.

4. Pickands 表現 (Pickands, 1981) $d=2$ のとき, 2.4 節で詳述する Pickands 従属関数 A を用いると, すべての極値接合分布は次のように表せる.

$$C(\mathbf{u})=\exp\left(\log(u_1u_2)A\left(\frac{\log u_2}{\log(u_1u_2)}\right)\right).$$

2.3 単純多変量極値分布

次に周辺分布として, 標準 Fréchet 分布 $F(x) = \exp(-1/x)$, $x > 0$, を選ぶ. 次の定理が基本的である. これは Pickands, de Haan and Resnick, Coles and Tawn たちによって, 1980-90 年代に確立された.

$(Y_{1n}, \ldots, Y_{dn})_{n=1}^{\infty}$ が i.i.d. 確率変数列で, すべての 1 変量周辺分布が標準 Fréchet 分布に従うとする.

$$\mathbf{M}_n^* = (M_{1n}^*, \ldots, M_{dn}^*) = (\max_{1\le i\le n} Y_{1i}/n, \ldots, \max_{1\le i\le n} Y_{di}/n),$$

とする. 吸引係数 $a_{j,n} = n, b_{j,n} = 0$, は Fréchet 分布が自分自身に収束するときの吸引係数である.

$$P\{\mathbf{M}_n^* \le \mathbf{x}\} \xrightarrow{d} G(\mathbf{x}),$$

で G の周辺分布は標準 Fréchet 分布である. 最大値安定性から

$$G^t(t\mathbf{x}) = G(\mathbf{x}),\ t > 0,$$

である. 標準 Fréchet 分布を周辺分布とする多変量極値分布を**単純多変量極値分布** simple multivariate extreme value distribution, SMvEV , と呼ぶ.

定理 5 (指数測度とスペクトル測度)**.** 単純多変量極値分布を

$$G(\mathbf{x}) = \exp(-V(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}_{>0}^d, \tag{2.8}$$

と表す. $\mathbb{R}_{>0}, \mathbb{R}_{\ge 0}$ により, それぞれ正または非負の実数全体を表す. すると,

$$V(x_1, \ldots, x_d) = \int_{S_d} \max_{1\le j\le d}(w_j/x_j) dH(\mathbf{w}), \tag{2.9}$$

である. S_d は $d-1$ 次元単位単体

$$S_d = \{\mathbf{w} : w_j \ge 0, \sum_{j=1}^{d} w_j = 1\},$$

H は S_d 上の非負測度で,

$$\int_{S_d} w_j dH(\mathbf{w}) = 1, \quad j = 1, \ldots, d, \tag{2.10}$$

従って $H(S_d) = d$, を満たすものである.

注意

1. V を**指数測度** (exponent measure), H を**スペクトル測度** (spectral measure) と呼ぶ. 後者は言わば極座標表示の回転成分で, より一般のノルムで定義できる.

2. $d^{-1}H$ は S_d 上の確率分布である.

3. (2.10) の制約条件は, G の周辺分布関数が $\exp(-\int_{S_d}(w_i/x_i)dH(\mathbf{w}))$ となるから,
$$\exp(-\int_{S_d}(w_i/x_i)dH(\mathbf{w})) = \exp(-1/x_i)$$
より必要となる.

4. V, (2.9), は $V(x_1/a,\ldots,x_d/a) = aV(x_1,\ldots,x_d)$ を満たすから, G, (2.8), は $G^n(n\mathbf{x}) = G(\mathbf{x}), \quad n = 2,3,\ldots$ を満たす.

5. (2.9) 式の積分の意味は次の通りである.
$$w_j/x_j = t, \quad j = 1,\ldots,d,$$
を満たす $\mathbf{w} = (w_1,\ldots,w_d)$ は, $\sum w_j = t\sum x_j = 1$ から, w_j に関する制約は領域
$$\{\mathbf{w}:\ w_j \geq x_j/\sum x_i\}, \quad j = 1,\ldots,d,$$
に関する積分となる. たとえば w_1 に対応する領域は $\mathbf{w} = (1,0,\ldots,0)$ を頂点とする凸多角形, 凸多面体, である. 与えられた $\mathbf{x}$ にたいして $\mathbf{x}$ に依存する領域での積分を行うことになる.

6. 定理の初等的なアプローチを 5.3 数学的説明で述べる.

2.3.1 独立性と従属性

SMvEV が独立であれば
$$G(\mathbf{x}) = \exp(-1/(\prod_{j=1}^{d} x_j)), \quad \mathbf{x} \in \mathbb{R}^d_{>0},$$
である.

SMvEV が完全に従属であるとすると, 確率分布は直線 $x_1 = \cdots = x_d$ 上の 1 次元分布に退化し, 分布関数は

$$G(\mathbf{x}) = \exp(-1/(\min_{1\le j\le d} x_j)), \quad \mathbf{x} \in \mathbb{R}^d_{>0},$$

となる. H は単位単体の中心 $w_1 = \cdots = w_d = 1/d$ に点測度 d をもつ.

定理 6 (Takahashi, 1994)**.** ある $G \sim$ SMvEV の成分が対ごとに独立であれば, 独立である. さらに, ある $\mathbf{y} \in \mathbb{R}^d$ が存在して

$$G_{ij}(y_i, y_j) = G_i(y_i)G_j(y_j), \quad 1 \le i < j \le d, \quad 0 < G_j(y_j) < 1,$$

であれば G は独立である,

実際, 周辺分布が等しい多変量極値分布 G が独立であるとし, $F \in MDA(G)$ の周辺分布も等しいとすると, F に関する次の条件は同等である. 最後の条件 (d) は 2.5 節で述べる裾の漸近独立性である.

(a) $\lim_{n\to\infty} F^n(a_n\mathbf{x} + b_n\mathbf{1}) = \prod_{i=1}^d G_1(x_i)$.

(b) $\lim_{n\to\infty} F^n_{k,l}(a_n x_k + b_n, a_n x_l + b_n) = G_1(x_k)G_1(x_l), \quad \forall\, 1 \le k < l \le n.$

(c) $\lim_{n\to\infty} nP\{X_{1,k} > a_n x_k + b_n, X_{1,l} > a_n x_l + b_n\} = 0,$
$\forall\, 1 \le k < l \le n,\ \forall\, x_k, x_l, G_1(x_k), G_1(x_l) > 0.$

(d) $\lim_{t\to x^*} P\{X_{1,k} > t \mid X_{1,l} > t\} = 0,\ x^* = \sup\{x : F(x) < 1\},\ \forall\, 1 \le k < l \le n.$

2.4 単純 2 変量極値分布

$d = 2$ の場合の SMvEV は, $S_2 = \{(w_1, w_2) : w_1 \ge 0,\ w_2 \ge 0,\ w_2 + w_2 = 1\}$ となるために, より単純となる.

定理 7 (2 変量単純極値分布)**.** 標準 *Fréchet* 分布を周辺分布とする *2* 変量極値分布関数 $G(x_1, x_2)$ は次のように表せる.

$$G(x_1, x_2) = \exp(-V(x_1, x_2)), \quad x_1 > 0, \quad x_2 > 0, \tag{2.11}$$

$$V(x_1, x_2) = \int_0^1 \max\left(\frac{w}{x_1}, \frac{1-w}{x_2}\right) dH(w). \tag{2.12}$$

$H/2$ は S_2 上の分布関数で平均値

$$\frac{1}{2}\int_0^1 w\, dH(w) = \frac{1}{2}, \tag{2.13}$$

のものである. たとえば $w = 1/2$ に関して対称であれば十分である. □

$d = 2$ のときは, 次の **Pickands 従属関数** (dependence funtion) A が見やすい. 関数 A を

$$V\left(\frac{1}{v_1}, \frac{1}{v_2}\right) \equiv (v_1 + v_2)A\left(\frac{v_1}{v_1 + v_2}\right) \tag{2.14}$$

により定義する. 書き換えると

$$A(t) := \int_0^1 \max(t(1-q), (1-t)q)dH(q). \quad t \in [0,1] \tag{2.15}$$

関数 A は

$$\max(t, 1-t) \le A(t) \le 1,$$

を満たす凸関数で, 上限は独立, 下限は comonotone の場合に相当する.

(2.14) 式の定義で右辺を $(v_1 + v_2)A(v_2/(v_1 + v_2))$ とすることもある. つまり (2.15) 式の左辺を $A(1-t)$ とすることになる. $A(t)$ を拡張した安定裾従属性関数 $\varphi(\mathbf{v}) := V(\mathbf{1}/\mathbf{v})$ も同様な性質をもっている.

単純 2 変量極値分布に対応する極値接合分布関数は次のようになる. 2.2 節参照.

$$C(u_1, u_2) = \exp\left(\log(u_1u_2)A\left(\frac{\log u_2}{\log(u_1u_2)}\right)\right).$$

例として次のようなものがある. 最初の 2 つは 2 次, 3 次の多項式で表わされるすべての $A(t)$ である.

1. mixed model
   $A(t) = \theta t^2 - \theta t + 1, \quad 0 \le \theta \le 1, \quad V(x_1, x_2) = \frac{1}{x_1} + \frac{1}{x_2} - \frac{\theta}{x_1+x_2}$.

2. asymmetric mixed model
   $A(t) = \phi t^3 + \theta t^2 - (\theta + \phi)t + 1, \quad \theta \ge 0,\ \theta + 2\phi \le 1,\ \theta + 3\phi \ge 0,$
   $V(x_1, x_2) = \frac{1}{x_1} + \frac{1}{x_2} - \frac{(2\theta+\phi)x_1+(\theta+\phi)x_2}{(x_1+x_2)^2}$.

3. logistic model
   $A(t) = ((1-t)^r + t^r)^{1/r}, \quad r \ge 1, \quad V(x_1, x_2) = (x_1^{-r} + x_2^{-r})^{1/r}$.

2 変量, 多変量のパラメトリック極値分布について, たとえば Kotz and Nadarajah (2000) を見よ.

2.5 裾の漸近的独立性・従属性

一般に連続対称 2 変量分布関数 $F(x_1, x_2), F_1(x) = F_2(x) = F(x, \infty)$, をもつ確率変数 (X_1, X_2) において

$$\lambda = \lim_{x \to x^*} P\{X_2 > x | X_1 > x\}, \quad x^* = \sup\{x : F(x) < 1\},$$

が存在するならば, これは (X_1, X_2) の裾における従属性の自然な尺度である. 対称性により X_1, X_2 を交換しても λ は同じである. F の接合関数 C , 生存接合関数 $\hat{C}$ を用いて書き換えると

$$\lambda = 1 - \lim_{x \to x^*} \frac{F_1(x) - F(x, x)}{1 - F_1(x)} = 2 - \lim_{u \to 1} \frac{1 - C(u, u)}{1 - u} = 2 - \lim_{u \to 0} \frac{\hat{C}(u, u)}{u}.$$

$\lambda = 0$ (あるいは $\lambda > 0$) のとき F または (X_1, X_2) の**裾が漸近的に独立** (あるいは**従属**) であるという.

2 変量正規分布では, 相関係数 ρ が $\rho < 1$ であれば漸近的に独立である, Sibuya (1960). 裾が漸近的に独立であると, 多変量極値分布に基づく推論には困難を生じる. 小規模データからは, 弱い従属性があると想定するために, データの外に補外すると偏りを生じる. より詳しくはたとえば Fougères (2004) 参照.

3 移動最大値過程

3.1 強定常時系列

時系列 $(X_t)_{t \in \mathbb{Z}_{>0}}$ が**強定常** (strictly stationary), つまり任意の $t_1, \ldots, t_n, k, n \in \mathbb{Z}_{>0}$ にたいして,

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_1+k}, \ldots, X_{t_n+k})$$

であるとし, (X_t) の周辺分布関数を F とする. $(\tilde{X}_t)_{t \in \mathbb{Z}_{>0}}$ を同じ分布関数 F に従う i.i.d. 系列とし,

$$M_n = \max(X_1, \ldots, X_n), \quad \tilde{M}_n = \max(\tilde{X}_1, \ldots, \tilde{X}_n),$$

とする.

定理 8.　任意の $i < j < i' < j'$ と, ある係数列 $(a_n > 0)_n, (b_n)_n$ にたいして, $\max(X_i, \ldots, X_j)$, $\max(X_{i'}, \ldots, X_{j'})$ が $i' - j \to \infty$ のときに漸近的に独立ならば

$$\lim_{n\to\infty} P\{(\tilde{M}_n - b_n)/a_n \le x\} = H(x),$$
$$\iff \quad \lim_{n\to\infty} P\{(M_n - b_n)/a_n \le x\} = H^\theta(x), \quad \exists\theta \in (0, 1].$$

強定常時系列の例としては　(1). 定常ガウス時系列.　(2). 自己回帰時系列 $X_t = Z_t + cZ_{t-1}$,　ただし $Z_0, Z_1, \ldots$ i.i.d. $Var(Z_t) < \infty$, がある. (Brockwell and Davis, 1991).

強定常時系列では, 大きな X_t に続く値が大きくなりがちであるため, 極大値の塊りが生じる (clustering). これを十分に長い区間に分けて区間ごとの最大値を択ぶと, これらが近似的に独立となる. そのために M_n は $\tilde{M}_n$ と同じ H に同じ係数で収束する. $u_n := a_n x + b_n$ とすると,

$$P\{\tilde{M}_n \le u_n\} \approx H(x), \quad P\{M_n \le u_n\} \approx H^\theta(x),$$

となるから, 有効なサンプル・サイズが θ 倍に減少する. これは区間内の利用可能なデータが平均 $(1/\theta)$ 個あるのが無視されるためである. パラメータ θ を**極値指数** (extremal index) と呼ぶが, 裾指数 tail index γ を extreme value index とも呼ぶのと紛らわしい, 吉原 (2004).

$(Z_t)_{t\in\mathbb{Z}}$ を標準 Fréchet 分布に従う i.i.d. 確率変数列, $\alpha_i > 0$ を有限または可算の定数列とすると, 時系列

$$X_t := \max_i \alpha_i Z_{t-i}, \quad \alpha_i \ge 0, \ \sum_i \alpha_i = 1, \tag{3.16}$$

は強定常で X_t の周辺分布は標準 Fréchet 分布である. これを**移動最大値過程** (moving maximum, MM process) と呼ぶ, Deheuvels (1983).

$$X_t := \max(\alpha X_{t-1}, (1-\alpha)Z_t), \quad 0 \le \alpha < 1,$$

もその特別な場合となり **ARMAX** (maximum autoregression process) と呼ばれる. MM process の準備としての 5.3.3 超単純多変量極値分布を参照.

3.2　M 3 / M 4

(3.16) をさらに拡張することができる. 標準 Fréchet 変数の独立 2 重系列 $(Z_{l,t})$, および非負係数の 2 重系列 $(\alpha_{l,k})$ を導入し, 新しい添え字について

さらに最大値をとることにより, 多様な重み付けができる. この強定常時系列を **maxima of moving maxima, M3**, process と呼ぶ.

$$X_t := \max_{\ell\geq 1}\max_{k}\ \alpha_{\ell,k}Z_{\ell,t-k},\quad \alpha_{\ell,k,}\geq 0,\quad \sum_{\ell\geq 1}\sum_{k}\alpha_{\ell,k}=1.$$

これはさらに多変量時系列 $(\mathbf{X}_t)_{t\in\mathbb{Z}}, \mathbf{X}_t=(X_{1t},\ldots,X_{dt})$ に拡張できる. 変数はそのまま, 係数を 3 重数列とし, 成分ごとに M3 系列を作る. これを **multivariate maxima of moving maxima, M4**, process と呼ぶ.

$$X_{it} := \max_{\ell\geq 1}\max_{k}\ \alpha_{\ell,k,i}Z_{\ell,t-k},\quad \alpha_{\ell,k,i}\geq 0,\quad \sum_{\ell\geq 1}\sum_{k}\alpha_{\ell,k,i}=1,\ 1\leq i\leq d. \tag{3.17}$$

多変量強定常時系列の収束と extremal index について, 1 変量強定常時系列と同様なことが言えるが, extremal index は定数ではなく, 分布関数の引数に依存する. このような M4 にたいして

$$P\{X_{it}\leq x_{it}, 1\leq t\leq T; 1\leq i\leq d\}=\exp\left(-\sum_{\ell=1}^{\infty}\sum_{s=-\infty}^{\infty}\max_{1-s\leq k\leq T-s}\max_{1\leq i\leq d}\frac{\alpha_{\ell,k,i}}{x_{s+k,i}}\right).$$

これはデータから推定できる. また extremal index は

$$\theta(x_1,\ldots,x_d)=\frac{\sum_{\ell}\max_k\max_i\alpha_{\ell,k,i}x_i}{\sum_{\ell}\sum_k\max_i\alpha_{\ell,k,i}x_i}$$

により定まる. ファイナンスへの応用について次節を参照.

4 いくつかの応用例

1. **保育園のデザート戦争**

   Sibuya (1960) の動機となった問題を漫画化すると次の通りである. ある保育園で n 人の幼児が昼食を取る. 食べ終わると続いてデザートを食べる. 早い子が両方食べ終わったとき, 遅い子が昼食も終わっていないと残っているデザートを巡って騒動が起こる. 幼児 k の昼食を終える時間, デザートを終える時間をそれぞれ X_k, Y_k とする. 騒動が生じる確率は

   $$P\{\min_{1\leq k\leq n}(X_k+Y_k)<\max_{1\leq k\leq n}X_k\}$$

   X_k, Y_k が独立であると仮定しても, $(X_k+Y_k), X_k$ は独立でない. つまり $(\max X_k, \max(-(X_k+Y_k)))$ の同時分布が必要である. 同時分布が

漸近的独立なので, 周辺分布の最大値分布の差の分布により, 上記確率分布を近似できることを裏付けた.

2 変量極値分布の研究は, 統計学の辺境で, 独立に, ほとんど同時に行われ, 同じ結果が導かれた. Geffroy (1958/59), Tiago de Oliveira (1958), Sibuya (1960).

2. 諸応用

Kotz and Nadarajah (2000), には Gumbel の初期の応用を始め, いくつかの紹介がある. パラメトリックモデルとして双ロジスティック分布が代表的であり, Coles and Tawn (1991) はイギリス東海岸 Immingham, Lowestoft, Sheerness, における高波 (surge) の毎時測定値に, Joe, Smith and Weissman (1992) は酸性雨中の硫酸塩, 硝酸塩濃度に双ロジスティック分布を当てはめている.

3. 自然災害

de Haan and de Ronde (1998) は, オランダの防波堤 Pettemer zeedijk を波が越える危険確率を推定した. 沖合い観測点における潮位, 波高, 波長について, 過去 13 年間に生じた 828 回の暴風雨の際の最高値がデータである. 沖合いの状況, 防波堤の状況, 可能な危険との間の関係は工学的知識で定まる. これから潮位・波高平面, あるいは潮位・波高・波長空間における危険領域が定まる. これらの暴風雨による事故はなく, 危険領域はデータの散布図からかなり離れている. SMvEV を推定し, その補外によって確率を推定する.

神田, 西嶋 (2004) は台風時の風速を解析し, スペクトル関数 $f_t(s), s \in [0,1], t \in T = \{1, \ldots, p\}$, の推定を行なっている.

4. ファイナンス

Smith (2004) は Pfizer, General Electric, Citibank の 1982–2001 年間の daily return に M4 モデルを当てはめている. 各社の第 t 日の daily return に GARCH $(1,1)$ モデルを適用し, その推定値により尺度の標準化を行う. 標準化データで閾値 $u = 1.2$ を越すものにたいして一般 Pareto 分布を当てはめて裾指数を推定する. 当てはめた分布を変換して標準 Fréchet 分布からの標本を構成する. ここまでは 3 社独立に計算を行う. その結果は, 3 社間の従属性があるだけでなく, 前日との弱い従属性が見られた. そこで M4 process を当てはめ, その妥当性を検討している.

Zhang and Huang (2006) は Dow Jones industrial average, NASDAQ, S&P 500 の daily return, 1972/12/14 - 2003/7/28 を, 上記と同じように M4 モデルで解析し, これらの指数から成るポートフォリオの評価を行なっている.

5 あとがき

5.1 信頼性・リスク管理のための極値理論

20 世紀における統計学発展の原動力は, “正規分布” および関連するモデルを中心とする推測の理論であった. それは正規分布の釣鐘型密度関数の図が統計学の教科書を飾っていることにも象徴されている. その背後には, ばらつき, ゆらぎのある個体集団の大多数を満足させる, あるいは管理することを要請する技術的・社会的要請が働くという基盤がある. このような枠組みでは, 観測値の箱ひげ図で大きく “外れる” データは解析する人間を当惑させる, 異物であり, これを “ごみ箱に捨てる” ことも容認されがちである.

しかし統計学の歴史では, 非常に稀な現象にたいする関心も古くから続いていた. “小数の法則” として Poisson 分布が研究された. しかし, Poisson では不十分で, 事故数・規模の統計では, 無数の小さな事故が起こる一方, 大規模な事故が非常に小さな頻度で生ずる. 言わば極値理論の離散版を要するが, 統一した方法論は確立していない. 今日でも経験的な “Zipf 法則” が繰り返し話題となり, 経営学で “The Long Tail” についての話題が続いている. 経済学での “power law” への関心とも連動している.

王立協会に叛旗を翻して Karl Pearson が 20 世紀初頭に創刊した Biometrika 誌を中心としたイギリス生物統計学にたいして, 北ヨーロッパでは保険学の伝統が強かった, 現在の Scandinavian Journal of Statistics 誌の母体である Scandinavian Actuarial Journal 誌は 1915 年の創刊で独自の確率統計を育んだ. ここでは古い時代からリスク理論すなわち損害保険の研究とみなされていたようである. 現状では損害保険以上に危険な金融商品が市場に溢れているだけでなく, 箪笥預金も危険にさらされる.

安全な社会は基盤構造物の適切な設計だけでなく, それを “発注する” 主体の判断に依存している. その方法は課題ごとに異なり, 単一の方法で多くの状況を覆うことは難しい. 極値理論とその応用の難しさと複雑さもこのような現実の反映である.

5.2 参考書とソフトウェア

参考書

1. Beirlant, Goegebeur, Segers and Teugels (2004)　分厚い (490p.) だけに推測理論に詳しい. データ解析例も豊富である. 11 章中最後の 3 章, 多変量, 時系列, ベイズ法, は別の著者による新しい話題の説明で, それまでのように体系的ではない. 本文で利用している S プログラム, データが出版社ウェッブサイトから利用可能である.

2. Coles (2001)　簡潔 (208p.) で読みやすい. 数理統計学を省いてあるが, 概念の意味は的確に説明しており, 本書に沿った S, R のパッケージが利用可能であり, その使用例も説明されている.

3. Falk, Hüsler and Reiss (2004)　数学的体系的である. 汎関数的小数法則を点過程に適用して, 極値理論を展開している.

4. Finkenstädt and Rootzén (2004)　Gothenburg で開かれたセミナーに招待された 7 人の, 諸トピックスについての講義ノート.

5. de Haan and Ferreira (2006)　極値統計の確率統計理論として, もっとも新しく体系的である. 著者自身が開発した正則変動関数理論に基づいている.

6. Kotz and Nadarajah (2000)　パラメトリックモデルが, 多変量極値分布も含めて数多く集められている.

7. McNeil, Frey and Embrechts (2005)　まえがきで述べた QRM に関する 10 章中に, 多変量モデル, 時系列, 接合関数と従属性, 極値理論が 1 章ずつ当てられている. それぞれオリジナリティが高い.

ソフトウェア

CRAN (The Comprehensive R Archive Network, cran.r-projeect.org) には, 次の名前の極値関係パッケージがある.

evd, evdbayes, evir, extRemes, fExtremes ismev, VaR

これらについて, また S その他のパッケージについては Stephenson and Gilleland (2006) の解説を参照.

5.3 数学的説明

5.3.1 von Mises 十分条件の証明

$F(x)$ の累積ハザード関数を $H(x) = -\log(1 - F(x))$, ハザード関数を $h(x) = (d/dx)H(x) = f(x)/(1 - F(x)) = 1/\phi(x)$ とする.

$$H(u+v) - H(u) = \int_0^v h(u+w)dw = \int_0^x h(u+s\phi(u))\phi(u)ds,$$
$$v = x\phi(u),\ w = s\phi(u).$$

被積分関数の逆数を変形する. 最後の等式は平均値の定理による.

$$\frac{1}{h(u+s\phi(u))\phi(u)} = \frac{\phi(u+s\phi(u))}{\phi(u)}$$
$$= 1 + \frac{\phi(u+s\phi(u)) - \phi(u)}{\phi(u)} = 1 + s\phi'(y), \quad u < y < u + s\phi(u).$$

上の積分に代入すると

$$\int_0^x \frac{ds}{1+s\phi'(y)} = \int_1^{1+x\phi'(y)} \frac{dt}{t}\frac{1}{\phi'(y)} = \frac{1}{\phi'(y)}\log(1+x\phi'(y)),$$
$$t = 1 + s\phi'(y),\ u < y < u + s\phi(u).$$

したがって

$$\frac{1 - F(u+x\phi(u))}{1 - F(u)} = \exp(-(H(u+x\phi(u)) - H(u)))$$
$$= (1+x\phi'(y))^{-1/\phi'(y)},\ u < y < u + x\phi(u).$$

$(a_n), (b_n)$ を $n(1-F(b_n)) = 1, a_n = \phi(b_n)$, により定めると, 定理の条件から $\phi'(b_n) \to \gamma(b_n \uparrow x^*)$ であり,

$$n(1 - F(b_n + a_n x)) = \frac{1 - F(b_n + a_n x)}{1 - F(b_n)} \approx (1 + \phi'(b_n)x)^{-1/\phi'(b_n)}$$
$$\to (1+\gamma x)^{-1/\gamma} = -\log G_\gamma(x), \quad (n \to \infty),$$
$$F^n(a_n x + b_n) \to G_\gamma(x), \quad (n \to \infty).$$

注意 この証明は, 水準超過値の一般 Pareto 分布への収束も同時に示している.

de Haan and Ferreira (2006) Chap. 1 に, この十分条件にたいする別のアプローチがある.

5.3.2 Hill 推定量

分布関数 F が Fréchet 分布，すなわち一般極値分布 G_γ，$\gamma > 0$，の最大値吸引領域に属している ($F \in MDA(G_\gamma)$) とする．このことは，次の条件

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma}, \quad \gamma > 0,$$

と同値である．つまり生存関数が指数 $-1/\gamma$ の正則変動関数である．この条件はまた

$$\lim_{t\to\infty} \frac{\int_t^\infty (1-F(x))\frac{dx}{x}}{1-F(t)} = \lim_{t\to\infty} \frac{\int_t^\infty (\log u - \log t) dF(u)}{1-F(t)} = \gamma,$$

と同値になる．上式の等号は部分積分による．

この漸近的な条件から γ の推定量を構成する．$Y_1, Y_2, \ldots, Y_n$ を F からの確率標本とし，$Y_{1,n} \le Y_{2,n} \le \cdots \le Y_{n,n}$ をその順序統計量とする．上式で t を $Y_{n-k,n}$，F を経験分布関数で置き換えると，Hill 推定量 (Hill, 1975)

$$\widehat{\gamma}_H = \frac{1}{k}\sum_{i=0}^{k-1} \log Y_{n-i,n} - \log Y_{n-k,n},$$

が得られる．

最尤法や平均超過関数を用いても Hill 推定量を導出することができる．上の説明から, 正則変動関数が本質的な他の分野でも, Hill 推定量を利用できる. Hill 推定量については，Embrechts et al. (1997)，Beirlant et al. (2004)，de Haan and Ferreira (2006) に詳しい説明がある．

Hill 推定量の性質

定理　$Y_1, Y_2, \ldots$ を分布 F からの確率標本とする．
(1) $F \in MDA(G_\gamma)$, $\gamma > 0$ で，$n \to \infty$，$k = k(n) \to \infty$，$k/n \to 0$ ならば

$$\widehat{\gamma}_H \xrightarrow{p} \gamma.$$

(2) 逆に，$n \to \infty$ で，$k = k(n) \to \infty$，$k(n)/n \to 0$，$k(n+1)/k(n) \to 1$ のとき

$$\widehat{\gamma}_H \xrightarrow{p} \gamma > 0,$$

ならば $F \in MDA(G_\gamma)$．
(3) 分布関数 F が，次の 2 次の性質

$$\lim_{t\to\infty} \frac{\frac{1-F(tx)}{1-F(t)} - x^{-1/\gamma}}{A\left(\frac{1}{1-F(t)}\right)} = x^{-1/\gamma}\frac{x^{\rho/\gamma}-1}{\gamma\rho},$$

(ここで，$\gamma > 0$，$\rho \leq 0$ で A は正または負の関数で $\lim_{t\to\infty} A(t) = 0$) を満たすならば，$n \to \infty$ で $k = k(n) \to \infty$，$k/n \to 0$ のとき

$$\sqrt{k}(\widehat{\gamma}_H - \gamma) \xrightarrow{d} \mathrm{N}\left(\frac{\lambda}{1-\rho},\, \gamma^2\right).$$

ここで，

$$\lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right) = \lambda \quad \text{有限.}$$

これらの結果から Hill 推定量は一致推定量であり，漸近的に正規分布をするが偏りがあることがわかる．Hill 推定量に関して，最適な k を決める，iid の条件をゆるめる，改良する等の研究がある．

5.3.3 超単純多変量極値分布

$Z_1, \ldots, Z_n$ を標準 Fréchet 分布に従う独立確率変数列, $A = (a_{i,j}), a_{i,j} > 0$ を $\sum_{j=1}^n a_{i,j} = 1,\ i = 1, \ldots, m$ を満たす $m \times n$ 正要素行列とし, $\mathbf{X} = (X_1, \ldots, X_m)$, $X_i = \max_{1\leq j\leq n} a_{i,j} Z_j$, $i = 1, \ldots, m$ とする. あるいは

$$\mathbf{X} = \max_{1\leq j\leq n} Z_j \mathbf{a}_j, \quad A = [\mathbf{a}_1, \ldots, \mathbf{a}_n].$$

$\mathbf{X}$ の分布関数は m 変量極値分布に従う. その分布関数は

$$\begin{aligned} F(\mathbf{x}) = F(x_1, \ldots, x_m) &= P\{\max_j a_{1,j} Z_j \leq x_1, \ldots, \max_j a_{m,j} Z_j \leq x_m\} \\ &= \prod_{j=1}^n P\{Z_j \leq \min(x_1/a_{1,j}, \ldots, x_m/a_{m,j})\} \\ &= \prod_{j=1}^n \exp(-1/\min(x_1/a_{1,j}, \ldots, x_m/a_{m,j})) \\ &= \exp(-\sum_{j=1}^n \max(a_{1,j}/x_1, \ldots, a_{m,j}/x_m)). \end{aligned} \tag{5.18}$$

と表せ, これが最大値安定であることが確かめられる:

$$F^r(rx_1, \ldots, rx_m) = F(x_1, \ldots, x_m),\ \forall r > 0.$$

$\sum_{j=1}^n a_{i,j} = 1$ の条件から, たとえば X_1 の周辺分布関数が,

$$F(x_1, \infty \ldots, \infty) = \exp(-\sum_{j=1}^n \max(a_{1,j}/x_1, 0, \ldots, 0)) = \exp(-1/x_1).$$

つまり標準 Fréchet 分布となり, F は本文で述べる単純 m 変量極値分布である. F を**超単純多変量極値分布** と呼ぶことにする. 単純多変量極値分布の中でももっとも単純な場合である. $F(\mathbf{x})$ のいくつかの成分を ∞ とした結果は同じ型の分布関数になるから, すべての低次元周辺分布が超単純多変量極値分布である.

(5.18) の最終式における各 j にたいする max でどの項が択ばれるかは, $\mathbf{x}/|\mathbf{x}|, |\mathbf{x}| = \sum_{j=1}^{m} x_j$, と $\mathbf{a}_j/|\mathbf{a}_j|$ の関係で定まる. m 次元単位単体を $\Delta_m = \{\mathbf{w} \in \mathbb{R}^m_{\geq 0} : |\mathbf{w}| = 1\}$ とすると, Δ_m の $\binom{m}{2}$ 本の辺および, $\mathbf{a}_j/|\mathbf{a}_j|$ と Δ_m の頂点とを結ぶ m 本の線分で作る m 個の多面体のどれに $\mathbf{x}/|\mathbf{x}|$ が入るかで定まる. たとえば $x_1 \ll x_i (i \neq 1)$ のときには面 $w_1 = 0$ の近くにあり, $F(\mathbf{x}) = \exp(-1/x_1)$ となる.

超単純多変量極値分布の性質

特に $m = n = 2$ のときの $\mathbf{X} = (X_1, X_2)$ の分布を理解するために, $W_{\mathbf{X}} = X_1/(X_1 + X_2)$ の分布関数を求める. $n = 2$ のときの $W_{\mathbf{Z}} = Z_1/(Z_1 + Z_2)$ は $(0, 1)$ 一様分布に従う. 一般性を失うことなく (成分番号の交換により) $a_{1,1}/a_{1,2} < a_{2,1}/a_{2,2}$ と仮定する. これは $0 < q_1 < q_2 < 1, q_i = a_{1,i}/(a_{1,i}+a_{2,i})$ と同等である.

$$W_{\mathbf{X}} = q_1 \iff Z_1\mathbf{a}_1 > Z_2\mathbf{a}_2 \iff W_{\mathbf{Z}}\mathbf{a}_1 > (1 - W_{\mathbf{Z}})\mathbf{a}_2,$$
$$\iff W_{\mathbf{Z}} > \max(a_{1,2}, a_{2,2}) = a_{1,2},$$

これから,

$$P\{W_{\mathbf{X}} = q_1\} = 1 - a_{1,2} = a_{1,1}.$$

同様の計算で $P\{W_{\mathbf{X}} = q_2\} = a_{2,2}$ である.

$\{q_1 < W_{\mathbf{X}} < q_2\} \iff \{X_1 = a_{1,2}Z_2 \ \&\ X_2 = a_{2,1}Z_1\}$ の確率は

$$1 - (a_{1,1} + a_{2,2}) = a_{1,2} - a_{2,2} = a_{2,1} - a_{1,1}.$$

さらにこのとき,

$$W_{\mathbf{X}} \leq w \iff \frac{X_1}{X_2} \leq \frac{w}{1-w} \iff \frac{Z_2}{Z_1} \leq \frac{w}{1-w}\frac{a_{2,1}}{a_{1,2}}$$
$$\iff W_{\mathbf{Z}} \geq \frac{(1-w)a_{1,2}}{(1-w)a_{1,2} + wa_{2,1}},$$

あるいは $W_{\mathbf{X}} = \mathbf{X}/|\mathbf{X}|$ の分布関数を $H(w)$ とすれば,

$$H(w) = \begin{cases} 0, & 0 < w < q_1, \\ wa_{2,1}/((1-w)a_{1,2} + wa_{2,1}), & q_1 \le w < q_2, \\ 1, & q_2 \le w < 1. \end{cases}$$

つまり $W_{\mathbf{X}}$ は 2 点 q_1, q_2 で離散確率, 区間 (q_1, q_2) で連続確率密度をもつ.

一般の (m, n) にたいする $W_{\mathbf{X}}$ の分布は複雑である. $\{\mathbf{a}_1/|\mathbf{a}_1|, \ldots, \mathbf{a}_n/|\mathbf{a}_n|\}$ の凸包の境界および内部に分布し, これらの点で離散確率, これらの点を結ぶ線分の上で 1 次元確率密度, など 1 次元から m 次元までの確率密度関数を含んでいる. A の各行は離散確率とみなせる. これを一般の確率分布とすることにより, 単純多変量極値分布 (2.8) が得られる. 逆に任意の単純多変量極値分布を, n を増やすことにより, 超単純多変量極値分布で近似できる.

Pickands の従属関数

$m = 2$ の場合に超単純多変量極値分布の指数部

$$V(\mathbf{x}) = \sum_{j=1}^{n} \max\left(\frac{a_{1,j}}{x_1}, \frac{a_{2,j}}{x_2}\right), \quad \mathbf{x} \in \mathbb{R}^m_{\ge 0} \setminus \mathbf{0},$$

の変数を逆数に変え, 次のように変形する. 再び一般性を失うことなく,

$$\frac{a_{1,1}}{a_{2,1}} < \cdots < \frac{a_{1,n}}{a_{2,n}} \iff q_1 < \cdots < q_n, \quad q_j = \frac{a_{1,j}}{a_{1,j} + a_{2,j}}$$

を仮定する. $\mathbf{t} = \mathbf{1}/\mathbf{x}$ とし,

$$|\mathbf{x}|V(\mathbf{x}) = |\mathbf{t}|^{-1}V(\mathbf{1}/\mathbf{t}) = |\mathbf{t}|^{-1}\sum_{j=1}^{n} \max(t_1 a_{1,j}, t_2 a_{2,j}),$$

の変数を, さらに $u = t_2/(t_1 + t_2) = x_2^{-1}/(x_1^{-1} + x_2^{-1})$ に変え, 記号を濫用して

$$A(u) = \sum_{j=1}^{n} \max((1-u)a_{1,j}, ua_{2,j})$$

とする.

$$(1-u)a_{1,j} \ge ua_{2,j} \iff u \le q_j = \frac{a_{1,j}}{a_{1,j} + a_{2,j}}$$

であるから

$$A(u) = \sum_{u \le q_j} (1-u)a_{1,j} + \sum_{u > q_j} u a_{2,j}$$

となる. u が 1 区画動くと, 項が $(1-u)a_{1,j}$ から $ua_{2,j}$ に変わるので折れ線の勾配が増加し $A(u)$ は凸関数である. $n = 3$ のとき

$$A(u) = \begin{cases} 1-u, & 0 \le u \le q_1, \\ ua_{2,1} + (1-u)(a_{1,2} + a_{1,3}), & q_1 \le u < q_2, \\ q_2(1-a_{2,3}) + (1-q_2)a_{1,3}, & u = q_2, \\ u(a_{2,1} + a_{2,2}) + (1-u)a_{1,3}, & q_2 < u < q_3, \\ u, & q_3 \le u \le 1. \end{cases}$$

$m = 2$ の場合には, 任意の Pickands 従属関数を, 超単純多変量極値分布の Pickands 従属関数で近似できる.

参考文献

[1] Akahira, M. and Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*, Lecture Notes in Statistics **7**, Springer.

[2] Balkema, A. A. and de Haan, L. (1974). Residual life time at great age, *Ann. Probab.* **2**, 792–804.

[3] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Exteremes*, John Wiley and Sons Ltd, England.

[4] Brockwell, P. J. and Davis, R. A.(1991). *Time Series: Theory and Methods*, 2nd ed., Springer.

[5] Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag, London.

[6] Coles, S. G. (2004). The use and misuse of extreme value models in practice, in Finkenstädt and Rootzén, Eds., *Extreme Values in Finance, Telecommunications, and the Environment*, 79–100.

[7] Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events, *Journal of the Royal Statistical Society*, **B 53**, 377–392.

[8] Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data, *Appl. Statist.*, **45**, 463–478.

[9] de Haan, L. (1990). Fighting the arch-enemy with mathematics, *Statistica Neerlandica*, **44**, 45–68.

[10] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory, An Introduction*, Springer, New York, NY.

[11] de Haan, L. and de Ronde, J. (1998). Sea and wind: Multivariate extremes at work, *Extremes*, **1**, 7–45.

[12] Deheuvels, P. (1983). Point processes and multivariate extreme values, *J. Multiv. Anal.*, **13**, 257–272.

[13] Embrechts, P. (2004). Extremes in economics and the economics of extremes, in Finkenstädt and Rootzén, Eds., *Extreme Values in Finance, Telecommunications, and the Environment*, 169–183.

[14] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer Verlag, Berlin Hidelberg.

[15] Falk, M., Hüsler, J. and Reiss, R.-D. (2004). *Laws of Small Numbers: Extremes and Rare Events*, 2nd. ed., Birkhäuser. (2010, 3rd ed.)

[16] Finkenstädt, B. and Rootzén, H. Eds. (2004). *Extreme Values in Finance, Telecommunications, and the Environment.* CRC/Chapman and Hall, Boca Raton.

[17] Fougères, A.-L. (2004). Multivariate extremes, in Finkenstädt and Rootzén, Eds., *Extreme Values in Finance, Telecommunications, and the Environment*, 373–388.

[18] Geffroy, J. (1958/59). Contributions à la théorie des valeurs extrêmes, *Publ. Inst. Statist. Univ. Paris*, **7/8**, 37-185.

[19] Gumbel, E. J. (1958). *Statistics of Extremes*, Columbia University Press. 河田竜夫, 岩井重久, 加瀬滋男監訳 「極値統計学」生産技術センター新社 (再刊版).

[20] Heffernan, J.E. and Tawn, J. A. (2001). Extreme value analysis of a large designed experiment: A case study in bulk carrier safety, *Extremes*, **4**, 359–378.

[21] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, **3**, 1163–1174.

[22] Joe, H., Smith, R. L., and Weissman, I. (1992). Bivariate threshold methods for extremes, *Journal of the Royal Statistical Society*, **B 54**, 171–183.

[23] Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions*, Imperial College Press.

[24] McNeil, J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management*, Princeton University Press.

[25] Pickands, J. (1981). Multivariate extreme value distributions, *Bulletin of the International Statistical Institute*, **53**, Buenos Aires, 859–878.

[26] Pickands, J. (1975). Statistical inference using extfeme order statistics, *Ann. Statist.* **3**, 119–131.

[27] Sibuya, M. (1960). Bivariate extreme statistics, *Ann. Inst. Statist. Math.*, **11**, 195–210.

[28] Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, **72**, 67–90.

[29] Smith, R. L. (2004). Statistics of extremes, with applications in environment, insurance, and finance, in Finkenstädt and Rootzén, Eds., *Extreme Values in Finance, Telecommunications, and the Environment*, 1–78.

[30] Stephenson, A. and Gilleland, E. (2006). Software for the analysis of extreme events: The current state and future directions, *Extremes*, **8**, 87–109.

[31] Takahashi, R. (1994). Asymptotic independence and perfect dependence of vector components of multivariate extreme statistics, *Statist. Probab. Lett.*, **19**, 19–26.

[32] Takahashi, R. and Sibuya, M. (2002). Metal fatigue, Wickwell transform and extreme values, *Applied Stochastic Models in Business and Industry*, **18**, 301–312.

[33] Tiago de Oliveira, J. (1958). Extremal distributions, *Rev. Fac. Ciencias. Lisboa*, Ser. A, **7**, 215–227.

[34] von Mises, R. (1936). La distribution de la plus grande de n valeurs, *Rev. Math. Union Interbalcanique*, **1**, 141–160.

[35] Zhang, Z. and Huang, J. (2006). Extremal financial risk models and portfolio evaluation, *Computational Stat. & Data Analysis*, **51**, 2313–2338.

[36] 神田順, 西嶋一欽 (2004). 多変量極値分布を用いた多地点強風および地震危険度解析, 統計数理, **52**, 151–173.

[37] 渋谷政昭, 華山宣胤 (2004). 年齢時代区分データによる超高齢者寿命分布の推測, 統計数理, **52**, 117–134.

[38] 塚原英敦 (2007). 接合分布関数の理論と応用, 本書別章.

[39] 吉原健一 (2004). 弱従属性をもつデータに基づく極値統計の最近の話題, 統計数理, **52**, 25–43.

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2011 年 10 月
第 2 部 生物・健康と統計科学

第5章　分子進化の統計科学

岸野洋久[1]

(東京大学大学院農学生命科学研究科 教授)

生命現象を研究する生命科学において，生物の適応と進化のメカニズムを知ることは，究極の目標で，かつ困難な課題であると言える．適応進化の背景には，突然変異がある．遺伝子配列を種間で比較することにより，突然変異のうち集団に定着し，進化の過程で生き残ったものを推測することができる．突然変異は確率的な現象であり，進化には時間の要素が欠かせない．したがって，分子進化は確率過程として捉えることができる．病原菌とウイルスへの抵抗性に関るリボヌクレアーゼ多重遺伝子族の解析を通して，分子進化速度の変遷を辿ることにより，適応進化の背後にある分子メカニズムを推測する方法を紹介する．

[1]kishinoa@lbm.ab.a.u-tokyo.ac.jp

1　突然変異と分子進化，分子進化速度

自然界には 20 種類のアミノ酸が存在する．アミノ酸が次から次に連なり，複雑に折りたたまれることにより，タンパク質がつくられる．さまざまなタンパク質が互いに相互作用し，また化学物質と結びつきながら，生物の組織や体が形づくられ，生命活動が繰り広げられる．したがって，生物の適応の背後にある分子メカニズムを突き詰めていくと，さまざまな環境に応じて，さまざまな種類のタンパク質を，適当な分量だけ生産することが，本質であることが了解される．

その設計図は，チミン (T)，シトシン (C)，アデニン (A)，グアニン (G) の 4 種類の塩基が連なって構成されるゲノムに刻み込まれている．そこには，タンパク質を生産する鋳型となる遺伝子領域が散在する．タンパク質を生成するコード領域においては，塩基 3 つが，コドンと呼ばれるひとつの単位になって，アミノ酸に翻訳される．たとえば，核ゲノムにおいては，TTT はフェニルアラニンに翻訳され，TTA はロイシンに翻訳される．TAA，TAG，および TGA は，翻訳の終了を支持する終止コドンである．遺伝子領域以外の部分も，ゲノムをコンパクトに折りたたむことやタンパク質の生産量，あるいはそれらの相互作用などに関して大きな役割を果たしていることが，次第に明らかになってきている．

ゲノムは，アデニンとチミン，およびシトシンとグアニンが水素結合で相補的に結びつき，二重らせん構造をしている．これがほどけると，一方が鋳型となって再び二重らせんが作られる．こうしてゲノムが複製され，世代をまたいで継承されていく．複製の際のエラーが突然変異である．確率的に起きる突然変異が集団に多様性をもたらす．そしてこの多様性が，集団が環境の変化に柔軟に適応することを可能にする．ゲノムが突然変異により次第に変化していくことを，分子進化という．配列データに基づき，分子進化の履歴を推定することができる．

突然変異はサイコロを振るようにランダムに起きる．突然変異の中には，タンパク質の折りたたみ構造を不安定にしたり，タンパク質の生産量に不具合を生じ，生体内のバランスを壊すものも多く含まれる．こうした有害な突然変異は，やがて集団から脱落する．全体からすると稀ではあるが，適応度に勝れた突然変異も生ずる．こうした突然変異は集団に定着する傾向がある．まわりの遺伝子に対する相対的な適応度で，集団への固定確率が決まる．

私たちは，集団から脱落した突然変異は観測することはできない．すなわち，遺伝子配列を種間で比較して推測することができる進化の歴史は，集

団から脱落せずに残った突然変異の蓄積である．突然変異率を ν，突然変異が集団中の他の遺伝子を押しのけて固定する確率を f とすると，N 個体からなる集団（2 倍体の生物では遺伝子数は $2N$ 個）の分子進化速度 r は，

$$r = 2N\nu f$$

となる．集団の大きさ $2N$ が大きいときには淘汰が強く働き，適応した突然変異はすぐに集団に固定し，有害な変異は集団から脱落する．これに対して，集団の大きさが小さいときには不確実性の要素が大きくなり，突然変異が固定したり脱落したりするまでに要する時間が長くなる．場合によっては，少々有害な突然変異もたまたま集団に固定する可能性も出てくる．こうしたことから，突然変異の多くが有害な突然変異である場合には，分子進化速度は集団の大きさと負の相関を持つことになる (Ohta (1972))．

分子進化の中立説 (Kimura (1968)) は，分子レベルの突然変異の多くは個体の生存力や繁殖力に影響せず，まわりの遺伝子と優劣のつかない突然変異であるとする．中立説が妥当する場合には，固定確率は $f = 1/2N$ となる．集団あたりの突然変異率は $2N\nu$ であるので，分子進化速度は $r = 2N\nu\frac{1}{2N} = \nu$ と，突然変異率に等しくなる．したがって，もしも突然変異率が変わらなければ，集団の大きさは系統の間で大きく振れても，分子進化速度も一定であることが期待される．これを分子時計という．分子時計が働いていないときは，突然変異率 ν が変化したか，あるいは集団に固定する突然変異の数の期待値 $2Nf$ が変化したことになる．

1 世代あたりの突然変異率は変わらなくても，世代の長さ（子供の誕生と次の世代の子供の誕生の時間間隔）が変化すると，単位時間当たりの突然変異率は変化する．世代あたりの突然変異率の変化，世代の長さの変化のいずれにしても，それらはゲノム全体にわたり（単位時間当たり）突然変異率に影響を与えるだろう．これに対して，環境の淘汰圧とそれへの適応は，これに大きく関る遺伝子における突然変異の集団への固定確率を変化させる．したがって，分子進化速度の変化がゲノム全体で共通に見られるときには，まずは時間当たり突然変異の変化が疑われる．これに対して，分子進化速度の変化のパターンがゲノム上の部位により大きく異なる場合には，環境適応と関連した機能の変化が示唆されることになる．

2　分子進化の統計モデル

突然変異には，いくつかの種類がある．かつてチミンであったものがアデニンに入れ替わるといった塩基置換，ゲノム断片の挿入や欠失，ある部

分が別の部分に移る転座，異なる部分が互いに入れ替わる逆位，そして遺伝子やゲノムの重複である．ここでは，最もよく観察される突然変異である塩基置換に即して，分子進化の統計モデルを紹介する．

2.1 確率過程

図 1 は，2 回の種分化を通じて 3 つの種 A, B, C に分かれた集団で，それの持つ遺伝子配列の変化を模式的に表したものである．第 1 の種分化の時点で $\cdots$CCG$\cdots$ であった配列が塩基置換を起こし，種 C では現在 $\cdots$CTG$\cdots$ になっている．もう一方の系統でも塩基置換が起きていて，第 1 の種分化後 t_1 時間を経て起きた第 2 の種分化では，$\cdots$TCG$\cdots$ に変化している．その後種 A にいたる系統では t_2 時間を経て現在に至るまで変化が見られないが，種 B では，現在 $\cdots$TCA$\cdots$ である．

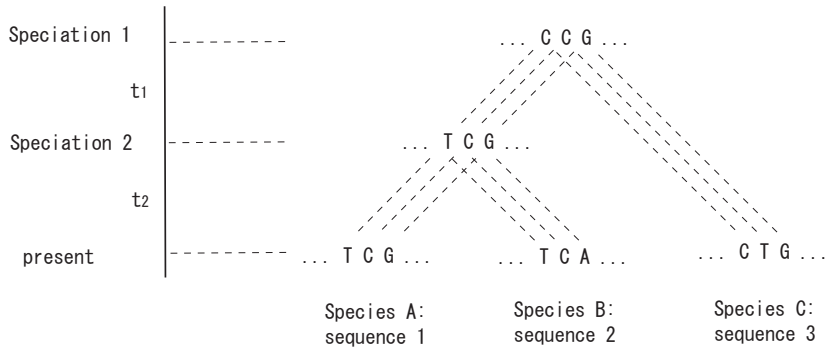


図 1: 遺伝子配列の分子進化

種 A, B, C から得られた 3 つの配列のデータ

$$\begin{array}{lll} \text{配列 1} & \cdots \quad \text{TCG} & \cdots \\ \text{配列 2} & \cdots \quad \text{TCA} & \cdots \\ \text{配列 3} & \cdots \quad \text{CTG} & \cdots \end{array} \tag{2.1}$$

は，多変量カテゴリカルデータ（この例では 3 変量）である．サンプルを構成する要素は，サイトと呼ばれる．変量の間の相関関係は，祖先を共有することに由来する．

突然変異とそれの集団への定着は確率的な現象である．その結果，個々のサイトにおいて，塩基が確率的に変化する．確率的に変化するプロセス

を確率過程という．図1は，確率過程の母集団からの標本とみることができる．それぞれの標本は，標本過程と呼ばれる．配列データは，標本過程の現在におけるクロスセクションを集めたものである．

2.2 マルコフ過程と推移確率

分子進化は，世代を跨いで遺伝情報が受け継がれつつも，変異を受け，次第に変化する様子を表現している．第t世代の状態X_tは1世代前の状態X_{t-1}のみで決まり，それよりも前の世代の影響は受けない．すなわち，

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots) = P(X_t = x_t | X_{t-1} = x_{t-1})$$

という関係を満たす．こうした性質を持つ確率過程をマルコフ過程という．そして右辺の条件付確率を，状態x_{t-1}から状態x_tへの推移確率という．1世代の間で変異を受ける確率は微小である．また，変異を受けたとしても，親世代の塩基と似た性質のものの方がそうでないものよりも集団への固定確率は高い傾向がある．たとえば，チミンとシトシンはともにピリミジン（化学式$\mathrm{C_4H_4N_2}$）と同一の骨格を持つピリミジン誘導体で，分子量も比較的小さい．これに対して，アデニンとグアニンはともにプリン（化学式$\mathrm{C_5N_4H_4}$）の誘導体で，比較的分子量も大きい．したがって，チミンからシトシンへの変異は，アデニンへの変異に比べて集団への固定確率が高い．$\mathrm{T} \leftrightarrow \mathrm{C}$，$\mathrm{A} \leftrightarrow \mathrm{G}$の塩基置換をトランジション，それ以外の塩基置換をトランスバージョンという．

2世代隔てたものの間の推移確率も，マルコフ性より

$$\begin{aligned} & P(X_t = x_t | X_{t-2} = x_{t-2}) \\ = & \sum_{x_{t-1}=\mathrm{T,C,A,G}} P(X_t = x_t, X_{t-1} = x_{t-1} | X_{t-2} = x_{t-2}) \\ = & \sum_{x_{t-1}=\mathrm{T,C,A,G}} P(X_t = x_t | X_{t-1} = x_{t-1}) P(X_{t-1} = x_{t-1} | X_{t-2} = x_{t-2}) \end{aligned}$$

と，1世代の推移確率を用いて表現することができる．推移のパターンは4×4通りあるが，T, C, A, Gをそれぞれ1, 2, 3, 4に対応付け，状態iから状態jへの推移確率$p_{ij} = P(X_t = j | X_{t-1} = x_i)$を$ij$要素に持つ行列$\mathbf{P}$を定義すると，見通しがよくなる．上で求められた2世代を経たときの推移確率行列$\mathbf{P}(2)$は，1世代の推移確率行列の積$\mathbf{P}^2$と表現されることがわかる．

同様の考え方を演繹的に進めていくと，一般に $k = 1, 2, \ldots$ 世代を経たときの推移確率行列について

$$\mathbf{P}(k) = \mathbf{P}^k$$

が成り立つことがわかる．

ところで，分子進化の研究では，通常数万から数千万世代にわたる変化を問題にする．一方で，1 世代において突然変異が起きる確率は微小である．こうした場合には，連続時間近似を行うことが適当である．たとえば，時間の単位を 100 万年にとって，微小時間 dt における推移確率を

$$\mathbf{P}(dt) = \mathbf{I} + \mathbf{R}dt$$

とモデル化する．ここで $\mathbf{I}$ は単位行列である．$\mathbf{R}$ を速度行列と呼ぶ．その非対角要素は塩基置換率を表現しており，対角要素は行和が 0 になるように調整されている．このとき，一般に時間 t を経た推移確率行列 $\mathbf{P}(t)$ は

$$\mathbf{P}(t) = \lim_{n\to\infty} \left(\mathbf{P}\left(\frac{t}{n}\right)\right)^n = \lim_{n\to\infty} \left(\mathbf{I} + \frac{t}{n}\mathbf{R}\right)^n = \exp(t\mathbf{R})$$

と，速度行列の指数関数をとることにより得られる．

$\mathbf{R}$ の固有値を $\lambda_1, \lambda_2, \lambda_3, \lambda_4$，対応する固有ベクトルを $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4$ とし，固有ベクトルを並べて行列 $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4)$ を用意すると，

$$\exp(t\mathbf{R}) = \mathbf{U} \exp(t\Lambda) \mathbf{U}^{-1}$$

という関係式が成り立つ．$\exp(t\Lambda)$ は，固有値の指数関数 $\exp(t\lambda_1)$, $\exp(t\lambda_2)$, $\exp(t\lambda_3)$, $\exp(t\lambda_4)$ を対角要素に持つ対角行列である．一度固有値・固有ベクトルを求めておけば，任意の時間 t について推移確率行列が求められることがわかる．なお，たとえば，トランジションとトランスバージョンの間の速度の違いを考慮に入れ，それぞれの塩基置換速度を α, β とする木村のモデルでは固有値と固有ベクトルをこれらのパラメータの関数として陽に求めることができ，時間 t を経た後の推移確率は

$$\begin{aligned}
p_{\mathrm{TT}}(t) = p_{\mathrm{CC}}(t) = p_{\mathrm{AA}}(t) = p_{\mathrm{GG}}(t) & \\
= \frac{1}{4} + \frac{1}{4}\exp(-4\beta t) + \frac{1}{2}\exp(-2(\alpha+\beta)t) & \\
p_{\mathrm{TC}}(t) = p_{\mathrm{CT}}(t) = p_{\mathrm{AG}}(t) = p_{\mathrm{GA}}(t) & \\
= \frac{1}{4} + \frac{1}{4}\exp(-4\beta t) - \frac{1}{2}\exp(-2(\alpha+\beta)t) & \\
p_{\mathrm{TA}}(t) = p_{\mathrm{TG}}(t) = p_{\mathrm{CA}}(t) = p_{\mathrm{CG}}(t) = p_{\mathrm{AT}}(t) = p_{\mathrm{AC}}(t) = p_{\mathrm{GT}}(t) = p_{\mathrm{GC}}(t) & \\
= \frac{1}{4} - \frac{1}{4}\exp(-4\beta t) &
\end{aligned} \tag{2.2}$$

となる．

2.3 配列データの尤度

3 本の配列の例 ((2.1) 式) に戻って，尤度関数を求めてみよう．種分化により生殖的に隔離された集団は，それぞれの系統で独立に分子進化を経験するとする．このとき，仮に図 1 のように，種分化の時点における祖先配列が直接観察できたとすると，図中表示されている相次ぐ 3 つのサイトの尤度 L_h, L_{h+1}, L_{h+2} はそれぞれ

$$\begin{aligned} L_h &= \pi_{\mathrm{C}} p_{\mathrm{CT}}(t_1) p_{\mathrm{TT}}(t_2) p_{\mathrm{TT}}(t_2) p_{\mathrm{CC}}(t_1+t_2) \\ L_{h+1} &= \pi_{\mathrm{C}} p_{\mathrm{CC}}(t_1) p_{\mathrm{CC}}(t_2) p_{\mathrm{CC}}(t_2) p_{\mathrm{CT}}(t_1+t_2) \\ L_{h+2} &= \pi_{\mathrm{G}} p_{\mathrm{GG}}(t_1) p_{\mathrm{GG}}(t_2) p_{\mathrm{GA}}(t_2) p_{\mathrm{GG}}(t_1+t_2) \end{aligned}$$

と表される．ここで，$\pi_{\mathrm{T}}, \pi_{\mathrm{C}}, \pi_{\mathrm{A}}, \pi_{\mathrm{G}}$ は平衡確率，$p_{ij}(t)$ は塩基態 i が時間 t を経て塩基 j に移る推移確率である．

実際には，ほとんどの場合祖先配列は観察されない．そこで，再び図 1 に戻って種分化の時点での配列を伏せて，第 h サイトの状態を過去に辿ってみると，種 A, B においてともに T であるので，種分化 2 の時点でも T であったと想像するのは自然だろう．ところが，種分化 1 の時点においては，種 C では第 h サイトは C であるとはいえ，現在とは時間 t_1+t_2 も隔たっている．むしろ種分化 2 と時間 t_1 しか隔たっていないことから，このときの状態 T であったとみることもできる．その場合には，種分化 1 と種分化 2 の間には状態の変化はなく，種 C への系統で T から C へ置換が起きており，このサイトの尤度は

$$L'_h = \pi_{\mathrm{T}} p_{\mathrm{TT}}(t_1) p_{\mathrm{TT}}(t_2) p_{\mathrm{TT}}(t_2) p_{\mathrm{TC}}(t_1+t_2)$$

となる．より複雑な分子進化が起きた可能性も否定できない．そこで，2 つの種分化における祖先配列に対する不確実性を考慮に入れ，

$$L_h = \sum_{j_1=\mathrm{T,C,A,G}} \sum_{j_2=\mathrm{T,C,A,G}} \pi_{j_1} p_{j_1 j_2}(t_1) p_{j_2 T}(t_2) p_{j_2 T}(t_2) p_{j_1 C}(t_1+t_2)$$

と尤度表現する．

塩基置換はそれぞれのサイトで独立に起きるとすると，配列データの尤度 L は

$$L = \cdots \times L_h \times L_{h+1} \times L_{h+2} \times \cdots$$

と，サイトの尤度を掛け合わせることにより求められる．ところで，(2.2) 式からもわかるように，推移確率は速度と時間の積の関数として表現される．したがって，分子進化速度が系統間で異なることを許す場合には，

$$\sum_{i \neq j} \pi_i r_{ij}^0 = 1$$

のように期待塩基置換速度が 1 となるように規準化した速度行列 $\mathbf{R}_0$ とを用いて，推移行列を $\mathbf{P}(t) = \exp(tr\mathbf{R}_0) = \exp(b\mathbf{R}_0) \equiv \overline{\mathbf{P}}(b)$ と，期待置換数 $b = tr$ の関数として表すのが自然である．

一般に，系統樹 T であらわされる系統関係を持つ長さ n の s 本の配列

$$\begin{array}{lcccccc}
\text{配列 1} & X_{11} & \cdots & X_{1h} & \cdots & X_{1n} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{配列 p} & X_{p1} & \cdots & X_{ph} & \cdots & X_{pn} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{配列 s} & X_{s1} & \cdots & X_{sh} & \cdots & X_{sn}
\end{array}$$

について，$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ と行列表現する．このときも同様に，対数尤度は

$$l(\boldsymbol{\theta}|\mathbf{X}) = \sum_{h=1}^{n} \log f(\mathbf{X}_h|\boldsymbol{\theta}) \tag{2.3}$$

と，各サイトの対数尤度 $\log f(\mathbf{X}_h|\boldsymbol{\theta})$ $(h = 1, \ldots, n)$ の和として書き下される．

多くの配列を扱うときは，系統樹も大きく，複雑になってくる．このときは，各サイトの対数尤度は部分木の条件付尤度を通して再帰的に定義しておくと見通しが良い．また，アルゴリズムの工夫にも馴染んでいる (Felsenstein (2004))．たとえば，図 2 のように，データ $\mathbf{X}_h$ が二つの部分木に対応して $\mathbf{X}_h^{(1)}$ と $\mathbf{X}_h^{(2)}$ に分解されているとする．このとき，このサイトの尤度は

$$\begin{aligned}
f(\mathbf{X}_h|\boldsymbol{\theta}) &= \sum_{j_0=\mathrm{T,C,A,G}} \pi_{j_0} f^{(1)}(\mathbf{X}_h^{(1)}|\boldsymbol{\theta}) f^{(2)}(\mathbf{X}_h^{(2)}|\boldsymbol{\theta}) \qquad (2.4) \\
&= \sum_{j_0=\mathrm{T,C,A,G}} \pi_{j_0} \left(\sum_{j_1=\mathrm{T,C,A,G}} \overline{P}_{j_0 j_1}(b_1) L^{(1)}\left(\mathbf{X}_h^{(1)}|j_1\right) \right) \times \\
&\qquad \left(\sum_{j_2=\mathrm{T,C,A,G}} \overline{P}_{j_0 j_2}(b_2) L^{(2)}\left(\mathbf{X}_h^{(2)}|j_2\right) \right)
\end{aligned}$$

と表される．ここで，$L^{(1)}\left(\mathbf{X}_h^{(1)}|j_1\right)$ と $L^{(1)}\left(\mathbf{X}_h^{(1)}|j_1\right)$ はそれぞれ，部分木の根の状態で条件付けした尤度である．$\boldsymbol{\theta}$ は，分子進化速度行列を規定するパラメータ $\boldsymbol{\theta}_0$ と系統樹を構成する枝の長さ $\mathbf{b}$ からなる．尤度を最大化することにより，系統樹の形や枝の長さを最尤推定する．通常，まずは枝ごとに期待置換数を自由パラメータとして割り当てられたモデルを解析し，それらを枝の長さとして分子系統樹を描く．種分化後，両系統で現在にいたるまでの時間は同じであるため，2 つの下流の枝の長さを比較することにより分子進化速度の違いが検出される．

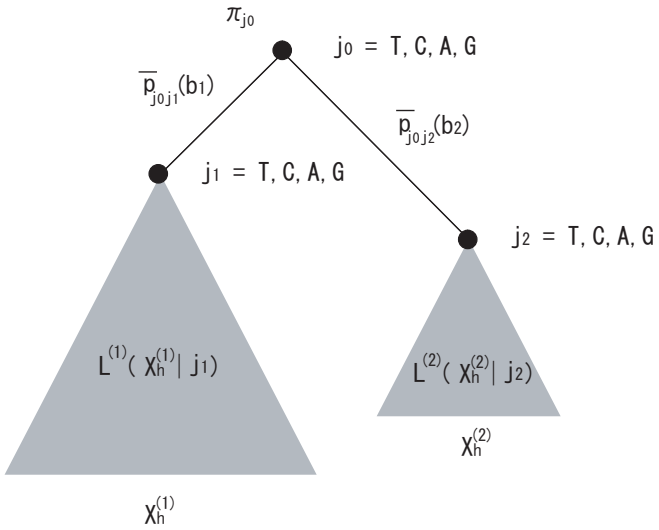


図 2: 部分木と条件付尤度

進化的に遠く隔たった種の配列を比較するときは，個々のサイトで多くの塩基置換が蓄積していることから，ノイズの中にシグナルが埋もれ，進化の履歴を推測することが困難となる．こうしたときは，たとえばタンパク質の進化を扱うような場合には，塩基配列を比較する代わりにアミノ酸配列を比較する．この場合も，推移確率行列が 20×20 の行列に変わるだけで，基本的な枠組みは同じである．なお，要素数が多いことから，個別のデータ解析ごとにアミノ酸置換の速度行列を精度良く推定することはできない．ゲノムデータベースに登録された数多くのタンパク質配列の比較から経験的に求められたものが利用されており，JTT モデルなどいくつかが知られている．

タンパク質の内部で骨格を形作る部分では，構造を不安定にする突然変

異は有害であるため，容易に変わることは許されず，分子進化速度は遅い．これに対して，タンパク質表面は，他のタンパク質や化学物質と直接結合する部位を除いては，比較的構造の自由度が高い．このため，分子進化速度は比較的速い傾向がある．さらに，後に図 4 でも見られるように，タンパク質は複雑に折りたたまれているため，構造で近接しているサイトが配列でも近接しているとは限らない．そこで，枝の長さについては混合モデルが多く使われる．すなわち，固定効果 $\mathbf{b}$ に期待値 1，サイト間で独立同分布する確率変数 λ を乗じ，各サイトで周辺尤度

$$\overline{f}(\mathbf{X}_h|\boldsymbol{\theta}) = \int f(\mathbf{X}_h|\boldsymbol{\theta}_0, \lambda\mathbf{b})g(\lambda|\alpha)d\lambda \tag{2.5}$$

をとることにより，配列内の不均質性を考慮に入れて最尤推定することができる．また，ある部位での置換速度は隣接する塩基やアミノ酸に影響されることがある．たとえばタンパク質の進化のモデリングでは，立体構造上近接するアミノ酸の間の相互作用を考慮に入れた統計モデルが開発されている．

3　多重遺伝子族の進化：リボヌクレアーゼと EDN と ECP

アミノ酸置換などの点突然変異に加えて，ゲノムは遺伝子重複を通じてレパートリーを広げていく．遺伝子重複後のコピー遺伝子の運命については，大野の新機能化仮説が有名である．一方の遺伝子はそれまでの機能を引き継ぐが，他方は機能的な制約から解放され，多くの変異を受ける．やがて機能を失うものがほとんであるが，稀に運よく新たな機能を獲得する，というものである．

これに対して，あるいはこれを補償する形で，部分機能化仮説も広く受け入れられている．これは遺伝子の発現パターンに注目するものである．先に述べたように，DNA は RNA に転写され，これがアミノ酸に翻訳されてタンパク質を生み出す．したがって，転写と翻訳の効率が，タンパク質の生産量を左右する．この効率を決める部位は主としてアミノ酸をコードする領域外にいくつかあり，すべてを同定することは一般に困難である．遺伝子重複はこうした部位の変異を促し，異なる組織や時間の環境に応じて，二つの遺伝子が特異的に発現することを可能にした，と説明する．

二つのコピーの分子進化は確率的に起きるが，両者に非対称性が存在することに着目し，これを指示する実証データを得た研究もある．非対称性の原因としては，ゲノムの環境が均質でないことが考えられる．局所的な組換え価が小さい領域では，そこにおける多様度は低く，突然変異は定着しやすい．これに対して，組換え価の大きな領域は，有害な突然変異はほぼ確実に排除される．したがって，コピー遺伝子の置かれた位置におけるゲノムの局所環境により，どちらの遺伝子が多くの変異を受けやすいか，予め運命づけられているのである．

同一ゲノムの中に似た遺伝子配列が複数あると，それらは遺伝子重複によって生まれた遺伝子であるとみなすことができる．これらの遺伝子をまとめたセットを，多重遺伝子族という．特に，数多くのメンバーを持つ多重遺伝子族は，相次ぐ遺伝子重複で，コピー遺伝子の多くが集団から脱落することなく残っている．これは，この遺伝子が作り出す機能や形質に，強い多様化圧がかかっていることを意味する．

こうした多重遺伝子族のひとつに，リボヌクレアーゼ (RNase) 多重遺伝子族がある．生体内で触媒活動を行うタンパク質を酵素というが，リボヌクレアーゼは，タンパク質を作り出す中間体である RNA を分解する酵素である．生物はこれとともに，RNase と強力に結合してその働きを阻害する RNase 阻害タンパク質を持っており，両者のバランスにより，生体内の環境に応じて触媒反応などの一連の化学反応が適切に制御されている．このさじ加減の重要性が多様化圧となって現れる．

図 3(a) は，ゲノムデータベースから多重遺伝子族のメンバーであるヒト EDN 遺伝子と高い相同性を持つ配列を集めて得られた RNase 分子系統樹である．これらの配列には，ヒトの遺伝子のほかに，サルやマウスなど，さまざまな生物から得られた配列も含まれる．したがって，分子系統樹は種分化と遺伝子重複の歴史を表現している．図からわかるように，大雑把に分けると 10 ほど (RNase 1-RNase 10) のグループに分けられるが，ここでは RNase 2 と RNase 3 に着目し，その適応進化を探ることにする．

3.1 遺伝子重複後の進化速度の加速

これら二つのグループは，それぞれ好酸球由来神経毒 (EDN, eosinophil-derived neurotoxin)，好酸球塩基性タンパク質 (ECP, eosinophil-cathionic protein) と呼ばれている．ヒトにウイルスや病原菌が侵入すると，抹消血白血球のひとつである好酸球が増加し，EDN と ECP が彼らの RNA を分解

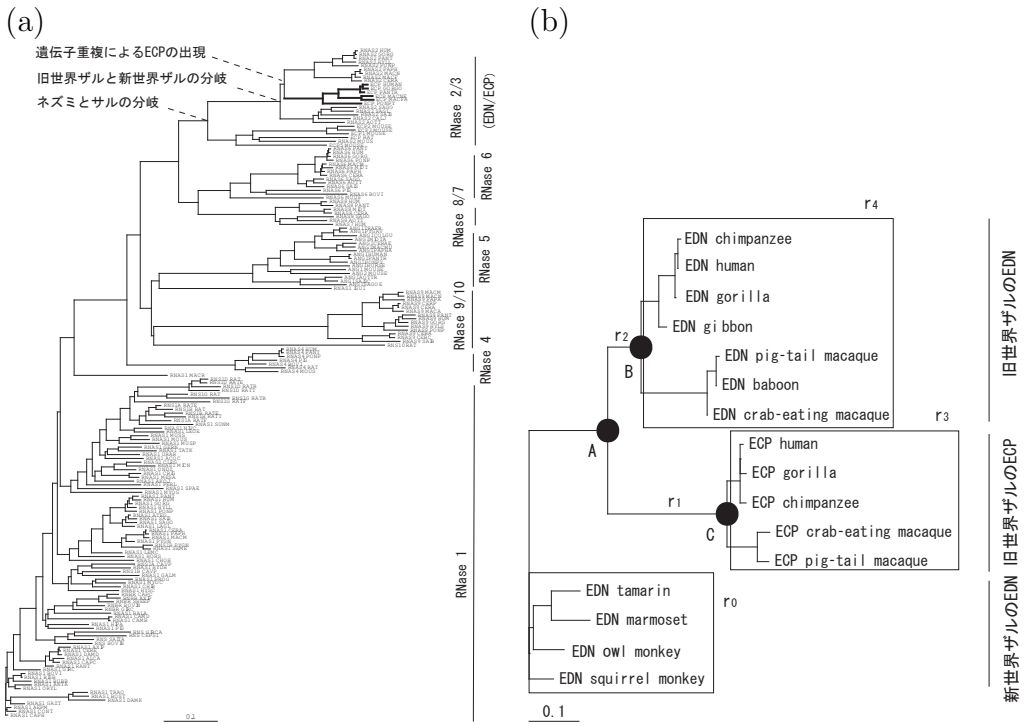


図 3: リボヌクレアーゼ多重遺伝子族：(a) ヒト EDN 遺伝子と高い相同性を持つ配列を集めて得られた全体像, (b) EDN (RNase 2) と ECP (RNase 3) にフォーカスを当てる．A は遺伝子重複，B, C は旧世界ザルの中からヒト上科（ヒト，チンパンジー，ゴリラ，テナガザル）が分岐した時点に対応している．

して死滅に追い込む．発現に不具合が起きると，アレルギー疾患などを引き起こし，私たち自身を痛めてしまうが，生体防御の重要な手段として注目されている．EDN は弱塩基性で，HIV-1 など RNA ゲノムを持つウイルスを撃退する．これに対して ECP は強塩基性で，EDN より活性は低いが，病原菌をやっつける．これらはネズミなどのげっ歯類からサルが分かれた後に遺伝子重複により誕生した．

3100 万年前にローラシア大陸が旧世界（現在のヨーロッパ・アジア・アフリカ）と新世界（アメリカ・オーストラリア）に分断されるのに伴い，旧世界ザル（ヒト，チンパンジー，ゴリラ，オランウータン，マカクなど）と新世界ザル（タマリン，マーモセット，リスザルなど）に分かれる．EDN はすべてのサルに存在するが，ECP は旧世界ザルにしか見られない．図 3(a)

からは，旧世界ザルと新世界ザルが分かれた直後か，間もない時期に，旧世界ザルにおいて遺伝子重複が起き，ECP が誕生したことが読み取れる．これにより，ウイルスに加えて病原菌に対する抵抗性を獲得した．加えて，旧世界ザルの持つ EDN は新世界ザルのそれよりも，活性が高い．

そこで，遺伝子重複後，両遺伝子に何が起きたか，見ていく．以下の解析は Zhang らの 2 つの研究 (Zhang *et al.* (1998), Zhang and Rosenberg (2002)) を情報量統計科学の視点で定量化して再解析し，重複遺伝子の適応進化を詳細に調べたものである．図 3(b) は，サルの持つ EDN/ECP アミノ酸配列の最尤系統樹である．図中 A は遺伝子重複，B および C は旧世界ザルの中からヒト上科（ヒト，チンパンジー，ゴリラ，テナガザル）が分岐した時点に対応している．A から B にいたるまでの進化時間と A から C にいたるまでの進化時間は等しいにも係らず，後者の枝の方がかなり長い．このことは遺伝子重複後早い時期に，ECP において数多くのアミノ酸が起きたことを示唆している．機能的な制約が緩んだり，あるいは多様化圧がかかると，分子進化速度が上がる．そこで，分子進化速度の系統の間の変動のパターンから，適応進化の様相を推測する．

遺伝子重複直後の ECP の枝における速度を r_1，EDN の枝における速度を r_2，それらの下流における速度をそれぞれ r_3, r_4 として，局所的な分子時計を利用して分岐年代と進化速度を最尤推定した (Yang and Yoder (2003))．図 3(a) から伺われるように，ここでは大まかに捉え，遺伝子重複は 3100 万年前に旧大陸と新大陸に分かれた直後に起きたとした．また，旧世界ザルからヒト上科（ヒト，チンパンジー，ゴリラ）が分かれた年代を 2500 万年前とする先行研究に基づき，B と C の分岐年代を 2500 万年前とした．(2.3) 式における枝の長さ b_j をその枝における分子進化速度 $\tilde{r}_j$ と時間間隔 δt_j の積で表現した対数尤度を最大化する．規準化されたアミノ酸置換速度行列としては JTT モデルを用い，配列内のサイトの間で速度がガンマ分布に従い分布することを考慮に入れた（(2.5) 式）．

表 1 は，進化速度の変化に 4 通りのシナリオを考え，最大対数尤度と AIC (Akaike (1974))，および旧世界ザルの EDN と ECP について分子進化速度の最尤推定値を示している．モデル 1 は進化速度が一定とするものである．ここでは旧世界ザルの EDN および ECP の分子進化に注目しているため，それらの間の系統の外に位置する新世界ザルの EDN の分子進化速度の推定に影響され偏りを生じないよう，r_0 と r_1 $(= r_2 = r_3 = r_4)$ は別扱いにした．したがって，推定すべき未知パラメータとしては r_0 と r_1，12 個の分岐年代（旧世界ザルの EDN および ECP の間の分岐年代，新世界ザルの EDN の間の分

	モデル 1	モデル 2	モデル 3	モデル 4
	$r_1 = r_2 = r_3 = r_4$	$r_1 = r_2 \neq r_3 = r_4$	$r_1 \neq r_2 \neq r_3 = r_4$	$r_1 \neq r_2 \neq r_3 \neq r_4$
最大対数尤度	-1427.11	-1399.28	-1395.11	-1394.42
パラメータ数	15	16	17	18
AIC	2884.22	2830.56	2824.22	2824.84
分子進化速度 ($\times 10^{-8}$/site· 年)				
r_1	0.512 ± 0.071	2.590 ± 0.528	4.189 ± 0.976	4.206 ± 0.978
r_2			1.150 ± 0.519	1.142 ± 0.519
r_3		0.319 ± 0.053	0.318 ± 0.052	0.266 ± 0.061
r_4				0.373 ± 0.077

表 1: EDN と ECP の分子進化速度

岐年代，そしてそれらの直近の共通祖先の年代），そしてガンマ分布の形状を規定するパラメータの，合計 15 個のパラメータがある．モデル 2 は，遺伝子重複直後に両コピーとも同様に速度が変化し，しばらくして機能が獲得されると，両コピーとも同じ速度になるとする．すなわち，$r_1 = r_2 \neq r_3 = r_4$ を仮定する．モデル 3 は，遺伝子重複直後の分子進化速度の変化については，二つのコピーの間で異なることを許す ($r_1 \neq r_2$)．モデル 4 は，r_3 と r_4 が等しいことも仮定しない．

進化速度は一定とするモデル 1(AIC=2884.22) に対して遺伝子重複後の加速 (モデル 2, AIC=2830.56) は明確に支持されている．さらに，遺伝子重複後の両遺伝子の進化速度を比較すると，EDN への系統においては 1 年にサイトあたり $1.150 \pm 0.519 \times 10^{-8}$ 回 ($\pm$ の後は標準誤差）の速さでアミノ酸置換が起きたのに対し，ECP への系統では 1 年にサイトあたり $4.189 \pm 0.976 \times 10^{-8}$ 回と推定され，4 倍近く速い (モデル 2, AIC=2824.22)．それ以降は 1 年にサイトあたり $0.318 \pm 0.052 \times 10^{-8}$ 回と，遺伝子重複直後に比べて 1 桁遅く，適応した機能が成熟した後は変化に対して制約がかかっていることが読み取れる．そして，モデル 3 の AIC は，両系統の進化速度の違いを許したモデル 4 の AIC=2824.84 と大差なく，いまでは機能的な制約の強さはほとんど同程度になっていることが伺われる．

3.2 祖先配列の復元と速度変化のサイト

遺伝子重複して比較的早い時期に，何かが起きたことは確かである．アミノ酸配列の急激な変化の中でいったい何が適応進化に結びついたか，その正体を突き止めるために，遺伝子重複直後のアミノ酸の置換を詳細に追ってみる．各サイト $h = 1, \ldots, n$ の尤度は (2.4) 式で表現されているが，ベイ

ズの定理より内部節におけるアミノ酸の祖先形質の事後確率が

$$P\left(j_0|\mathbf{X}_h,\boldsymbol{\theta}\right)\frac{\pi_{j_0}f^{(1)}(\mathbf{X}_h^{(1)}|\boldsymbol{\theta})f^{(2)}(\mathbf{X}_h^{(2)}|\boldsymbol{\theta})}{f(\mathbf{X}_h|\boldsymbol{\theta})}$$

のように得られる．配列データに基づくパラメータの最尤推定量 $\hat{\pi}_j$, および $\hat{\boldsymbol{\theta}}$ を代入することにより，経験ベイズによる事後確率を得ることが出来る．そこで，この事後確率が 90%以上になる祖先形質に基づき，図 3(b) で A と B を結ぶ枝，および A と C を結ぶ枝における変化を追うことにする．

まず，ECP の系統（図 3(b) で A と C を結ぶ枝）における大きな加速は，多少のサイトの置換では説明できず，タンパク質全体が変化していることを想像させる．実際，全 133 アミノ酸残基のうちほぼ 4 分の 1 である 33 残基がアミノ酸置換を起こしていた．変異の部位はタンパク質全体にわたり，図 4(a) に見られるように，EDN と似た形を保ちながらも，凹凸に変化をもたらしている．この数多くの変異が触媒作用に係る領域（図 4(b)）にも影響し，活性が下がったと思われる．ただそのうち 13 の変異が非常に塩基性の強いアルギニンへの変化であった．病原菌の細胞表面は負に帯電していることから，こうした変異により，病原菌細胞膜への接着が促され，RNase 加水分解により細孔を開けたと推察される (Zhang *et al.* (1998))．

EDN の系統（図 3(b) で A と B を結ぶ枝）における速度も，ECP の系統ほどではないが，加速している．祖先配列を復元してみると，図 4(b) に見られるように，9 つのアミノ酸残基で置換が起きていた．触媒部位における変化は見られない．その代わり，これに隣接する第 132 アミノ酸残基の変異は，活性に重要であることを過去の実験結果が示していた．そこで Zhang and Rosenberg (2002) は，このアミノ酸残基とこれに隣接する第 64 アミノ酸残基に注目し，これら 2 つのサイトにおけるアミノ酸置換がともに変化することにより，はじめて活性が大幅に上がることを確認した．図 4(a) と同図 (b) を対比し，触媒部位近くのタンパク質の形状の変化を見てみると，第 64 および第 132 アミノ酸残基とさらにこれに隣接する第 60 および第 108 アミノ酸残基の計 4 サイトで相次ぐ置換が起きた結果，全体としてこの領域が隆起している．図から見て取れるように，配列の上では遠く離れているものも，立体構造上では近接することがしばしばある．

ところで，9 つの置換には，二つのコピー遺伝子の間の機能の多様化に本質的に貢献したものと，偶然に起きた置換で特に有害ではないためそのまま集団に定着したものとが混在している．分子進化の統計的モデリングを通して，前者を選び取ることができよう．一般に，機能に直接結びついているアミノ酸残基は構造上の制約が強く，たとえ変異が起きても多くは個体

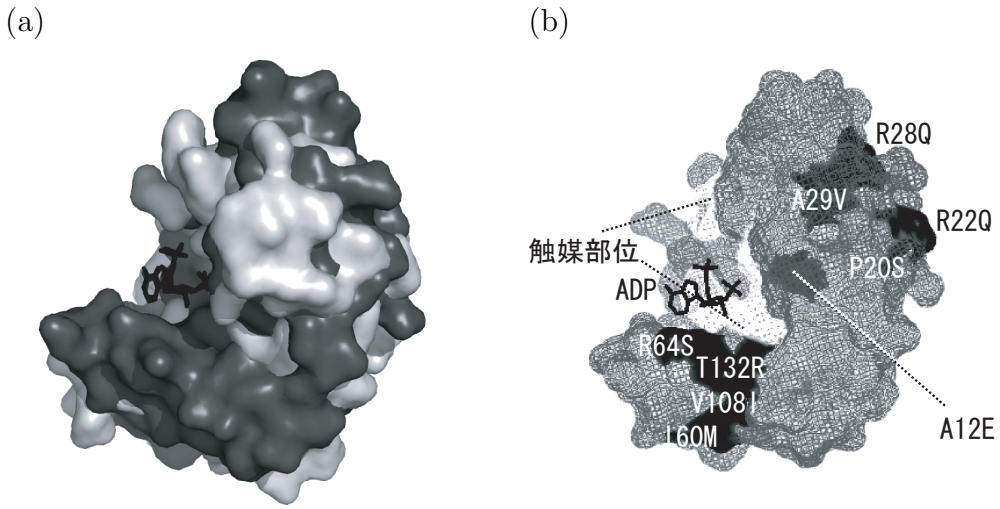


図 4: 遺伝子重複後のアミノ酸置換と微細構造の変化．(a) ヒトの EDN(暗灰色) と ECP(明灰色) を，結合のターゲットとなる ADP を中心に位置併せして重ね合わせたもの (PDB コード：1HI3 (EDN),1H1H (ECP)), (b) EDN の触媒部位 (白色：Leonidas *et al.* (2001)) と遺伝子重複後，類人猿にいたる系統で起きた 9 つのアミノ酸置換（黒色）．タンパク質の裏側も見えるように，メッシュでタンパク質表面を表示している．たとえば，V108I は 108 番目のアミノ酸残基がバリン (V) からイソロイシン (I) に変わったことを意味する．

の生存力を著しく下げる．そこで，3.1 項では遺伝子配列全体の平均的な速度の変化を解析したが，同様の解析をアミノ酸残基ごとに行うことにより，EDN と ECP の間で機能的な制約の強さが変化したアミノ酸残基を検出することができる．さらに重要なのは，現在では EDN と ECP の両方で機能的な制約が強く，保存されているにも係らず，EDN と ECP ではアミノ酸が異なるサイトである．そこで，塩基性・酸性や疎水性・親水性により 20 のアミノ酸を大きくグループ分けしてアミノ酸置換の速度を

$$r_{ij} = \begin{cases} r_w & \text{(同一グループ内の置換)} \\ r_b & \text{(グループをまたぐ置換)} \end{cases}$$

と，グループ内の置換とグループ間の置換に分解する (Gu (2006))．A と B を結ぶ枝で起きた置換から推定される速度パラメータが，データベースに登録されたアミノ酸配列全体からの平均値と大きくずれていた場合には，偶

然を超えてアミノ酸の形質を変化させることにより新たな機能に結びついた置換であると予測する．

こうして浮かび上がってきたアミノ酸残基は，実は上で見た 4 つのサイトではなく，第 12 番目のアミノ酸残基であった．このアミノ酸は触媒部位の裏側にあるため，直接的な関係はなさそうにも見える．しかし，詳細に見てみると，触媒部位の 1 つである第 14 アミノ残基のグルタミンと 3Å 程度と非常に近接しており，相互作用をしている可能性がある．小さなアラニン (A) が，1.5 倍程度の大きさを持つグルタミン酸 (E) に置換することにより，内側から触媒部位を押し出し，活性を上げたのではないか，という推論も成り立つ．触媒部位自身の変異は活性に対する影響力が強く，さじ加減が難しい．これに対して，近接する部位の変異，あるいはさらにこの近接する部位に影響を与える部位の変異が，間接的に触媒部位における微細構造を変化させることもある．その効果は触媒部位自身の変異よりも弱いであろうが，自由度が高いだけに，微調整が容易であったのであろう．

3.3 探索から仮説，反証可能性

ここで見られた遺伝子重複直後の多様化圧の背景に思いを馳せてみよう．生命体の中では，RNA を加水分解する RNase と，これと結合し分解を防ぐ RNase 阻害タンパク質がバランスし，生命活動が維持されている．遺伝子重複によりコピー遺伝子を持った個体では，そのままでは RNase が過剰に発現してバランスが崩れ，恐らく生存力が低下したであろう．生命体にはこうした事態を回避する機構が備わっている．遺伝子重複や人工的な遺伝子導入によりゲノムが複数のコピーを獲得すると，遺伝子の発現が抑制されるのである．タンパク質に翻訳されない遺伝子配列は，機能的な制約から解放され，その後多くの変異を受けることになる．

現在私たちが観測している分子進化は，長い歴史の中で集団に定着した変異である．普通の環境では，変異を受けたコピーは，仮に遺伝子としての資格を失い偽遺伝子とならなくても，他を凌駕し，集団を埋め尽くすまでにはいたらないであろう．旧大陸に生息する集団が疫病の大流行などの強い淘汰圧を受けて，絶滅の危機にさらされたとすると，話は別である．活性が低いながらも病原菌への抵抗性を持ち，淘汰圧に耐える突然変異体が現れると，こうした個体のみが生き延び，集団を支えて行ったであろう．実際に，環境や疫病の淘汰圧が遺伝子の組成を急速に変化させることは，ヒトの病気関連遺伝子の地理分布や殺虫剤抵抗性を持つ昆虫，治療を受ける

患者の体内におけるウイルス集団など，さまざまな場面で認められている．ただ，こうした推論が科学的仮説としての資格を持つためには，実験や調査のデータに照らし合わせて実証あるいは反証できなければならない．反証可能であり，さらに普遍的な現象を説明する仮説は，科学的な価値を持つ．

ここでは必要条件としての実証・反証可能性について，さらに検討したいと思う．遺伝子重複に伴うリボヌクレアーゼの過剰発現と有害性については，恐らく実験で確認することができるであろう．他方，大陸分断後の疫病の大流行については，直接検証することは困難かもしれない．ただし，ゲノム比較を通じて間接的に検証する可能性は残されている．これは，分子進化速度と集団の大きさの間の負の相関関係に注目するものである．

突然変異の多くは，まわりと同等か，あるいは有害であると信じられている．そしてこのことは将来，直接定量的に検証されるであろう．このとき，もしも集団が強い淘汰圧を受けて絶滅の危機にさらされるほどに細ったとすると，この間分子進化速度が上がることになる．ただし，分子進化速度は集団の大きさに影響されるのみならず，遺伝子にかかる機能的な制約の強さが変化したり，世代の長さが変わったりしても変化する．機能的な制約の強さについては，遺伝子ごとにばらばらであろう．また大陸分断と前後してこの時期のみに世代の長さが変化するとは考えにくい．したがって，ゲノム上の数多くの遺伝子配列や遺伝子領域外の配列を分子系統解析して，それらに共通して，大陸分断直後の時期で進化速度が加速していることが認められれば，ここで想定した仮説は支持されることになる．

一般に，(2.3)–(2.4) 式で表される配列の尤度は，各枝における速度とその間の時間間隔の積の関数である．したがって，時間と速度という二つの要素を個別に推定することはできない．すなわち，分子進化速度の変化を推定するためには，何らかの制約条件を課す必要がある．この制約条件は緩やかなものであることが望ましい．Thorne *et al.* (1998) は，分子進化速度の対数をとったものがブラウン運動するという事前分布を導入し，ベイズの枠組みで分子進化速度と分岐年代を推定する方法を提案した．速度変化の揺れの大きさについては予め固定せず，不確実性を考慮に入れることにより，頑健な推定を可能にした．Huelsenbeck *et al.* (2000) は，ポアソン過程によるジャンプ型速度変化の確率モデルを開発した．

現在のところ分岐年代の推定には前者が多用されているが，これらの手法は，速度変化を平滑化することにより，パラメータ推定の識別可能性と頑健性を調和させている，という点で共通している．集団が絶滅の危機に瀕しながらも生きながらえることのできる時間は，何百・何千万年という

オーダーを問題にする進化的時間スケールにおいては，ほんの瞬間であろう．図5は，こうした状況を想定し，分子進化速度が小刻みに変化する過程で瞬間的に大きく加速する状況をシミュレートしている．通常に平滑化を行っても，全体としての傾向は掴めるものの，瞬間的な加速は捉え切れていないことがわかる．検出力が低い分析手法では，実証にも反証にも効力を発揮しない恐れがある．ある枝における局所的な変化を高い感度で検出する手法が開発されてはじめて，反証可能性が担保されるのである．Kitazoe *et al.* (2007) は，分子進化の解析においては平均速度は調和平均をとることが自然であることに注目し，速度の逆数の平滑化が短期間の強い加速を感度良く検出することを示した．今後，分子進化に対する科学的仮説の検証・反証可能性を高めるという観点からも，進化速度の推定手法がさらに研ぎ澄まされていくことが期待される．

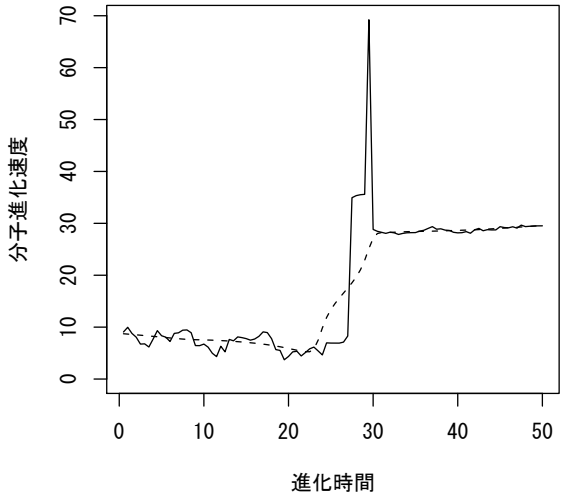


図 5: 速度変化の連続変化と不連続変化のシミュレーションと平滑化 (破線)

3.4 分離した仮説としての系統樹とモデルの検討

最後に，分子系統樹の推定における不確実性と偏りについて，触れることにする．分子系統樹の推定精度は，多くの場合サイトをリサンプリングするブートストラップにより評価される．あるいは，最近ではベイズにより事後分布として表現する方式を用いることも多い．. 数多くの種についてゲノムが解読されてくると，個々の遺伝子を超えて，ゲノム全体を解析す

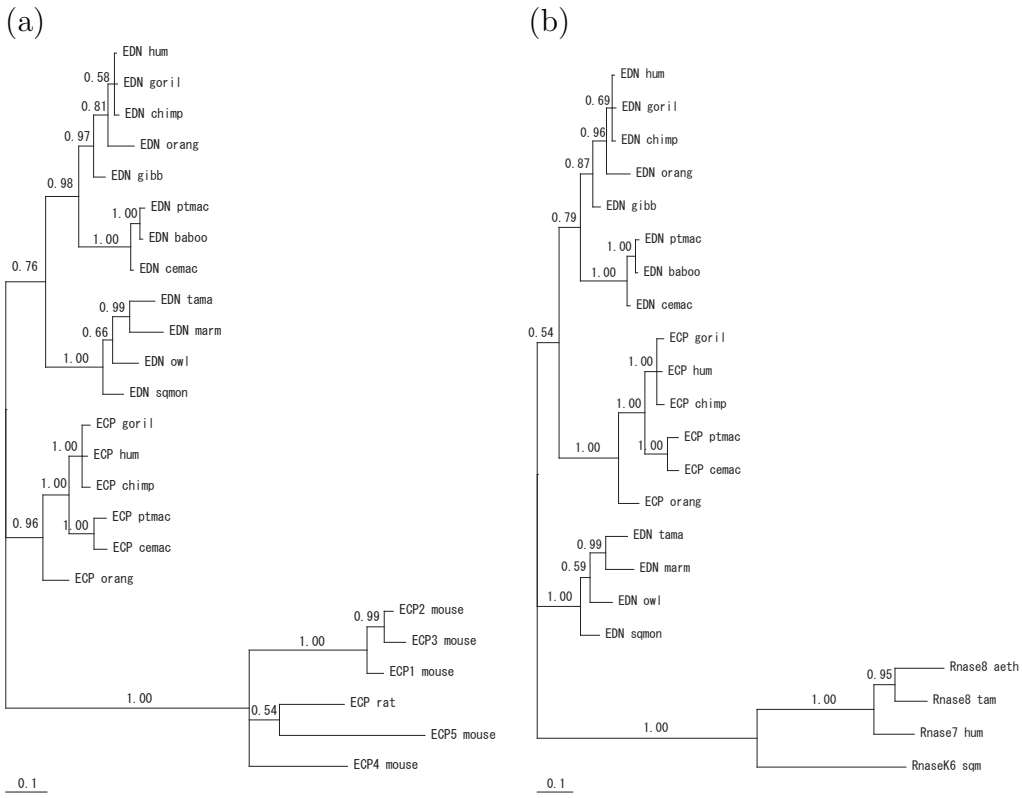


図 6: EDN と ECP のベイズ分子系統樹：(a) ネズミの RNase を外群にした場合, (b) RNase7 と RNase8 を外群にした場合

ることも日常的になろうとしている．こうした場合には，遺伝子間の分散を考慮に入れた評価が必要になってきた．分子進化のプロセスを記述する統計モデルは，改良が重ねられてきている．しかしモデルはあくまでも現象を簡潔に近似するものであるため，しばしば先行研究と照らし合わせて不自然な結果を得ることがある．

これまで新世界ザルの EDN を外群にして，旧世界ザルにおいて遺伝子重複後 EDN と ECP の系統で起きた変異を調べてきた（図 3(b)）．これはリボヌクレアーゼ多重遺伝子族全体から眺めた鳥瞰図（図 3(a)）と矛盾していない．そこで，旧世界ザルと新世界ザルの間で何が起きたか，詳細に追うために，最も近いネズミの RNase を外群にして，サルの EDN と ECP の分子系統樹をベイズ推定すると，図 6(a) のようになった．枝の上の数字は，事後確率である．これによると，旧世界ザルと新世界ザルが分かれる前に遺伝子重複が起き，その後新世界ザルにおいて ECP が抜け落ちたことになる．

ネズミの RNase の代わりに，次に近い関係にある RNase 7, 8 を外群にとると，再び上で解析した系統樹と矛盾しない系統樹が得られる（図 6(b)）．

特にコピー遺伝子が同一染色体上に近接して存在したりしている場合には，両遺伝子が似ているために，減数分裂により DNA 情報を受け渡す際に一方を他方に読み誤ることがある．この現象は遺伝子変換と呼ばれる．こうしたことが起きると，二つの遺伝子は，見かけ上協調して変化して行くことになる．現在の進化系統樹の推定はすべて，遺伝子が分岐した後は独立に変化することを仮定しているため，得られた系統樹は偏っている可能性がある．

そこで，これら二つの系統樹は統計的に有意に食い違っているものか，尤度比を通して調べてみよう．分岐の順番を示す系統樹の形をトポロジーという．このトポロジーを決めると，(2.3) 式により対数尤度が表現される．分岐年代や進化速度は，未知パラメータである．トポロジーが異なれば，未知パラメータの意味するところも異なってくる．すなわち，一方のモデルが他方のモデルを含むという関係にはなく，分離した仮説 (Cox (1962)) となっている．

二つのトポロジーの対数尤度をそれぞれ

$$l_i(\boldsymbol{\theta}_i|\mathbf{X}) = \sum_{h=1}^{n} \log f_i(\mathbf{X}_h|\boldsymbol{\theta}_i) \qquad i = 1, 2$$

とする．$\mathbf{X}_h,\ h = 1, \ldots, n$ の真の分布を $g(\cdot)$ で表すと，二つのモデルの真の構造からの遠さは Kullback-Leibler 情報量

$$I(g(\cdot) : f_i(\cdot|\boldsymbol{\theta}_i^*)) \equiv \mathrm{E}_{\mathbf{X}_1}\left[\log \frac{g(\mathbf{X}_1)}{f_i(\mathbf{X}_1|\boldsymbol{\theta}_i^*)}\right] \qquad i = 1, 2$$

により表現される．$\boldsymbol{\theta}_i^*$ は $I(g(\cdot) : f_i(\cdot|\boldsymbol{\theta}_i)$ を最小にするパラメータの値である．したがって，

$$\begin{aligned} d_{12} &= I(g(\cdot) : f_1(\cdot|\boldsymbol{\theta}_1^*)) - I(g(\cdot) : f_2(\cdot|\boldsymbol{\theta}_2^*)) \\ &= -\left(\mathrm{E}_{\mathbf{X}_1}\left[\log f_1(\mathbf{X}_1|\boldsymbol{\theta}_1^*)\right] - \mathrm{E}_{\mathbf{X}_1}\left[\log f_2(\mathbf{X}_1|\boldsymbol{\theta}_2^*)\right]\right) \end{aligned}$$

の符号でモデルの優劣が評価される．

母平均を標本平均で近似することにより，

$$\hat{d}_{12} = -\frac{1}{n}\left(l_1(\hat{\boldsymbol{\theta}}_1|\mathbf{X}) - l_2(\hat{\boldsymbol{\theta}}_2|\mathbf{X})\right)$$

系統樹	最大対数尤度	対数尤度比	標準誤差	p 値
T_1	-2601.9	5.4	10.0	0.589
T_2	-2607.2			

表 2: マウスの RNase を外群にし，サルの EDN/ECP については図 6(a) のトポロジーを仮定する系統樹 T_1 と、サルの EDN/ECP については図 6(b) のトポロジーを仮定する系統樹 T_2 の比較

を計算する．$\hat{\boldsymbol{\theta}}_i$ $(i = 1, 2)$ は最尤推定量である．その分散は，サイトの対数尤度比の標本分散により

$$\hat{\mathrm{V}}\left[\hat{d}_{12}\right] = \frac{1}{n(n-1)} \sum_{h=1}^{n} \left\{ \log \frac{f_1(\mathbf{X}_h|\hat{\boldsymbol{\theta}}_1)}{f_2(\mathbf{X}_h|\hat{\boldsymbol{\theta}}_2)} - \hat{d}_{12} \right\}^2 \tag{3.6}$$

により大まかに見積もることが出来る．中心極限定理により，配列の長さがある程度長い通常の場合では，この分布は正規分布で近似できる (Kishino and Hasegawa (1989))．比較するトポロジーが予め固定されたものではなく，推定されたものであることを考慮することが不可欠な場合もある (Shimodaira and Hasegawa (1999))．

マウスの RNase を外群にし，サルの EDN/ECP の部分については図 6(a) のトポロジーを仮定する最尤系統樹 T_1 と，サルの EDN/ECP の部分については図 6(b) のトポロジーを仮定する最尤系統樹 T_2 を比較した（表 2）．その結果 p 値は 58.9%となり，両者に有意差はないことがわかった．

今の場合は幸い事なきを得たが、ゲノムデータが充実するにつれて、ノイズの寄与率は減少し、単純なモデルとは有意に矛盾する不可解な現象に数多く遭遇することになるであろう．モデルの改良と、妥当性の検証のためのさらに研ぎ済ませれた道具立てが必要になることが予想される．

参考文献[2]

[2]生命科学における学術論文の多くは、刊行後しばらくすると、インターネットから無料でダウンロードできるため、できるだけ読者が原典に触れることができるよう、配慮した．また、モデルと解析手法の開発者たちは、恥ずかしながら著者を例外として、次々とソフトウエアを開発し、一般に公開している．読者は、本章における解析に用いた `DIVERGE`, `MrBayes`, `PAML`, `PHYLIP`, `PYMOL`, `R`, `Treeview` の他、各種ソフトウエアを駆使することにより、最新の手法の有効性と限界を見抜き、さらに改良された方法を開発することであろう．

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Contr*, **AC-19**, 716–723.

[2] Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Royal Stat. Soc. Ser. B*, **24**, 406–424.

[3] Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc. Sunderland, Massachusetts.

[4] Gu, X. (2006). A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol. Biol. and Evol.*, **23**, 1937–1945.

[5] Huelsenbeck, J. P., Larget, B. and Swofford, D. L. (2000). A compound Poisson process for relaxing the molecular clock. *Genetics*, **154**, 1879–1892.

[6] Kimura, M. (1968). Evolutionary rate at molecular level. *Nature*, **217**, 624–626.

[7] Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data. *J. Mol. Evol.*, **29**, 170–179.

[8] Kitazoe, Y., Kishino, H., Waddell, P. J., Nakajima, N., Okabayashi, T., Watabe, T., and Okuhara, Y. (2007). Robust time estimation reconciles views of the antiquity of placental mammals. *PLoS ONE*, **4**, e384.

[9] Leonidas, D. D., Boix, E., Prill, R., Suzuki, M., Turton, R., Minson, K., Swaminathan, G. J., Youle, R. J., and Acharya, K. R. (2001). Mapping the ribonucleolytic active site of eosinophil-derived neurotoxin (EDN). *J. Biol. Chem.*, **276**, 15009?-15017.

[10] Ohta, T. (1972). Population size and rate of evolution. *J. Mol. Evol.*, **1**, 307–314.

[11] Ronquist, F. and Huelsenbeck, J. P. (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

[12] Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. and Evol.*, **16**, 1114–1116.

[13] Thorne, J. L., Kishino, H. and Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. and Evol.*, **15**, 1647–1657.

[14] Yang, Z. and Yoder, A. D. (2003). Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst. Biol.*, **52**, 705–716.

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2011 年 10 月
第 2 部 生物・健康と統計科学

第6章　生命システムネットワークを明らかにするための統計的モデリング

井元清哉[1]

(東京大学医科学研究所ヒトゲノム解析センター 准教授)

生命科学において，統計科学的アプローチが有効な例の一つとして遺伝子発現データからの遺伝子制御ネットワークの推定問題を取り上げる．遺伝子発現データとは，ある状況下において細胞内の各遺伝子が生成しているメッセンジャー RNA の量のデータであり，マイクロアレイや次世代シークエンサーによる RNA-seq によって計測される．遺伝子制御ネットワークとは，生命システム機構の一部である．このネットワークによる情報伝達を経て，遺伝子は必要なときに必要なだけのメッセンジャー RNA を生成し，それを基にタンパク質が合成され生命は維持されている．遺伝子発現データに基づく遺伝子制御ネットワークの推定は，観測データに基づくグラフィカルモデルの構造推定に帰着される．この問題に対して，本章では，ノンパラメトリック回帰を用いた変数間の非線形構造探索法，ベイズアプローチ

[1]imoto@ims.u-tokyo.ac.jp

による統計的モデリング，モデル評価を行うための情報量規準という一連の統計手法の有用性を紹介する．

1　はじめに

1.1　バイオインフォマティクス

バイオインフォマティクス (Bioinformatics) という単語は造語である．容易に想像できるように，生物学 (Biology) と情報学 (Informatics) の 2 つの単語を合わせて作られた．このような言葉が現れた背景には，1990 年代初頭からスタートした国際ヒトゲノム計画がある．この国際ヒトゲノム計画は，ヒトの全 DNA 配列を決定することを目的としたものである．ヒトゲノム計画以前にも生物学者は，自分の興味のある遺伝子や関連する少数の遺伝子について DNA 配列を調べ，それらの機能 (細胞内における遺伝子の役割) を小〜中規模に調べていた．ヒトゲノム計画は，2003 年のヒトゲノム配列決定の宣言により一通りの成果をあげ，終了した．この宣言により，ヒトゲノムに関する研究は終わったとしばしば誤解されることがある．しかしながら，実際は，ヒトにおける分子生物学の研究は，ヒトゲノム配列決定によってようやくスタート地点に立つことができたばかりである．

ヒトゲノム配列決定により，ヒトには 2 万から 2 万 5000 個程度の遺伝子領域とよばれるタンパク質をコードしている領域が存在することが明らかになった．この遺伝子コード領域をひな形として，約 10 万種類のタンパク質が合成され我々の生命を維持している．DNA 配列の決定により，SNP (Single Nucleotide Polymorphism，一塩基多型) に代表されるようなヒトの個体差や疾患リスクに関与する DNA 配列のバリエーション，細胞内において各遺伝子が生成するメッセンジャー RNA の網羅的計測，選択的スプライシングによるメッセンジャー RNA のバリエーション，タンパク質間の相互作用，タンパク質立体構造，タンパク質局在情報など様々な計測データを高速かつゲノムワイドに（もしくはゲノムワイドに近い網羅的なスケールで）得ることができるようになった．このような膨大な計測データや先験的な知識をまとめ，整理するためのデータベースの構築やデータ解析による知識抽出・発見がバイオインフォマティクス分野に期待されている．

1.2　マイクロアレイデータ

遺伝子からタンパク質が合成される過程は，2 段階に分けることができる．まず，遺伝子をひな形としてメッセンジャー RNA が生成され，そのメッセンジャー RNA がタンパク質へと翻訳される．マイクロアレイにより計測されるのは，各遺伝子が生成しているメッセンジャー RNA の量であ

る．その原理については，ここでは詳しくは述べないが，著者の web ページ (http://bonsai.hgc.jp/~imoto/index_j.htm) に日本語の解説記事があるので興味のある読者は参考にされたい．

記号の導入を行おう．今，j 番目の遺伝子について計測したメッセンジャー RNA の量を x_j と表す．p 個の遺伝子について計測したのであれば $j = 1, 2, ..., p$ となる．つまり，$x_1, x_2, ..., x_p$ が 1 枚のマイクロアレイにより計測されるデータということになる．しかしながら，1 枚のマイクロアレイからでは病気に関連する遺伝子の予測などの統計解析は難しいため，通常は複数のマイクロアレイを用意する．具体的な例で説明しよう．今，正常細胞と肺癌の細胞を遺伝子の発現データを用いて区別したいとする．つまり，統計科学におけるいわゆる判別の問題を考える．このとき，n_1 人の正常細胞から n_1 枚のマイクロアレイデータ，n_2 人の肺癌細胞から n_2 枚のマイクロアレイデータを得ると合計 $n_1 + n_2$ 枚のマイクロアレイデータを得る．このマイクロアレイデータを用い，判別分析における変数選択を行うことにより，正常細胞，肺癌細胞を判別するために有効な遺伝子のサブセットを同定することができる．また，正常細胞，癌細胞というデータにラベルが付いた判別分析ではなく，ラベルの付いていない，いわゆるクラスタリングはマイクロアレイデータの解析に最もよく用いられている統計手法の一つである．図 1 にはマイクロアレイデータのクラスタリングの例を挙げた．このように通常マイクロアレイは複数枚準備するため，添え字を一つ追加し，x_{ij} は i 番目のマイクロアレイにより計測された j 番目の遺伝子の発現データとする．図 2 はマイクロアレイデータを Microsoft Excel で表示したものである．行に遺伝子の名前 (このデータは CodeLink 社のマイクロアレイデータであるため CodeLink 社の遺伝子 ID が書かれている)，列にマイクロアレイの ID が書かれている．マイクロアレイの種類としては，上の例で用いた CodeLink 社のもの以外に，Affymetrix 社，Agilent 社，illumina 社，日本では株式会社 DNA チップ研究所などがそれぞれ独自の技術によるマイクロアレイを販売している．このマイクロアレイデータには，必ずシステムノイズ，観測ノイズが含まれる．従って，このようなマイクロアレイデータからの情報抽出において統計科学的手法は必須の技術である．

1.3 生命システムネットワーク

生命をシステムとして理解することを目的とした研究分野は，特に，システムバイオロジーと呼ばれる．本節では，このシステムバイオロジーに

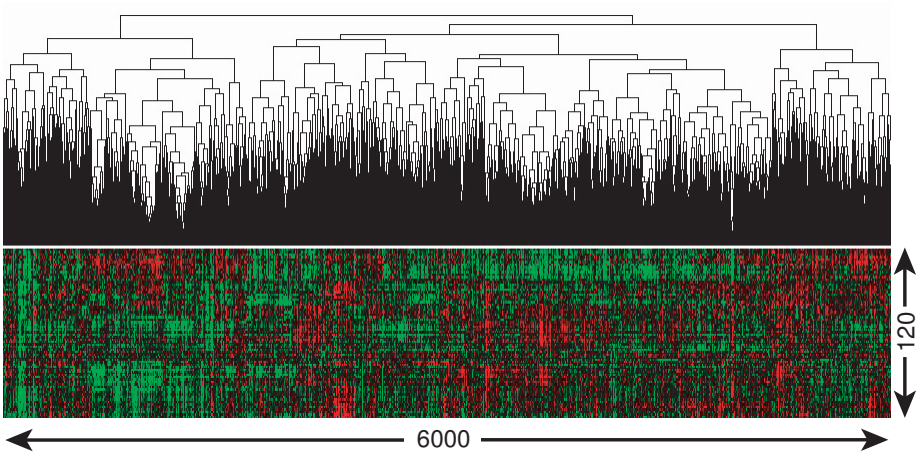


図 1: 出芽酵母遺伝子のクラスタリング例．120 枚のマイクロアレイデータに基づき，約 6000 個の出芽酵母遺伝子の階層型クラスタリングを行った．様々な状況下で計測した発現データが似たパターンを示す遺伝子たちは，その機能も共通であろうという予測の元で，クラスタリングにより似た発現データのパターンを示す遺伝子群を見つけ出し，その中に機能未知の遺伝子があればその機能予測を行うというものである．

おいて本質的な情報であるネットワークに注目する．前述した遺伝子たちは細胞内においてそれぞれの役割を担っている．そして，必要なときに働けるように遺伝子たちはネットワークを形成している．例えば，エネルギーを生み出すための代謝パスウェイは，酵素反応のネットワークである．また，シグナル伝達パスウェイでは，遺伝子から生成されたタンパク質同士の相互作用により細胞外からの刺激を核へと伝達する役割を担う．核内においては，転写因子と呼ばれる遺伝子たちが多くの遺伝子を制御しており，この転写因子の活性に従い，被制御遺伝子たちがタンパク質をどの程度生成するかが決定される．また，遺伝子間だけではなく，これら紹介したネットワーク間においても情報伝達は行われている．このような生命システムネットワークの中でも，本章では，遺伝子制御ネットワーク (遺伝子ネットワーク) に着目し，それをゲノムワイドに計測されたマイクロアレイデータに基づき推定する問題を考える．

簡単に遺伝子間の制御システムを説明しよう．まず，ある遺伝子 A からメッセンジャー RNA が生成され，タンパク質へ翻訳されたとする．このタンパク質が，遺伝子 A が制御する遺伝子 B のある特定の部分 (遺伝子の上

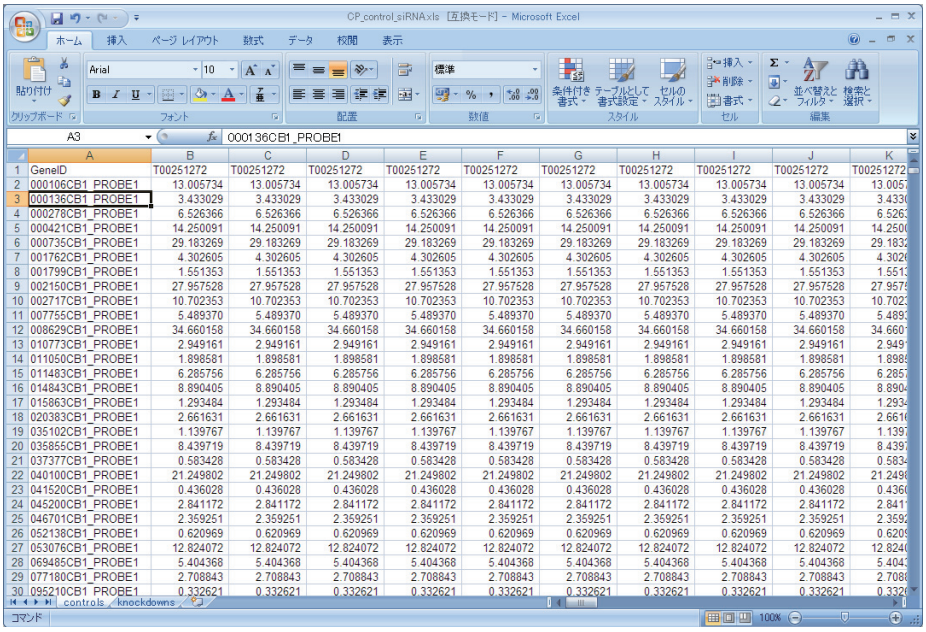


図 2: 遺伝子発現データの例．複数のマイクロアレイにより計測された遺伝子発現データは表としてまとめることができる．ここでは Microsoft Excel を用いて表示した．各セル内の値が遺伝子発現量 (x_{ij}) を表す．

流配列に存在する結合配列と呼ばれる短い配列) に結合することで，遺伝子 B はメッセンジャー RNA の生成を開始する．このようなシステムが複雑に入り組んでいるものが遺伝子ネットワークである．

この問題に対して，著者らの研究グループはベイジアンネットワークとノンパラメトリック回帰を組み合わせた統計モデルを構築し，そのネットワーク構造を推定する研究を行ってきた．各遺伝子を確率変数とみなし，マイクロアレイデータはその実現値であると考えることで，ベイジアンネットワークによる遺伝子ネットワークの推定問題は，ベイジアンネットワークの構造推定の問題として定式化することができる．このとき，どのような構造を選択すべきかは統計的モデル選択の問題となり，赤池情報量規準，ベイズ型情報量規準など様々なモデル選択基準を用いることができる．本章では，その数理的な基礎を紹介し，この遺伝子ネットワーク推定技術にどのような発展性・可能性があるのかを議論したい．

本章の構成は次の通りである．2 節ではベイジアンネットワークとノンパラメトリック回帰を用いた統計モデルについて紹介し，その推定法につい

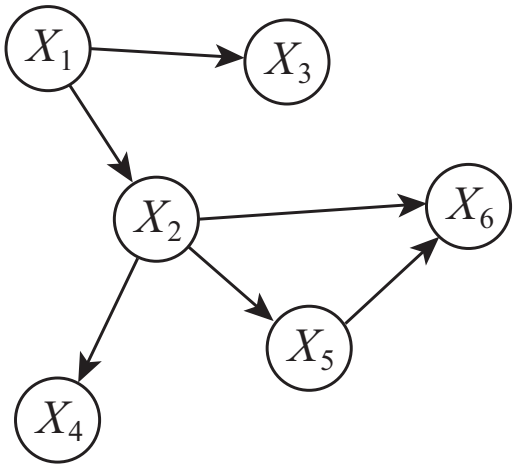


図 3: 非閉路有向グラフの例.

て述べる．3 節では，実際のマイクロアレイデータの解析例を示す.

2 ベイジアンネットワークによる遺伝子ネットワーク推定

2.1 遺伝子ネットワーク推定のための統計モデル

2.1.1 ベイジアンネットワーク

ベイジアンネットワークは，多数の確率変数間の依存関係を捉えるためのグラフィカルモデルの一つである．今，p 個の確率変数 $X_1, X_2, ..., X_p$ 間の依存関係に着目する．X_i と X_j が関係があるか否かを判定するためのもっとも直感的な方法は X_i と X_j が独立か否かを調べることであると考えられる．すなわち $\Pr(X_i, X_j) = \Pr(X_i)\Pr(X_j)$ が成り立つか否かを判定すればよい．しかしながら，多数の確率変数がある場合，例えば X_i と X_j の関係に対して別の確率変数 X_k が介在し X_i と X_j の間に見せかけの関係を生じることがある．これは見せかけの相関（偽相関）と呼ばれる．この場合，介在する確率変数 X_k の値が与えられたとき，X_i と X_j が独立になる．これを，X_i と X_j は X_k を与えた下で条件付き独立という．数式で表すと，$\Pr(X_i, X_j|X_k) = \Pr(X_i|X_k)\Pr(X_j|X_k)$ が成り立つことをいう.

次に，条件付き確率と有向グラフの関係について説明する．今，確率変数間に非閉路有向グラフで表される依存関係があるとする．例えば，図3で表される非閉路有向グラフの関係が6つの確率変数 $X_1, ..., X_6$ にあるとする．また，各確率変数は，その非閉路有向グラフにおける直接の親確率変数にのみに依存依存すると仮定する．このとき，各 $X_1, ..., X_6$ の同時確率が

$$\begin{aligned}\Pr(X_1, ..., X_6) &= \Pr(X_1)\Pr(X_2|X_1)\Pr(X_3|X_1)\Pr(X_4|X_2) \\ &\quad \times \Pr(X_5|X_2)\Pr(X_6|X_2, X_5)\end{aligned}$$

と分解されることが分かる．一般には，$\mathcal{X} = \{X_1, ..., X_p\}$ を確率変数の集合とし，それらの間に非閉路有向グラフ G で表される依存関係があるとする．このとき，同時確率の分解は

$$\Pr(\mathcal{X}) = \prod_{j=1}^{p} \Pr(X_j|Pa(X_j)) \tag{2.1}$$

と表すことができる．ただし，$Pa(X_j)$ は確率変数 X_j の G 上での直接の親に対応する確率変数の集合である．ここで，$\mathcal{X} = \{X_1, ..., X_p\}$ に対して，その依存関係を表す非閉路有向グラフ G が与えられると，その分解は一意である．その分解に従って，確率変数間の条件付き独立性は与えられる．例えば，図3では X_2 を与えた下では X_1 と X_5 は条件付き独立となる．すなわち $\Pr(X_1, X_5|X_2) = \Pr(X_1|X_2)\Pr(X_5|X_2)$ が成り立つ．

本章では，このような同時確率の分解を与える非閉路有向グラフの構造を，データから推測することを主目的とする．与えられた非閉路有向グラフのもとで，事後確率 $\Pr(X_j|\boldsymbol{y})$ ($\boldsymbol{y}$ は $\mathcal{Y} \subset \mathcal{X}$ に対して得られた観測値) を求めるための Belief Propagation などの確率推論法の議論は，本章では取り扱わない．このような確率推論や確率変数に対する作為的な介入における因果推論については，黒木・宮川 (2002)，宮川 (2004) などの優れた研究論文，教科書があるので参考にされたい．

一方，確率変数間の条件付き独立性から議論を開始したとき，非閉路有向グラフは一意に決まるであろうか？実は，条件付き独立性からは非閉路有向グラフが決まらない場合がある．簡単な例を示そう．今，3つの確率変数 A, B, C があるとする．2つの非閉路有向グラフ

$$\begin{aligned}G_1 &: \quad A \to B \to C \\ G_2 &: \quad A \leftarrow B \leftarrow C\end{aligned} \tag{2.2}$$

を考える．この 2 つのグラフは異なる同時確率の分解

$$\begin{aligned} G_1: &\ \Pr(A,B,C) = \Pr(A)\Pr(B|A)\Pr(C|B) \\ G_2: &\ \Pr(A,B,C) = \Pr(A|B)\Pr(B|C)\Pr(C) \end{aligned} \tag{2.3}$$

を得る．しかしながら，この 2 つの分解は共に

$$\Pr(A,C|B) = \Pr(A|B)\Pr(C|B)$$

を満たす．この例が示しているものは，今，確率変数の集合 $\mathcal{X} = \{X_1, ..., X_p\}$ に関して観測データ D を計測したとき，非閉路有向グラフ G_k $(k = 1, 2)$ が同じ条件付き独立性を示すならば，非閉路有向グラフ G_k の与える分解に従うモデルの尤度 $L(D|G_k)$ は $L(D|G_1) = L(D|G_2)$ となる．上述した 3 つの確率変数 A, B, C の例では，各確率変数が R 個の値を取り得る離散型確率変数であるとき，例えば $\Pr(B = s|A = t) = \theta_{st}^{B|A}$ をパラメータとみなすと二つのモデルは同数のパラメータを持つことになる．従って，2 つのモデルの尤度が同じ値であり，パラメータ数も同じであれば，例えば，情報量規準 AIC や BIC も同じ値となり，観測データに基づきどちらのグラフ構造が妥当かというグラフ構造の選択が出来ない状況となる．つまり，離散型確率変数に基づくベイジアンネットワークを用いて確率変数の依存関係を探索する際，可能なものは確率変数間の条件付き独立性の探索であり，それを満たすグラフ構造は複数存在する可能性がある．

それでは，確率変数間の非閉路有向グラフの構造に興味がある場合はどうすればよいのであろうか？　離散型確率変数ではなく，連続型確率変数の場合でも，確率測度を条件付き密度に置き換えることで (2.1) 式の分解は成り立つ．つまり，確率変数の集合 $\mathcal{X} = \{X_1, ..., X_p\}$ に関して観測データ $\boldsymbol{x} = (x_1, ..., x_p)^T$ が与えられたとき，

$$f(\boldsymbol{x}) = \prod_{j=1}^{p} f_j(x_j|\boldsymbol{pa}_j) \tag{2.4}$$

と表せる．ただし，$\boldsymbol{pa}_j$ は $\boldsymbol{x}$ における $Pa(X_j)$ に対応する観測データであり，$\boldsymbol{x}^T$ はベクトル $\boldsymbol{x}$ の転置を表す．例えば，図 3 において $j = 6$ とすると，$\boldsymbol{pa}_6 = (x_2, x_5)^T$ である．この枠組みでは，条件付き密度をどのように構成するかが本質的となる．一般には，条件付き密度 $f_j(x_j|\boldsymbol{pa}_j)$ は x_j の任意の点に対して f_j の値を定める必要があり，条件付き密度は $\int f_j(x|\boldsymbol{pa}_j)dx = 1$ の条件の下で無限個のパラメータ (x_j の任意の点に対する f_j の値) を推定する問題となる．ここでは，f_j は $f_j(x_j|\boldsymbol{pa}_j, \boldsymbol{\theta}_j)$ のように有限次元のパラメータ $\boldsymbol{\theta}_j$ によって規定されているとする．この条件付き密度の推定は，回帰モデルの構築に他ならない．

2.1.2 回帰モデリング

それでは，次に条件付き密度の構成について議論しよう．$f_j(x_j|\boldsymbol{pa}_j, \boldsymbol{\theta}_j)$ の構成を考える．このための最も基本的な統計モデルは，正規線形モデルであろう．つまり，

$$x_j = \beta_0 + \boldsymbol{pa}_j^T \boldsymbol{\beta} + \varepsilon_j$$

で表されるモデルであり，x_j と $\boldsymbol{pa}_j$ の間に線形の構造を仮定したものである．ただし，β_0，$\boldsymbol{\beta}$ は係数パラメータ，誤差項 ε_j は平均 0，分散 σ^2 の正規分布に従うとする．より理解がやさしいものは，X_j が唯一の親確率変数 X_k を持つ場合の $x_j = \beta_0 + \beta_1 x_k + \varepsilon_j$ であろう．この式は傾きが β_1，切片が β_0 の直線を表すことは明らかである．この正規線形モデルのもとでは，条件付き密度は

$$f_j(x_j|\boldsymbol{pa}_j, \beta_0, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x_j - \beta_0 - \boldsymbol{pa}_j^T \boldsymbol{\beta})^2}{2\sigma^2} \right\} \tag{2.5}$$

と表せる．

先ほどは簡単のため，一組の p 次元観測値 $\boldsymbol{x}$ を用いて説明したが，(2.5) 式のパラメータ β_0，$\boldsymbol{\beta}$，σ^2 を推定するためには通常複数組の観測値を用意する．ここでは，n 組の観測値が得られたとして，それらを $\{\boldsymbol{x}_i;\ i = 1, ..., n\}$ と表すことにする．ただし，$\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ と表す．x_{ij} は X_j に対する i 番目の観測値である．このとき，パラメータの尤度関数は $L(\beta_0, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n f_j(x_{ij}|\boldsymbol{pa}_{ij}, \beta_0, \boldsymbol{\beta}, \sigma^2)$ と表せる．また，最尤推定量 $\hat{\beta}_0$，$\hat{\boldsymbol{\beta}}$，$\hat{\sigma}^2$ は $L(\beta_0, \boldsymbol{\beta}, \sigma^2)$ を最大にするものである．モデルのパラメータを最尤法により推定したとき，推定されたモデルの良さは赤池情報量規準 AIC (Akaike, 1973; 1974)，ベイズ型情報量規準 BIC (Schwarz, 1978) などにより評価することができる．赤池情報量規準は

$$\text{AIC} = -2\log(\text{尤度の最大値}) + 2(\text{モデルに含まれるパラメータ数})$$

で定義され，AIC が最も小さくなるモデルが良いモデルと結論づけることができる．AIC や BIC など種々のモデル評価基準については，小西・北川 (2004) を参照されたい．

この正規線形回帰モデルを用いると (2.2) 式の二つのグラフは識別できるであろうか? (2.3) 式で表される 2 つの分解に従い条件付き密度を (2.5) 式の要領で構築し，そのパラメータを上述した最尤法により推定する．その推定されたモデルの良さを評価すれば 2 つのモデルの優劣が評価できると考

えるのが自然であろう．しかしながら，この正規線形回帰モデルでは，やはり (2.2) 式の二つのグラフは識別できない．離散型データの場合と同じく，二つのモデルの尤度の最大値は同じ値となり，モデルに含まれるパラメータ数も同じであるため AIC，BIC ともに同じ値となってしまう．この問題を解決するためには (1) 非正規分布 (2) 非線形モデルを用いるという二つの方法，もしくはその両方が考えられる (狩野・宮村，2006)．

2.1.3 ノンパラメトリック回帰

それでは，確率変数間の非線形な関係を捉え，条件付き密度関数を構築するためのノンパラメトリック回帰について説明する．前節と同じく $f_j(x_j|\boldsymbol{pa}_j, \boldsymbol{\theta}_j)$ の構成を考えよう．一般の加法ノイズを想定したノンパラメトリック回帰は

$$x_j = m(\boldsymbol{pa}_j) + \varepsilon_j \tag{2.6}$$

と表せる．ただし，$m(\cdot)$ は $q_j = |Pa(X_j)|$ としたとき $\mathcal{R}^{q_j}$ から $\mathcal{R}$ への平均関数 $m(\boldsymbol{pa}_j) = E[X_j|\boldsymbol{pa}_j]$，$\varepsilon_j$ は誤差項を表し，平均 0，分散 σ^2 とする．ここでは，親確率変数間の交互作用は考えず，主効果からなるノンパラメトリック加法回帰モデル

$$x_j = m_1(pa_{j1}) + \cdots + m_{q_j}(pa_{jq_j}) + \varepsilon_j \tag{2.7}$$

を考える．ここで，$m_k(\cdot)$ は k 番目の親確率変数に対する関数であり，$\boldsymbol{pa}_j = (pa_{j1}, ..., pa_{jq_j})^T$ と表した．親確率変数間の交互作用とは，例えば 2-way ANOVA の交互作用を思い出していただければ理解の助けになるであろう．遺伝子ネットワークにおいては，この交互作用が無いわけではない．回帰モデルにおける交互効果を一つ一つモデリングすることも可能かと思われる．しかしながら，遺伝子ネットワーク推定においては，構築する非閉路有向グラフに含まれる確率変数（遺伝子数）は数百から千程度になることもある．後で述べるグラフ構造の学習においては，(1) 各 X_j に対して $Pa(X_j)$ を選ぶ，(2) X_j と $Pa(X_j)$ 間の関係を定める，(3) モデルの評価を行いより良い $\{X_j, Pa(X_j)\}$ $(j = 1, ..., p)$ を求めるという作業が必要となる．この (2) のステップにおいて詳細にモデリングすることは大切ではあるが，実際の計算時間との拮抗を考えて行う必要がある．

それでは次に，関数 $m_k(\cdot)$ の構築法について説明する．この関数は，遺伝子ネットワーク推定においては，二つの遺伝子間の関係を表す関数である．この「関係」というのは，親確率変数に対応する遺伝子がその子供確

率変数に対応する遺伝子を制御している関係である．マイクロアレイで計測しているものは，各遺伝子から生成されるメッセンジャー RNA である．遺伝子間の制御は，実際は制御する側の遺伝子から生成されるタンパク質が制御される側の遺伝子のある特定の部分に結合することにより，その遺伝子の転写を促進したり抑制したりする．この制御するタンパク質と制御される遺伝子が生成するメッセンジャー RNA の関係は，簡単な線形式では書き表すことができず，非線形な関係であることが知られている．この背景を基にすると，いま，統計的モデリングの対象はメッセンジャー RNA とメッセンジャー RNA の関係であるが，その関係を線形に限ることなく非線形関係を捉えることのできるモデリングを行うことは自然である．もし，制御する側の遺伝子から生成されるメッセンジャー RNA の量と制御される側の遺伝子から生成されるメッセンジャー RNA の量の関係が，ある非線形なパラメトリックモデルによって書けるならば，そのパラメトリックモデルを用いてモデリングすることが妥当である．しかしながら，遺伝子間の制御に対して，そのようなパラメトリックモデルを構築できるまでに知見は整っていないため，ノンパラメトリックに非線形関係を観測データに基づいて探索することが必要となる．そこで，本章では，関数 $m_k(\cdot)$ は，特定の関数形を仮定せず，B-スプラインに基づく基底関数展開法により柔軟に構築する．

基底関数展開法について簡単に述べる．今，対象とする関数を $y = m(x)$ とする．この関数 $m(x)$ を任意に与えた M 個の基底関数 $\{b_1(x), ..., b_M(x)\}$ の線形結合

$$m(x) = \sum_{m=1}^{M} \gamma_m b_m(x) \tag{2.8}$$

により構築する方法は基底関数展開法と呼ばれる．基底関数としては，本章で用いられる B-スプライン，動経基底関数，フーリエ級数，ウェーブレットなど様々なものが提案されている．極めて複雑な構造を有する関数に対しても，基底関数の取り方，およびその個数 M を適切に定めることにより任意の精度で近似できることが知られている．図 4 に B-スプラインを用いた基底関数展開法による関数構成の例をあげた．係数パラメータ γ_m $(m = 1, ..., M)$ の値を変えることで柔軟に関数を構成できることが分かる．B-スプラインによるノンパラメトリック回帰については，Eilers & Marx (1996)，井元・小西 (1999a, b)，Imoto & Konishi (2003) も参照されたい．

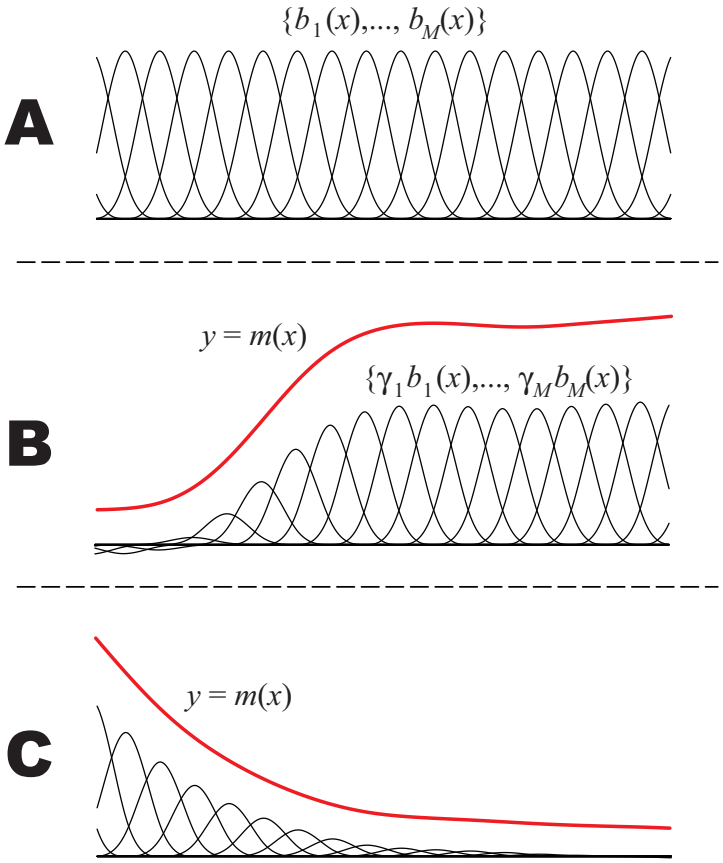


図 4: B-スプラインを用いた基底関数展開法による様々な曲線の構成例．(A) $M = 20$ とした基底関数（等間隔に配置したもの）．(B)，(C) 係数 $\gamma_1, ..., \gamma_{20}$ を変えることで様々な関数を構成できることを示す例．

2.2 遺伝子ネットワーク統計モデル評価のための情報量規準

2.2.1 ベイズアプローチ

構築した統計モデルを評価し，最適なグラフ構造を得るための情報量規準を導出するための準備をしよう．前節で紹介した B-スプラインを用いたノンパラメトリック回帰に基づくベイジアンネットワークは，確率変数間にある非閉路有向グラフ構造を仮定することでモデルを立式できる．このモデルは，非常に柔軟なモデルであるため，そのパラメータ（係数パラメータ）の推定に対しては最尤法は有効に働かない．つまり，最尤法はモデルのデータへの適合度を最大化するため，柔軟なモデルに対してはデータへの

過適合を生じてしまう．そこで，最尤法を修正し，データへの過適合を避けるための方法として罰則付き最尤法，または正則化最尤法と呼ばれる方法がある．この方法は，尤度関数にモデルの複雑さを表す項を加えた関数をパラメータ推定に利用する方法である．つまり，$L(\boldsymbol{\theta})$ を尤度関数，$R(\boldsymbol{\theta})$ をモデルの複雑さを表す項としたとき

$$l_{\mathrm{P}}(\boldsymbol{\theta}|\lambda) = \log L(\boldsymbol{\theta}) - \lambda R(\boldsymbol{\theta}) \tag{2.9}$$

を最大化するパラメータを推定する．ここで，λ $(\lambda \geq 0)$ は平滑化パラメータ，または正則化パラメータと呼ばれるパラメータであり，データへの適合度とモデルの複雑さのトレード・オフをコントロールする役割を担う．つまり，大きな λ を与えると，$l_{\mathrm{P}}(\boldsymbol{\theta}|\lambda)$ の最大化は第二項の最小化と等価となり，シンプルなモデルが推定される．逆に，λ を小さくし，0 とすると最尤推定により推定されるモデルとなる．

この罰則付き最尤法は

$$\begin{aligned} l_{\mathrm{P}}(\boldsymbol{\theta}|\lambda, D) &= \log L(\boldsymbol{\theta}|D) - \lambda R(\boldsymbol{\theta}) \\ &= \log \left[L(\boldsymbol{\theta}|D) \exp\{-\lambda R(\boldsymbol{\theta})\} \right] \end{aligned}$$

の簡単な式変形により $\exp\{-\lambda R(\boldsymbol{\theta})\}$ をパラメータ $\boldsymbol{\theta}$ の事前分布 $p_{\mathrm{prior}}(\boldsymbol{\theta}|\lambda)$ とみなすと

$$L(\boldsymbol{\theta}|D) p_{\mathrm{prior}}(\boldsymbol{\theta}|\lambda) \propto p_{\mathrm{post}}(\boldsymbol{\theta}|D, \lambda)$$

の関係から $\boldsymbol{\theta}$ の事後分布に比例することが分かる．すなわち，罰則付き最尤法は，事後確率を最大化するパラメータ値 (Maximum *a posteriori*; MAP 解) を求めていると言える．$p_{\mathrm{post}}(\boldsymbol{\theta}|D, \lambda)$ を最大にする $\hat{\boldsymbol{\theta}}$ は平滑化パラメータ λ に依存するため，$p_{\mathrm{post}}(\boldsymbol{\theta}|D, \lambda)$ をパラメータ $\boldsymbol{\theta}$ に関して積分した周辺尤度 $\int p_{\mathrm{post}}(\boldsymbol{\theta}|D, \lambda) d\boldsymbol{\theta}$ の最大化によりその値を決定する．ここで，周辺尤度は，平滑化パラメータ λ のみの関数となっており，λ の尤度とみなすことができる．つまり λ に関して最尤推定を行っていることに相当する．

2.2.2 ベイズ型情報量規準

パラメータ推定からグラフ構造の評価に話を戻そう．上述したようなベイズアプローチに従うと，最適なグラフ構造とは，観測データ D が与えられたもとでのグラフ構造 G の事後確率 $p_{\mathrm{post}}(G|D)$ を最大にするもの，つ

まり $\hat{G} = \mathrm{agr\,max}\{p_{\mathrm{post}}(G|D)\}$ として定義される．この事後確率は，条件付き確率の計算により

$$\begin{aligned} p_{\mathrm{post}}(G|D) &= \frac{p_{\mathrm{prior}}(G)p(D|G)}{p(D)} \\ &\propto p_{\mathrm{prior}}(G)p(D|G) \end{aligned}$$

と表せる．ただし，$p_{\mathrm{prior}}(G)$ はグラフ構造 G に関する事前確率，$p(D|G)$ は，グラフ構造を与えたもとでの観測データ D の周辺尤度，

$$p(D) = \sum_{G}\{p_{\mathrm{prior}}(G)p(D|G)\}$$

は規格化定数である．この事後確率 $p_{\mathrm{post}}(G|D)$ を最大にする G を選ぶ方法に対しては，周辺尤度 $p(D|G)$ の計算と，事前確率 $p_{\mathrm{prior}}(G)$ の設定を行う必要がある．まず，周辺尤度 $p(D|G)$ から説明しよう．尤度を前節で説明したベイジアンネットワークにより計算する際，モデルにはパラメータ $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p)^T$ が含まれている．このパラメータに対して尤度関数の周辺化を行い，周辺尤度 $p(D|G)$ は計算される．つまり，ベイズアプローチにより，パラメータ $\boldsymbol{\Theta}$ もある確率分布 (事前分布) $p_{\mathrm{prior}}(\boldsymbol{\Theta}|G, \boldsymbol{\lambda})$ に従う確率変数とする．ここで，$\boldsymbol{\lambda}$ は事前分布を規定するパラメータ (ハイパーパラメータ) であり (2.9) 式の平滑化パラメータに相当する．すると，

$$\begin{aligned} p(D|G) &= \int p(D, \boldsymbol{\Theta}|G)d\boldsymbol{\Theta} \\ &= \int p(D|\boldsymbol{\Theta}, G)p_{\mathrm{prior}}(\boldsymbol{\Theta}|G, \boldsymbol{\lambda})d\boldsymbol{\Theta} \\ &= \int \prod_{i=1}^{n} f(\boldsymbol{x}_i|G, \boldsymbol{\Theta})p_{\mathrm{prior}}(\boldsymbol{\Theta}|G, \boldsymbol{\lambda})d\boldsymbol{\Theta} \\ &= \int \prod_{i=1}^{n}\prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{pa}_{ij}, G, \boldsymbol{\theta}_j)p_{\mathrm{prior},j}(\boldsymbol{\theta}_j|G, \boldsymbol{\lambda}_j)d\boldsymbol{\theta}_j \end{aligned}$$

と式変形される．ここで，事前分布 $p_{\mathrm{prior}}(\boldsymbol{\Theta}|G, \boldsymbol{\lambda})$ に対して

$$p_{\mathrm{prior}}(\boldsymbol{\Theta}|G, \boldsymbol{\lambda}) = \prod_{j=1}^{p} p_{\mathrm{prior},j}(\boldsymbol{\theta}_j|G, \boldsymbol{\lambda}_j)$$

なる分解が成り立つと仮定した．すなわち，グラフ構造の事後確率 $p_{\mathrm{post}}(G|D)$ は $\mathrm{sub}.G_j$ を非閉路有向グラフ G における $\{X_j, Pa(X_j)\}$ からなる部分グ

ラフとしたとき

$$
\begin{aligned}
&p_{\mathrm{post}}(G|D) \\
&\quad = \prod_{j=1}^{p} \left\{ p_{\mathrm{prior}}(\mathrm{sub}.G_j) \int \prod_{i=1}^{n} f_j(x_{ij}|\boldsymbol{pa}_{ij}, G, \boldsymbol{\theta}_j) p_{\mathrm{prior},j}(\boldsymbol{\theta}_j|G, \boldsymbol{\lambda}_j) \boldsymbol{\theta}_j \right\} \\
&\quad = \prod_{j=1}^{p} \mathrm{score}(\mathrm{sub}.G_j)
\end{aligned}
$$

と分解される．ただし，

$$
p_{\mathrm{prior}}(G) = \prod_{j=1}^{p} p_{\mathrm{prior}}(\mathrm{sub}.G_j)
$$

の分解を仮定した．事後確率の計算は

$$
\begin{aligned}
&\mathrm{score}(\mathrm{sub}.G_j) \\
&\quad = p_{\mathrm{prior}}(\mathrm{sub}.G_j) \int \prod_{i=1}^{n} f_j(x_{ij}|\boldsymbol{pa}_{ij}, G, \boldsymbol{\theta}_j) p_{\mathrm{prior},j}(\boldsymbol{\theta}_j|G, \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j \quad (2.10)
\end{aligned}
$$

を各 j $(j = 1, ..., p)$ において求めればよいことになる．

(2.10) 式により，問題は周辺尤度に含まれる高次積分をいかに計算するかという問題に帰着される．この積分計算については，マルコフ連鎖モンテカルロ法などの数値的解決法も適用可能であるが，後で述べるグラフ構造の最適化アルゴリズムには膨大な時間がかかるため，できるだけ高速にこの積分値を求めることが要求される．そこで，本章ではラプラス近似を用いて積分を計算する方法を用いる．(2.10) 式に含まれる高次積分のラプラス近似は，簡単のために添え字の j を省略すると

$$
\begin{aligned}
&\int \prod_{i=1}^{n} f(x_i|\boldsymbol{pa}_i, G, \boldsymbol{\theta}) p_{\mathrm{prior}}(\boldsymbol{\theta}|G, \boldsymbol{\lambda}) d\boldsymbol{\theta} \\
&\qquad = \int \exp\left\{ n l_\lambda(\boldsymbol{\theta}|D) \right\} d\boldsymbol{\theta} \\
&\qquad = \frac{(2\pi/n)^{r/2}}{|J_\lambda(\hat{\boldsymbol{\theta}})|^{1/2}} \exp\left\{ n l_\lambda(\hat{\boldsymbol{\theta}}|D) \right\} \left\{ 1 + O_p(n^{-1}) \right\}
\end{aligned}
$$

と表せる．ただし，r はパラメータ $\boldsymbol{\theta}$ の次元であり，$\log p_{\mathrm{prior}}(\boldsymbol{\theta}|G, \boldsymbol{\lambda}) = O(n)$

と仮定し

$$
\begin{aligned}
l_\lambda(\boldsymbol{\theta}|D) &= \frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\boldsymbol{pa}_i, G, \boldsymbol{\theta}) + \frac{1}{n}\log p_{\text{prior}}(\boldsymbol{\theta}|G, \boldsymbol{\lambda}) \\
J_\lambda(\hat{\boldsymbol{\theta}}) &= -\frac{\partial^2 l_\lambda(\boldsymbol{\theta}|D)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}
\end{aligned}
$$

と表記した．ここで，$\hat{\boldsymbol{\theta}}$ は

$$
\hat{\boldsymbol{\theta}} = \arg\max\{l_\lambda(\boldsymbol{\theta}|D)\} \tag{2.11}
$$

である．前節で説明したパラメータの MAP 推定を思い出していただくと，(2.11) 式で定義される $\hat{\boldsymbol{\theta}}$ はパラメータ $\boldsymbol{\theta}$ の MAP 解であり，ラプラス近似は，被積分関数のモードである $\hat{\boldsymbol{\theta}}$ の周りで積分を近似していることが分かる．ラプラス近似によるベイズモデリングについて詳しく知りたい読者は，小西・北川 (2004)，Konishi *et al.* (2004) を参照されたい．

(2.10) 式で定義されるグラフ構造 G の局所的スコア $\text{score}(\text{sub}.G_j)$ はハイパーパラメータベクトル $\boldsymbol{\lambda}_j$ に依存する量である．そこで，このハイパーパラメータベクトルは局所スコアの最大化に基づいて決定し，その値を周辺尤度に代入した値を局所スコアの値とする．すなわち，$\hat{\boldsymbol{\lambda}}_j = \arg\max\{\text{score}(\text{sub}.G_j)\}$ を用いる．

いま，B-スプラインに基づくノンパラメトリック回帰を用いた統計モデルに対して，パラメータの事前分布として平滑化事前分布を仮定する．つまり，無情報事前分布とは異なり，各確率変数間の関係を表す関数に対して，ある程度の滑らかさを仮定し，その情報を推定に対して積極的に使おうとする方法である．簡単のため，ノンパラメトリック回帰の節で用いた $m(x) = \sum_{m=1}^{M}\gamma_m b_m(x)$ を再び用いて説明する．ノンパラメトリック回帰の枠組みにおいては，曲線 $y = m(x)$ の複雑さを表す量として $\int\{m''(x)\}^2 dx$ がよく用いられる．今，B-スプラインの節点を等間隔に取り，次数が 3 のスプラインにより構築すると，

$$
\int\{m''(x)\}^2 dx \approx \sum_{k=3}^{M}(\gamma_k - 2\gamma_{k-1} + \gamma_{k-2})^2 \tag{2.12}
$$

という関係が知られている (Eilers & Marx，1996)．そこで，(2.9) 式において $R(\boldsymbol{\theta}) = \sum_{k=3}^{M}(\gamma_k - 2\gamma_{k-1} + \gamma_{k-2})^2$ とおくことにより曲線の複雑さを考慮に入れたパラメータ推定が可能となる．このとき，λ によってどの程度曲線の複雑さを許容するかをコントロールすることができる．λ の逆数は，この

事前分布の分散に対応する．つまり，λ を大きくすると分散が小さくなり，曲線は滑らかであるという先験的知識を積極的にパラメータ推定に反映させることになる．一方，λ を小さく 0 に近く設定すると分散は大きくなり，事前分布は一様分布に近く，パラメータ推定に関してなんら情報は与えない無情報事前分布となる．Imoto *et al.* (2002) では，この方法により計算される B-スプラインノンパラメトリック回帰に基づくベイジアンネットワークの構造を評価するためのベイズ型情報量規準を導出し，マイクロアレイデータからの遺伝子制御ネットワークの推定に用いている．

2.3 遺伝子ネットワークの構造推定アルゴリズム

これまでに，遺伝子ネットワークの統計的モデリングについて，ある非閉路有向グラフ G が与えられた下での統計モデルの構築法と，情報量規準に基づくその評価方法について説明した．それでは，次に，どのようにして事後確率を最大とするグラフ構造 $\hat{G}$ を得るのか，その手順について説明する．グラフ構造 G は，前節において説明したグラフ構造の事後確率 $p_{\mathrm{post}}(G|D)$ の最大化に基づいて推定される．すなわち

$$\hat{G} = \arg\max\{p_{\mathrm{post}}(G|D)\}$$

である．この事後確率の最大化に対しては，グラフ構造を表す隣接行列の各成分が 0 か 1 かを判定することに帰着される問題であるため，ニュートン法などの数値最適化の手法を用いることはできない．考えられる最も実直な方法は，全てのグラフ構造 $G_1, ..., G_L$ を枚挙し，その事後確率 $p_{\mathrm{post}}(G_1|D), ..., p_{\mathrm{post}}(G_L|D)$ を計算し，最も事後確率が大きくなるグラフ構造を選択する，いわゆる枚挙の方法である．しかしながら，この方法は，候補となるグラフ構造の個数を考えると現実的ではない．詳しい紹介は省くが，たった 9 個のノードからなる非閉路有向グラフの個数でさえ，約 1.21×10^{15} 個ある．もし，1 秒間に 10,000 個のグラフ構造の事後確率が計算できても 3,800 年程度かかる (井元，2007)．実際，非閉路有向グラフの個数は，含まれるノードの個数に対して超指数のオーダーで増えていくことが知られている．そこで，多くの遺伝子数を含むようなネットワークの構造推定に対しては，少しずつネットワーク構造を情報量規準の値をみながら変化させていくという，いわゆるグラフ構造の学習を行う．そのためのアルゴリズムが数多く提案されており，その中でもここでは greedy アルゴリズムと呼ばれる方法を紹介する．

greedy アルゴリズムの操作は次の通りである．

(1) あるノード X_i に着目する．

(2) そのノードに対して，

(2-1) $Y \notin Pa(X_i)$ なる Y に対しては Y を $Pa(X_i)$ に追加する．

(2-2) $Y \in Pa(X_i)$ なる Y に対しては Y を $Pa(X_i)$ から除外する．

(2-3) X_i の子供である Y に対しては矢印を反転させ Y を $Pa(X_i)$ に追加する．

(3) (2-1)–(2-3) の操作において，事後確率を最も大きくする操作を一つだけ実施する．

(4) (3) の操作が情報量規準の値を改善すればその修正を採択し，改善しなければ破棄する．

(5) (1) から (4) の操作を各 X_i に対して行い，事後確率が大きくならなくなった時点で学習をストップする．

なお，(4) において，グラフ構造に対して変更を施す場合は，その変更後もグラフ G が非閉路有向グラフであることを確認する必要がある．また，数百など非常に多くの遺伝子を含むネットワークを推定する際，あらかじめ各 X_i に対して $Pa(X_i)$ の候補となる遺伝子集合 $\mathcal{X}_i \subset \mathcal{X}$ を選んでおき，$Pa(X_i) \subset \mathcal{X}_i$ と制限する．この $\mathcal{X}_i$ の構成法には様々なものがあり，例えば，相関係数や，先ほどのスコア $\mathrm{score}(\mathrm{sub}.G_i)$ を用いる方法がある．例えば，X_i に対して $\{X_j\} = Pa(X_i)$ $(j \neq i)$ としたときの $\mathrm{score}(\mathrm{sub}.G_i)$ は X_j の選び方に依るため，特に $\mathrm{score}(\mathrm{sub}.G_i) = \mathrm{score}_i(X_j)$ と書くと

$$\mathrm{score}_i(X_{i_1}) > \cdots > \mathrm{score}_i(X_{i_{p-1}})$$

という順序を得る．もし，X_i に対して，上位 20 個の候補を用いるのであれば $\mathcal{X}_i = \{X_{i_1}, ..., X_{i_{20}}\}$ とできる．また，ベイジアンネットワークよりも高速にグラフィカルモデルの構築ができるような手法を用い，X_i と条件付き独立性でない可能性のある遺伝子により $\mathcal{X}_i$ を構築する方法も考えることができる．高速にグラフィカルモデルを推定できる手法としては，L_1 ペナルティ (Lasso 型) を用いたグラフィカル・ガウシアンモデル (Shimamura *et al.*, 2007) がある．小規模な遺伝子ネットワークであれば，アルゴリズムを改良することにより厳密な MAP 解を得ることができる．Ott *et al.* (2004) では，ダイナミックプログラミングを基にしたアルゴリズムを開発した．ま

た，Perrier et al. (2008) は，super-structure と呼ばれる無向グラフを用い，推定するネットワークのスケルトンがあらかじめ設定された super-structure の部分グラフとなるという制約下での最適ネットワーク構造探索アルゴリズム (Constraint Optimal Search; COS) を提案し，super-structure が比較的疎なとき (平均次数が 2 程度) 50 程度のノードを有するネットワークの最適学習が出来ることを示した．また，Kojima et al. (2010) では, COS を拡張したアルゴリズムを提案し，平均次数が 4 程度でも 400 程度のノードを含むネットワークの条件付き最適解を求めることができることを示した．Tamada et al. (2011) では，Ott et al. (2004) の開発した最適アルゴリズムに対する並列アルゴリズムを提案し，2011 年時点で正解最大となる 32 ノードネットワークの完全探索を達成している．

2.4 ベイズアプローチに基づく他の生物学的情報の融合

マイクロアレイデータに基づく遺伝子制御ネットワークの推定について，ベイジアンネットワークとノンパラメトリック回帰に基づく方法を説明した．しかしながら，遺伝子制御ネットワークの推定に用いることのできる観測データはマイクロアレイデータだけではない．被制御遺伝子にある，ある特定の遺伝子が生成したタンパク質が結合する DNA 配列などは遺伝子制御ネットワークの大きな情報となる．また，既に高精度な生物学実験により確認されている遺伝子制御の関係はデータベースに蓄積されている．このような情報を上手く利用することで，モデルのマイクロアレイデータへの過適合を回避し，より有効にマイクロアレイデータでは欠損している情報を補うことができる．このための統計学的方法論研究が進んでいる．その鍵となるのが，グラフ構造の事前確率 $p_{\mathrm{prior}}(G)$ である．この事前確率を，生物学的情報を基に構築し，情報事前分布を用いることで上記の役割を果たすことができる．その理論と具体的な方法に関しては，井元 (2007)，Imoto *et al.* (2004)，Tamada *et al.* (2003)，Nariai *et al.*, (2005) などを参照されたい．

3 適用例

本節では，これまでに説明したベイジアンネットワークとノンパラメトリック回帰に基づく方法を用いて，マイクロアレイデータから遺伝子制御ネットワークを推定した適用例を紹介する．

本節では，実際のマイクロアレイを用いた適用例を紹介するが，本章で取り上げたような新しい統計モデル，およびそのモデリングを提案する際には，人工的に生成したデータを基にシミュレーションを行い，その性能を評価することが必要である．本節において紹介したベイジアンネットワークとノンパラメトリック回帰を用いた方法に対しては，Imoto *et al.* (2003a) にそのシミュレーションがあり，前節で述べた他の情報を付与した際のシミュレーションは，Imoto *et al.* (2004)，Tamada *et al.* (2003)，Imoto *et al.* (2006b) にある．興味ある読者は合わせて参照されたい．

本章で紹介したベイジアンネットワークとノンパラメトリック回帰に基づく遺伝子ネットワーク推定手法は，計算量が非常に多いため，数千という遺伝子を含むネットワーク推定に用いることは現時点の計算機の性能では十分な解を得ることは不可能である．また，現在，遺伝子ネットワークに利用できるマイクロアレイデータの量も，数千という遺伝子の依存関係を明らかにするには十分ではない．そこで，例えば，細胞内における何らかの機能に特化した遺伝子たちや，ある薬剤に反応する遺伝子たちの部分ネットワークを推定するという解析を行っている．前者の例としては，出芽酵母の細胞周期に関与する遺伝子ネットワーク (Imoto *et al.*, 2002，Nariai *et al.*, 2005)，後者の例としては，抗真菌薬 griseofulvin に影響を受ける遺伝子ネットワーク (Imoto *et al.*, 2003b) や高脂血症薬 fenofibrate によって影響を受ける遺伝子ネットワーク (Imoto *et al.*, 2006a) があげられる．

3.1 出芽酵母において抗真菌薬 griseofulvin 投与により影響を受ける遺伝子とその制御パスウェイの推定

図 5 は，出芽酵母に抗真菌薬 griseofulvin を投与し，影響を受けたと予測された遺伝子ネットワークの一部である．下段に配置した遺伝子は，griseofulvin から直接影響を受けたと予測された遺伝子である．出芽酵母に griseofulvin を投与し計測したマイクロアレイデータと投与していないコントロールのマイクロアレイデータを比較することにより直接影響を受けた候補遺伝子を同定した．その同定のために Imoto *et al.* (2003b) では仮想遺伝子法と呼ばれる手法を用いている．仮想遺伝子法を簡単に説明する．薬剤をネットワーク上の仮想的なノードと見なし，投薬により大きく変動した遺伝子をその直下にまず配置する．次に遺伝子破壊株マイクロアレイデータを用いて，仮想ノードの直下にある遺伝子同士の制御関係を推定する．もし，遺伝子 A と遺伝子 B が仮想ノードの直下にあり，かつ遺伝子 A が遺伝

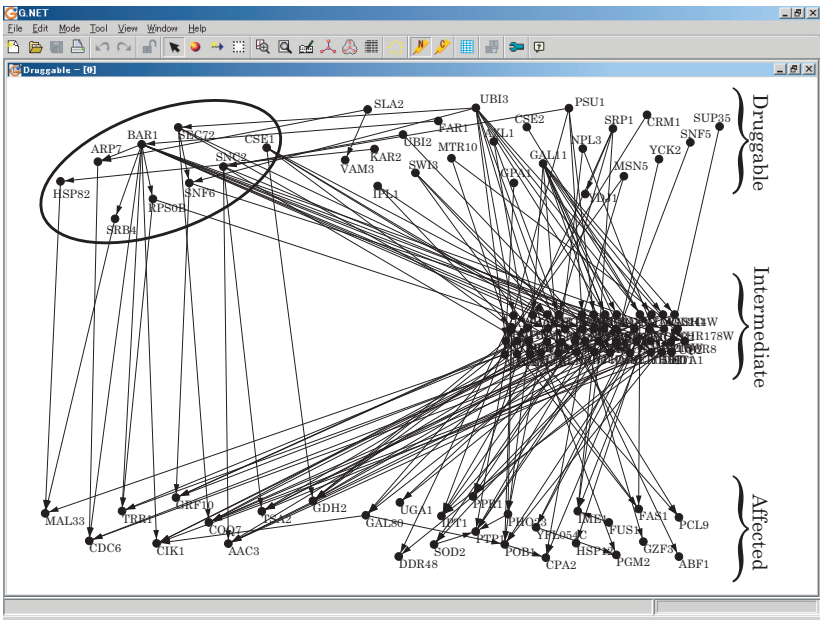
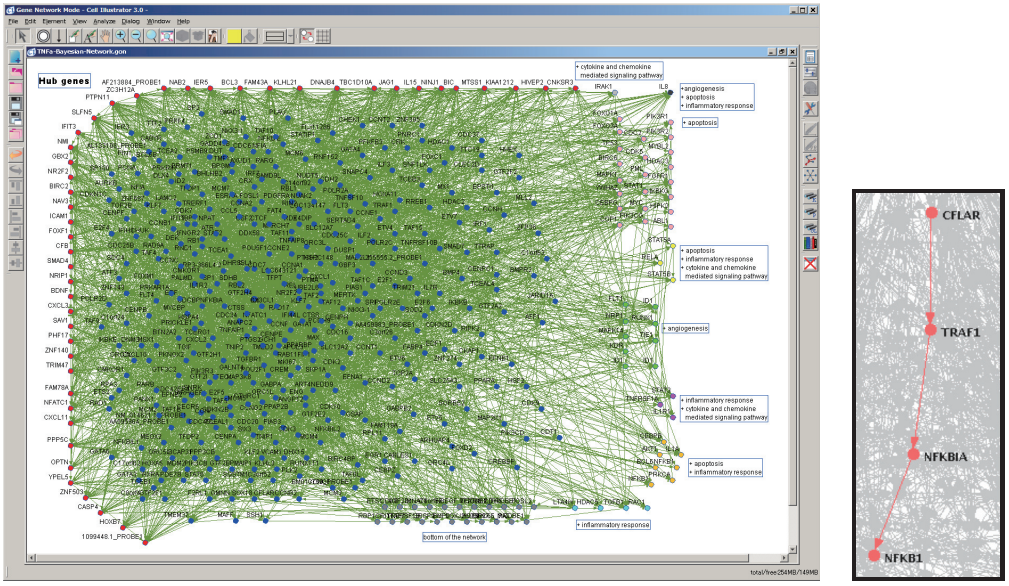


図 5: 出芽酵母に抗真菌薬 griseofulvin を投与したマイクロアレイデータを用い，griseofulvin に直接影響を受けると予測された遺伝子 (下段) と影響を受ける遺伝子を制御していると予測された遺伝子 (上段).

子 B を制御しているならば，遺伝子 B は遺伝子 A からの制御により薬剤からの影響を受けたと考えられる．したがって，遺伝子 B は薬剤によって直接影響を受けたとは言えないため仮想ノードから遺伝子 B への矢印は取り除く．

次に，新たな薬剤ターゲット遺伝子同定のために，仮想遺伝子法により薬剤から直接影響を受けたと予測された遺伝子を含む遺伝子ネットワークを推定した．ネットワーク推定に用いた遺伝子は，上述した薬剤から直接影響を受けたと予測された遺伝子，転写因子および既知の制御遺伝子，核内受容体からなる 735 遺伝子である．これら 735 遺伝子のネットワークを 120 枚のマイクロアレイデータから推定した．このネットワーク推定に前述したベイジアンネットワークとノンパラメトリック回帰を組み合わせた方法を用いた．推定されたネットワークにおいて，薬剤により直接影響を受けたと予測された遺伝子とその上流にある核内受容体からなる部分ネットワークを図 5 に示している．核内受容体は薬のターゲットとして考えることができるため，この遺伝子ネットワークを手がかりに新たな創薬標的遺伝子を発見できる可能性がある．



(a) (b)

図 6: (a) ヒト血管内皮細胞において TNFα 投与により影響を受けると予測された 482 遺伝子のネットワークをベイジアンネットワークとノンパラメトリック回帰を用いて 351 枚のマイクロアレイデータにより推定した結果．(b) 4 遺伝子 (*CFLAR*, *TRAF1*, *NF$\kappa\beta$IA*, *NF$\kappa\beta$1*) からなる部分ネットワーク．遺伝子ネットワークの表示には Cell Illustrator (`http://www.cellillustrator.com`) を用いた．

3.2 ヒト血管内皮細胞における炎症性サイトカイン TNF により誘導される遺伝子ネットワークの推定

図 6(a) に示したのは，ヒト血管内皮細胞の遺伝子に対して TNFα というタンパク質を曝露した際に影響を受ける遺伝子のネットワークをベイジアンネットワークとノンパラメトリック回帰に基づく方法により推定したものである (Hurley *et al.*, 2011)．TNFα (Tumor Necrosis Factor alpha) は，代表的な炎症性サイトカインであり，生体防御炎症反応を司る役割を有する．この TNFα に反応する遺伝子ネットワークは，炎症反応やアポトーシス (細胞死)，血管形成に関連し，血管内皮細胞の中心的役割を担うネットワークと考えることができる．このネットワークは 482 個の遺伝子を含み，それらは TNFα 曝露によって影響を受けたと推測された遺伝子からなる．このネットワークから様々な生物学的仮説を構築・考察することができる．例えば，

図 6(b) には 4 つの遺伝子 (*CFLAR*，*TRAF1*，*NFκβIA*，*NFκβ1*) からなる部分ネットワークを示した．この中で，*TRAF1* 遺伝子は，TNF シグナルを中継する役割を有する．また，TNF 存在下において MAPK8，JNK，NF$\kappa\beta$ のパスウェイを制御することが知られている．この部分ネットワークにより，*TRAF1* は *CFLAR* から制御を受けることが示唆されるが，TRAF1 タンパク質と CFLAR タンパク質に相互作用が存在するという事実からその予測はサポートされる．また，*TRAF1* の制御遺伝子には *NFκβIA* や *NFκβIE* など多くの血管炎症に関わる遺伝子が存在した．この二つの遺伝子はタンパク質レベルでの TRAF1 からのシグナル伝達により制御されることが知られている．また “*NFκβIA*→*NFκβ1*” は，*NFκβ1* の転写因子としての活性を NF$\kappa\beta$IA タンパク質が制御しているというタンパクレベルの知見から解釈可能である．転写因子 *NFκβ1* のさらに下流，つまり，*NFκβ1* から制御されると予測された遺伝子たちに関しては，井元 (2007) p.115 図 4.20 に結合配列解析を合わせた考察がある．合わせて参照されたい．

推定された遺伝子ネットワークにおいて，多くの遺伝子を制御している，すなわち，多くの子供遺伝子をもつ遺伝子 (ハブ遺伝子) は，TNFα 関連遺伝子，アポトーシス関連遺伝子であり，重要な役割を担うということが知られていた．このハブ遺伝子の中から EST (Expressed Sequence Tag: 遺伝子転写産物 (RNA) の一部の短い配列．転写産物の目印として用いられる) としてしか情報のない遺伝子を 2 つ選び，その EST が実際に炎症反応に対して重要な役割を担っていることを実験によって確認した．この例を通して，ベイジアンネットワークとノンパラメトリック回帰を用いた遺伝子ネットワーク推定は，ネットワークを基礎とした生命システムの理解，新たな生物学的発見において本質的な役割を果たすことを実証した．

4　終わりに

マイクロアレイによって計測された遺伝子発現データから，遺伝子制御のネットワークを推定するための統計科学的手法を紹介した．先にも述べたが，遺伝子制御ネットワークは，生命システムを形成する一つの重要なネットワークである．この遺伝子制御ネットワークや代謝パスウェイ，シグナル伝達経路など多様な役割を担うネットワークを合わせて解明することが，生命のシステム的理解を達成するために必要不可欠である．この目的達成のためには，観測されたデータに基づく情報抽出を行う統計科学の担う役割は大きい．今後も引き続き生命科学においては，新たな計測機器，技

術革新に伴い様々な観測データが大量に得られ続ける．このような多様な観測データに基づくデータ駆動型科学の発展のためには，統計科学のより大きな発展が必要不可欠である．

本章では簡単にしか触れることができなかったが，マイクロアレイデータに加えて多様な生物学的情報を融合し情報抽出を行うアプローチは，生命システムネットワークを解明する上で必要不可欠である．このためには，ベイズ的な情報統合手法が強力な武器となる．詳しくは，井元 (2007) を参照されたい．

最近では次世代シークエンサーを用いた個人ゲノムデータの取得が可能となっている．国際ヒトゲノム計画では，10 年の月日をかけて一人のヒトゲノム配列を決定したが，次世代シークエンサーを用いることで 2011 年現在で数週間で一人のヒトゲノム配列が決定できる．この次世代シークエンサーにより，個人ゲノム時代の幕が開けたと言えるであろう．マイクロアレイによるメッセンジャー RNA の網羅的計測，SNP アレイによる 100 万近い箇所における一塩基多型の有無に加え，次世代シークエンサーを用いた個人のゲノム配列，ヒストン修飾の網羅的計測のデータが利用できるようになった．これからは，そのような多様なデータを統合的に解析し，個人の生命システムの特徴を捉える研究が必要となる．この個人の生命システムの違いにより，疾患に対するリスクが異なり，また，薬剤に対する感受性の違いが明らかになることが期待される．特に，がんのようなゲノムに生じた複数の変異が引き起こしているような複雑な疾患を理解し，制御するためにはこのような生命システムに基づく解析が必要になって来るであろう．そのためには，本章で解説したような統計科学的手法が重要な役割を担う．

参考文献

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *2nd Inter. Symp. on Information Theory* (Petrov, B.N. & Csaki, F. eds.), Akademiai Kiado, Budapest, 267–281. (Reproduced in *Breakthroughs in Statistics*, Volume 1, Kotz, S. & Johnson, N.L. eds., Springer-Verlag, New York, (1992))

[2] Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Contr.*, **AC-19**, 716–723.

[3] Eilers, P. & Marx, B. (1996) Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

[4] Hurley, D., Araki, H., Tamadaz, T., Dunmore, B., Sanders, D., Humphreys, S., Affara, M., Imoto, S., Yasuda, K., Tomiyasu, Y., Tashiro, K., Savoie, C., Cho, V., Smith, S., Kuhara, S., Miyano, S., Charnock-Jones, D.S., Crampin, E.J., Print, C.G. (2011) Gene network inference and visualization tools for biologists: application to new human transcriptome datasets, *Nucleic Acids Research*, in press.

[5] 井元 清哉，小西 貞則 (1999a) 情報量規準に基づく B-スプライン非線形回帰モデルの推定．応用統計学会誌，**28**，137–150.

[6] 井元 清哉，小西 貞則 (1999b) B-スプラインによる非線形回帰モデルと情報量規準．統計数理，**47**，359–373.

[7] Imoto, S., Goto, T. & Miyano, S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pacific Symposium on Biocomputing*, **7**, 175–186.

[8] Imoto, S. & Konishi, S. (2003) Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics*, **55**, 671–687.

[9] Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003a) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology*, **1**, 231–252.

[10] Imoto, S., Savoie, C.J., Aburatani, S., Kim, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003b) Use of gene networks for identifying and validating drug targets. *Journal of Bioinformatics and Computational Biology*, **1**, 459–474.

[11] Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. & Miyano, S. (2004) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of Bioinformatics and Computational Biology*, **2**, 77–98.

[12] Imoto, S., Tamada, Y., Araki, H., Yasuda, K., Print, C., Charnock-Jones, S., Sanders, D., Savoie, C., Tashiro, K., Kuhara, S. & Miyano, S. (2006a) Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. *Pacific Symposium on Biocomputing*, **11**, 559–571.

[13] Imoto, S., Higuchi, T., Goto, T. & Miyano, S. (2006b) Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks. *Statistical Methodology*, **3**, 1–16.

[14] 井元 清哉 (2007) ベイズモデルによる遺伝子制御ネットワークの推定. (樋口知之 (編) 統計数理は隠された未来をあらわにする－ベイジアンモデリングによる実世界イノベーション－), pp.85–117, 東京電気大学出版.

[15] 狩野 裕, 宮村 理 (2006) 統計的因果推論と因果探索. 第 1 回データマイニングと統計数理研究, SIG-DMSM-A601.

[16] Konishi, S., Ando, T. & Imoto, S. (2004) Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**, 27–43.

[17] 小西 貞則, 北川 源四郎 (2004) 情報量規準. シリーズ・予測と発見の科学, 朝倉書店.

[18] Kojima, K., Perrier, E., Imoto, S., Miyano, S. (2010) Optimal search on clustered structural constraint for learning Bayesian network structure, *Journal of Machine Learning Research*, **11**, 285–310.

[19] 黒木 学, 宮川 雅巳 (2002) 線形構造方程式モデルにおける同時介入効果の線形回帰母数による表現, 応用統計学会誌, **31**, 107–121.

[20] 宮川 雅巳 (2004) 統計的因果推論–回帰分析の新しい枠組み. シリーズ・予測と発見の科学, 朝倉書店.

[21] Nariai, N., Tamada, Y., Imoto, S. & Miyano, S. (2005) Estimating gene regulatory networks and protein-protein interactions of *Saccharomyces cerevisiae* from multiple genome-wide data. *Bioinformatics*, **21**, ii206–ii212.

[22] 日本バイオインフォマティクス学会 (編) (2006) バイオインフォマティクス事典. 共立出版.

[23] Ott, S., Imoto, S. & Miyano, S. (2004) Finding optimal models for small gene networks. *Pacific Symposium on Biocomputing*, **9**, 557–567.

[24] Perrier, E., Imoto, S., Miyano, S. (2008) Finding optimal Bayesian network given a super-structure, *Journal of Machine Learning Research*, **9**, 2251–2286.

[25] Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

[26] Shimamura, T., Imoto, S. Yamaguchi, R. & Miyano, S. (2007) Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics*, **18**, in press.

[27] Tamada, Y., Imoto, S., Miyano, S. (2011) Parallel algorithm for learning optimal Bayesian network structure, *Journal of Machine Learning Reseach*, **12**, 2437–2459.

[28] Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, ii227–ii236.

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2008 年 5 月
第 II 部 生物と健康の統計科学

第7章　統計科学と健康科学の相互寄与

吉村功 (東京理科大学工学部)

統計学には、健康科学から研究課題を出され、それにこたえる形で学問の内容を深めたという歴史がある。本章では、その相互寄与の関係が今後も続くかどうか、続くと考えられるならば、その内容をより豊かにするためにどのようなことに留意すべきか、を考察する。考察においては近年の新たな課題を、臨床試験、薬剤の市販後調査、遺伝子解析、動物実験代替法、インシリコ試験、等の事例で紹介する。

統計科学と健康科学の相互寄与

吉村功

1　はじめに

統計という用語と健康という用語は，共に，中学や高校の教科書に登場するし，新聞・雑誌・テレビといったものにもしきりに出てくる．しかし，これに「科学」をつけた「統計科学」や「健康科学」は，あまり見慣れない，聞き慣れない用語であろう．そこでまずこの用語の説明をしよう．

人によって意見は違うが，筆者は統計科学や健康科学を特定の理論・学問体系というふうには考えていない．そうではなくて，統計科学は，統計的方法論を確立し，かつそれを現実問題に適用しようという指向性を持つ科学的営為の総体，健康科学は，生物の健康を維持・推進しようという指向性を持つ科学的営為の総体，と考えている．英語で言うと,「The statistical science is a complex of scientific works with the aim of applications of statistical methodolgy and the health sicence is the one with the intention to establish and promote the health of livings.」という感じである．

世間的な学問分類でいうと，統計科学の骨格は数理統計学と応用統計学であり，健康科学の心棒は,（遺伝学や疫学を含む）医学，薬学，生物学である．

2　統計科学と健康科学の古いつきあい

統計科学が確立されたのはいつかというのは，簡単に答えが出ない問題である．しかし，統計科学を主課題とする団体が設立され，その機関誌が発行されていれば，統計科学は社会的に認知されていたと言える．そういう見方をすると，英国王立統計協会（Royal Statistical Society) の設立と Journal of the Royal Statistical Society (JRSS) の発行は，統計科学の社会的認知の証拠であろう．その JRSS の第 1 巻発行は 1839 年である．

JRSS に約 60 年遅れた 1901 年に，Biometrika という学術誌が K. Pearson や W. F. R. Weldon らによって創刊された．発刊の辞に，“memoirs on variation, inheritance, and selection in animals and plants, based on the examination of statistically large numbers of species ...” （統計的に多くの種を調べた上での動植物の多様化，遺伝，淘汰の記録）とあり (Sowan, 1982)，第 1 号に “On the inheritance of the duration of life, and on the intensity of natural selection” というような遺伝関係論文が多く載っているから，Biometrika は健康科学を主対象とした統計科学の雑誌であると言えよう．この学術誌のその後の内容を見ると，健康科学が統計科学に研究課題を提出し，統計科学の側がそれに答えようとして，研究成果を発表していたことが見て取れる．

Biometrika についで歴史の古い同系統の学術誌は米国統計学会が 1945 年に創刊した Biometrics (当初は Biometrics Bulletin) である．この雑誌は，Biometrika より直接的に生物現象を認識・把握することを重視している．創刊の辞がそれを，“The BIOMTRICS BULLTIN is designed primarily for biologists who see in statistics a potent tool for their work.”と明記している．健康科学が統

計科学に課題を出して統計科学が発展することも大事だが，統計科学が健康科学に寄与することはもっと大事だ，という問題意識が明確である．

統計科学の推測的側面を強調した R. A. Fisher は，統計学の雑誌に多くの論文を発表すると同時に，遺伝学の雑誌にも多くの論文を発表している．Fisher の古典的に有名な著書 “Design of Experiments” (1935) には例題として，健康科学関連のものがあふれている．遺伝現象の解明，生態の測定・把握，病気の治療法の開発などを効率的に行えるようなデータ取得と，そのための統計解析法の考案・体系化を通して，統計科学と健康科学は相互に寄与しあってきたというのが両者の歴史であろう．

この相互寄与の歴史は，ときどきに関連の強さを変えながら現在まで続いている．そして近年は，その相互寄与が一段と強まっている．その状況を読者に伝えるために，健康科学の側からどのような課題が出され，統計科学の側がそれにどのように答えているかを，§3 臨床試験，§4 市販後調査，§5 遺伝子解析，§6 動物実験代替法，§7 インシリコ試験 (*in silico* assay)，などの事例で紹介してみよう．

3 臨床試験

3.1 国際的統計ガイドラインの確立

さまざまな薬害や無効治療の認識・経験の下で,「根拠に基づく治療」(evidence based medicine; EBM) が必要である，という認識が医学関係者の間で定着したのは 1995 年頃である．

根拠として主要なものはデータである．Chambers の英英辞書によれば，データというのは，“Data are facts given (quantities, values, names etc) from which other information may be inferred; such facts, in the form of numbers or characters, which can be input to a computer.（数量，数値，名前などの，与えられた事実で，それから情報が引き出せるもの；コンピュータに入力できる数・文字形式の事実）” のことである．

データの中でも，根拠としての証拠能力が最も大きいのは，ヒトに計画的に投薬・治療を施した結果のデータ，すなわち臨床試験のデータである.（カタカナで書いてある「ヒト」は生物種としての人間のことである.）

臨床試験は投薬治療だけでなく，治療一般，さらには食品の健康寄与効果や安全性についても用いられるが，本節では，簡単のために，新しい薬の候補物質（被験薬）の効果を対照薬（現在使われている標準薬あるいはプラセボ）に比べて比較する臨床試験，すなわち「治験」に焦点を絞って話を進める．プラセボとは，薬理作用としての薬効は全くないが，見かけは被験薬と区別がつかない薬物のことで，薬を飲んだから直るはず，という心理効果を被験薬の評価において差し引くために用いるものである．以下では，被験薬と対照薬を「治験薬」と総称し，治験を行う主体を「治験者」ということにする．

治験では，例えば 25 歳以上 60 歳未満で，被験薬の対象となる疾患の患者であること，というように被験者の選択規準が定められている．その条件を満たす被験者を必要人数だけ集めて，ランダム (random) に 2 群に分け，一方に被験薬，他方に対照薬を対応づけ，被験者には割り付けられた群の治験薬を投与する．しかもそのとき，2 重盲検法 (double masking)，すなわち被験者にも医師にも当該被験者がどちらの群に含まれているか分からないやり方を用いる．

予め定めておいた薬効評価のための主要評価変数，例えば 1ヶ月で治癒したら 1 そうでなかったら 0 という値を取る変数，を測定し，測定値の平均の群間差が統計的に有意で，被験薬群の方の

値が大きかったら被験薬に薬効があると評価する．これが臨床試験における標準的な判断法である.（ランダムとは，どちらの群に割り付けるかを等確率で偶然に委ねることである.）

この判断法は EBM の視点で客観的で適切と思われるが，実際の運用には注意すべきことが少なくない．その一つは事後解析・後知恵解析 (*post hoc* analysis) の乱用である．例えば次のようなやり方である．

- 60 歳未満という選択規準があるのに 65 歳の人がいるというように，選択規準を満たさない被験者が治験に入っていて，この被験者を含めて検定すると統計的に有意差があるのに除外すると有意差が無くなるというとき，治験者が前者を採用する．
- 被験者を男性と女性に分けると，男性では有意差がないが女性では有意差があるというとき，治験者がこの被験薬を女性に有効なものとする．
- t 検定，ウエルチ検定，ウイルコクソン検定というように，同じデータに適用できる複数の検定手法があり，得られたデータについてはウエルチ検定のみで有意差があるというときに，治験者がウエルチ検定をもって被験薬には薬効があるとする．

一見 EBM に沿っているように見えるこの種の結論の出し方は，実は，良いところ取りの論理，いわゆる後知恵解析で，統計科学の視点で誤った結論を導きやすいものである．実際，薬として売り出された後で，薬効がないとして製造販売が取り消された抗痴呆薬のようなものには，この種の解析に依っていたものが多かった．

これらは，データをどのように取得し利用し結論を導くかということであり，統計科学が答えるべきことなので，関係者は世界的な規模で議論を行った．そして臨床試験のための統計ガイドラインを作ろうという流れができた．

ガイドラインは，当初，米国，欧州（EU），日本と個別に作成されたが（厚生省医薬局審査管理課長, 1992; CPMP, 1995），やがて米国を先頭とした臨床試験の国際的共通化の試み,「日米 EU 医薬品規制調和国際会議」(International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; ICH) によって統一化され，1998 年に「臨床試験のための統計的原則」(Statistical Principles for Clinical Trials) が作られた (厚生省医薬安全局審査管理課長, 1998; Lewis *et al.*, 2001)．現在はこれが日本を含めた世界各国で標準となっている．

ICH では，統計科学に限らず，新薬承認に関係する多くのガイドラインを作っている．これらは一般に「ICH ガイドライン」あるいは「ICH ガイダンス」と呼ばれている．統計科学に関係するものでは，適切な対照薬の選択についてのガイドラインも存在している（厚生労働省医薬局審査管理課長, 2001; http://www.pmda.go.jp/）.

「臨床試験のための統計的原則」では，

- 臨床試験のデータ解析で最も重視すべきことは，偏りを最小にして，かつ精度を最大にすることである．
- 試験のやり方とデータの解析法は事前に計画書に詳しく書き，それを変更するときにはその正当性を明確に記しておかなければならない．
- 被験者の割付はランダムにすべきであり，主要評価変数はできるだけ一つにして，しかも客観的なものにすべきである．

- データ解析計画も盲検解除の前，すなわちどの被験者がどの群に割り付けられているか明らかにする前に解析法を定めるべきである．

といったことを指摘して，治験者がデータを自分の都合の良い方に恣意的にゆがめて利用することを厳しく戒めている．これはその後の臨床試験の質を高めるための統計科学からの寄与であると筆者は考えている．

3.2 非ランダム割付法

「臨床試験のための統計的原則」は，その有用性が広く認められているが，制約が強すぎて臨床試験の効率を損なうことがある，という批判がないわけではない．その批判の上で提案・検討されている一つの手法を紹介しよう．それは被験者をランダムに 2 群に割り付けるのではなく，バランスを良くするように系統的に割り付けようという方法で，一般に「最小化法」(minimization method) と呼ばれているものである．

臨床試験で被験薬の有効性を確認するためには，試験計画書で事前に定められた選択規準を満たし，しかも同様に定められた除外規準に抵触しない被験者をできるだけ偏りなく募集する．事前に対照薬に対して薬効がどれだけ大きければ薬として承認すべきか，という有効サイズ（薬効差/標準偏差）を定めておき，これに対応する対立仮説を（例えば有意水準 2.5%の）仮説検定で,（例えば検出力 80%で）検出できるように,（例えば各群 180 人の）被験者数を設定する．その数だけの被験者をそれぞれの群に逐次，ランダムに割り付けるのが単純ランダム割付である．逐次というのは，全被験者を一度にではなく，同意を得られた被験者を次々と 1 人づつということである．例えば毎月 10 人のペースで，3 年間かかって約 360 人の被験者を集める，といったことである．

治験者は，このようにして治験に参加してもらった被験者に，医師を通して，試験計画書に指定されたやり方で治験薬を投与し，主要評価変数を測定する．得られた結果の群平均を 2 群間で比較し被験薬の有効性を評価する．このようなやり方の試験を,「ランダム化 2 重盲検並行群間比較試験」(randomized, double-masked, parallel-group controlled trial) という．

このやり方は，群間比較に確率論的原理を適用して仮説検定を行うためのフィッシャー 原則を応用したものである (Fisher, 1935; 佐藤俊哉, 1995; 椿広計他, 1999).

この原理は一見良さそうであるが，これを単純に適用すると，以下に述べるように，偏り無く精度の良い比較が必ずしも実現できない，という問題が生じる．

被験者はある疾患の患者であるが，環境条件と遺伝因子を均質にして飼育した実験動物とは違って，非常に多種多様で個性的である．性，年齢，遺伝因子，食事の質，肥満度，疾患の重症度等がまちまちである．これに加えて，治療を行う医療施設，医師の違いも無視できない．

これらの因子の中には，治療の結果を大きく左右する因子，すなわち予後因子 (prognostic factor) となるものが少なくない．これらの因子の影響を無視してランダム割付で 2 群比較を行うのは，これらの因子による被験者の反応の違いをすべて誤差と見なすことになり，誤差を非常に大きなものとする．

単に誤差を大きくするだけでなく，結果として明らかな偏りが実現してしまうことも稀でない．例えば全被験者 200 人の中に，治療効果が劣る重症患者が 30 人いたとしよう．ランダムに 100 人ずつの 2 群に被験者を割り付けたとき，重症患者が一方の群にどれくらい偏っては含まれるかを調べるとその確率分布（超幾何分布）は表 1 になる．

12 人以下対 18 人以上，というように人数が偏る確率が 1/3 である．こういうことが起こると，当然，18 人割り付けられた群の方が治療の結果が平均的に悪くなる．真の治療効果とは別の原因

表 1: 被験者 200 人重症患者 30 人のランダム割付における対照群の重症患者数の確率分布

人数 (人)	9	10	11	12	13	14	15	16	17	18	19	20	21
確率 (%)	1.1	2.5	4.8	7.9	11.4	14.2	15.2	14.2	11.4	7.9	4.8	2.5	1.1

が交絡 (confound) して真の状態とは異なった偏った結果をもたらすのである．これが分かっていながら，ランダムに割り付けたのだからその差は偏りでなく偶然誤差である，というのは無理である．結果としてこうなった場合には，予後因子の影響を調整して偏りが少なくなるように，推測を行うべきである．この調整のことを共変量調整 (covariate adjustment) という．

共変量調整には，何らかの仮定，すなわち確率モデルの想定が必要である．そしてそこに，想定したモデル次第で結論が異なるという問題が生じる．

フィッシャー (Fisher, 1935) はこの問題に対して，明らかに影響する因子を層別因子として扱い，層別ランダム化でその影響が偏りをもたらさないようにすることを提案した.「層別ランダム割付」である．よく知られているように，層別ランダム割付は群間比較の精度を上げるための優れた工夫である．

しかしながら臨床試験では，層別ランダム割付がきわめて利用しにくい．総被験者数がたかだか 200〜300 人なのに，考慮しなければならない層の数が 10 を超えることが稀でないからである．実際，医療施設を層として考慮すると，各層での被験者数が多くても 10 人程度，少ないときは，4 人程度になる．これに例えば重症度というような因子を含めて層別ランダム化を行うことは実際上不可能である．

このような状況の下で，層別ランダム割付よりは実用的で，ランダム割付よりは確実に予後因子のバランスを図ることができる割付法が提案され，使用されている.「最小化法」と通称される方法がそれで，典型はポコック・サイモン法 (Pocock and Simon, 1975) である．

最小化法は，ある種の癌など，被験者数を多く集めるのが困難でしかも影響の大きい予後因子が存在する分野で，積極的に採用される傾向があった．やがてそれが，その必要性がない臨床試験にまで用いられるようになり，欧州の規制当局 (Committee for Proprietary Medicinal Products; CPMP) は,「臨床試験のための統計的原則」と絡めて最小化法を強く否定 (strictly discourage) する文書，“Point to Consider” (CPMP, 2004）を発行した．その結果，最小化法の是非についての論議が激しく交わされるようになった (McEntegart, 2003; Buyse, 2004a; Senn, 2004; Buyse, 2004b).

Point to Consider は，その根拠・理由があまり明確にされていない．これが議論を混乱させているのであるが，擁護者はその論議の中で，系統的にバランスを持たせることの利点がほとんど無い（minimum）のに予見可能性が生じることで結果が偏るから用いるべきでない，と主張している．そこで研究として必要なことは，予見可能性がほとんど生じないバランス割付法があるかということと，系統的であってもバランスをとることに利点があるかどうかということ，を明らかにすることであった．

前者については，ポコック・サイモンの原法などに偏コイン法 (biased coin method) を導入することが考えられる (Efron, 1971)．すなわち，決定論的にバランスを取るのではなく，ある確率でバランスを取りやすくする方法の提案である．癌の臨床試験の条件でその方法を提案し，シミュレーション実験（モンテカルロ法）を通してその有用性を示したのが，例えば萩野篤司らの研究である (Hagino *et al.*, 2004).

後者については，計量的予後因子のバランス割付を提案した西次男ら (Nishi-Takaichi, 2003) の

割付法に，遠藤輝ら (Endo *et al.*, 2006a; 2006b) が，新しい評価規準と確率化を導入する研究を行い，バランス割付を採用することが予後因子調整に関する薬効評価の頑健性をもたらすことをシミュレーション実験で示した．これは 2 群だけでなく 3 群の場合にも適用でき，予後因子に質的因子と計量的因子が混じっているときでも適用できる方法なので，割付法の選択肢を増やしたものである．

3.3 複数評価変数への対処

「臨床試験のための統計的原則」は，検証的臨床試験における主要評価変数を，できるだけ一つに絞ることを勧めている．しかし現実には，複数の評価変数を用いざるを得ない疾患が少なくない．米国の規制当局である食品薬品庁 (Food and Drug Administration; FDA) がやむを得ないとしているものには，アルツハイマー病，慢性閉塞性呼吸器疾患等の 20 疾患がある (Offen *et al.*, 2007).

このようなときに問題になるのは被験者数の設計法 (sample size design) である．臨床試験では，必要な結論が明確に出せるという条件の下で，可能な限り少ない被験者の臨床試験を行うことを倫理的な観点から原則としている (佐藤俊哉, 1995)．より具体的には，被験薬の実際の効果サイズ，あるいは対象疾患で最低限必要な効果サイズに対して，一定の有意水準（「臨床試験のための統計的原則」では 2.5%）の検定で一定の検出力が保持されるような被験者数を計算し，臨床試験からの脱落確率を考慮に入れて被験者数を設定することになっている．

ところが複数評価変数の場合に現実に採用されている設計法は，評価変数ごとに必要被験者数を設計してその大きい方を採用するという方法と，各評価変数を独立なものとして多変量的に被験者数を設計するという方法であった．これらの設計法は，原則的な視点での被験者数とは異なった被験者数を設定している．

主要評価変数が複数の場合には，多重性の調整無しにすべての変数で有意差がついたときのみ被験薬の有効性を認めるのが標準である．この原則で被験者数を設計するには，変数が正規分布に従う場合でも，多変量 t 分布ではなく，多変量正規分布の確率をウィシャート分布で重み付けて積分する必要がある．その視点と計算方法が臨床試験家の間では最近まで確立していなかった．

これに対して寒水孝司らは (Sozu *et al.*, 2006) ，ウィシャート分布に関するモンテカルロ積分が確かな計算結果を与えることを示し，そのための計算プログラムを用意した．モンテカルロ積分というのは，積分を系統的な数値計算で求めるのではなく，ランダムに点を選ぶことで積分を計算する技法である．寒水らがモンテカルロ積分を利用したのは，分散・共分散が局外母数となる関係で積分が，3 次元（2 変数）あるいは 6 次元（3 変数）というように高次元になり，系統的な数値積分で必要な精度が確保できなくなるためである．この寒水らの提案は，従来見過ごされてきた症例数設計の問題を解決したものとして，評価されている．

参考のために，ある実例で用いられた方法で採用された被験者数と，寒水らの提案によって計算された必要被験者数が，実際にどの程度違うかをかを示すと，表 2 のようになる．検出力を 80%あるいは 90%としたとき，実例では 77 人あるいは 102 人と計算されたが，寒水らの提案では 83〜94 人，あるいは 109〜118 人が必要になる．実例では検出力不足の臨床試験を設計したことになる．臨床試験に協力してもらう患者数を必要十分な適正数に定めることは，患者のみでなく試験を実施する側にとっても非常に重要なことなのである．

表 2: 主要評価変数が 2 つのときの実例に則した必要被験者数（人）の試算値
有意水準 2.5%, 効果サイズ $\delta_1/\sigma_1 = 1.0/2.00, \delta_2/\sigma_2 = 0.5/1.10$

想定	実例	試算例における相関係数				
検出力	計算値	-0.8	-0.4	0.0	0.4	0.8
0.80	77	94	94	93	90	83
0.90	102	118	118	117	115	109

3.4 適応的試験計画の提案

本当に良い治療・薬剤をできるだけ短期間で開発するには，ICH ガイドラインに定めているような硬いやり方ではなく，もっと柔軟にやり方を変更して試験を実施した方がいいという声が，がん治療や生殖医療など先端性のある分野で大きくなった．標準的なやり方は時間と手間がかかり，良い治療・薬剤を早く普及させることの障害になっているというわけである．

この種の要求はやがて，適応的試験計画 (adaptive design, flexible design) というキャッチフレーズの下で大流行することとなった (Chou and Chang, 2007; Röhmel, 2006)．この 3,4 年での関連する研究論文数は，他のカテゴリーのものに比べて群を抜いて多い．

適応的試験計画の必要性は,「臨床試験のための統計的原則」を制定する過程でも議論されていた．そのときのおおよその合意は，検証的臨床試験，つまり治療や薬剤の有効性を明確に確認するための臨床試験では，事前に推定していた局外母数の値の調整というような，比較的限られた場合のみにするべきだということであった．ガイドラインの文章としては，§4.4 「必要な被験者数の調整」で次のように述べている．

> （被験者数の計算根拠となる仮定の）確認は，試験計画の詳細が予備的情報もしくは不確実な情報，又はその両方に基づいている場合，特に重要であろう．盲検下のデータを用い中間での確認を行うことにより，それまでの試験全体での，反応の分散，イベントの発生率又は生存状況が予期していた状況と異なることが明らかにされる場合がある．その場合，適切に修正した仮定に基づいて被験者数の再計算を行うこととなるが，その正当性を明らかにし，治験実施計画書の改訂及び総括報告書に記録しなければならない．

これに対して，現在の議論はこのような水準をはるかに超えて，柔軟性を極限まで追求したいというものである．すなわち，試験の途中で盲検を解除してデータを調べ，それに基づいて以下に列挙するような点について，試験計画を変更するというものである．

- 被験者数を計算し直して変更する．
- 計画されていた総被験者数はそのままにして各群への割付の比率を変更する．
- 複数の群のいくつかを途中で停止させて，残りの群のみで試験を継続する．
- 選択規準を，例えば重症者のみ，というように変更する．
- 当初に設定していた有効性の主要評価変数を，例えば当初の副次評価変数でおきかえる．
- 単純に平均が大きいことの検定を行う優越性試験を，下駄 (non-inferiority margin) を履かせた検定を行う非劣性試験におきかえる．

このような試験法の変更は本来なら，実施場面での必要性と妥当性が先導して，それに問題がないかどうかを統計科学的に吟味するべきものである．ところが現在進行している状態は逆である．このような適応的試験計画を採用すると，第 1 種の過誤がどの程度不安定になるか，一部のデータが開示された後の条件付き検出力を制御する被験者数計算はどのようにすればよいか，途中で被験者数を変更したときの最終的な有効性評価はどのように行うべきか，ということなどについての数学的モデル上の議論・研究が活発なのである．被験者母集団の変更のような，どのように数学的仮定が変わるのか統計科学的に想定し難いことまで，単純な数学モデルで議論されているのが現状である．

筆者は，このような統計科学独走の研究は，研究としては成立するが，実際の臨床試験の発展に必ずしも寄与しないのではないかと考えている．もっと丁寧にかつ実際的に，臨床試験関係者と統計科学の研究者が協同で議論・検討・理解・協調・合意を試みるべきであると考えている．

4　市販後データの利用

4.1　薬剤の市販後調査の必要性

薬事法上の「薬」は臨床試験で有効性と安全性が確かめられて市販が許されたものである．しかし臨床試験は非常に限られた条件でしか行うことができない．実際，臨床試験では，(1) 疾患の重症度を選択規準で指定する，(2) 併用薬の使用は原則として禁止する，(3) 合併症を持つ患者は除外する，(4) 事前に別の治療が行われている患者は除外する，(5) 妊娠期の女性と幼少年は除外する,… というように，選択規準と除外規準を細かく定めている．さらに担当医師は原則として専門医であり，治療施設はある程度以上の技術・施設水準を持っているのが普通である．

倫理的理由によって被験者数は多くても 1,000 人くらいまでであり，有効性の吟味が主眼とされるために，安全性を保証する条件が十分でないのが普通である．現実に使用される多種多様な条件下で多数の患者でのみ発見できる稀な副作用についての情報は，臨床試験で得ることができない．安全性をより詳しく調べ，問題が起こっていないかどうか監視するには市販後調査が必須である．

市販後調査のデータには，(1) 製薬企業が使用成績を調べる調査の結果，(2) 製薬企業に自発的に送られてくる報告，(3) 行政当局に自発的に送られてくる報告，(4) 臨床試験として計画される市販後試験の結果，等いくつかの種類がある．このどれであるかによってデータ解析での困難点や要工夫点が異なるので，統計科学に対しては，それぞれに応じた解析法の開発と適用が求められている．

困難点を一般的に列挙すると，(1) 対象母集団（被験薬が投与される対象疾患患者の全体）の大きさや特徴が不明確で，得られた結果の適用範囲が確かでない，(2) 情報に偏りがあるにも関わらずその程度・大きさが評価し難い，(3) 交絡している予後因子・共変量が無数にあり，そのどれが真に結果に影響しているかが調べにくい，(4) 例えば現在論議の的になっているタミフルと少年の異常行動のように，副作用と疑われるものが見いだされても，その因果関係を確認する手段が乏しい，(5) 報告の精度・正確さが確かめ難い，等がある．

市販後調査担当者はこのような困難の中で，可能な限り早く，知られている副作用の頻度の推定，知られていない稀な副作用の発見に努めている．近年はそのための技法の研究が，データマイニング手法を応用したシグナル検出法という形で進められている (Matsushita *et al.*, 2007).

4.2　統計モデルを導入した使用成績調査データの解析例

市販後調査データは，臨床試験データと違って，整然とした条件に従っていないので，解析では，単純に頻度分布を比較する程度にするのが一般である．しかし，特定の疾患と特定の薬剤の組み合わせに注目する使用成績調査では，その特徴をモデル化して有用な情報を集約することが可能な場合もある．その一例が福島彰らの研究である (Fukushima *et al.*, 2006)．以下にその要点を紹介しよう．

対象薬剤は抗がん剤 TS-1 である．抗がん剤は，がん細胞を攻撃すると同時に正常細胞である白血球や赤血球も破壊する．その結果，副作用としての血液毒性が現われる．したがって抗がん剤治療では，投与スケジュールの管理が重要になる．血液毒性が重篤な有害作用を発現させない範囲で，投与と休薬とをクール (cours) という単位で繰り返し，腫瘍縮小効果が最大になるように用法・用量を定めることになっている．(クールというのは，ラテン語起源の用語で，英語の course と同じものである．)

抗ガン剤治療のこのスケジュールは，開発過程での検討に従って市販承認時に定められるが，その妥当性は市販後の使用成績調査で確認されなければならない．そこで TS-1 の投与が行われた患者に対して定期的な血液検査が行われた．血液毒性は検査でしか調べられないからである．

図 1 は，そのような検査結果で発見された血液毒性の初発の頻度分布の例である．このデータの特徴は，検査日によって観察される頻度が左右されることである．実際には，その検査日以前に毒性が発現しているのであろうが，その日を特定できないから，いわゆる打ち切りデータ (censored data) になっている．このようなデータでは，単純な頻度比較で有用な情報を得ることができない．だからといって，例えば 1 週間単位に頻度を集約したのでは，治療開始後のどの時期にどのように副作用が発現しているかということについて，精度の良い情報が得られない．情報の欠損が大きくなりすぎるからである．

[図 1 の挿入]

これについて福島らは，有害事象の初発までの時間の分布（生存時間分布）について確率モデルを想定して，最尤法で適合度を評価し，妥当と思われたモデルを前提にして有害事象発現プロファイルを評価した．ここでの有害事象は，白血球の数，赤血球の数等がある基準値以下になる事象のことである．

候補としてのモデルには，多少の試行錯誤の後で，後に示す式で表される，「滑り混合ワイブルモデル」 (slip-mixed Weibull model) と「滑り混合対数ロジスティックモデル」 (slip-mixed log-logistic model) が用いられた．ここで $h(t)$ はハザード関数，すなわち時点 t で生存している個体が有害事象を発現する瞬間確率，$\lambda_1, \lambda_2, \gamma_1, \gamma_2$ は分布のパラメータ，w は第 1 クールと第 2 クール間の有害事象発現についての違いを表す重みパラメータ，$I(\bullet)$ は命題 "$\bullet$" が真のとき 1，そうでないとき 0 という値を取る指示関数 (indicator function) である．時間 t の単位は日で，第 2 クールの始まりが 42 日目であるため，第 2 クールについては時間を 42 日ずらしてある．このずらしを入れたことが「滑り」という形容詞を用いた理由である．

滑り混合ワイブルモデルのハザード関数

$$h(t) = W(t)\gamma_1\lambda_1(\lambda_1 t)^{\gamma_1 - 1} + (1 - W(t))\gamma_2\lambda_2(\lambda_2(t-42))^{\gamma_2 - 1} I(42 < t)$$

ただし，$W(t)$ は次式の値である．

$$W(t) = \frac{w \exp\left(-(\lambda_1 t)^{\gamma_1}\right)}{w \exp\left(-(\lambda_1 t)^{\gamma_1}\right) + (1-w)(1 - (1 - \exp\left(-(\lambda_2(t-42))^{\gamma_2}\right))I(42 < t))}$$

滑り混合対数ロジスティックモデルのハザード関数

$$h(t) = W(t)\frac{\gamma_1\lambda_1(\lambda_1 t)^{\gamma_1-1}}{1+(\lambda_1 t)^{\gamma_1}} + (1-W(t))\frac{\gamma_2\lambda_2(\lambda_2(t-42))^{\gamma_2-1}}{1+(\lambda_2(t-42))^{\gamma_2}}I(42<t)$$

ただし，$W(t)$ は次式の値である．

$$W(t) = \frac{w\frac{1}{1+(\lambda_1 t)^{\gamma_1}}}{w\frac{1}{1+(\lambda_1 t)^{\gamma_1}} + (1-w)(1-\frac{(\lambda_2(t-42))^{\gamma_2}}{1+(\lambda_2(t-42))^{\gamma_2}}I(42<t))}$$

これらのモデルを実際のデータに当てはめたところ，滑り混合対数ロジスティックモデルの方がデータによく当てはまったので，福島らは，このモデルに基づいて有害事象発現プロファイルを推定して薬理学及び実際の医療現場に接している担当者に確かめた．その結果この結論は妥当なものであるという評価を得ることができた．統計科学的検討によって有用な知見が得られた例と言える．

5　遺伝子解析の進展が提起する統計科学の問題

5.1　マイクロアレイデータの特徴

ヒトの全塩基配列を同定するというヒトゲノム計画が 2000 年に一段落したとき，非専門家の人たちの中には，これで個性がすべて計算機で調べられることになる，と誤解した人が少なくなかったであろう．(英語の genome を日本語でゲノムと書くかジェノムと書くかは好みの問題のようで，genome にゲノムという用語を使っているのに，gene をジーンと書く人が稀でない．)

実際はそうでなかった．30 億対の塩基配列が記号として分かったからといって，それでその人の個性が分かるわけではないからである．アナロジーで言えばそれは，30 億ステップのコンピュータプログラムを目の前に出されたからといって，そのプログラムが何をするか分かるわけでないのと同じである．現在のわれわれにできることは，その一部分に注目して，その部分に少し違いがあるプログラム（塩基配列）を並列的に流して（測定して），それによる結果の違いから，そのプログラム部分（対立遺伝子，allele）の役割を推測することである．プログラムについてのことであればこれを行うのがプログラム解析であり，塩基配列についてのことであればこれを行うのが統計的遺伝子解析である．

遺伝子解析では，ある遺伝子座 (locus) における対立遺伝子 (allele) の遺伝子型 (genotype) の違いを取り上げる場合と，ある一塩基 (single nucleotide) の位置（所番地）における一塩基多型 (single nucleotide polymorphism; SNP) の違いをとりあげるときがある．どちらの場合でも，その型の違いが薬剤への反応の違いになっているとき，これを「薬剤感受性がある」と言い，その違いが疾患のかかりやすさの違いになっているとき，これを「疾患感受性がある」と言う．そういう用語法でいうと，遺伝子解析とは，薬剤感受性や疾患感受性のある遺伝子や SNP を突き止め，それがどのような情報伝達経路で，表現型 (phenotype) の発現につながるかを検討することとなる．

薬剤（あるいは疾患）感受性遺伝子（あるいは SNP）を調べるのには，網羅的接近法が多く用いられている．例えばマイクロアレイと呼ばれる技術を用いて，数千あるいは数万個の遺伝子のそれぞれから産生されるリボ核酸 (RNA) などをプレート上の網目に捉え，各網目におけるその量（発現量）を調べるのである．その量が薬剤を投与された動物群（被験群）と投与されなかった動物群（対照群）で有意に異なるかどうかで，遺伝子の薬剤感受性を評価するのである．

ここで有意に異なるというのは，例えば有意水準 5%の t 検定での有意差というものではない．仮説検定では感受性が検出できないからである．なぜかというと，マイクロアレイで一度に捉える遺伝子は数千から数万である．検定をするとすると，英語表現では ‘1000s or 10000s’ というように多くの仮説検定を行うことになる．こんなに多くの検定を同時に行うと，検定の多重性によって第 1 種の過誤確率が極端に上昇し，感受性という現象を検出しているのか，第 1 種の過誤を検出しているのか分からなくなる．

検定の多重性とは，第 1 種の過誤，すなわち帰無仮説が真であるときに誤ってそれを棄却する確率を有意水準（例えば 5%）以下にしておいても，そういう検定を 100 回行えば 5 回くらいは誤りを犯すというように，誤りの確率が検定の数と共に大きくなる現象のことである．

遺伝子解析では，費用，手間，倫理性，プライバシー等の関係で，サンプルサイズが，英語表現では，‘10s or 100s’ という程度に少ない．実際，筆者が今手許に持っているデータは，各群 4〜6 匹の 3 群のマウスについての，12,489 個の遺伝子のマイクロアレイデータである．

このような小さいサンプルサイズでは，真の感受性遺伝子を検定で検出する確率がきわめて小さくなる．そのようなときに多重比較法で多重性を調整するのは無謀であるから，仮説検定ではない手法を用意することが統計科学に求められることになる．

5.2 感受性遺伝子同定法の例

前項で述べたように，超多変量小標本のマイクロアレイデータで感受性遺伝子を検出しようとしたとき，伝統的な有意水準指定の検定手法は無力である．統計科学上の工夫が必要である．いろいろ出されている提案の一つを例として紹介しよう．

マイクロアレイ上で N 個の遺伝子の発現量が計測されたとする．被験群と対照群それぞれが n 匹の動物（例えばラット）からなっているとすると，各動物個体について，1 枚分のマイクロアレイデータが得られる．したがって，感受性遺伝子を調べることはサンプルサイズ n の 2 標本の違いを評価する問題，いわゆる 2 標本問題になる．

発現量の差について，例えば t 統計量のような適当な検定統計量を用意すると，一つの遺伝子に一つの帰無仮説と対立仮説が対応する．検定ということであれば N 個の仮説検定を同時に行うことになる．

仮に 1 万個の遺伝子があると，その中に 50 個というオーダーで薬剤感受性遺伝子があるというのが，遺伝学研究者の口にすることである．その感受性遺伝子の一つ一つが異なる強度の感受性を持っているから，対立仮説はある分布を持って存在している．これでは検出力という概念が無意味で，対立仮説分布とその存在割合に応じて，棄却限界値を定めるという攻め方が合理的となる．

こういう背景の下で，第 1 種の過誤確率と第 2 種の過誤確率のバランスで棄却限界値を決めるのでなく，有意とされる遺伝子の中で誤って有意とされるものの割合，すなわち偽検出率 (false discovery rate; FDR) を制御しようという考え方が出てきた．例えば Tusher *et al.* (2001) が提案している “significance analysis of microarrays (SAM)” がそれである．

第 1 種の過誤確率であれば，帰無仮説のみが確率計算に必要で，分布族を適当に想定すれば，例えば t 分布で棄却限界値が計算できる．しかし FDR の場合にはその計算が感受性遺伝子の分布に依存する．それにもかかわらずその分布は不明である．何らかの方法で FDR をデータから推定しなければならない．この問題を並べ替え確率を用いることで答えようとしたのが Tusher らの工夫である．

検定統計量を別のものにしたらどうか，対立仮説の分布に何らかの想定をしたらどうか，とい

うことで SAM の改変版がいろいろと提案されている．平川晃弘ら (Hirakawa *et al.*, 2006) の提案はその一つである．

平川らは，検定統計量として t 型統計量を用い，FDR を指定値以下にするような棄却限界値 (cut-off value) を混合正規モデルの下で推定することを提案した．これは個々には他の研究で提案されているものであるが，両者を同時に用いたときに良い性能の感受性遺伝子検出法ができるというものである．t 型統計量というのは，通常，統計量の分母におく標準誤差 (SE) の推定量に，ある意味でジェームス・スタイン流の原理 (James and Stein, 1961) を導入して，他の遺伝子から求めた標準誤差の情報を加味するもので，式で書けば次のようなものである．

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/n} + a_0}$$

この式で，x, y はそれぞれ，被験群と対照群の測定値を意味し，分子は測定値の平均の差，分母の平方根の中は不偏分散の和，最後に加えてある a_0 は，N 個の不偏分散の上側 10%点である．分母にこのように a_0 を加えるのは，サンプルサイズがわずか n =4 というオーダーであることから来る不偏分散の不安定さを，他の遺伝子の情報を利用して安定化させるもので，従来の統計科学の枠組みからはなかなか思いつかない工夫であろう．実際，シミュレーション実験で調べると，この工夫はかなり良い性能を解析法に与えているのである．

6　動物実験代替法のバリデーション研究

6.1　代替法の 3Rs 原則

ヒトの健康のためだとしても動物を虐待するのはよくないという社会風潮が，ヨーロッパを起点にして世界に広がった．最初のターゲットは化粧品の安全性の吟味に動物を使うことの禁止であったが，現在ではこれが，一般的な生命倫理 (bioethics) の課題となってきている．2007 年 8 月には，3Rs 原則を普及するための第 6 回国際会議 (6th World Congress on Alternatives & Animal Use in the Life Sciences) が日本で開催されている．

3Rs 原則というのは，

Replacement: 動物を使わない試験法の採用

Refinement: 動物の苦痛の少ない試験法の採用

Reduction: 使用動物数の少ない試験法の採用

であり，毒性試験の分野ではそのための試験法が次々と開発されている．日本ではこれらを一括して「動物実験代替法」あるいは単に「代替法」と呼んでいる．

代替法にはどのようなものがあるかというと，例えば人の皮膚を侵す皮膚腐食性を評価するために「3 次元ヒト皮膚モデル」が市販されている．これはヒトの皮膚表面の 3 層をヒトの細胞を培養増殖させることでインビトロ的な実験材料としたものである．他には，紫外線等の光にさらされることで皮膚に炎症が起こる現象を助長する光皮膚刺激性を評価するのに，酵母菌の増殖が光を当てた場合と当てない場合でどれくらい違うかを調べる試験法も日本で開発されている．さらには，発癌性を調べる「細胞形質転換試験」等々，多くの試験法が世界中で開発されつつある．

参考のために言うと，薬理試験，毒性試験，安全性試験の生物的側面は，臨床薬理試験を除けば，ヒト以外の生物あるいは生物由来の実験材料を用いて行われる．その試験には, インビボ試験

(*in vivo* test)，インサイチュウ試験 (*in situ* test)，インビトロ試験 (*in vitro* test) の区別があって，費用，時間，労力，精度，を勘案してそのどれを用いるかがおおよそ定まっている．*vivo* は生体，*situ* は存在しているその場という意味である．*vitro* はガラスを意味するラテン語であるが，ここでは試験管を意味している．これらにはそれぞれ独自の実験計画とデータ解析法の統計科学的問題があって，例えば大森崇ら (Omori *et al.*, 2002), 松永信人ら (Matsunaga *et al.*, 2002), セクら (Soek *et al.*(2006) などの研究もあるが，これについての詳細は割愛する．

代替法の開発が進む中で統計科学の課題が次々と登場してきている．例えば，開発された試験法が確かにヒトへの安全性を担保するか，予測性能の保証をどのように行えばよいか，というバリデーション研究の方法論の確立である．

一般論として言えば，経済開発協力機構 (Organization for Economic Co-Operation and Development; OECD) がバリデーションの規準をガイドライン “OECD series on Testing and Assessment No. 34: Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Sssessment” にまとめている．代替法のバリデーションは世界的にこのガイドラインに沿って行なわれている．このガイドラインが，個々の代替法の妥当性がどの水準に達しているかを評価するための一里塚 (modular approach) として出しているのが次の項目である．

(i) Test definition (including purpose, need and scientific basis);（定義）

(ii) Intra-laboratory repeatability and reproducibility;（施設内再現性）

(iii) Inter-laboratory transferability;（技術易移転性）

(iv) Inter-laboratory reproducibility;（施設間再現性）

(v) Predictive capacity (accuracy);（予測性能）

(vi) Applicability domain; and,（適用対象範囲）

(vii) Performance standards.（標準性能）

これらの各段階では常に，統計科学的評価が必要である．(i) では，毒性に関する陽性，擬陽性，陰性の判定規準 (criteria) ，あるいは毒性の強さの指標が適切に定義されていることが必要であり，これが統計科学的に妥当なものであることが根拠を持って示されていなければならない．(ii) では，同じ実験施設で実験をすればその代替法が，許容可能な程度のばらつきで同じ結果を与えるかどうかを実験データから評価することが求められる．(iii),(iv),(v) でも同様な課題が提起されている．

6.2 バリデーション研究における実験計画

ある試験法がある毒性試験の代替法として提案されると，その施設間再現性を確認するという問題が生じる．試験すべき被験化合物は典型的なものでも 100 種以上あり，研究をする労力・施設・機材・時間などはボランティアに依存しているから，バリデーション研究のために行う実験は必要最小限にしなければならない．

そのため，いろいろな制約下で最適な研究計画を作ることが統計科学の課題として持ち込まれ

表 3: LLNA-DA 法のバリデーション研究における制約と割付の例
10 施設に 9 被験物質を，最下行の数を指定して割り付ける．C1 と C2 は溶媒を共通にする関係で同じ施設で実験する．C5 と C6 も同様.

この表に示した割付は，与えられた制約条件下で各被験物質の毒性評価が比較的効率よく行える例である．

被験	施設									
物質	1	2	3	4	5	6	7	8	9	10
C1	⊕	⊕	⊕							
C2	⊕	⊕	⊕							
C3	◯			◯			◯			
C4	◯						◯	◯		
C5				⊙	⊙	⊙				
C6				⊙	⊙	⊙				
C7		◯			◯			◯		
C8			◯			◯			◯	
C9							◯	◯		◯
物質数	4	3	3	3	3	3	3	3	1	1

る．これは 100 年前に農場試験のやり方について統計科学が提起された課題と似ている．違うのは制約が多種多様で，釣り合い不完備ブロック計画 (balanced incomplete block design; BIB) という類の単純なやり方が役に立たないことである．

施設間再現性を調べるのであるから，多くの施設に実験を依頼する必要があるが，各施設それぞれの事情があり，すべてに同じ数の被験化合物，実験反復数，時期を求めることはできない．

例えば表 3 は，実際に皮膚感作性，すなわち皮膚に被験物質が塗布されることでアレルギー反応が生じやすくなる性質，を調べる LLNA-DA 法のバリデーション研究で生じた制約である．10 施設がボランティア的に研究に参加することとなったが，研究施設とそれに費やせるマンパワーの関係で，表の最下段にある被験物質しか実験を行うことができなかった．各被験物質は少なくとも 3 施設で実験を行いたいが，表中の被験物質 C1 と C2，及び C5 と C6 は溶媒を同じにしなければならない．被験物質名を隠して実験をする関係で，共通溶媒の被験物質は同じ施設に割り当てなければならない，という制約も生じた．その制約下でどのような割付が可能か，というのが統計科学に問われたわけである．

このような場合考えられることは，現実上の制約を満たす割付計画を全部列挙し，適当な最適化規準に関して指標値を計算し，指標値が比較的良好な割付をバリデーション研究に用いることである．例えば 高沼ら (Takanuma *et al.*, 2006) はこのやり方で皮膚感作性試験代替法のひとつである LLNA-DA 法の実験計画を検討したが，もっと多種多様な制約が現実にはあって，そのあらゆる場合に対処できる方法論の確立が必要になっている．このような問題に対する一般的方法論の確立は今後の課題であろう．

7　インシリコ試験の利用

7.1　インシリコ試験とは

従来はインビボ試験とインビトロ試験がほとんどであった毒性試験，薬理試験の分野に，最近インシリコ試験 (*in silico* assay) が割り込んで来て，活用されるようになっている．

silico は元素の珪素 (silicon) のことであり，計算機の主要部の IC や LSI が珪素を基材としている関係で，計算機を意味するのに使われている．つまり計算機の中での試験ということである．

インシリコ試験は，何万とある開発候補化合物から目的に適している化合物を篩い分けする際に，計算機内で化合物同士の結合しやすさを評価し，有望な物質とそうでない物質を区分けする技術である．これに試験というラベルを貼るのは，有望性の評価の際にいろいろな条件設定をして評価値を求め，それを試験での測定値とみなし，その測定値に基づいて篩い分けを行うからである．

インシリコ試験法の本体をなす道具は，被験化合物の構造や属性を計算機に入力したときに評価値を出力するソフトウエアである．すでに多くのインシリコ試験法が市販されているが，その開発者は主として計算機技術者と化学者であり，統計科学の専門家はあまり関与していないようである．出力をどのように活用するかというマニュアルに統計科学的センスがないからである．

インシリコ試験に入力される変数・データは，例えば受容体の立体構造，被験化合物の立体構造，大きさ，酸・塩基の化学的特性など，かなり多い．それらを用いてソフトウエアが受容体と被験化合物の結合しやすさを評価値とするとき，実験者がオプションとして指定すべきパラメータは，シミュレーション回数，許容距離限界等沢山ある．それをどういう値に設定したらよいかというのが統計科学に問われる問題である．

7.2　インシリコ試験データの解析法の開発例

角元慶二ら (Kakumoto *et al.*, 2004) は Dock というインシリコ試験法を用いている実験者から依頼を受けて，最適なオプションパラメータの定め方の研究を行った．

研究の当初では，最適なパラメータの同定という形で問題を設定していたが，どのようにパラメータを設定しても，インビトロ試験あるいはインビボ試験の結果を予測する性能があまり良くないという結果を得て，これがこのインシリコ試験法の限界であろうと考えていた．

しかしあるとき，特定の最適設定を求めるのではなく，複数のパラメータのそれぞれで実験値を得て，それを多変量的に利用するというアイデアを思いつき，変数選択を行って最適予測式を構成してみたところ，これが非常によい性能を持つことが分かった．

考えてみれば当然で，受容体と被験化合物は，どちらも分子量が非常に大きい有機化合物である．結合のしやすさは 1 次元的であり得ないから，被験化合物の性質に応じて，異なる結合関係で他と同じ結合の強さ（活性）を持つことがあり得る．したがってスクリーニングでは，できるだけ多くの側面を評価し得る多変量解析的な接近法が有効なのである．分かってみれば当たり前なコロンブスの卵である．

角元らは，一つの試験法での出力を多変量的に利用することを考えたが，同様な利用法は，複数の試験法の結果を利用するときにも考えられる．実際，林真ら (Hayashi *et al.*, 2005) は三つのインシリコ試験法をある決定樹 (decision tree) にしたがって組み合わせて用いることが，遺伝毒性の評価に有効であることを報告している．インシリコ試験法のデータは多変量的に利用するのが良いようである．

8　未来に向けての考察

8.1　短期的流行と永続的発展

前節までに述べたように，統計科学は，確率モデルでの定式化を超えた問題，母集団的なものが曖昧な状況，超多変量小標本のデータ，統計科学的帰無仮説での過誤や検出力では評価できない誤りの評価，データの多変量的利用，といったところで，健康科学から提出された課題に，いろいろな新しい試みで答えようとしている．それは反射的に健康科学に十分な寄与を与えていると筆者は考えている．

このような近年の相互寄与と同時発展は，流行的な短期間のものだろうか，それとも長く将来に継続するものだろうか？仮に継続するものだとしたら，その推進においてはどのような点を注意すべきだろうか？

他の分野で見ると，過去に無数というべき論文が流行的に出されたのに，最終的には社会から消滅したものが稀でない．例えば，1980 年代にフィーバーが起きた「常温核融合」，通産省が大金を費やした大型プロジェクトの「電磁流体発電」や「ごみの完全自動処理」がそれである．

他方で「information technology」や「micro-technology」は，流行がそのまま新しい技術革命をもたらし，社会のありよう全体に巨大な影響をもたらしている．生殖技術はこれらに裏付けられて革命的な人工操作を可能とし，結果として親子関係や兄弟姉妹の関係が定義困難という状況まで創造してしまっている．

どういう要因が短期的流行と永続的発展を分けるのか，あるいは左右するのかという考察が統計科学に関して必要と思われる．これは，技術史，技術論の主題の一つであり，気楽に答えられるようなことではないが，統計科学と健康科学の関係については，参考になりそうな面白い文献がある．Cordell (2002) である．

8.2　遺伝学と統計科学

遺伝学にはエピスタシス (epistasis) という概念がある．ある遺伝子によって異なった座にある遺伝子の発現が抑止され，上位の遺伝子の発現が優先される現象である．発現型が現われる遺伝子を上位の (epistatic), 抑止されるものを下位の (hypostatic) 遺伝子という．これについて Cordell (2002) は次のように書いている．

> The term ‘epistatic’ was first used in 1909 by Bateson (1909) to describe a masking effect whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect. This was seen as an extension of the concept of dominance for alleles within the same allelomorphic pair, i.e., at a single locus.
> （エピスタティックという用語がはじめて使われたのは，1909 年に Bateson がある遺伝子座での変異が他の遺伝子座の変異の発現を妨げるという遮蔽効果を記したときである．これは一つの遺伝子における対立遺伝子対間の優性・劣性の概念の拡張と見なされた.）

表 1 はエピスタシスの一つの表現例である．もし，上位の遺伝子 G が存在しなければ，あるいはそれが g/g という型であれば，下位の遺伝子 B によって髪の色が白か黒になる．遺伝子型が b を含めば白，そうでなければ B が優性なので髪の毛の色が黒になる．ところが相互作用をもつ上

位の遺伝子 G があって，その型が g/G あるいは G/G のときは白と黒が覆われて，髪の毛の色が灰色になる．このような現象がエピスタシスである．

表 4: Bateson が例示したエピスタシス現象

	遺伝子 G の表現型		
遺伝子 B の表現型	g/g	g/G	G/G
b/b	白	灰色	灰色
b/B	黒	灰色	灰色
B/B	黒	灰色	灰色

エピスタシスという現象・概念が何故提起されたかというと，それはメンデルの法則が成り立つという仮説の下で観察データを集めていたら，うまくいく例とうまくいかない例があったことであろう．これにエピスタシスという仕組みを仮定すれば，統計的に観察された後者の現象がうまく説明できることになる．もしそうだとすれば，これは 100 年前に統計科学が遺伝学に対して提起した仮説であったに違いない．

Cordell はこれについて，次のように述べている．推測通りと考えて良いであろう．

> The situation has been confused further by the fact that in quantitative genetics, following a paper by Fisher in 1918, the term ‘epistatic’ has been generally used in yet another different sense from its original usage. In Fisher’s 1918 definition, epistasis refers to a deviation from additivity in the effect of alleles at different loci with respect to their contribution to a quantitative phenotype.

DNA 配列が同定され，遺伝子座が具体的に指定できるようになった現在，統計科学と遺伝学の共同作業として，具体的にどの遺伝子座の間でエピスタシスが起こっているかが遺伝子ー遺伝子相互作用，ということで追求されている．そしてそこでは，先に述べた関連性評価における多重性が，統計的検出力を乱すという困難をもたらし，統計科学はこの困難に，例えば “Multifactor dimensionality reduction” というようなソフトウエアで挑戦している (Moore, 2005; Ritchie, 2005).

100 年近くの間をおいて今，昔の仮説の追求が再開されているのであるが，この間に起こったことは遺伝子解読という革命的技術の発展である．つまり，統計科学と健康科学（に限るわけではないが）の一方の進展は他方に大きな課題を提起し，それがある時間をかけて解決されると，次にはそれが別の課題を返してくるという関係があることが，研究を単なる流行で終わらせない要因ではなかろうか．

かって筆者は,「統計学は良薬を創るのに役立つだろうか」(吉村功, 1995) という問題提起を行ったが，それは他の分野に影響を与えられるほどでなければ，相互発展に寄与することはないという視点に基づいたものである．これは現在の多く行われている健康科学関連の統計科学の研究の将来を決めるものではなかろうか．

8.3 未来の相互寄与・発展を願って

レーメル (Röhmel, 2006) の適応的試験計画についての文章は，筆者にとって大変示唆的で興味深いものである．

> If popularity of a scientific concept were measured by the number of recently published manuscripts, one could only agree that “adaptive design” must be a very popular method.… the theoretical progress is far ahead of practical applications.
> （近年に公刊された論文の数で概念の普及ぶりを評価するなら，適応的試験計画は抜群に普及した概念ということになる．しかし実際は，理論が駆け足で走っていて実用がはるかに引き離されてしまっている.）

ここでいうところの論文は，統計科学の雑誌に載ったものであって健康科学の雑誌に載ったものではない．統計的に検証するのはほぼ不可能であるが，今のところ適応的試験計画という概念は健康科学の分野であまり受け入れられていないようである．これは単に理論が実用を引き離しているというだけでなく，統計科学が健康科学に寄与する内容を提供していないことを意味しているのではないだろうか.

第 4 節に例示した福島彰らの研究結果は，当初，TS-1 の開発・営業担当者から，強い異議を受けた．結論が実感に合わないというのである．研究の過程で福島らは，その異議が薬剤投与初期でのモデルとデータ間の不適合であることをつきとめ，当初用いていたワイブルモデルに加えて対数ロジスティックモデルを候補に入れ，どちらがデータにより適合するかを検討した．結果として，後者の方が適合するという統計的結論が得られ，その結果が異議を唱えた担当者からも受け入れられ，社内での検討会で実際の営業現場でも活用できるということになった.

この例は，統計科学での研究結果を健康科学の分野に提示し，健康科学の分野からの疑問を受け，共同研究を進めることで，統計科学の方法論としての提案を健康科学の成果にできたものと言える.

同様なことは，佐藤泰憲ら (Sato *et al.*, 2004; 2006), 菅波秀規ら (Suganami *et al.*, 2007) の研究についても言える．実際，遺伝子解析の佐藤らの研究は，検出した SNPs が医学的な意味で，確かに疾患感受性 SNPs であるかどうか，医学的に吟味する必要があるということで，医学者から発表の保留を求められた．しかしその後，医学的にも妥当性があるということで，その分野の研究論文として研究結果が投稿される段階に来ている.

これらの例を通して筆者が感じることは，健康科学は次々と統計科学的課題を提出し解答を要求してくる．それに対して統計科学側は自らの方法論を駆使して解答を試みる．その結果を健康科学側に提示してその受容可能性を問う．もし，受容可能性が低いならば，健康科学の発展に寄与する形で解答の改善を試み相手に受け入れられる内容にまで結果を昇華させる.

もし，その受け入れが，例えば後知恵解析の否定のように，正しい原則であるにもかかわらず健康科学側が受け入れを渋っているのであれば，説得性のある説明を通してその普及を図る．このような過程を通して，統計科学への信頼感を健康科学に持たせられるならば，未来の相互寄与は確かなものとなるであろう．統計科学の方もそれを通して新しい方法論を作らざるを得なくなり，ひいてはそれが統計科学の発展になるということである.

飛躍を承知で言うならば，過去の歴史では，統計科学の研究者が学のアイデンティティの確立にこだわらないで健康科学にとって何が求められているかを追求し，提出された課題に解答を出したところ，結果としてそれが統計科学自体の確立，発展をもたらした，と言えるのではないだ

ろうか．現在，統計科学の中に健康科学が提起した課題が大きな位置を持っているのは過去のそういう営為の結果であるから，それを続けていくことが未来の相互発展をもたらすと筆者は考えている．

謝辞

本稿をまとめるに当たっては山本拓氏の適切な問題提起と，国友直人氏の寛容なる激励を受けた．本稿の内容を日本統計学会誌に掲載したときには，査読者から適切な改善のための助言を得た．例示した各論文の著者からは，その内容に関する詳細な資料の提供を受けた．これらの方々に心から謝意を表したい．学術書の慣例にしたがって文中では敬称を省略した．お許し頂きたい．

参考文献

[1] Bateson, W. (1909). “Mendel’s Principle of Heredity,” *Cambridge University Press*, Cambridge. (注：筆者は未入手であるが，Cordell (2002) に引用されているので，読者の参考のためリストに入れておく．)

[2] Buyse, M. and D. McEntegart (2004a). “The CPMP’s position discouraging dynamic allocation techniques is unfair,” *Applied Clinical Trials*, Letters to Editor, May 1, 20004.

[3] Buyse, M. and D. McEntegart (2004b). “More nonSENNse about balance in clinical trials,” *Applied Clinical Trials*, Letters to Editor, July 1, 2004.

[4] Chou, S.C. and M. Chang (2007). “Adaptive Design Methods in Clinical Trials,” *Chapman & Hall*.

[5] Cordell, H.J. (2002). “Epistatis: what it means, what it doesn ’t mean, and statistical methods to detect it in humans,” *Human Molecular Genetics*, **11**, 2463-2468.

[6] CPMP Working Party on Efficacy of Medicinal Products (1995). “Biostatistical methodology in clinical trials in application for marketing authorisations for medicinal products,” *Statistics in Medicine*, **14**, 1659-1682.

[7] CPMP (2004). “Points to consider on adjustment for baseline covariates,” *Statistics in Medicine*, **23**, 701-709.

[8] Efron, B. (1971). “Forcing a sequential experiment to be balanced,” *Biometrika*, **58**, 403-417.

[9] Endo, A., F. Nagatani, C. Hamada and I. Yoshimura (2006a). “Minimization method for balancing continuous prognostic variables between treatment and control groups using Kullback-Leibler divergence,” *Contemporary Clinical Trials*, **27**, 420-431.

[10] Endo, A., C. Hamada and I. Yoshimura (2006b). “An allocation method for balancing continuous prognostic variables among treatment groups using the Kullback-Leibler information,” *Japanese Journal of Biometrics*, **27**, 1-16.

[11] Fisher, R.A. (1935). “The Design of Experiments,” Oliver & Boyd.

[12] Fisher, R.A. (1918). “The correlation between relatives on the supposition of Mendelian inheritance,” *Transactions of the Royal Society of Edinburgh*, **52**, 399-433. (注：現在筆者の手元にないが，Cordell (2002) に引用されているので，読者の参考のためリストに入れておく．)

[13] Fukushima, A., W. Kashiwagi, M. Sano, C. Hamada and I. Yoshimura (2006). “Estimating a hazard function for each of four items of adverse event induced by the anti-cancer drug TS-1 – Application of slip-mixed log-logistic model for interval censored data –,” *Japanese Journal of Pharmacoepidemiology*, **11**, 9-21.

[14] Hagino, A., C. Hamada, I. Yoshimura, Y. Ohashi, J. Sakamoto and H. Nakazato (2004). “Statistical comparison of random allocation methods in cancer clinical trials,” *Controlled Clinical Trials*, **25/6**, 572-584.

[15] Hayashi, M., E. Kamata, A. Hirose, M. Takahashi, T. Morita and M. Ema (2005). “In silico assessment of chemical mutagenesis in comparison with results of salmonella microsome assay on 909 chemicals,” *Mutation Research*, **588**, 129-135.

[16] Hirakawa, A., Y. Sato, T. Sozu, C. Hamada and I. Yoshimura (2006). “Estimating the false discovery rate using a normal mixture distribution with microarray data,” poster in the *XXIIIrd International Biometric Conference*, July 16-21, 2006, Montreal, Canada.

[17] James, W., C. Stein (1961). “Estimation with quadratic loss,” In *Proc. 4th Berkley Symp. Mathematical Statistics and Probability*, **1**, 361-379.

[18] Kakumoto, K., S. Yamanaka, C. Hamada and I. Yoshimura (2004). “A statistical analysis of an effective method to conduct *in silico* screening for active compounds,” *Chem-Bio Informatics Journal*, **4**, 121-132.

[19] 厚生省医薬安全局審査管理課長 (1998). “「臨床試験のための統計的原則」について,” http://www.nihs.go.jp/dig/ich/eindex.html, last access 2007 年 1 月.

[20] 厚生省薬務局新医薬品課長 (1992). “「臨床試験の統計解析に関するガイドライン」について,” 薬新薬第 20 号.

[21] 厚生労働省医薬局審査管理課長 (2001). “「臨床試験における対照群の選択とそれに関連する諸問題」について,” http://www.nihs.go.jp/dig/ich/eindex.html, last access 2007 年 1 月.

[22] Lewis, J., J. Römel, B. Huitfeldt, I. Yoshimura, T. Sato, T. Uwoi, H. Uesaka, R. O’Neill, S. Ellenberg, B. Louv and S. Ruberg (1999). “Statistical principles for clinical trials,” *Statistics in Medicine*, **18**, 1905-1942.

[23] Matsunaga N, J. Kanno, and I. Yoshimura (2002). “A statistical method for judging synergism: Application to an endocrine disruptor animal experiment,” *Environmetrics*, **14**, 213-222.

[24] Matsushita, Y., Y. Kuroda, S. Niwa, S. Sonehara, C. Hamada and I. Yoshimura (2007). "Criteria revision and performance comparison of three methods of signal detection applied to the spontaneous reporting database of a pharmaceutical manufacturer," *Drug Safety*, **30**, 715-726.

[25] McEntegart, D.J. (2003). "The pursuit of balance using stratified and dynamic randomization techniques: An overviews," *Drug Information Journal*, **37**, 293-308.

[26] Moore, J.H. (2005). "A global view of epistasis," *Nature Genetics*, **37**, 13-14.

[27] Nishi, T. and A. Takaichi (2003). "An extended minimization method to assure similar means of continuous prognostic variables between treatment groups," *Japanese Journal of Biometrics*, **24**, 43-55.

[28] Offen, W., C. Chuang-Stein, A. Dmitrienko, G. Littman, J. Maca, L. Meyerson, R. Murhead, P. Stryszak, A. Boddy, K. Chen, K. Copley-Merriman, W. Dere, S. Givens, d. Hall, D. Henry, J.D. Jackson, A. Krishen, t. Liu, S. Ryder, A.J. Sankoh, J. Wang, C.H. Yeh (2007). "Multiple co-primary endpoints: Medical and statistical solutions; A report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America," *Drug Information Journal*, **41**, 31-46.

[29] Omori, T., M. Honma, M. Hayashi, Y. Honda and I. Yoshimura (2002). "A new statistical method for evaluation of L5178Y tk+/- mammalian cell mutation data using microwell method," *Mutation Research*, **517**, 199-208.

[30] Pocock, S.J. and R. Simon (1975). "Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial," *Biometrics*, **31**, 103-115.

[31] Ritchie, M.D. (2005). "A global view of epistasis," *Neurosurg Focus*, **19**, October, 1-4.

[32] Röhmel, J. (2006). "Adaptive designs: Expectations are high," *Biomedical Journal*, **48**, 491-492.

[33] 佐藤俊哉 (1995). "ヘルスサイエンスのための統計科学: サンプルサイズの設計," 医学のあゆみ, **173/13**, 1041-1046.

[34] Sato, Y., H. Suganami, C. Hamada, I. Yoshimura, T. Yoshida and K. Yoshimura (2004). "Designing a multistage SNP-based genome screen for common diseases," *Journal of Human Genetics*, **49**, 669-676.

[35] Sato, Y., H. Suganami, C. Hamada, I. Yoshimura, H. Sakamoto, T. Yoshida and K. Yoshimura (2006). "The confidence interval of allelic odds ratios under the Hardy-Weinberg disequilibrium," *Journal of Human Genetics*, **51**, 772-780.

[36] Senn, S. (2004). "Unbalanced claims for balance," *Applied Clinical Trials*, Letters to Editor, July 1, 20004.

[37] Seok, K.J., H. Wanibuchi, K. Morimura, Y. Totsuka, K. Wakabayashi, I. Yoshimura and S. Fukushima (2006). "Existence of a no effect level for MelQx hepatocarcinognicity on a background of thioacetamide-induced liver damage in rats," *Cancer Science*, **37**, 453-458.

[38] Sowan, B. (1982). "Biometrika," *Encyclopedia of Statistical Sciences*, **1**, 248-251.

[39] Sozu, T., T. Kanou, C. Hamada and I. Yoshimura (2006). "Power and sample size calculations in clinical trials with multiple primary variables," *Japanese Journal of Biometrics*, **27**, 83-96.

[40] Suganami, H., K. Kano, Y. Kuwayama, C. Hamada and I. Yoshimura (2007). "Comparison of methods for parameter estimation in a circular linear mixed effect model incorporating the diurnal variation for evaluating the treatment effects of glaucoma therapy," *Japanese Journal of Biometrics*, **28**, 1-17.

[41] 高沼正幸, 寒水孝司, 大森崇, 浜田知久馬, 吉村功 (2006). "動物実験代替法バリデーション研究における被験物質割付の最適性に関する検討," 日本動物実験代替法学会第 20 回大会要旨集, 119-120.

[42] 椿広計, 藤田利治, 佐藤俊哉 (1999). "これからの臨床試験," 朝倉書店.

[43] Tusher, V.G., R. Tibshirani and G. Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of National Academy of Sciences*, **98**, 5116-5121.

[44] 吉村功 (1995). "統計学は良薬を創るのに役立つだろうか," 数理科学, **389**, 43-49.

第 8 章

時空間統計解析の理論と応用

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2011 年 11 月
第 III 部　時空間統計解析の理論と応用

矢島美寛[*1]

(東京大学経済学研究科教授)

時空間統計解析とは, 時間の推移とともに様々な地点において実験や観測によって採取されるデータに対して, その時空間的相互作用を明確に考慮した統計モデルを構築し, このモデルに基づいてデータに内在する時空間的変動メカニズムを明らかにする統計解析を意味する. 近年新たな発展を遂げている統計科学諸分野の中でも, 最も注目を浴びているテーマの一つであり, 関係する学問分野は自然科学から人文・社会科学に至るまで広範囲に渡っている. 本章では時空間統計解析に用いられる統計モデル, これらのモデルに対する推測理論, 応用例について紹介する. ただし「究極の統計解析」とも言うべき時空間統計解析には, 今後解決すべき問題が山積している. 最後にそのいくつかについても言及する

[*1] yajima@e.u-tokyo.ac.jp

8.1　序

すべてのデータは, 時間の推移とともに様々な地点において実験や観測によって採取される時空間データ (spatio-temporal data) である. したがって従来から統計科学は時空間データの解析に携わってきたと言える. しかし本章で言う時空間統計解析 (Spatio-temporal Statistical Analysis) とは, データに内在する時空間的相互作用を明確に考慮した統計モデルを構築し, このモデルに基づいてデータの時空間的変動メカニズムを明らかにすることを目的とした解析プロセスを意味する.

近年新たな発展を遂げている統計科学諸分野の中でも, 時空間統計解析は最も注目を浴びている分野の一つと言える. その理由としては地球温暖化, オゾン層の減少, 地震の発生, 鳥インフルエンザの伝播, 動植物の植生・生態の変化, 欧州統合に象徴される経済活動の国際化など, 環境学・疫学・経済学を含む広範な学問分野においてグローバルな視点からその時空間的変動メカニズムを緊急に解明しなければならない問題に我々が直面していることにある.

他方これらのデータを収集, 解析するためのインフラストラクチャーは, リモート・センシング, 全地球測位システム (Global Positioning System, GPS), 地理情報システム (Geographical Information System,GIS) など科学技術の発達, また統計解析ソフトウエアの普及, さらには大規模データが瞬時に解析可能な計算機の開発などにより整備され, 充実してきている.

このような時代状況を意識しつつ,　本章では時空間データに対する統計モデルの構築法, これらのモデルに対する推測理論, 応用例について説明する. ただし以下の点をあらかじめ断っておく. まずこの分野の研究は既に確立されているわけではなく, 今後解決すべき問題は多々ある. それらの若干については最後に言及する. 次に専門用語の訳語のなかにはまだ定着していないものもある. また筆者の興味, 力量などに依存して, トピックは選択的であり網羅的ではない. そのため 個々のトピックに関する叙述は質量ともに均等には構成されていない.

本章を読了後, 時空間統計解析について興味を持たれた読者が, さらに深くこの分野を学ぶための優れた参考書をここで列挙しておく. 洋書としては Cressie(1993), Haining(1990), Stein(1999), Finkenstädt(2007) がある. 最後の本はまだ発刊後日も浅く, この分野の碩学達が若手研究者向けに, 入門的な内容から最近の発展に至るまで平易に解説している. また理論を実際の時空間データ解析に応用する際の留意点を知る上では, Haining(2003) が有用で

ある. 和書としては間瀬・武田 (2003) がある.

前述のように時空間統計解析と密接に関連する学問分野は多岐に渡る. 経済データが与えらたとき, その変動の時空間的な相互関係あるいは空間的特性を解明する分野として時空間計量経済学 (Spatio-temporal Econometircs) や空間計量経済学 (Spatial Econometircs) がある (Paelinck and Klaassen(1979)). これらの分野を概観する上で便利な文献としては Anselin et al.(2004), Arbia(2006), Getis et al.(2004) がある. なかでも Arbia(2006) はコンパクトな文献で, この分野の成り立ちから最近までの発展, 今後研究を進めるべきトピックが平易に解説されている. 次にリモート・センシングによる環境データの採取方法および解析手法を解説した参考書として清水 (2002) がある. 最後に疫学では感染性疾病の伝播メカニズムの解明, 特定地域への集積の検出が重要な目的となる. この分野の参考書としては丹後ら (2007) がある.

8.2 時空間データの種類

本節ではまず時空間データをどのように数学的に表現するかについて説明する. 次にこの表現に基づいて分類される 3 種類のデータについて解説する.

8.2.1 データの数学的表現

実数の全体を $\boldsymbol{R} = (-\infty, \infty)$ とし, その $d(= 1, 2, \ldots)$ 次元ユークリッド空間は $\boldsymbol{R}^d$ と表す. また整数の全体を $\boldsymbol{Z} = \{0, \pm 1, \pm 2, \ldots\}$ とし, その d 次元直積集合 $\boldsymbol{Z} \times \cdots \times \boldsymbol{Z}$ は $\boldsymbol{Z}^d$ と表す. 両者を統一的に表す場合には $\boldsymbol{K}^d$ とする.

次に観測地点 (site) を $\boldsymbol{s}(\in \boldsymbol{K}^d)$ とし, そこで観測されるデータは確率変数 $Y(\boldsymbol{s})$ と表す. $Y(\boldsymbol{s})$ がスカラーのときは一変量データ, ベクトルのときは多変量データである. 本章では一変量データについて解説する. 多変量データに対するモデルに関しては Arbir(2006), Banerjee et al.(2005), Cressie(1993) を参照されたい. 例えば $d = 2$ のときには, $\boldsymbol{s}$ は 2 次元ベクトルで第 1,2 座標は各々緯度, 経度を示し, $Y(\boldsymbol{s})$ はその地点における地価などを考えればよい. また $d = 3$ であれば, $\boldsymbol{s}$ は 3 次元ベクトルで, 第 1,2,3 座標は各々緯度, 経度, 高さを示し, $Y(\boldsymbol{s})$ はその地点における気温などとする. $\boldsymbol{s}$ の動く領域を $D(\subset \boldsymbol{K}^d)$ としたとき, データの全体を $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ と書き, 確率場 (random field) と言う.

さらに観測時点も考慮する場合には, 次元 d を一つ大きくし $\boldsymbol{s}$ の最後の座

標が時点を表すとすればよい. 先ほどの気温の例では $d = 4$ として, $\boldsymbol{s}$ の第 4 座標を時点に取ればよい. 時点を強調したいときには第 4 座標のみ分離して $t(\in \boldsymbol{K})$ と記し, データを $Y(\boldsymbol{s}, t)$ と表す.

8.2.2　データの種類

領域 D の定義に依存して, 時空間データは 3 種類に大別される.

(a) 地点参照データ

D が正の体積を持つ d 次元直方体を含む $\boldsymbol{R}^d$ の部分集合であり, $\boldsymbol{s}$ が D 上を連続的に変化するとき, $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ を地点参照データ (point-referenced data) と言う. 前節で例に挙げた気温のデータや風速, 風向データなどがこのカテゴリーに分類されるが, あくまでデータに固有の構造や性質に基づいた定義である. 理論的には空間上を連続的に変動していくが, 実際の観測値は有限個の地点および時点で得られる.

(b) 格子データ

D は高々可算個の点からなる $\boldsymbol{R}^d$ の部分集合のとき, 格子データ (lattice data) あるいは地域データ (areal data) と言う. 観測地点の間隔が規則的な場合と不規則な場合がある. 前者の場合は D を $\boldsymbol{Z}^d$ あるいはその部分集合で表す.　各格子に画素が与えられた画像データなどがこれに当たる. 後者の例として, 図 8.1 は 2001 年の首都圏における公示地価の観測地点を表している (松田・矢島 (2007)). なお地域データにおいては $Y(\boldsymbol{s})$ が地点 $\boldsymbol{s}$ におけるデータ自身を意味する場合もあるが, ある行政地区における集計データなどではその地区の中心都市のデータとして割り当てることもある. その場合 $\boldsymbol{s}$ は観測地点ではなく地域を代表する地点と解釈する方が自然である. 例えば都道府県別の失業率をその都庁, 道庁, 府庁, 県庁所在地に割り当てる場合などである.

(c) 点配置データ

D そのものが確率変数となるデータを点配置データ (point pattern data) と言う. 例えばある事象が生起した地点のデータを解析する場合である. いま地点 $\boldsymbol{s}$ で地震が起きたときには $W(\boldsymbol{s}) = 1$, 起きなかったときには $W(\boldsymbol{s}) = 0$ とし, 地震の震度を $Y(\boldsymbol{s})$ とする. このとき $D = \{\boldsymbol{s} | W(\boldsymbol{s}) = 1\}$ が地震の起きた地点の全体になり, 領域 D および観測値 $Y(\boldsymbol{s})$ の特性について解析する. 起きた時点まで考慮すれば $D = \{(\boldsymbol{s}, t) | W(\boldsymbol{s}, t) = 1\}$ となる.

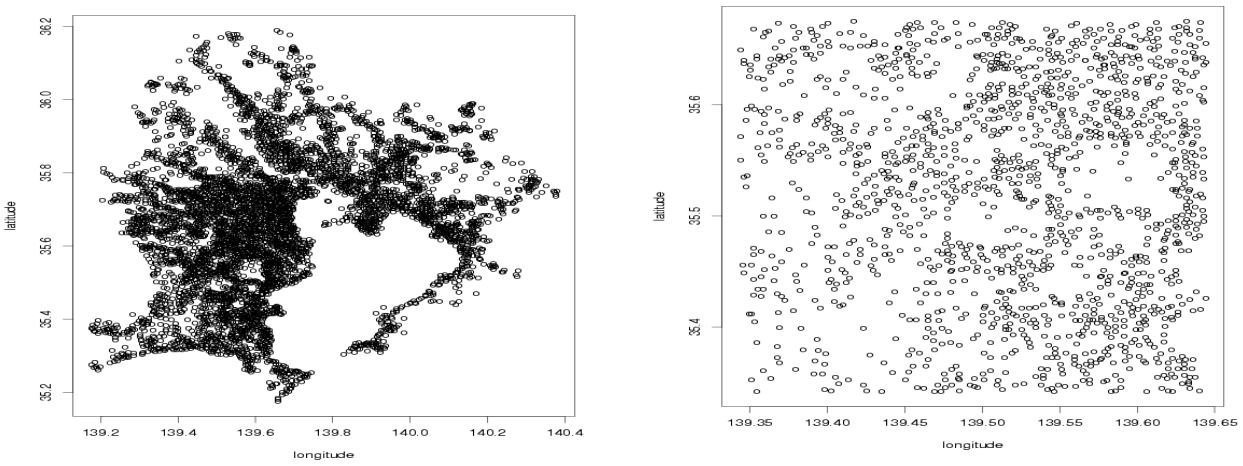


図 8.1　首都圏公示地価観測地点とその拡大図

8.3　定常確率場とそのモデル

本節以降では, 8.2 節で説明したデータの分類に対応して, その特性を表現する様々なモデルを紹介する. 本章で省略する点配置データに対する統計モデルおよび他の時空間モデルについては Cressie((1993), Finkenstädt et al.(2007), Lesage and Pace(2004), 間瀬・武田 (2003) および本シリーズ所収の尾形 (2008) を参照されたい.

まず本節では地点参照データおよび観測地点が規則的に並んでいる格子データの解析において中心的な役割を果たす定常確率場について説明する.

8.3.1　定常確率場の定義

確率場 $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \boldsymbol{K}^d\}$ が次の 2 条件を満たすとき, (弱) 定常確率場 (weakly stationary random field) と言う.

(i) 期待値は $\boldsymbol{s}$ に依存せず一定の値 $E(Y(\boldsymbol{s})) = \mu$ を取る. 以下では簡単のため断りのない限り $\mu = 0$ とする.

(ii) 共分散はベクトル差 $\boldsymbol{t}-\boldsymbol{s}$ のみに依存し, $C(\boldsymbol{t}-\boldsymbol{s}) = E(Y(\boldsymbol{t})Y(\boldsymbol{s}))$, $\boldsymbol{t}, \boldsymbol{s} \in \boldsymbol{K}^d$ となる.

特に $d = 1$ の場合は時系列解析における弱定常過程になる (8.10.1 付論 定常過程・ARMA モデルを参照されたい). $\{C(\boldsymbol{h}),\ \boldsymbol{h} \in \boldsymbol{K}^d\}$ を弱定常過程の場合と同様に自己共分散関数 (autocovariance function) と呼ぶ.

このとき $Y(\boldsymbol{s})$ および $C(\boldsymbol{h})$ はスペクトル表現が可能で, それぞれ

$$Y(\boldsymbol{s}) = \int_{T^d} \exp(i\boldsymbol{s}'\boldsymbol{\lambda}) dM(\boldsymbol{\lambda})$$

$$C(\boldsymbol{h}) = \int_{T^d} \exp(i\boldsymbol{h}'\boldsymbol{\lambda}) dF(\boldsymbol{\lambda}) \tag{8.1}$$

と表すことが出来る. ここで i は虚数単位 $i^2 = -1$, $\boldsymbol{s} = (s_1, \ldots, s_d)'$, $\boldsymbol{h} = (h_1, \ldots, h_d)'$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)'$, $'$ は転値ベクトルを意味する. $\boldsymbol{K} = \boldsymbol{Z}$ のときは $T = (-\pi, \pi]$, $\boldsymbol{K} = \boldsymbol{R}$ のときは $T = (-\infty, \infty)$ である. $M = \{M(\boldsymbol{\lambda}), \boldsymbol{\lambda} \in T^d\}$ は d 次元複素直交増分過程, $F(\boldsymbol{\lambda})$ は d 次元非負測度でスペクトル分布関数 (spectral distribution function) と各々呼ばれ, 任意のボレル集合 Δ, Δ_1, $\Delta_2 (\Delta_1 \cap \Delta_2 = \phi) \subset T^d$ に対して, $E|M(\Delta)|^2 = F(\Delta), E(M(\Delta_1)\overline{M(\Delta_2)}) = 0$ が成立する. スペクトル分布関数が絶対連続な場合には, その密度関数 $f(\boldsymbol{\lambda})$ をスペクトル密度関数 (spectral density function) と言う (Guyon(1995), Stein(1999), Yaglom(1987)).

8.3.2　定常確率場に対するモデル

統計モデルの導入法は大別して 2 つに分けられる. ひとつは $Y(\boldsymbol{s})$ 自身に対して直接導入する方法であり, 他方は自己共分散関数あるいはスペクトル密度関数に対して導入する方法である.

(a)$Y(\boldsymbol{s})$ 自身に対するモデル

(a1) ARMA モデル

定常過程 ($d = 1$) に対して応用上頻繁に用いられる統計モデルとして, 自己回帰移動平均モデル (Autoregressive Moving Average モデル, ARMA モデル) がある（8.10.1 節 付論 定常過程・ARMA モデルを参照されたい). このモデルを定常確率場に拡張しようとする場合に, 一つ根本的な問題が生じる. それは時系列データ ($d = 1$) の場合には過去, 現在, 未来と言う時間の推移に基づく自然な順序を導入できるが, 空間データ ($d \geq 2$) の場合にはこのような自明な順序は存在しないことである. 例えば「自己回帰」という言葉は, 時系列データの現在の値を自分自身の過去の値に回帰することを意味するが, 時空間データではもはやこうした意味を持たなくなる. この問題を解決し時空間データに順序を導入するための数学的概念として, 半空間 (half space) がある (Guyon(1995)).

簡単のため $d = 2$ とする. いま集合 $S \subset \boldsymbol{Z}^2$ が以下の 3 条件 (i)$S \cup (-S) = \boldsymbol{Z}^2(ii)S \cap (-S) = \{\boldsymbol{0}\}$ (iii) $\boldsymbol{s}, \boldsymbol{t} \in S \Rightarrow \boldsymbol{s} + \boldsymbol{t} \in S$ を満たすとき半空間と言

う. ここで $-S = \{\boldsymbol{s}| -\boldsymbol{s} \in S\}$ とする. S を用いてパラメータ間 $\boldsymbol{s}$, $\boldsymbol{t}$ に $\boldsymbol{s} \leq \boldsymbol{t} \Leftrightarrow \boldsymbol{t} - \boldsymbol{s} \in S$ と言う大小関係を定義すれば, この大小関係は反射法則, 対称法則, 推移法則を満足するので $\boldsymbol{Z}^2$ 上の全順序になる. 代表的なものとしては, 辞書式順序 (lexicographic order)$S_{lex} = \{(m,n); m > 0 \text{ or } m = 0, n \geq 0\}$ および α を無理数として $S_\alpha = \{(m,n); m\alpha + n \geq 0\}$ がある.

半空間 S を利用すれば, $d \geq 2$ に対しても自己回帰モデルを定義できる. $d = 2$ のときは $\boldsymbol{s} = (s_1, s_2)$ とおき, 後退作用素 (Backward Sift Operator)B_1, B_2 を $B_1 Y(s_1, s_2) = Y(s_1 - 1, s_2), B_2 Y(s_1, s_2) = Y(s_1, s_2 - 1)$ によって定義する. また $\{\epsilon(\boldsymbol{s}) : \boldsymbol{s} \in \boldsymbol{Z}^2\}$ が 2 次元のホワイト・ノイズ, すなわち $E(\epsilon(\boldsymbol{s}))^2 = \sigma^2$, $E(\epsilon(\boldsymbol{t})\epsilon(\boldsymbol{s})) = 0 (\boldsymbol{s} \neq \boldsymbol{t})$ を満たすとき, ARMA モデルは

$$P(B_1, B_2)Y(s_1, s_2) = Q(B_1, B_2)\epsilon(s_1, s_2)$$

によって定義される. ここで $P(B_1, B_2) = \sum_{(k,l)\in V} a_{kl} B_1^k B_2^l$, $Q(B_1, B_2) = \sum_{(k,l)\in W} b_{kl} B_1^k B_2^l$ とする. ここで V, W は有限個の要素からなる $\boldsymbol{Z}^2$ の部分集合とし, $(0,0)$ を必ず含むとともに $a_{0,0} = b_{0,0} = 1$ を満たし, かつ残りの要素はすべて S に含まれる.

$Q(B_1, B_2) \equiv 1$, $P(B_1, B_2) \equiv 1$ のときは, 各々時系列解析 ($d = 1$) における自己回帰モデル (AR モデル), 移動平均モデル (MA モデル) に対応する. ARMA モデルのスペクトル密度関数 $f(\lambda_1, \lambda_2)$ は, $d = 1$ の場合に対応して 2 変数 (λ_1, λ_2) の有理関数

$$f(\lambda_1, \lambda_2) = \frac{\sigma^2}{4\pi^2} \frac{|Q(e^{i\lambda_1}, e^{i\lambda_2})|^2}{|P(e^{i\lambda_1}, e^{i\lambda_2})|^2} \tag{8.2}$$

によって与えられる.

このように自己回帰移動平均モデルは $d \geq 2$ の場合にも数学的に正当化されるモデルではあるが, 必ずしもそれが現実のデータ解析においても有用なモデルとはならないことに注意が必要である. なぜならば半空間 $\boldsymbol{S}$ によって導入された順序が, データ間の時空間的相互依存関係を的確に表現しているか否かにモデルの有用性は大きく依存するからである.

次にパラメータの推定方法について説明する. 実際の解析では, 多くの場合データの期待値が地点・時点とともに変化する非定常性を考慮して, $Y(\boldsymbol{s})$ を説明変数 $g_i(\boldsymbol{s})(i = 1, \ldots, k)$ と定常ランダムフィールドに従う誤差項からなる回帰モデルによって表現する場合が多い. いま $Y(\boldsymbol{s})$ が,

$$Y(\boldsymbol{s}) = g(\boldsymbol{s})'\beta + \epsilon(\boldsymbol{s}) \tag{8.3}$$

によって定義されるとする．ここで $g(\boldsymbol{s}) = (g_1(\boldsymbol{s}), \ldots, g_k(\boldsymbol{s}))'$ は説明変数ベクトル, $\beta = (\beta_1, \ldots, \beta_k)'$ は回帰係数ベクトルとし, $\{\epsilon(\boldsymbol{s})\}$ は期待値 0 の ARMA 過程に従うと仮定する．$\sigma^2, a_{kl},\ b_{kl}$ をまとめて母数ベクトル θ とする．また回帰係数ベクトルと合わせて $\phi = (\beta', \theta')'$ とおく．

観測値 $Y(\boldsymbol{s}_i)(i = 1, \ldots, n)$ が与えられたとき, $\boldsymbol{Y}_n = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$ とおく．$\{Y(\boldsymbol{s})\}$ を任意の有限次元確率分布が多変量正規分布に従う正規定常確率場と仮定した場合には, 定数項を除去した ϕ の対数尤度関数は

$$L_n(\phi; \boldsymbol{Y}_n) = -\frac{1}{2}\log\det(V_n) - \frac{1}{2}(\boldsymbol{Y}_n - G_n\beta)'V_n^{-1}(\boldsymbol{Y}_n - G_n\beta)$$

になる (詳しくは 8.10.2 付論 正規過程に対する最尤推定法を参照されたい)．ここで G_n は $n\times k$ 行列でその j 列 $(j = 1, \ldots, k)$ は $g_j = (g_j(\boldsymbol{s}_1), \ldots, g(\boldsymbol{s}_n))'$ である．また V_n は $n \times n$ 共分散行列で, その (i, j) 成分は $\sigma(\boldsymbol{s}_i, \boldsymbol{s}_j; \theta) = Cov(\epsilon(\boldsymbol{s}_i), \epsilon(\boldsymbol{s}_j))$ である．

理論的には $L_n(\phi; \boldsymbol{Y}_n)$ を最大にする ϕ が最尤推定量であるが, 行列式 $\det(V_n)$ や逆行列 V_n^{-1} の計算が複雑になるので, 近似解ではあるがより効率的な計算方法が Arbia(2006), Basu and Reinsel(1994), Besag(1974), Mardia and Marshall(1984), Stein et al.(2004), Vecchia(1988), Wikle and Cressie(1999) により提案されている．

ここでは Vecchia(1988) によって提案され, 後に Stein et al.(2004) によって一般化された方法を説明する．まず $\boldsymbol{Y}_n$ を b 個の部分ベクトル $\boldsymbol{Y}_{jn}(j = 1, \ldots, b)$ に分割し, $\boldsymbol{Y}_n = (\boldsymbol{Y}'_{1n}, \ldots, \boldsymbol{Y}'_{bn})'$ と表す．また $\boldsymbol{Y}_{(jn)} = (\boldsymbol{Y}'_{1n}, \ldots, \boldsymbol{Y}'_{jn})'$ とする．このとき $L_n(\phi; \boldsymbol{Y}_n)$ は条件付き対数尤度関数の和

$$L_n(\phi; \boldsymbol{Y}_n) = L_n(\phi; \boldsymbol{Y}_{1n}) \sum_{j=2}^{b} L_n(\boldsymbol{Y}_{jn}|\phi; \boldsymbol{Y}_{(j-1,n)})$$

によって表現できる．ここで $L_n(\boldsymbol{Y}_{jn}|\phi; \boldsymbol{Y}_{(j-1,n)})$ は $\boldsymbol{Y}_{(j-1,n)}$ が与えられたときの $\boldsymbol{Y}_{jn}$ の条件付き対数尤度関数である．

さらに $\boldsymbol{S}_{(jn)}$ を $\boldsymbol{Y}_{(jn)}$ の部分ベクトルとして,

$$L_n(\phi; \boldsymbol{Y}_n) \approx L_n(\phi; \boldsymbol{Y}_{1n}) \sum_{j=2}^{b} L_n(\boldsymbol{Y}_{jn}|\phi; \boldsymbol{S}_{(j-1,n)})$$

により対数尤度関数を近似する．計算は簡便になるがその近似精度は $\boldsymbol{Y}_n$ の分割の仕方, $\boldsymbol{S}_{(jn)}$ の選択などに依存する．

もしデータ $Y(s_1, s_2)$ が直方体 $P_n = [1, 2, \ldots, n_1] \times [1, 2, \ldots, n_2]$ 上で観測されるときには, $d = 1$ の場合に用いられる近似最尤法を一般化した方法を利用できる．

まず総標本数は $n = n_1 n_2$ となる. 簡単のため $\beta = \mathbf{0}$ すなわち $Y(\boldsymbol{s}) = \epsilon(\boldsymbol{s})$, $\phi = \theta$ の場合を考える. 連続な関数 $h(u) : [0,1] \to [0,1]$（テイパー (taper) と言う）を用いて,$\underline{h}(\boldsymbol{s}) = \prod_{i=1}^{2} h(s_i/n_i)$, $H_n = \sum_{\boldsymbol{s} \in P_n} \underline{h}^2(\boldsymbol{s})$ とおき, ピリオドグラムを

$$I_n^T(\boldsymbol{\lambda}) = \frac{|\sum_{\boldsymbol{s} \in P_n} \exp(-i\boldsymbol{s}'\boldsymbol{\lambda}) \underline{h}(\boldsymbol{s}) Y(\boldsymbol{s})|^2}{(2\pi)^d H_n}$$

によって定義する. そして対数尤度関数の (-2) 倍である $-2L_n(\theta; \boldsymbol{Y}_n)$ を

$$L_n^*(\theta; \boldsymbol{Y}_n) = \int_{(-\pi,\pi]^d} \left(\log f(\boldsymbol{\lambda}, \theta) + \frac{I_n^T(\boldsymbol{\lambda})}{f(\boldsymbol{\lambda}, \theta)} \right) d\boldsymbol{\lambda}$$

によって近似し, $L_n^*(\theta; \boldsymbol{Y}_n)$ を最小にする θ を推定量に採用する. この推定量は Whittle 推定量と呼ばれている. Whittle 推定量は一致性をもち, 極限分布は正規分布になることが証明されている (Dahlhaus and Künsch(1987), Heyde and Gay(1993), Ludeña and Lavielle(1999)).

ただし $d = 1$ の場合には $h(u) \equiv 1$ とおいた生のピリオドグラムを利用できるが, $d \geq 2$ の場合には極限分布が正規分布にならないことに注意が必要である. 理由は生のピリオドグラムは標本自己共分散関数のフーリエ変換であるが, 標本自己共分散関数のバイアスが d とともに大きくなることにある. 標本自己共分散関数は $\hat{C}(\boldsymbol{h}) = \sum_{\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h} \in P_n} Y(\boldsymbol{s}) Y(\boldsymbol{s}+\boldsymbol{h})/n$ によって定義されるが, 標本数が $n_i (i = 1, 2) \to \infty, n_2/n_1 \to \alpha \in (0, \infty)$ と言う仮定のもとで増加すると, バイアスは $E(\hat{C}(\boldsymbol{h})) - C(\boldsymbol{h}) = O(n^{-1/d})$ になる. これはデータの境界が $d = 1$ の場合は点, $d = 2,\ 3$ の場合は線分, 面と言うように次元がだんだん大きくなることに起因する. 推定量の極限分布は通常 $1/n^{1/2}$ のオーダーになる. したがって $d \geq 2$ のときには漸近的に無視できない量になる. この現象は端効果 (edge effect) と呼ばれ, 時系列データの解析では生じない時空間統計解析に特有な問題である.

(a2)CAR モデル

ここでも簡単のため $d = 2,\ D = \boldsymbol{Z}^2$ とし, $\{Y(\boldsymbol{s})\}$ は正規定常確率場に従うとする. 正規性の仮定から $\{Y(\boldsymbol{t}) : \boldsymbol{t} \in \boldsymbol{Z}^2,\ \boldsymbol{t} \neq \boldsymbol{s}\}$ が与えられたときの $Y(\boldsymbol{s})$ の条件付き期待値は線形和 $\sum_{\boldsymbol{t} \neq \mathbf{0}} \alpha(\boldsymbol{t}) Y(\boldsymbol{s} - \boldsymbol{t})$ になり, 誤差項を $e(\boldsymbol{s})$ とおけば,

$$Y(\boldsymbol{s}) = \sum_{\boldsymbol{t} \neq \mathbf{0}} \alpha(\boldsymbol{t}) Y(\boldsymbol{s} - \boldsymbol{t}) + e(\boldsymbol{s})$$

となる. $e(\boldsymbol{s})$ は定義より $Cov(Y(\boldsymbol{t}), e(\boldsymbol{s})) = 0 (\boldsymbol{s} \neq \boldsymbol{t})$ をみたし, 自己共分散関数の対称性 $C(\boldsymbol{h}) = C(-\boldsymbol{h})$ より, $\alpha(\boldsymbol{t}) = \alpha(-\boldsymbol{t})$ が成立する.

また $\{Y(\boldsymbol{s})\}$ のスペクトル密度関数は

$$f(\boldsymbol{\lambda}) = \frac{\sigma_e^2}{1 - \sum_{\boldsymbol{t}\neq\boldsymbol{0}} \alpha(\boldsymbol{t}) \exp(i\boldsymbol{t}'\boldsymbol{\lambda})} \tag{8.4}$$

によって与えられる. ここで $\sigma_e^2 = Var(e(\boldsymbol{s}))$ である. さらに $\{e(\boldsymbol{s})\}$ も正規定常確率場であり, そのスペクトル密度関数は

$$f_e(\boldsymbol{\lambda}) = \sigma_e^2(1 - \sum_{\boldsymbol{t}\neq\boldsymbol{0}} \alpha(\boldsymbol{t}) \exp(i\boldsymbol{t}'\boldsymbol{\lambda}))$$

になる. ただし

$$Cov(e(\boldsymbol{s}), e(\boldsymbol{s}+\boldsymbol{t})) = \begin{cases} \sigma_e^2, & \boldsymbol{t} = \boldsymbol{0} \\ -\sigma_e^2\alpha(\boldsymbol{t}), & \boldsymbol{t} \neq \boldsymbol{0} \end{cases}$$

が成立するので, $\{e(\boldsymbol{s})\}$ はホワイト・ノイズではないことに注意を必要とする.

次に有限個の要素からなる部分集合 R が存在して, $\alpha(\boldsymbol{t}) \neq 0,\ \boldsymbol{t} \in R$, $\alpha(\boldsymbol{t}) = 0,\ \boldsymbol{t} \notin R$ が成立するとしよう. このとき $\{Y(\boldsymbol{t}) : t \in \boldsymbol{Z},\ \boldsymbol{t} \neq \boldsymbol{s}\}$ が与えられたときの $Y(\boldsymbol{s})$ の条件付き分布は, $\{Y(\boldsymbol{t}) : \boldsymbol{t} \in R\}$ のみが与えられたときのそれに等しい. この性質は確率過程のマルコフ性を確率場に拡張した概念に相当する. このような確率場を一般にマルコフ確率場 (Markov Random Field, MRF) と言う (Rue and Held(2005)). 特に上述のモデルを条件付き自己回帰モデル (Conditional Autoregressive Model, CAR モデル) と言う.

ここで (a1) で説明した AR モデルと CAR モデルの関係について調べよう. (8.2) の $|P(e^{i\lambda_1}, e^{\lambda_2})|^2$ を展開すれば,

$$f(\boldsymbol{\lambda}) = \frac{\sigma^2}{4\pi^2\beta(\boldsymbol{0})(1 - \sum_{\boldsymbol{t}\neq\boldsymbol{0}} \alpha(\boldsymbol{t}) \exp(i\boldsymbol{t}'\boldsymbol{\lambda}))}$$

となる. ここで $\boldsymbol{t} = (t_1, t_2)',\ \beta(\boldsymbol{t}) = \sum_{k-k'=t_1, l-l'=t_2} a_{kl}a_{k'l'},\ \alpha(\boldsymbol{t}) = \beta(\boldsymbol{t})/\beta(\boldsymbol{0})$ である. したがって $\sigma_e^2 = \sigma^2/(4\pi^2\beta(\boldsymbol{0}))$, $R = \{\boldsymbol{s} - \boldsymbol{t} | \boldsymbol{s},\ \boldsymbol{t} \in V\}$ とおけば, (8.4) より AR モデルは常に CAR モデルであることが分かる. ただし分散の関係からも分かるように $\epsilon(\boldsymbol{s}) \neq e(\boldsymbol{s})$ である.

しかし CAR モデルは必ずしも AR モデルにはならないことが示されている (Guyon(1995)).

(b) 自己共分散関数（スペクトル密度関数）に対する統計モデル

(b1) 等方型モデルと分離型モデル

現時点までに提案されているモデルの中で代表的なものとして等方型モデル (Isotropic Model) と分離型あるいは乗法型モデル (Separable Model) がある.

等方型モデルは, 自己共分散関数 $C(\boldsymbol{h})$ が距離 $\| \boldsymbol{h} \| = \sqrt{\sum_{i=1}^{d} h_i^2}$ のみに依存し, 方向には無関係なモデルである. すなわち $C_0(x)(x \in \boldsymbol{R})$ を 1 変数正定値関数としたとき,$C(\boldsymbol{h})$ は, $x = \| \boldsymbol{h} \|$ を代入して

$$C(\boldsymbol{h})) = C_0(\| \boldsymbol{h} \|)$$

によって表現される. しかし注意することは $C_0(x)$ が正定値関数であっても, 必ずしも上式で定義される $C(\boldsymbol{h})$ が $d(\geq 2)$ 変数正定値関数にはならないことである (Cressie(1993)). 任意の d に対して $C(\boldsymbol{h})$ が正定値になるための必要十分条件は既に示されており, 有界な単調非減少関数 $G(u)$ が存在して $C_0(x)$ が

$$C_0(x) = \int_0^{\infty} \exp(-x^2 u^2) dG(u)$$

を満たすことである. 証明は Cressie(1993), Stein(1999) などを参照されたい.

$C_0(x)$ に対するパラメトリック・モデルのなかで, 応用上頻繁に用いられるモデルは Matérn 族 (Matérn(1960), (1980)) である. Matérn 族に関する詳しい歴史については Guttorp and Gneiting(2006) を参照されたい. Matérn 族の一般形は

$$C_0(x) = \frac{\pi^{1/2}\phi}{2^{\nu-1}\Gamma(\nu + 1/2)\alpha^{2\nu}} (\alpha|x|)^{\nu} \mathcal{K}_{\nu}(\alpha|x|)$$

によって表現される. ここで α, ν, ϕ は正の定数また $\mathcal{K}_{\nu}$ は変形された Bessel 関数 (modified Bessel function) と呼ばれる特殊関数である. ϕ が共分散の大きさ, α と ν が形状および $x \to \infty$ のときの 0 への収束速度を規定するパラメータである. 例えば $\nu = 1/2$ のときは $\mathcal{K}_{1/2}(x) = \sqrt{\pi/(2x)} \exp(-x)$, $\nu = 3/2$ のときは $\mathcal{K}_{3/2}(x) = \sqrt{\pi/(2x)}(1 + x^{-1}) \exp(-x)$ になる. したがって自己共分散関数は $\nu = 1/2$ のときは $C_0(x) = \pi\phi\alpha^{-1} \exp(-\alpha|x|)$, $\nu = 3/2$ のときは $C_0(x) = \frac{1}{2}\pi\phi\alpha^{-3} \exp(-\alpha|x|)(1 + \alpha|x|)$ になる. この自己共分散関数に対応するスペクトル密度関数は $\omega = \| \boldsymbol{\lambda} \|$ のみに依存し

$$f(\omega) = \phi(\alpha^2 + \omega^2)^{-\nu - 1/2}$$

によって表現される.

次に分離型モデルでは, $C(\boldsymbol{h})$ が 2 個以上の正定値関数の積として表現される. 例えば $\boldsymbol{h}$ を 2 つのベクトル $\tilde{\boldsymbol{h}}_1 = (h_1, \ldots, h_m)'$, $\tilde{\boldsymbol{h}}_2 = (h_{m+1}, \ldots, h_d)'$ に分割し, $C(\boldsymbol{h})$ は 2 つの正定値関数 $C_1(\tilde{\boldsymbol{h}}_1)$, $C_2(\tilde{\boldsymbol{h}}_2)$ の積 $C(\boldsymbol{h}) = C_1(\tilde{\boldsymbol{h}}_1)C_2(\tilde{\boldsymbol{h}}_2)$ によって定義される.

この統計モデルを周波数領域から眺めれば,$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)'$ を 2 つのベクトル $\tilde{\boldsymbol{\lambda}}_1 = (\lambda_1, \ldots, \lambda_m)'$, $\tilde{\boldsymbol{\lambda}}_2 = (\lambda_{m+1}, \ldots, \lambda_d)'$ に分割し, スペクトル密度関数 $f(\boldsymbol{\lambda})$ は 2 つの正の値を取る関数 $f_1(\tilde{\boldsymbol{\lambda}}_1)$, $f_2(\tilde{\boldsymbol{\lambda}}_2)$ の積 $f(\boldsymbol{\lambda}) = f_1(\tilde{\boldsymbol{\lambda}}_1)f_2(\tilde{\boldsymbol{\lambda}}_2)$ によって表現される. したがって例えば時系列解析におけるスペクトル密度関数の d 個の積を考えれば, 任意の d 次元ランダム・フィールドに対するモデルを構成できる.

(b2) バリオグラム

時空間統計解析では自己共分散関数とともに確率場を特徴づける基本的な量としてバリオグラム (variogram) がある. 自己共分散関数とは異なりバリオグラムは定常確率場だけではなく非定常な確率場に対しても定義することが可能である.

いま確率場 $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ が任意の $\boldsymbol{h}$ に対して以下の 2 条件

$$E(Y(\boldsymbol{s}+\boldsymbol{h}) - Y(\boldsymbol{s})) = 0$$
$$E[Y(\boldsymbol{s}+\boldsymbol{h}) - Y(\boldsymbol{s})]^2 = 2\gamma(\boldsymbol{h})$$

を満たすとき, $\{Y(\boldsymbol{s})\}$ を固有定常確率場 (intrinsic stationary random field), $2\gamma(\boldsymbol{h})$ をバリオグラム, $\gamma(\boldsymbol{h})$ を半バリオグラム (semivariogram) と言う (Matheron(1963)). ただし以下では簡単のため半バリオグラムもバリオグラムと呼ぶ. 定常確率場は固有定常確率場であり, 次の等式

$$\begin{aligned} & E|Y(\boldsymbol{s}+\boldsymbol{h}) - Y(\boldsymbol{s})|^2 \\ &= Var(Y(\boldsymbol{s}+\boldsymbol{h})) + Var(Y(\boldsymbol{s})) - 2Cov(Y(\boldsymbol{s}+\boldsymbol{h}), Y(\boldsymbol{s})) \\ &= 2\{C(\boldsymbol{0}) - C(\boldsymbol{h})\} \end{aligned}$$

に着目すると,

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h}) \tag{8.5}$$

と言う関係が成立する.　したがって $\| \boldsymbol{h} \| \to \infty$ のとき $C(\boldsymbol{h}) \to 0$ であれば, $\lim_{\|\boldsymbol{h}\| \to \infty} \gamma(\boldsymbol{h}) = C(\boldsymbol{0})$ となる.

定義から $\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$, $\gamma(\boldsymbol{0}) = 0$ である. また $\| \boldsymbol{h} \| \to 0$ のとき $E|Y(\boldsymbol{s}+\boldsymbol{h}) - Y(\boldsymbol{s})|^2 \to 0$(平均 2 乗連続 ($L_2$-continous) と言う) であれば, $\gamma(\boldsymbol{h}) \to 0$ が成立する. しかし実際のデータから推定したバリオグラムは $\gamma(\boldsymbol{h}) \to c_0 (> 0)$ となる場合が報告されている. この現象はナゲット効果 (Nugget Effect) と呼ばれており, 解釈には 2 通りある.　ひとつは, 実際のデータ解析においては $\| \boldsymbol{h} \|$ がある閾値より小さいとき バリオグラムを測定できないが, その閾値よりミクロなスケールの確率場が存在すると言う解釈で

ある. 他方は測定誤差によって生じる影響と言う解釈である. この不連続性と整合的な理論モデルとしては, 各地点 $\boldsymbol{s}$ において真値 $Y(\boldsymbol{s})$ に分散が c_0 であるホワイト・ノイズ $\epsilon(\boldsymbol{s})$ が加わって観測される統計モデルが考えられる. したがって実際の観測値 $Y^*(\boldsymbol{s})$ は

$$Y^*(\boldsymbol{s}) = Y(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$

によって表現されると仮定する. $\{Y^*(\boldsymbol{s})\}$ のバリオグラムおよび自己共分散関数を各々 $\gamma^*(\boldsymbol{h})$, $C^*(\boldsymbol{h})$ とおけば,

$$\gamma^*(\boldsymbol{h}) = \begin{cases} \gamma(\boldsymbol{h}) + c_0, & \boldsymbol{h} \neq \boldsymbol{0} \\ 0, & \boldsymbol{h} = \boldsymbol{0} \end{cases}$$

$$C^*(\boldsymbol{h}) = \begin{cases} C(\boldsymbol{h}), & \boldsymbol{h} \neq \boldsymbol{0} \\ C(\boldsymbol{0}) + c_0, & \boldsymbol{h} = \boldsymbol{0} \end{cases}$$

となる. このとき $\gamma^*(\boldsymbol{h}) \to c_0 (\| \boldsymbol{h} \| \to 0)$ であるから, $\boldsymbol{h} = \boldsymbol{0}$ において不連続になる.

(b3) 推定法

まずバリオグラムの推定方法について説明する. 定常時系列データでは通常自己共分散関数を推定するが, 時空間統計解析ではバリオグラムを推定する場合が多い. 主な理由は自己共分散関数とは異なり, 期待値が未知でも推定の必要がないことにある.

いま観測値 $Y(\boldsymbol{s}_i)(i = 1, 2, \ldots, n)$ が得られたとする. そして $S(\boldsymbol{h}) = \{(i,j) | \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j\}$ とおく. すなわちベクトル差がちょうど $\boldsymbol{h}$ に等しい 2 地点のインデックスから成る集合である. また $S(\boldsymbol{h})$ の要素数を $N(\boldsymbol{h})$ とおく. ただし不規則な格子データなどではちょうどベクトル差が $\boldsymbol{h}$ に等しい 2 地点は数少なくなるので, 差が $\boldsymbol{h}$ に近い地点も含める.

このとき標本バリオグラムは

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{(i,j) \in S(\boldsymbol{h})} (Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j))^2$$

によって定義される. $\hat{\gamma}(\boldsymbol{h})$ は外れ値に影響されやすいので, よりロバストな推定量が提案されているが, その一つの例として

$$\bar{\gamma}(\boldsymbol{h}) = \left[\frac{1}{2N(\boldsymbol{h})} \sum_{(i,j) \in S(\boldsymbol{h})} (Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j))^{1/2} \right]^4 / (0.457 + 0.494/N(\boldsymbol{h}))$$

が有効とされている (Cressie(1993)).

パラメトリックなバリオグラムを推定する方法もいくつ提案されている. いま母数ベクトルを θ, それに対応するバリオグラムを $\gamma(\boldsymbol{h};\theta)$ とおく. 例えば Matérn 族ならば, $\sigma^2,\ c_0,\ \alpha,\ \phi,\ \nu$ などが θ の成分となる. 一つの方法としては最小 2 乗法があり, 残差平方和

$$RSS(\theta)=\sum_{i=1}^{b}[\hat{\gamma}(\boldsymbol{h}_i)-\gamma(\boldsymbol{h}_i;\theta)]^2$$

を最小にする θ を推定量とする. しかし通常実際のデータ解析では $\parallel \boldsymbol{h}_i \parallel$ が大きくなるにつれて $N(\boldsymbol{h}_i)$ は小さくなるので, $\hat{\gamma}(\boldsymbol{h}_i)$ の信頼性は低くなる. そこで $\boldsymbol{h}_i(i=1,\ldots,b)$ としては, $\parallel \boldsymbol{h}_i \parallel$ が小さい方から b 個のみを使用する. 経験的には $\parallel \boldsymbol{s}_i-\boldsymbol{s}_j \parallel$ の最大値の半分程度に $\boldsymbol{h}_b$ を設定することが推奨されている.

しかし $\hat{\gamma}(\boldsymbol{h})$ の分散は不均一で近似的には $\gamma(\boldsymbol{h};\theta)^2/N(\boldsymbol{h})$ に比例する. その場合には最小 2 乗推定量より加重最小 2 乗法の方が優れているので, ウエイトを分散の逆数 $N(\boldsymbol{h}_i)/\gamma(\boldsymbol{h}_i;\theta)^2$ にとり, 加重残差平方和

$$WRSS(\theta)=\sum_{i=1}^{b}N(\boldsymbol{h}_i)[\frac{\hat{\gamma}(\boldsymbol{h}_i)}{\gamma(\boldsymbol{h}_i;\theta)}-1]^2$$

を最小にする θ を推定量とすることも考えられる.

ただし問題は $\hat{\gamma}(\boldsymbol{h})$ を構成する際に $\boldsymbol{h}$ と $\boldsymbol{s}_i-\boldsymbol{s}_j$ の差をどこまで許容するか客観的な基準を設定することが難しいことにある. また $\{Y(\boldsymbol{s})\}$ が正規確率場に従う場合には, 最尤推定量も構成できるがその計算は非常に複雑になる.

これらの難点を回避するために Curriero and Lele(1999) は複合尤度法 (Composite Likelihood Approach) と言う, $\hat{\gamma}(\boldsymbol{h})$ を用いずかつ最尤法より計算が簡便な方法を提案している.　この方法は $Y(\boldsymbol{s}_i)-Y(\boldsymbol{s}_j)(i\neq j)$ が正規分布にしたがい,　その密度関数は

$$f(v_{ij},\theta)=\frac{1}{\sqrt{2\pi}\sqrt{2\gamma(d_{ij};\theta)}}\exp\left(-\frac{(Y(\boldsymbol{s}_i-Y(\boldsymbol{s}_j)^2}{4\gamma(d_{ij};\theta)}\right)$$

であることを利用する. ここで $v_{ij}=Y(\boldsymbol{s}_i)-Y(\boldsymbol{s}_j)$, $\gamma(d_{ij},\theta)=E(Y(\boldsymbol{s}_i)-Y(\boldsymbol{s}_j))^2/2$ である. このとき複合尤度関数は, すべての (i,j) のペアに対するこれらの密度関数の積

$$CL(\theta,\boldsymbol{V})=\prod_{i=1}^{n-1}\prod_{j>i}f(v_{ij};\theta)$$

によって定義される. ここで $\boldsymbol{V}$ は v_{ij} を成分とするベクトルである. このとき対数複合尤度関数の (-1) 倍である $-\log CL(\theta, \boldsymbol{V})$ は定数項を無視すれば

$$\sum_{i=1}^{n-1}\sum_{j>i}\left(\frac{(Y(\boldsymbol{s}_i - Y(\boldsymbol{s}_j)^2}{2\gamma(d_{ij};\theta)} + \log(\gamma(d_{ij};\theta))\right)$$

になる. この関数を最小にする θ を複合最尤推定量と言う.

自己共分散関数の推定量は

$$\hat{C}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})}\sum_{(i,j)\in S(\boldsymbol{h})}(Y(\boldsymbol{s}_i) - \overline{Y})(Y(\boldsymbol{s}_j) - \overline{Y})$$

によって定義される. ここで $\overline{Y}$ は標本平均 $\overline{Y} = \sum_{i=1}^{n} Y(\boldsymbol{s}_i)/n$ である. この推定量は時系列解析 $(d = 1)$ における標本自己共分散関数の一般化である. ただし理論式 (8.5) と異なり, 一般には $\hat{\gamma}(\boldsymbol{h}) \neq \hat{C}(\boldsymbol{0}) - \hat{C}(\boldsymbol{h})$ であることに注意が必要である.

(b4) 検定法

等方性の検定方法としては Sherman et al.(2003) がバリオグラムを用いた方法を, 分離可能性の検定方法としては Fuentes(2005) がピリオドグラムを用いた方法を, Mitchell et al.(2005) が尤度比統計量を用いた方法を各々提案している.

8.4　地域データ (Areal Data) に対する SAR モデルと CAR モデル

8.2.2 節で説明した地域データのなかで, 地域を代表する地点にデータが割り当てられている場合には, 地点間の距離や方向に基づいて相関関係が規定される定常確率場は統計モデルとしてあまり有効ではない. このようなデータに対しては定常確率場を代替するモデルとして, 同時自己回帰モデル (Simultaneous Autoregressive model, SAR モデル) と条件付き自己回帰モデル (Conditional Autoregressive model, CAR モデル) が提案されている.

いま n 個の地域のデータ $Y(\boldsymbol{s}_i)(i = 1, 2, \ldots, n)$ が与えらているとし, これらの同時分布は n 次元正規分布に従うと仮定する. 期待値は $E(Y(\boldsymbol{s}_i)) = \mu_i$ とおく. このとき SAR モデルは

$$Y(\boldsymbol{s}_i) = \mu_i + \sum_{j=1}^{n} b_{ij}(Y(\boldsymbol{s}_j) - \mu_j) + \epsilon(\boldsymbol{s}_i),\ i = 1, \ldots, n \tag{8.6}$$

によって定義される (Whittle(1954)). ここで $\boldsymbol{\epsilon}_n = (\epsilon(\boldsymbol{s}_1), \epsilon(\boldsymbol{s}_2) \ldots, \epsilon(\boldsymbol{s}_n))'$ は n 次元正規分布 $N(\mathbf{0}, \Lambda)$ に従い, b_{ij} は定数で $b_{ii} = 0 (i = 1, \ldots, n)$ とする. $\epsilon(\boldsymbol{s}_i) (i = 1, \ldots, n)$ が互いに相関を持つ場合には, Λ が非対角行列になる.

$\boldsymbol{Y}_n = (Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2), \ldots, Y(\boldsymbol{s}_n))'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ とおき, I_n を $n \times n$ 単位行列, $B = (b_{ij})$ とすれば, (8.6) は行列表現により

$$(I_n - B)(\boldsymbol{Y}_n - \boldsymbol{\mu}) = \boldsymbol{\epsilon}_n$$

となる. したがって $I_n - B$ が正則行列のときは, $\boldsymbol{Y}_n$ の分布が多変量正規分布 $N(\boldsymbol{\mu}, (I_n - B)^{-1}\Lambda(I_n - B')^{-1})$ になる.

他方 CAR モデルにおいては, $Y(\boldsymbol{s}_j) (j = 1, \ldots, n, j \neq i)$ が与えられたときの $Y(\boldsymbol{s}_i)$ の条件付き分布が, $N(\mu_i + \sum_{j=1}^{n} c_{ij}(Y(\boldsymbol{s})_j) - \mu_j), \tau_i^2)$ によって定義される. ここで c_{ij} は定数で $c_{ii} = 0$ とし, τ_i^2 は条件付き分散である (Besag(1974)).

ところで前節で説明したように定常確率場に対する CAR モデルはスペクトル密度関数を用いてその存在が理論的に正当化されているが, ここで定義された条件付き分布を持つ n 次元分布が存在するか否かは自明ではない. ただし実際にはその存在性が肯定的に示されている. $C = (c_{ij})$, T を対角行列 $T = \mathrm{diag}(\tau_1, \tau_2, \ldots, \tau_n^2)$, $Q = T^{-1}(I_n - C)$ とおく. このとき Q の (i, j) 成分は

$$q_{ij} = \begin{cases} 1/\tau_i^2, & i = j \\ -c_{ij}/\tau_i^2, & i \neq j \end{cases}$$

となる. もし $c_{ij}/\tau_i^2 = c_{ji}/\tau_j^2$ が成立し, さらに $I_n - C$ が正定値行列であれば, $\boldsymbol{Y}$ の分布は $N(\boldsymbol{\mu}, Q^{-1}) (Q^{-1} = (I_n - C)^{-1}T)$ になる. 最初の条件が Q および Q^{-1} の対称性を, 2 番目の条件が正定値性を保証している. 証明は Rue and Held(2005) などに示されている.

ここでこれらのモデルについていくつか注意点を述べておく. 第 1 点として SAR モデルと CAR モデルの関係について説明する. SAR モデルにおいて $\Lambda = \sigma^2 I_n$ のときには, $C = B + B' - BB'$, $T = \sigma^2 I_n$ とおけば CAR モデルになることが分かる. 逆に CAR モデルにおいて $(I_n - C)^{-1}$ をコレスキー分解 (Cholesky decomposition) し, $LL' = (I_n - C)^{-1}$ とおく. $T = \sigma^2 I_n$ のときには, $B = I_n - L$ とおけば SAR モデルになる. しかし B の意味づけが一般には困難になる (Haining(1990)).

第 2 点として実際のデータ解析においては SAR モデルおよび CAR モデルの係数行列 B, C をどのように取るべきかが問題になる. 従来良く用いられ

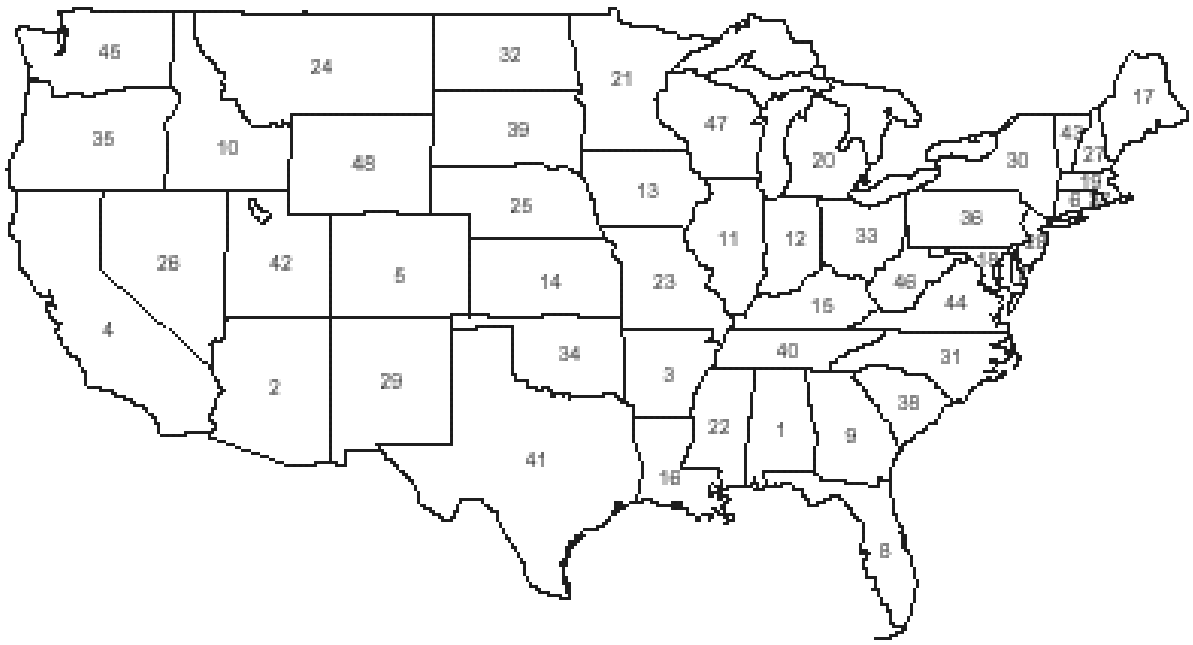


図 8.2　アメリカ合衆国地図

てきた方法では, まず隣接行列 (connectivity matrix)$W = (w_{ij})$ を

$$w_{ij} = \begin{cases} 1, & \boldsymbol{s}_i と \boldsymbol{s}_j (i \neq j) が共通の境界を持つ \\ 0, & i = j \\ 0, & その他 \end{cases}$$

によって定義する. この W を用いて $B = \rho_s W$, $C = \rho_c W$ とおく. ここで ρ_s, ρ_c は空間相関 (spatial correlation) あるいは空間従属 (spatial dependence) の強さを規定するパラメータである.

実際にアメリカ合衆国の州を示した図 8.2 を例にとって説明する. 番号は州名をアルファベット順に並べたときの順番を意味している. Alabama 州 (番号 1. 以下同様) と隣接する州は, Florida(8), Georgia(9), Mississippi(22), Tennessee(40) の 4 州である. したがって Alabama に対応する行では, これら 4 州に対応する列が 1 その他の列 0 がとなる.

しかしこの W を用いると, 地域の形状が非常に不規則な場合には, 行ごとに 1 となる成分数のばらつきが大きくなる. この欠点を是正するためには, 行ごとの成分の和が皆 1 となるように基準化した行列 $W^* = (w^*_{ij})$ を用いる. ここで $w^*_{ij} = w_{ij}/w_{i+}$, $w_{i+} = \sum_{j=1}^{n} w_{ij}$ とする. CAR モデルに対しては, 共分散行列 $(I_n - \rho_c W)^{-1} T$ は正値対称行列になる必要があるので, $w_{ij}/\tau_i^2 = w_{ji}/\tau_j^2$ および $\rho_c \omega_i < 1 (i = 1, \ldots, n)$ が成立しなくてはならない.

ここで ω_i は W の固有値である．W^* を用いた場合も同じ条件を満たさなければならない．

隣接行列を用いて B, C を定義することは簡便であり有効な場合も多いが, 多分に従来の慣行によるところもある．したがって個々のデータに対しては, 理論的に正当化されかつそのデータに適合する B, C の他の決定方法も常に可能性が残されている．例えば地点 $\boldsymbol{s}_i$ と $\boldsymbol{s}_j (i \neq j)$ の間のユークリッド距離や交通手段を利用したときの時間距離の関数を採用することも考えられる．

第 3 点としてしばしば誤解に陥りやすいことは, 時系列解析の AR(1) モデルにおいては, ρ_s, ρ_c が現時点と 1 時点前の観測値間の相関係数となるが, SAR モデルおよび CAR モデルでは必ずしもそのような含意を持っていないことに注意が必要である．SAR モデルでは $I_n - \rho_s W$ が正則行列, CAR モデルでは $I_n - \rho_s W$ が正定値対称行列になることを保証する ρ_s, ρ_c であれば良く, $|\rho_s| < 1$, $|\rho_c| < 1$ の範囲に限定する必要はない．

第 4 点として Wall(2004) は $\rho_s = \rho_c$ が成立したとしても, SAR モデルと CAR モデルでは地域データ間の相関構造がかなり異なっていることを指摘している．

最後に SAR モデルと CAR モデルの優劣については様々な見解がある．社会科学においては空間相関が直接的に表現できる SAR モデルの方が好ましいと言う見解がある一方で, 時系列解析における AR モデルの一般化と言う観点からは CAR モデルの方がより自然なモデルであるとする見解もある．これらの議論に関しては Cressie(1991), Haining(2003) を参照されたい．

8.5　クリギング

クリギング (Kriging) とは, 様々な地点で観測された値を用いて他の地点の値を予測する手法である．南アフリカの鉱山技師 Krige(1951) が実際のデータ解析に応用したことに因んで命名された (Cressie(1993), Matheron(1963), 間瀬・武田 (2003))．例えばある地点での鉱石の埋蔵量を知りたいとき, いくつかの別の地点でボーリングを行い埋蔵量を観測し, これらの観測値に基づいて 目的の地点の埋蔵量を予測する．統計学的に言えば, 予測方法は平均 2 乗誤差を最小にする線形不偏予測量を構成することに他ならない．現在では鉱山学に限らず環境学, 森林学, 水産学などにおける予測方法として基本的な方法である．

各地点の値 $Y(\boldsymbol{s})$ が一定の期待値 μ とランダムな項 $\epsilon(\boldsymbol{s})(E(\epsilon(\boldsymbol{s}) = 0)$ の和,

$$Y(\boldsymbol{s}) = \mu + \epsilon(\boldsymbol{s})$$

によって定義されるモデルを考える.

いま地点 $\boldsymbol{s}_i (i = 1, \ldots, n)$ における値 $Y(\boldsymbol{s}_i)$ が観測されたと仮定し, 他の地点 $\boldsymbol{s}_0$ における値 $Y(\boldsymbol{s}_0)$ を予測したいとする. まず期待値 μ が既知の場合から始める. 予測量としては, 線形予測量

$$P(Y; \boldsymbol{s}_0) = \mu + \sum_{i=1}^{n} l_i (Y(\boldsymbol{s}_i) - \mu)$$

を用いる. このとき平均 2 乗誤差 $E(Y(\boldsymbol{s}_0) - P(Y; \boldsymbol{s}_0))^2$ を最小にする係数は

$$l' = (l_1, l_2, \ldots, l_n) = C' \Sigma^{-1}$$

で与えられる. ここで Σ は $n \times n$ 共分散行列で, その (i, j) 成分は $Cov(\epsilon(\boldsymbol{s}_i), \epsilon(\boldsymbol{s}_j))$ であり, また C は n 次元ベクトルでその i 成分は $Cov(\epsilon(\boldsymbol{s}_0), \epsilon(\boldsymbol{s}_i))$ である. したがって最良線形予測量は

$$P(Y; \boldsymbol{s}_0)^* = \mu + C' \Sigma^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_n)$$

によって与えられる. ここで $\boldsymbol{\mu}_n = (\mu, \mu, \ldots, \mu)', \boldsymbol{Y}_n = (Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2), \ldots, Y(\boldsymbol{s}_n))'$ とする. この予測方法は単純クリギング (simple kriging) と呼ばれている.

次に μ が未知の場合は線形予測量

$$P(Y; \boldsymbol{s}_0) = \sum_{i=1}^{n} l_i Y(\boldsymbol{s}_i)$$

を用いる. ただし不偏予測量すなわち $E(P(Y; \boldsymbol{s}_0)) = \mu(\sum_{i=1}^{n} l_i) = \mu$ を満たす予測量を構成したいので, $\sum_{i=1}^{n} l_i = 1$ が制約条件となる.

いまラグランジュ未定乗数法により

$$E(Y(\boldsymbol{s}_0) - P(Y; \boldsymbol{s}_0))^2 - \lambda(\sum_{i=1}^{n} l_i - 1)$$

を最小にする $l_i (i = 1, 2, \ldots, n)$ を求めると,

$$l' = \left(C + J \frac{1 - J' \Sigma^{-1} C}{J' \Sigma^{-1} J} \right)' \Sigma^{-1}$$

が得られる. ここで $J = (1, 1, \ldots, 1)'$ である. 確かに制約条件

$$\sum_{i=1}^{n} l_i = l' J = C' \Sigma^{-1} J + (1 - C' \Sigma^{-1} J) = 1$$

が成立している. この予測方法は普通クリギング (ordinary kriging) と呼ばれている.

ところで実際には $\{\epsilon(\boldsymbol{s})\}$ の共分散構造が未知である. $\{\epsilon(\boldsymbol{s})\}$ が定常確率場に従う場合には, 前述の自己共分散関数あるいはバリオグラムの推定量を真の値に代入する.

次に回帰モデル (8.3) に対する最良線形予測量の構成法を一般クリギング (universal kriging) と言う. さらに非線形な予測量も既にいくつか提案されてはいるが (例えば Cressie(1993) を参照されたい), 線形予測量ほどにはまだ理論的に確立されていない.

8.6　時空間自己回帰移動平均モデル

定常確率場はパラメータ $\boldsymbol{s}$ の最後の座標が時点を表すと見なせば, 時空間モデルになる. 他方時空間自己回帰移動平均モデルは, 8.2.1 節で説明した時点を明示的に表現するモデルとして定義される. ただし定常確率場の自己回帰移動平均モデルとは異なることに注意を必要とする.

いま地点 $\boldsymbol{s}_i (i = 1, \ldots, n)$, 時点 $t = 1, \ldots, T$ において $Y(\boldsymbol{s}, t)$ が観測されたとし, $\boldsymbol{Y}(t) = (Y(\boldsymbol{s}_1, t), \ldots, Y(\boldsymbol{s}_n, t))'$ とおく. $\boldsymbol{Y}(t)$ が

$$\boldsymbol{Y}(t) = \sum_{k=0}^{p} \sum_{j=1}^{\lambda_k} \xi_{kj} W_{kj} \boldsymbol{Y}(t-k) - \sum_{l=0}^{q} \sum_{j=1}^{\mu_l} \phi_{lj} V_{lj} \boldsymbol{\epsilon}(t-l) + \boldsymbol{\epsilon}(t)$$

を満たすとき, $\{Y(\boldsymbol{s}, t))\}$ は時空間自己回帰移動平均モデル (Space-Time Autoregressive Moving Average model, STARMA モデル) に従うと言う. ここで W_{kj}, V_{lj} は $n \times n$ ウエイト行列であり, 地点間の隣接の度合いにより階層的に定義される. 例えば W_{k1} は 8.4 節で定義した隣接行列 W とする. 次に W_{k2} は W_{k1} で 1 を成分に持つ地域の隣接行列とする. このときもとの地域から見れば隣接する地域がさらに隣接する地域に対応する W_{k2} の成分が 1 となる. 8.4 節で説明したアメリカ合衆国の場合, Alabama 州に隣接する Mississippi 州を例に取ればこの州が隣接するのは Arkansas 州 (3), Louisiana 州 (16) である. したがって W_{k2} の Alabama に対応する行においては, Arkansas, Louisiana に対応する列が 1 の値を取る. 以下 $W_{kj} (j = 3, \ldots, \lambda_k)$ も同様に定義する.

また λ_k, μ_l は空間方向の自己回帰項および移動平均項の遅れ (lag) の次数を, p, q は時間方向の遅れの次数を各々意味する. $\boldsymbol{\epsilon}(t) = (\epsilon(\boldsymbol{s}_1, t), \ldots, \epsilon(\boldsymbol{s}_n, t))'$, $t = 1, \ldots, T$ は期待値 0 の独立同一分布に従う n 次元のホワイト・ノイズである. 推定すべきパラメータは ξ_{kj}, ϕ_{lj} などである (Cressie(1993)).

観測地点の個数 n が固定されていれば, STARMA モデルは時系列解析における多変量自己回帰移動平均モデルの特殊モデルである. この事実に基づいて Cliff and Ord(1975), Nu and Tiao(1995), Pfiefer and Deutsch(1980) などにより定常性が成立するための条件, パラメータの推定法, さらには実際データへの応用についても論じられている.

検定に関しては, Mónica et al.(2006) が一般の自己回帰モデル対時空間自己回帰モデルと言う仮説検定問題を論じている.

8.7 非定常モデル

現実のデータは多くの場合非定常性を示す. 期待値が地点あるいは時点に依存して変化するデータを解析する統計モデルの一つとしては, これまでに説明した回帰モデルがある. さらに本節では共分散関数が非定常性示すデータを解析するためのモデルをいくつか紹介する. ただしこれらのモデルも推測理論を展開するために, またデータの安定的な構造を解明するために, 内在的には何らかの定常性を仮定していることを注意しておく.

8.7.1 変形法 (Deformation Approach)

Sampson and Guttorp(1992), Perrin and Senoussi(2000) などによって提案された手法で, 地点 $\boldsymbol{s}$ を関数により変換した後, 8.3.2 節で説明した自己共分散関数に対するモデルを当てはめる. 式で表現すると $g : \boldsymbol{R}^d \to \boldsymbol{R}^d$ を非線形関数とし, $C(\boldsymbol{h})$ は定常確率場の自己共分散関数として

$$Cov(Y(\boldsymbol{s}), Y(\boldsymbol{t})) = C(g(\boldsymbol{s}) - g(\boldsymbol{t}))$$

によって自己共分散関数をモデル化する.

自己共分散関数は正定値関数になるので理論的な正当性は保証されている. 実際のデータを解析する場合には, 関数 g として例えばスプライン関数と呼ばれる区分的多項式関数を採用し, 接続点では微分係数が一致するように滑らかにつないでいく方法が提案されている.

8.7.2 たたみ込み法 (Convolution Approach)

まず「確率場の集合族」$\{X_{\theta(\boldsymbol{t})}(\boldsymbol{s}), \boldsymbol{s} \in D\}$ を用意する. 各 $\theta(\boldsymbol{t})$ に ひとつの確率場が対応する. このときカーネル関数 $K(\boldsymbol{t})$ と確率場 $\{X_{\theta(\boldsymbol{t})}(\boldsymbol{s}), \boldsymbol{s} \in D\}$

とのたたみ込み

$$Y(\boldsymbol{s}) = \int_D K(\boldsymbol{s}-\boldsymbol{t}) X_{\theta(\boldsymbol{t})}(\boldsymbol{s}) d\boldsymbol{t}$$

により,$\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ が定義される. 一般に $K(\boldsymbol{t})$ は $\boldsymbol{t} = \boldsymbol{0}$ で最大値を取り, $\boldsymbol{0}$ から遠ざかるにつれ減少するので, $\{X_{\theta(\boldsymbol{t})}(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ は, 地点 $\boldsymbol{t}$ が $\boldsymbol{s}$ に近いほど $Y(\boldsymbol{s})$ への影響が強くなる. 実際のデータ解析では積分を有限個の和で近似し,

$$Y(\boldsymbol{s}) = \sum_{j=1}^{k} K(\boldsymbol{s}-\boldsymbol{t_j}) X_{\theta(\boldsymbol{t_j})}(\boldsymbol{s})$$

とする. ここで確率場 $\{X_{\theta(\boldsymbol{t_j})}\}$ としては, $\boldsymbol{t_j}$ を中心とする D の小地域で適合度の高い定常確率場を選択する. 小地域およびその個数 k などの選択が重要になるが, Fuentes(2002), Fuentes et al.(2007) は情報量規準 AIC, BIC などを用いて選択することを提案している.

8.8　実際の応用例

本節では実際のデータ解析への応用例として 2 つ紹介する. 8.1 節で述べたように時空間統計解析が応用されている分野は多岐に渡り, 数多くの興味深い応用例がある. これらについては参考文献に所収の例を参考にされたい.

最初の例は Nu and Tiao(1995) による STARMA モデルを用いた大気中のオゾン量の解析である. オゾン量の減少は, それが吸収している紫外線の増加をもたらし, 皮膚ガンの発症など人類に深刻な影響を与えると指摘されている. データは 1978 年 11 月から 1990 年 5 月までの月平均のオゾン量のデータである. 緯度については南北ともに 0 度から 70 度までを 10 度刻みに 7 個, 経度については 10 度刻みに 36 個の地域に各々分割し, 合計で $504(= 14 \times 36)$ 地域, 各地域ごと 139 個の時系列データからなる.

彼らは緯度を固定して, 各緯度ごとに適合度の高い STARMA モデルの構築を試みている. いま $\boldsymbol{Z}(\boldsymbol{s}_j, t)(j = 1, \ldots, 36)$ を緯度を固定したときの経度 j 時点 t におけるオゾン量として, 以下の式

$$\boldsymbol{Z}(\boldsymbol{s}_j, t) = C(\boldsymbol{s}_j, t) + R(\boldsymbol{s}_j, t) + Y(\boldsymbol{s}_j, t)$$

によって表現する. ここで $C(\boldsymbol{s}_j, t)$, $R(\boldsymbol{s}_j, t)$ は各々季節成分, トレンド成分

を表し

$$C(\boldsymbol{s}_j,t)=\beta_{1j}+\beta_{2j}\sin(2\pi t/12)+\beta_{3j}\cos(2\pi t/12)$$
$$+\beta_{4j}\sin(4\pi t/12)+\beta_{5j}\cos(4\pi t/12)$$
$$R(\boldsymbol{s}_j,t)=\beta_{6j}\frac{t}{12}$$

によって定式化する. 季節成分 $C(\boldsymbol{s}_j,t)$ は定数項と 1 年および半年周期の波からなる. トレンド成分は 1 年あたりの増減を示すように 12 で割っている. そして $Y(\boldsymbol{s}_j,t)$ は STARMA の特殊形である

$$Y(\boldsymbol{s}_j,t)=[\alpha_1 Y(\boldsymbol{s}_{j-1},t)+\theta_1 Y(\boldsymbol{s}_{j+1},t)]$$
$$+[\alpha_2 Y(\boldsymbol{s}_{j-2},t)+\theta_2 Y(\boldsymbol{s}_{j+2},t)]+\phi Y(\boldsymbol{s}_j,t-1)+\epsilon(\boldsymbol{s}_j,t)$$

によって定式化する. MA 項はなく空間方向, 時間方向の遅れが各々 2, 1 なので STAR(2,1) と略記する. また $\alpha_2=\theta_2=0$ のときは STAR(1,1), 他の場合も同様に表現する. なお経度は 360 度回転すると元の位置に戻るので $Z(\boldsymbol{s}_j,t)=Z(\boldsymbol{s}_{j+36},t)$ を仮定する.　パラメータは条件付き最尤推定法により推定する (条件付き対数尤度関数の詳細については原論文を参照されたい).

まず表 8.1 には, 緯度を固定した 14 の地域の中から例として北緯 10 度〜20 度と南緯 60〜70 度に対するモデルの条件付き対数尤度関数を (-1) 倍した数値が掲載してある. Sym-STAR は $\alpha_1=\theta_1$ あるいは $\alpha_2=\theta_2$ を仮定した経度方向に対称なモデルを意味する. 他方 Gen-STAR はパラメータに制約を置かない一般モデルである. この数値を最小にするモデルを最良のモデルと見なせば, 北緯 10 度〜20 度に対しては Gen-STAR(1,1) が, 南緯 60〜70 度に対しては Sym-STAR(2,1) モデルが各々最良のモデルである.

表 8.1　条件付対数尤度関数

緯度	Sym-STAR(1,0)	Sym-STAR(1,1)	Gen-STAR(1,1)	Sym-STAR(2,1)
北緯 10〜20 度	8,020	7,643	7,603	7,643
南緯 60〜70 度	13,195	13,162	13,160	13,068

次に $\hat{\beta}_{1j},\ \hat{\beta}_{6j}(j=1,\ldots,36)$ を, 各緯度ごとに最良のモデルのもとで計算した経度 j における $\beta_{1j},\ \beta_{6j}$ の推定量とする. 表 8.2 には 10 年間のオゾンの変化率 $(\%)w_j=100\times 10\hat{\beta}_{6j}/\hat{\beta}_{1j}$ の特徴を示す指標を掲載してある. 各緯度ごとに 36 個の $w_j(j=1,\ldots,36)$ が得られるので, それらの最小値, 平均, 最

大値, 標準偏差である. 南極大陸上空のオゾンホールが 1980 年代以降深刻な問題になっているので, 南半球の数値を取りあげた. 確かにどの指標からも, 緯度が高くなるほど減少率が大きくなっていることが分かる.

表 8.2　オゾン量変化率の推定値

緯度	最小値	平均値	最大値	標準偏差
南緯　0〜10 度	-.02	.25	.72	.85
南緯 10〜20 度	-.77	-.33	-.02	.47
南緯 20〜30 度	-2.04	-1.52	-.93	.56
南緯 30〜40 度	-4.43	-2.94	-1.97	.62
南緯 40〜50 度	-5.56	-4.37	-3.26	1.07
南緯 50〜60 度	-7.24	-6.58	-5.60	1.57
南緯 60〜70 度	-11.33	-10.09	-8.91	2.48

次の例は松田・矢島 (2007) による図 8.1 に示した地点における公示地価の解析例である. データは国土交通省「2001 年度国土数値情報地価公示」より首都圏第一種・第二種住居専用地域から抽出した 5573 個の地価 (円/m^2) である. 標本地点の緯度・経度を km 単位でそれぞれ縦軸・横軸に示してある. モデルとしては (8.3) の回帰モデルを採用する. 説明変数としては山手線の主要ターミナル駅からの時間距離, 商業・業務用地 (m^2) などを用い, $\epsilon(\boldsymbol{s})$ が定常確率場に従うと仮定する (回帰モデルの詳細については松田 (2004) を参照されたい). ただし地点 $\boldsymbol{s}$ の動く範囲は $\boldsymbol{R}^2$ の部分集合とみなすので, $\epsilon(\boldsymbol{s})$ の自己共分散関数は, (8.1) において $T^d = \boldsymbol{R}^2$ を代入して

$$C_\epsilon(\boldsymbol{h}) = \int_{\boldsymbol{R}^2} \exp(i\boldsymbol{h}'\boldsymbol{\lambda}) df_\epsilon(\boldsymbol{\lambda})$$

になる. $f_\epsilon(\boldsymbol{\lambda})$ には Vecchia(1988) によって提案された

$$f_\epsilon(\lambda_1, \lambda_2) = \sigma^2(\kappa^2 + \phi)^{-2}$$

を採用した. ここで

$$\kappa^2 = [\beta^{-1}(\lambda_1 \cos\alpha - \lambda_2 \sin\alpha)]^2 + [\beta(\lambda_1 \sin\alpha + \lambda_2 \cos\alpha)]^2$$

とし, 母数ベクトルは $\theta = (\sigma^2, \alpha, \beta, \phi)$ である.

このスペクトル密度関数の含意は以下の通りである．座標軸 (λ_1, λ_2) をまず α 回転させ，　新たな座標軸を

$$\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

とおく．このとき (ω_1, ω_2) 平面においてスペクトル密度関数の等高線は楕円形

$$\kappa^2 = \beta^{-2}\omega_1^2 + \beta^2\omega_2^2.$$

によって与えられる．$\alpha = 0,\ \beta = 1$ のときは $f_\epsilon(\lambda_1, \lambda_2)$ は $\lambda_1^2 + \lambda_2^2$ のみに依存するので, 8.3.2 節で説明した等方型モデルになる．

ただし θ の推定は, 地点が不等間隔に並んでいるので, 8.3.2 節で説明した Whittle 推定法を直接には応用できない．松田・矢島 (2007) では地点の配置パターンをモデル化することにより, 不等間隔データに対しても適用できるように Whittle 推定法を一般化した．この推定法を用いると, 推定量は制約条件 $\alpha = 0,\ \beta = 1$ もとでは $\hat{\sigma}^2 = 0.205,\ \hat{\phi} = 0.590$, 無制約のもとでは $\hat{\sigma}^2 = 0.208,\ \hat{\phi} = 0.596,\ \hat{\alpha} = -0.77,\ \hat{\beta} = 1.04$ となる．対数尤度比統計量は 0.703 となり χ^2 検定を行うと有意水準 5% では等方性を棄却できない．しかし推定された密度関数とその等高線を示した図 8.3 を見ると, 東西方向 (横軸) の方が南北方向 (縦軸) より 0 に早く収束している．すなわち高周波の成分が早く小さくなる．この現象を時間領域から見れば, 東西方向の方が南北方向より相関関数がより永続していることになる．一つの解釈として, 首都圏では東西方向の方が南北方向より鉄道網が普及しているため相関関数が急激には減衰しないことが考えられる．

8.9　今後の課題（さらなる発展に向けて）

冒頭に述べたように時空間統計解析は既に確立された分野ではない．今後の発展に向けて, 解決すべき課題をいくつか列挙し本章を締めくくる．

8.9.1　同定法の開発

時系列解析における Box-Jenkins 法のようなスタンダードな解析手順は,

まだ確立されていない．Pfiefer and Deutsch(1980) は STARMA モデルに Box-Jenkins 法の一般化を試みているが, 次数が高くパラメータ数の多い複雑なモデルに応用することは難しい．その主要な原因は前述のように時空間データに自然な順序を導入することが困難であり, したがって時系列解析にお

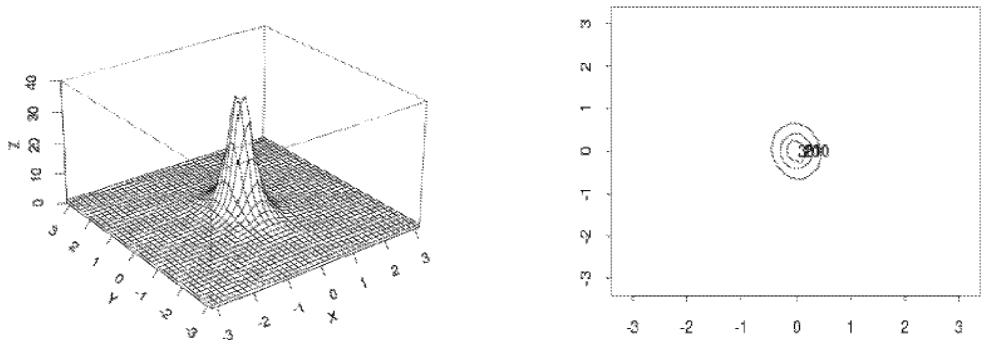


図 8.3　スペクトル密度関数とその等高線

ける標本 (偏) 自己相関関数のようなモデル同定において有用な統計量がまだ開発されていないことにある. 時空間データのモデル同定に有効な統計量の開発, またモデル選択規準の有効性の検証などが今後の課題として残されている.

8.9.2　非等方型モデル・非乗法型モデルの開発

データ解析における扱い易さもあって, 従来の時空間統計解析においては等方モデル, 乗法型モデルが実際のデータを解析する際には応用される場合が多かった.

しかし方向に依存する要因, 例えば交通網などの社会的インフラストラクチャー, 地勢・地形などの自然的条件, 風向などの気象条件が影響するデータに対しては, 等方性の仮定は妥当でないと考えられる.　このようなデータを解析するときの簡便な方法は, 8.7.1 節の変形法と類似の発想に基づき, 地点 $\boldsymbol{s}$ を線形変換し $\boldsymbol{s}^* = A\boldsymbol{s}$ とおき,$\boldsymbol{s}^*$ に等方型モデルを当てはめる方法である. 例えば $d = 2$ のとき

$$A = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} p & 0 \\ 0 & q \end{bmatrix}$$

とおけば, もとの座標軸における一つの軸の単位を p 倍, 他方の軸を q 倍したのち, α だけ回転させることに対応する. 実際の応用例としては前述の松田・矢島 (2007) がある.

次に天候の変化などのように空間相関が時間的遅れを伴って変化するようなデータの解析には, 非乗法型モデルが必要となる. そこで Cressie and

Huang(1999), Fuentes et al.(2007), Gneiting(2002) が非乗法型共分散関数の構成法を提案している．例えば Fuentes et al.(2007) は $d = 2$ のとき, $f(\lambda_1, \lambda_2)$ に対する統計モデルとして

$$f(\lambda_1, \lambda_2) = \gamma(\alpha^2\beta^2 + \beta^2|\lambda_1|^2 + \alpha^2|\lambda_2|^2 + \epsilon|\lambda_1|^2|\lambda_2|^2)^{-\nu}$$

を提案している．ϵ の大きさが非分離度を規定しており, $\epsilon = 1$ のときには

$$f(\lambda_1, \lambda_2) = \gamma(\alpha^2 + |\lambda_1|^2)^{-\nu}(\beta^2 + |\lambda_2|^2)^{-\nu}$$

となるので, 分離型モデルである．また $\epsilon = 0$ のときは

$$f(\lambda_1, \lambda_2) = \gamma(\alpha^2\beta^2 + \beta^2|\lambda_1|^2 + \alpha^2|\lambda_2|^2)^{-\nu}$$

となり, 3 節で説明した Matérn 族の一般化になる．さらに $\alpha = \beta$ のときは等方型モデルになる．

しかしこれらの非等方モデル, 非分離型モデルが実際のデータに対して持つ含意や汎用性はまだ明らかではなく, また他の代替的なモデルを開発することも可能であり, 今後解明すべき点が多く残っている．

8.9.3　ノンパラメトリックおよびセミパラメトリック・モデル

冒頭に述べたように, 近年の科学技術の発展や統計解析ソフトウエアの普及により, 大量のデータの採取および解析が可能になっている．このような状況においては, 特定化の誤りを回避できるとともに, 非線形性など変数間の複雑な関係も検出できるノンパラメトリック・モデルあるいはセミパラメトリック・モデルの理論的性質および実際のデータ解析への応用可能性なども重要なトピックである．例えば Arbia(2006), Gao et al.(2006), Hallin et al.(2004) により議論され始めているが, これらの議論をさらに発展させる必要があろう．

8.9.4　階層的ベイズモデル

時空間データの複雑な変動メカニズムを忠実にモデル化しようとすると, パラメータ数が膨大になる場合がある．この問題に対する一つの解決策は前節で述べたノンパラメトリック・モデルあるいはセミパラメトリック・モデルを構築することであるが, 別の方向の解決策としてはパラメータに事前分布を導入したベイズモデルを構築することである．このモデルのなかで特に事前分布を階層化したモデルをベイズ型階層モデルと言う (Banerjee et al.(2004), Wikle et al.(1998), Wikle et al.(2001))．従来からベイズモデルの欠点とし

て事後分布の導出に多大な計算量を要することが指摘されてきた. しかし近年になってマルコフ連鎖モンテカルロ法 (Markov Chain Monte Carlo 法, MCMC 法) と呼ばれる効率的な計算方法が考案され, その欠点も解消されつつある.　時空間データに対する階層的ベイズモデルおよび MCMC 法の有効性を解明することも今後の課題である. MCMC 法については本シリーズ所収の古澄 (2008) を参照されたい.

8.9.5　不等間隔データの解析

時系列データの解析においては, データが等間隔で観測されると言う仮定の下で議論されることが多い. 実際週次, 月次, 四半期, 年次ごとに公表されるデータは数多く存在し, 一定の現実性を持っている.　しかし時空間データ特に地点参照データでは地勢的な制約からこの仮定が現実性を欠く場合が多くある. 例えば地価などは河川, 湖沼がある地点では存在しない.　したがって不等間隔をどのようにモデル化するかまたそのモデルの下でどのようにパラメータを推定するかなどの問題を解決する必要がある. このような方向の論文としては, 例えば Fuentes(2007) および前述の松田・矢島 (2007) が格子データに対する Whittle 推定量を地点参照データの推定量に一般化し, 各々海面気温, 公示地価の解析に応用している.

8.9.6　時空間データに対する単位根検定・共和分分析

ここ 30 年ほど時系列計量経済学において, 理論, 応用ともに精力的に研究され, 著しい発展を遂げた分野として, 単位根検定 (unit root test) と共和分分析 (Cointegration analysis) がある. これらの分析手法を時空間相関を持つデータにどのように拡張すべきかについては, 最近 Baltagi et al.(2007), Mur and Trívez(2003), Paulauskas(2007) などによって展開され始めている.

8.10　付論

8.10.1　定常過程・ARMA モデル

本節では定常過程および定常過程に対する代表的なモデルである ARMA モデルについて説明する. 詳しくは本シリーズ所収の田中 (2008) を参照されたい.

まずこれらの過程やモデルの基礎となる確率過程について簡単に説明する.

T をパラメータの集合とし, それに含まれる要素を t とする. 任意の t に対してある確率変数 X_t が対応しているとする. このとき t を T 上でくまなく動かしたときにできる確率変数の集合 $\{X_t : t \in T\}$ を確率過程 (stochastic process) と言う. 時間の推移とともに変動する確率過程を考える場合には, T を $\boldsymbol{R}$, $\boldsymbol{Z}$ あるいは各々の部分集合とする. また本論で説明した時空間データを記述する確率過程を考える場合には, d を 2 以上の自然数として T を $\boldsymbol{R}^d$, $\boldsymbol{Z}^d$ あるいは各々の部分集合とする.

定常過程 (stationary process) は, 端的に言うと時間軸の平行移動に対して確率法則が不変な確率過程である. 具体的にどのような不変性を要請するかで 2 つの定義がある. 同時分布関数に不変性を課した定常過程を強定常過程 (strongly stationary process) あるいは狭義定常過程 (strictly stationary process) と言う. 他方 2 次までのモーメントに不変性を課した定常過程を弱定常過程 (weakly stationary process) あるいは共分散定常過程 (covariance stationary process) と言う. 以下では本論で必要になる弱定常過程の数学的な定義を与える.

$T = \boldsymbol{R}$ あるいは $T = \boldsymbol{Z}$ とする. 確率過程 $\{X_t : t \in T\}$ が次の 3 つの条件を満たすとき, 弱定常過程に従うと言う.

(i) $E|X_t|^2 < \infty$

(ii) 期待値は t に依存せず一定で $E(X_t) = \mu$ となる.

(iii) 任意の s, t に対して X_s と X_t の共分散は時間差 $s - t$ のみに依存する. したがって任意の l に対して

$$Cov(X_s, X_t) = Cov(X_{s+l}, X_{t+l})$$

が成立する.

このとき

$$C(h) = Cov(X_{t+h}, X_t) = Cov(X_h, X_0)$$

を遅れ (lag)h の自己共分散 (autocovariance) と言う. また h の関数と見なしたときは, 自己共分散関数 (autocovariance function) と言う. $C(0)$ は X_t の分散であるから, 期待値と同様に分散も t に依存せず一定の値である.

いま $T = \boldsymbol{Z}$ とする. このとき自己共分散関数 $C(h)$ は, 右連続単調非減少な関数 $F(\lambda)$ を用いてフーリエ表現

$$C(h) = \int_{(-\pi,\pi]} \exp(ih\lambda) dF(\lambda)$$

が可能である. $F(\lambda)$ が絶対連続な関数のときは, 非負の値を取る関数 $f(\lambda)$ が

存在して, $C(h)$ は

$$C(h) = \int_{(-\pi,\pi]} \exp(ih\lambda) f(\lambda) d\lambda$$

と表現できる. これを自己共分散関数のスペクトル表現と言い, $F(\lambda)$, $f(\lambda)$ を各々スペクトル分布関数 (spectral distribution function), スペクトル密度関数 (spectral density function) と呼ぶ. $dF(\lambda)$ あるいは $f(\lambda)$ は, 周波数 λ の波の平均的な強さを表す. $T = \boldsymbol{R}$ のときには, 積分範囲を $(-\pi, \pi]$ から $(-\infty, \infty)$ に変更すればよい.

$T = \boldsymbol{Z}$ のときには, 弱定常過程に対する代表的なモデルとして自己回帰移動平均モデル (Autoregressive Moving Average model, ARMA モデル) がある. 簡単のため期待値は $\mu = 0$ と仮定する. 弱定常過程 $\{X_t\}$ が

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots \theta_p \epsilon_{t-q}$$

を満たすとき, $\{X_t\}$ は次数 (p, q) の ARMA モデルに従うと言い, ARMA(p, q) と書く. ここで $\{\epsilon_t\}$ は期待値 0 分散 σ^2 の互いに無相関な確率変数列であり, 白色雑音 (White Noise) と呼ばれる.

特に $q = 0$ のときすなわち $\{X_t\}$ が

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t$$

を満たすとき自己回帰モデル (Autoregressive model, AR モデル) と呼び, AR(p) と書く. また $p = 0$ のときすなわち $\{X_t\}$ が

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots \theta_p \epsilon_{t-q}$$

を満たすとき移動平均モデル (Moving Average model, MA モデル) と呼び, MA(q) と書く.

ARMA(p, q) が弱定常過程になるための必要十分条件は, 複素数 z に関する p 次多項式 $\phi(Z)$ を

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

によって定義したとき, 方程式 $\phi(z) = 0$ の根が単位円周上に存在しないことである. すなわち $\phi(z) \neq 0$, $|z| = 1$ を満たすことである.

ARMA(p, q) のスペクトル密度関数は

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{i\lambda})|^2}{|\phi(e^{i\lambda})|^2}$$

によって与えられる有理関数である.

8.10.2 正規過程に対する最尤推定法

$T = \boldsymbol{R}^d$ あるいは $T = \boldsymbol{Z}^d$ とする. 確率過程 $\{X_t : t \in T\}$ から任意の有限個の時点の確率変数を取り出したとき, その同時分布が多変量正規分布に従うとき, $\{X_t : t \in T\}$ は正規過程に従うと言う.

したがって正規過程に対する最尤推定法は, 多変量正規分布に対する最尤推定法に帰着する. 観測値 $X_{t_i}(i = 1, \ldots, n)$ が与えられているとし, $\boldsymbol{X}_n = (X_{t_1}, \ldots, X_{t_n})'$ とおく. $\boldsymbol{X}_n$ が多変量正規分布 $N(\boldsymbol{\mu}_n, \boldsymbol{V}_n)$ に従うとき, その同時確率密度関数は

$$f(\boldsymbol{X}_n; \boldsymbol{\mu}_n, V_n) = \frac{1}{(2\pi)^{n/2}\det(V_n)} \exp\left(-\frac{(\boldsymbol{X}_n - \boldsymbol{\mu}_n)' V_n^{-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_n)}{2}\right)$$

となる. ここで $\boldsymbol{X}_n$ 固定し, $f(\boldsymbol{X}_n; \boldsymbol{\mu}_n, V_n)$ を $\boldsymbol{\mu}_n, V_n$ の関数と見なしたとき $f(\boldsymbol{X}_n; \boldsymbol{\mu}_n, V_n)$ を尤度関数と呼ぶ. また尤度関数を最大にする $\boldsymbol{\mu}_n, V_n$ を最尤推定量と言う. 多くの場合尤度関数よりも対数変換した対数尤度関数を最大にする方が計算が簡単になる. 多変量正規分布に対する対数尤度関数は, 定数部分を除去すると

$$L_n(\boldsymbol{\mu}_n, V_n; \boldsymbol{X}_n) = -\frac{1}{2}\log\det(V_n) - \frac{1}{2}(\boldsymbol{X}_n - \boldsymbol{\mu}_n)' V_n^{-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_n)$$

になる.

本論 8.3.2 節の対数尤度関数 $L_n(\phi; \boldsymbol{Y}_n)$ においては, $\boldsymbol{Y}_n$ が $\boldsymbol{X}_n$ に $G_n\beta$ が $\boldsymbol{\mu}_n$ に各々対応する.

参考文献

Anselin, L., Florax, R.J.G.M. and Rey,S.J.(2004). *Advances in Spatial Econometrics:Methodology. Tools and Applications.* Springer.

Arbia,G.(2006). *Spatial Econometrics:Statistical Foundations and Applications to Regional Convergence*. Springer.

Baltagi, B.H., Bresson, G. and Pirotte, A.(2007).Panel unit root tests and spatial dependence. *J.Applied Econometrics* **22** 339-360.

Banerjee,S., Carlin, B.P. and Gelfand,A.E.(2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Basu, S. and Reinsel,G.C.(1994).Regression models with spatially correlated errors. *J.Amer. Statist. Assoc.* **89**, 88-99.

Besag,J.(1974).Spatial interaction and the statistical analysis of lattice systems(with discussion). *J.Roy. Statist. Soc. Ser. B* **36** 192-236.

Cliff,A.D. and Ord, J.K.(1975).Space-time modeling with an application to regional forecasting. *Trans. the Inst. of British Geographers* **66** 119-128.

Cressie, N.(1993).*Statistics for Spatial Data*. rev. ed. Wiley.

Cressie, N. and Huang, H.-C.(1999).Classes of nonseparable, spatiotemporal stationary covariance functions. *J.Amer.Statist. Assoc.* **94** 1330-1340.

Curriero, F.C. and Lele, S.(1999).A composite likelihood approach to semivariogram estimation. *J. Agricultural, Biological, and Environmental Statist.* **4** 9-28.

Dahlhaus, R. and Künsch, H.(1987).Edge effects and efficient estimation for stationary random fields. *Biometrika* **74**, 877-882.

Finkenstädt, B., Held, L. and Isham, V.(2007).*Statistical Methods for Spatio-Temporal Systems*. Chapman and Hall/CRC.

Fuentes, M.(2002).Spectral methods for nonstationary spatial processes, *Biometrika*. **89**. 197-210.

Fuentes, M.(2006).Testing for separability of spatial-temoral covariance functions. *J. Statist. Planning and Inference*. **136**, 447-466.

Fuentes,M.(2007).Approximate likelihood for large irregularly spaced spatial data. *J.Amer.Statist.Assoc.* **102** 321-331.

Fuentes,M., Guttorp,P. and Sampson, P.D.(2007).Using transforms to analyze space-time process. *Statistical Methods for Spatio-Temporal Systems.* Finkenstädt et al. eds. Chapman and Hall/CRC.

Gao,J., Liu,Z. and Tjøstheim,D.(2006).Estimation in semiparametric spatial regression. *Ann.Statist.* **34** 1395-1435.

Getis, A., Mur, J. and Zoller, H.G.(2004). *Spatial Econometrics and Spatial Statistics*. Palgrave Macmillan.

Gneiting,T.(2002).Nonseparable, stationary covariance functions for space-time data. *J.Amer.Statist.Assoc.* **97**, 590-600.

Guttorp, P. and Gneiting,T.(2006).Studies in the history of probability and statistics XLIX. On the Matérn correlation family. *Biometrika* **93** 989-995.

Guyon,X.(1995).*Random Fields on a Network;Modeling, Statistics, and Applications*. Springer.

Haining,R.(1990).*Spatial Data Analysis in the Social and Environmental*

Sciences. Cambridge University Press.

Haining,R.(2003).*Spatial Data Analysis:Theory and Practice.* Cambridge University Press.

Hallin, M., lU, Z. and Tran, L.T.(2004).Local linear spatial regression, *Ann.Statist.* **32** 2469-2500.

Kriege,D.G.(1951).A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. Chemical Metallurgical and Mining Society of South Africa* **52** 113-139.

Lesage, J.P. and Pace,R.K.(2004).*Spatial and Spatiotemoral Econometrics* Advanced in Econometrics vol.18 Elsvier.

Ludeña C. and Lavielle, M.(1999).The Whittle estimator for strongly dependent stationary Gaussian fields. *Scand. J. Statist.* **26** 433-450.

Mardia, K.V. and Marshall,R.J.(1984).Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135-146.

Matérn, B.(1960). *Spatial Variation-Stochastic Models and Their Application to some Problems in Forest Surveys and other Sampling Investigations.* Stockholm: Medd. Statens Skogsforskningsinstitut. **49** no.5.

Matérn, B.(1986). *Spatial Variation.* 2nd ed. Springer.

Matheron,G.(1963). Principles of geostatistics. *Economic Geology* **58** 1246-1266.

Mitchell, M.W., Genton, M.G. and Gumpertz,M.(2005).Testing for separability of space-time covariances. *Environmetrics* **16** 819-831.

Mur,J. and Trívez, F.J.(2003).Unit roots and deterministic trends in spatial econometrics. *International Regional Science Review* **26** 289-312.

Mónica, A., Antunes,C. and Rao,T.S.(2006).On hypotheses testing for the selection of spatio-temporal models. *J.Time Ser.Anal.* **27** 765-791.

Niu, X. and Tiao,G.C.(1995).Modeling satellite ozon data. *J.Amer.Statist.Assoc.* **90** 969-983.

Paelinck, J.H. and Klaassen,L.H.(1979). *Spatial Econometrics.* Gower.

Paulauskas, V.(2007).On unit root tests for spatial autoregressive models. *J. Mult.Anal.* **98** 209-226.

Perrin, O. and Senoussi, R.(2000).Reducing non-stationary random fields to stationarity and isotropy usinf a space deformation. *Statist. and Probab. Letters* **48** 23-32.

Pfeifer, P.E. and Deutsch, S.J(1980).A three-stage iterative procedure for space-time modeling. *Technometrics* **22** 35-47.

Rue, H. and Held, L.(2005).*Gaussian Markov Random Fields:Theory and Applications*, Chapman and Hall/CRC Boca Raton.

Sampson, P.D. and Guttorp, P.(1992).Nonparametric estimation of nonstationary spatial covariance structure. *J.Amer.Statist.Assoc.* **87** 108-119.

Sherman, M., Guan,Y. and Calvin, J.A.(2003).Assesing spatial isotropy. *Recent Advances and Trends in Nonparametric Statistics* Akritas,M.G. and Politis, D.N.(eds) Elsvier.

Stein,M.L.(1999).*Interpolation of Spatial Data:Some Theory for Kriging*. Springer.

Stein,M.L., Chi, Z., and Welty, L.J.(2004).Approximating likelihoods for large spatial data sets. *J.R.Statist.Soc.B* **66** 275-296.

Vecchia, A.V.(1988).Estimation and model identification for continuous spatial processes. *J.R.Statisti.Soc.B* **50** 297-312.

Wall, M.M.(2004).A close look at the spatial structure implied by the CAR and SAR models. *J.Statist. Planning and Inference* **121** 311-324.

Whittle,P.(1954).On stationary processes in the plane. *Biometrika* **41** 434-449.

Wikle, C.K., Berliner, L.M. and Cressie,N.(1998).Hierachical Bayesian space-time models. *Environmental and Ecological Statist.* **5** 117-154.

Wikle, C.K. and Cressie,N.(1999).A dimension-reduced approach to space-time Kalman Filtering. *Biometrika* **86** 815-829.

Wikle, C.K., Miliff,R.F., Nychka,D. and Berliner,L.(2001).Spatiotemporal hierarchical Bayesian modeling:tropica ocean surface winds. *J.Amer.Statist.Assoc.* **96** 382-397. Yaglom, A.M.(1987).*Correlation Theory of Stationary and Related Random Functions.* Vol.1, Springer.

尾形良彦 (2008). 地震活動予測の統計科学. 21 世紀の統計科学第 2 巻所収. 東京大学出版会

古澄英男 (2008). マルコフ連鎖モンテカルロ法. 21 世紀の統計科学第 3 巻所収. 東京大学出版会

清水邦夫 (2002). 地球環境データー衛星リモートセンシング. データサイエンス・シリーズ 8 共立出版.

田中勝人 (2008). 時系列解析の展開. 21 世紀の統計科学第 3 巻所収. 東京大学出版会

丹後俊郎・横山徹爾・高橋邦彦 (2007). 空間疫学への招待. 疾病地図と疾病集積性を中心として. 医学統計シリーズ 7 朝倉書店

間瀬茂・武田純 (2003). 空間データモデリングー空間統計学の応用. データサイエンス・シリーズ 7 共立出版.

松田安昌 (2004). 非線形回帰モデルによるヘドニック・アプローチ. 住宅土地経済 **52** 29-35.

松田安昌・矢島美寛 (2007). 不等間隔時空間データに対するフーリエ解析. 応用統計学 **36** 1-14.

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2011 年 10 月

第9章　正定値カーネルによる統計的推論の方法

福水健次[1]

(情報・システム研究機構 統計数理研究所 教授)

高次元のデータや複雑な構造を持ったデータの解析を得意とする方法として，正定値カーネルないしは再生核ヒルベルト空間を用いたデータ解析の方法が，近年急速に発展を遂げている．本章では特に均一性検定や独立性検定といった多変量解析と関連する話題に関し，最近の研究までを含めて解説する．この分野は計算機科学分野の研究者が主体となって発展させてきたが，本章で論じるように統計学の問題と密接に関連しており，今後の興味ある展開が期待される．

1　はじめに

統計的学習理論ないしは機械学習の分野では，1990 年代半ばごろから，サポートベクターマシン（SVM）(Boser et al., 1992, Vapnik, 1998) と呼ばれ

[1]fukumizu@ism.ac.jp

る識別の方法が注目を集め，文字認識などさまざまな応用において高い能力を発揮することが示された．SVM は,（1）ベイズ最適を捨ててマージン最大化[2]という規準を用いることにより，数値解が容易に得られる 2 次計画問題による定式化を行った（2）正定値カーネルによってデータを非線形変換することにより，必要な計算を複雑にせずにデータの高次モーメントを取り込めた，という 2 つの点において新しい方法論であった．前者は，線形で解けるクラスを超えて，2 次計画や半正定値計画などの実用的な凸計画法をデータ解析に応用する研究を盛んにし，後者は，正定値カーネルないしは再生核ヒルベルト空間を用いたデータ解析の方法論である「カーネル法」[3]を生み出した (Schölkopf and Smola, 2002)．

最近になって，再生核ヒルベルト空間上の平均として分布を埋め込むことによって，変数間の任意の高次統計量を考慮し，独立性，条件付独立性，分布の同一性など，確率変数の性質を推定するための新しい方法論が提案されてきた (Bach and Jordan, 2002, Fukumizu et al., 2004, Gretton et al., 2007)．これらの方法は，確率分布の性質に関する一種のノンパラメトリック推定法と考えることができるが，再生核ヒルベルト空間を用いることにより，ガウス分布ないしは線形の関係に基づく方法を，計算をあまり複雑にすることなく，自然な形で一般の分布に拡張することが可能となる．

本稿は，カーネル法の一般的な解説をごく簡単に行った後，正定値カーネルによる確率分布の推論に関する最近の研究について解説する．

2　カーネル法の概要

カーネル法を一言で述べると，データの（非線形）変換を行う方法論である．データに何らかの変換を施した後に解析を行う手法は古くから存在するが，カーネル法の特徴は，特殊な内積を持つ関数空間（再生核ヒルベルト空間）への変換を用いることにより，変換後のデータに対する線形の処理が効率的計算によって実行可能な点にある．

本稿ではカーネル法の一般的な方法論に関しては簡単に触れるだけなの

[2] 2 クラス識別問題において，データが線形識別可能なとき，各クラスの最近データへの距離を最大にする識別平面を最適とする規準．本稿では SVM は議論しないので，興味ある読者は赤穂 (2008) や福水 (2010) などを見ていただきたい．

[3]「カーネル」という用語は，統計学においては，ノンパラメトリック推定におけるカーネル密度関数など，必ずしも正定値とは限らないカーネル関数を指すことが多いが，機械学習分野では正定値カーネルを用いるデータ解析の方法論を「カーネル法」と呼んでいるため，本稿でもこれにならった．

で，より詳しく知りたい読者は，例えばSchölkopf and Smola (2002)，赤穂 (2008)，福水 (2010) といった教科書を見ていただきたい．

2.1 正定値カーネルと再生核ヒルベルト空間

データの変換に用いる空間を導入するために，正定値カーネルとそれが定める再生核ヒルベルト空間についてまとめておく[4]．なお，以下では実数値の場合のみ説明する．

集合 Ω に対し，$k : \Omega \times \Omega \to \mathbb{R}$ が Ω 上の**正定値カーネル**であるとは，対称性 $k(x, y) = k(y, x)$ を満たし，かつ任意の n 個の点 $x_1, \ldots, x_n \in \Omega$ と 実数 $c_1, \ldots, c_n$ に対し，

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \tag{2.1}$$

が成り立つことをいう．すなわち，**グラム行列**と呼ばれる $n \times n$ 対称行列 $\big(k(x_i, x_j)\big)$ が常に半正定値となる対称な関数として定義される．

Ω 上の正定値カーネル k に対し，Ω 上の実関数からなる（実）ヒルベルト空間 $\mathcal{H}$ が存在し，以下の 2 つの性質を満たす．

(i) 任意の $x \in \Omega$ に対して $k(\cdot, x) \in \mathcal{H}$ であり，$\{k(\cdot, x) \in \mathcal{H} \mid x \in \Omega\}$ の張る線形空間は $\mathcal{H}$ で稠密である．ここで $k(\cdot, x)$ は，x を固定した第 1 変数に関する関数を表す

(ii) 任意の $f \in \mathcal{H}$ と $x \in \Omega$ に対し，再生性

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \tag{2.2}$$

が成り立つ．ここで $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ は $\mathcal{H}$ の内積を表す．

このようなヒルベルト空間のことを（k が定める）**再生核ヒルベルト空間**といい，再生核ヒルベルト空間と対応する正定値カーネルの組を $(\mathcal{H}, k)$ であらわす．(ii) の再生性は再生核ヒルベルト空間をデータ解析に応用する上で最も重要な性質であり，ヒルベルト空間内での内積計算を容易にする．例えば $f = \sum_{i=1}^{n} a_i k(\cdot, x_i)$ と $g = \sum_{j=1}^{m} b_j k(\cdot, y_j)$ という 2 つの $\mathcal{H}$ の要素の内積は

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j k(x_i, y_j)$$

[4] より詳しくは Aronszajn (1950) や 福水 (2010) を参照．

で与えられ，k の値の評価に還元される．これは，内積計算に積分を必要とする２乗可積分関数のなす関数空間などと大きく異なる点である．

k_n $(n = 1, 2, ...)$ を Ω 上の正定値カーネルとするとき，以下で定義される関数がまた正定値カーネルとなることは，比較的容易に示される．(i) 非負結合 $c_1k_1 + c_2k_2$ （$c_1, c_2 \geq 0$）, (ii) 積 k_1k_2, (iii) 各点収束先 $k(x_1, x_2) = \lim_{n\to\infty} k_n(x_1, x_2)$ （各点収束を仮定する）.

$\mathcal{X}$ 上の再生核ヒルベルト空間 $(\mathcal{H}_1, k_1), (\mathcal{H}_2, k_2)$ に対し，k_1+k_2 により定められる再生核ヒルベルト空間は，ベクトル空間として $f+g$（$f \in \mathcal{H}_1, g \in \mathcal{H}_2$）の形の関数（$f + g$ は関数値の和で定義する）からなることが知られている．これを $\mathcal{H}_1$ と $\mathcal{H}_2$ の直和といい，$\mathcal{H}_1 + \mathcal{H}_2$ で表す．また，$(\mathcal{H}_1, k_1), (\mathcal{H}_2, k_2)$ をそれぞれ $\mathcal{X}, \mathcal{Y}$ 上の再生核ヒルベルト空間とするとき，積 k_1k_2 の定める $\mathcal{X} \times \mathcal{Y}$ 上の再生核ヒルベルト空間はベクトル空間としての直積 $\mathcal{H}_1 \otimes \mathcal{H}_2$ と一致し，$\sum_{i=1}^{n} f_i(x)g_i(y)$ （$f_i \in \mathcal{H}_1, g_i \in \mathcal{H}_2$）の形の関数集合は $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$ で稠密である．

ユークリッド空間 $\mathbb{R}^m$ 上の正定値カーネルの代表的な例は，通常の内積 $k(x_1, x_2) = x_1^T x_2$ のほかに，多項式カーネル

$$k_{d,c}^{poly}(x_1, x_2) = (x_1^T x_2 + c)^d$$

（$c \geq 0$， $d \in \mathbb{N}$）や，ガウス RBF(Radial Basis Function) カーネル

$$k_\sigma^G(x_1, x_2) = \exp\Big(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\Big)$$

（$\sigma > 0$）などである．これらが正定値であることは，上で述べた３つの性質を用いると比較的容易に証明できる．また，再生核ヒルベルト空間の性質 (i) から，多項式カーネル $k_{d,c}^{poly}$ （$c > 0$）の定める再生核ヒルベルト空間は，ベクトル空間として d 次以下の多項式全体と一致することがわかる．ただし内積は $k_{d,c}^{poly}$ により与えられる．ガウス RBF カーネルが定める再生核ヒルベルト空間がベクトル空間としてどのようなものかは，正定値カーネルの豊富さと関連して，3.2 節で触れる．

2.2 正定値カーネルによるデータ解析の方法論

正定値カーネルおよび再生核ヒルベルト空間をデータ解析に用いる方法について述べる．データ $x_1, \ldots, x_n$ を集合 Ω の点とする．これに対して Ω

上の正定値カーネル k とそれが定める再生核ヒルベルト空間 $\mathcal{H}$ を用意し，変換

$$\Phi : \Omega \to \mathcal{H}, \qquad x \mapsto k(\cdot, x) \tag{2.3}$$

によって，関数データ $\{\Phi(x_i)\}_{i=1}^n = \{k(\cdot, x_i)\}_{i=1}^n$ を作成する．例えば，ガウス RBF カーネルを用いると，

$$\Phi(x_1) = e^{-\frac{1}{2\sigma^2}\|x-x_1\|^2}, \ldots, \Phi(x_n) = e^{-\frac{1}{2\sigma^2}\|x-x_n\|^2}$$

という関数データを得る．カーネル法の方法論の核心は，ユークリッド空間のベクトルデータに対して適用可能な手法を，関数データ $\{\Phi(x_i)\}_{i=1}^n$ に拡張するというものである．この方法論は線形手法のカーネル化と呼ばれ，主成分分析，フィッシャー (Fisher) 判別分析，正準相関分析など様々な手法のカーネル化がなされている．SVM も，マージン最大化を規準とする線形識別器のカーネル化として定義される (Schölkopf and Smola, 2002).

カーネル法の例：カーネル主成分分析

典型的な例として，カーネル主成分分析 (Kernel PCA, Schölkopf et al., 1998) を間単に紹介する．集合 Ω 内のデータ $x_1, \ldots, x_n$ に対し，Ω 上に正定値カーネル k を用意する．式 (2.3) の変換によって得られたデータ $\Phi(x_1), \ldots, \Phi(x_n)$ に対して，k で定まる再生核ヒルベルト空間 $\mathcal{H}$ 上で主成分分析を行うと，その第 1 主軸は以下のように与えられる．

$$\arg \max_{f \in \mathcal{H}, \|f\|_{\mathcal{H}}=1} \sum_{i=1}^n \langle f, \Phi(x_i) - \widehat{m} \rangle_{\mathcal{H}}^2 \tag{2.4}$$

ここで，$\widehat{m} \in \mathcal{H}$ は $\{\Phi(x_i)\}_{i=1}^n$ の標本平均

$$\widehat{m} = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) = \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i) \tag{2.5}$$

であり，$\|f\|_{\mathcal{H}}$ は $\mathcal{H}$ のノルムを表す.

関数 $f \in \mathcal{H}$ は一般に無限次元空間のベクトルであるが, $\{\Phi(x_1) - \widehat{m}, \ldots, \Phi(x_n) - \widehat{m}\}$ の張る高々 $n-1$ 次元部分空間の成分 f_0 とその直交補空間の成分 $f_\perp$ に分解して $f = f_0 + f_\perp$ と表示したとき，(2.4) 式の右辺が $f_\perp$ に依存しないことを用いると，結局

$$f = \sum_{j=1}^n a_j (\Phi(x_j) - \widehat{m})$$

(a_j は実数）という形で軸 f を探せばよいことがわかる．このとき (2.4) 式は

$$\arg \max_{a \in \mathbb{R}^n, a^T G a = 1} a^T G^2 a \tag{2.6}$$

の最大値問題の解 $a \in \mathbb{R}^n$ を求めることに還元される．ここで，$n \times n$ 行列 G は中心化グラム行列と呼ばれる対称行列で，

$$G_{ij} = k(x_i, x_j) - \frac{1}{n}\sum_{t=1}^{n} k(x_i, x_t) - \frac{1}{n}\sum_{s=1}^{n} k(x_s, x_j) + \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n} k(x_s, x_t) \tag{2.7}$$

により定義される．式（2.6）の解 a は固有値問題として容易に解くことができる．

以上の導出は，カーネル法の方法論の典型的なものである．この方法論では，変換された関数データに対し，内積や相関を基本とした解析手法を関数空間 $\mathcal{H}$ で適用する．解のベクトル f は一般には無限次元空間の元であるが，与えられた n 個の関数データ $\Phi(x_i)$ の張る有限次元部分空間内で解を求めれば十分であることが多く（Representer 定理, Schölkopf and Smola (2002), Section 4.2），その範囲で問題を書き直すと n 次元のグラム行列を用いた問題に還元される．

データに変換を施すことにより非線性を取り込む手法は古くからある．例えば，実数値データ X を $(X, X^2, X^3, ...)$ と冪により拡張した後に解析を行うことは可能である．しかしながら，例えば m 次元データを d 次までの冪によって拡張すると，m に対して指数的に次元が増大し，その後の解析の計算量に問題が生じる．例えば，500 次元のデータに対して 2 次までの冪変換を施すと，12 万以上の次元のデータを扱う必要がある．正定値カーネルを用いると，直接的な展開による内積計算を避けることができ，データ数の次元の計算で済む．したがって，高次元でデータ数が比較的少ない場合には計算上有利である．逆にデータ数が大きいと，計算量を実用的なレベルにまで引き下げる工夫が必要となる．

カーネル法の長所として，上で述べた計算量的なものに加えて，非ベクトル的なデータに対しても，正定値カーネルさえ定義されれば，全く同じ方法論が適用可能な点が挙げられる．この性質を活かして，グラフやストリングデータといった構造化された非ベクトルデータにカーネル法が積極的に適用されている (福水, 2010, 7 章).

具体的にどのカーネルを選択すべきかは重要な問題である．本稿では詳しい議論は行わないが，結果に対する何らかの評価基準を用いる必要がある．例えば SVM では，期待識別誤差に対するクロスバリデーションによって，カーネルまたはカーネルの持つパラメータを選択する場合が多い．

3　再生核ヒルベルト空間による確率分布に関する推論

SVM やカーネル PCA などカーネル法の多くの手法は，もとの空間で考えた場合に複雑な非線形手法となることが多い．以降では，もとの空間上での統計的な意味が明確な問題を，再生核ヒルベルト空間への変換によって解く方法を紹介する．特に，データを発生させる分布の，独立性，条件付独立性，同一性などの性質を推論する方法に関して述べる．

3.1　再生核ヒルベルト空間における期待値と共分散

一般にヒルベルト空間 $\mathcal{H}$ はボレル集合族によって可測空間と考えることにする．$\mathcal{H}$ に値をとる確率変数 F に対し，$\mathcal{H}$ 上の汎関数

$$f \mapsto E[\langle F, f\rangle_{\mathcal{H}}]$$

が連続であると仮定すると，リース (Riesz) の表現定理 (例えば Reed and Simon, 1980, Section II.2) により，

$$E[\langle F, f\rangle_{\mathcal{H}}] = \langle m_F, f\rangle_{\mathcal{H}} \qquad (\forall f \in \mathcal{H})$$

を満たす $m_F \in \mathcal{H}$ が一意的に存在する．この m_F のことを F の**平均**と呼ぶ．

いま，$(\mathcal{X}, \mathcal{B})$ を可測空間，X を $\mathcal{X}$ に値をとる確率変数とし，$\mathcal{X}$ 上の，可測な正定値カーネルを持つ再生核ヒルベルト空間 $(\mathcal{H}, k)$ を考える．以下では，確率変数と再生核ヒルベルト空間に対して

(A-1)　　$E[k(X, X)] < \infty$

という仮定をおく．

前章と同様，写像 $\Phi : \mathcal{X} \to \mathcal{H}$ を $\Phi(x) = k(x, \cdot)$ により定めると，$\Phi(X)$ は $\mathcal{H}$ に値を持つ確率変数である．このとき，

$$\|\Phi(X)\|_{\mathcal{H}}^2 = \langle k(X, \cdot), k(X, \cdot)\rangle_{\mathcal{H}} = k(X, X)$$

に注意すると，$|E[\langle f, \Phi(X)\rangle_{\mathcal{H}}]| \leq \|f\|_{\mathcal{H}} E[\sqrt{k(X, X)}]$ となり，仮定 (A-1) から $f \mapsto E[\langle \Phi(X), f\rangle_{\mathcal{H}}]$ は $\mathcal{H}$ 上の連続汎関数である．そこで，$\Phi(X)$ の平均を m_X とおくと[5]，

$$\langle f, m_X\rangle_{\mathcal{H}} = E[f(X)] \qquad (\forall f \in \mathcal{H}) \tag{3.1}$$

[5] m_X は k に依存するが，記法を簡単にするため省略する．

が成立する．$m_X \in \mathcal{H}$ を X の $\mathcal{H}$ における**平均**と呼ぶことにする．X の分布が P であるとき，その平均を m_P と書くこともある．平均 m_X を用いると，関数 $f \in \mathcal{H}$ によって得られる確率変数 $f(X)$ の期待値が m_X と f との内積によって計算できる．特に $f = k(\cdot, u)$ とおくと

$$m_X(u) = E[k(u, X)] = \int k(u, x) dP(x) \tag{3.2}$$

（P は X の分布）となり、関数 m_X はカーネル関数の期待値であることが分かる。

2.1 節で述べたように，d 次の多項式カーネル（$c > 0$）が定める再生核ヒルベルト空間 $\mathcal{H}$ はベクトル空間として d 次以下の多項式全体と一致する．したがって，$\mathbb{R}$ 上の確率変数 X に対し，その r 次モーメント（$0 \leq r \leq d$）が

$$E[X^r] = \langle X^r, m_X \rangle_{\mathcal{H}}$$

により計算される．ここでは簡単のため $\mathbb{R}$ の場合を述べたが，$\mathbb{R}^\ell$ の場合も同様である．この例からわかるように，平均 m_X は X の分布の高次モーメントの情報を持っている．

次に確率変数の共分散の概念を再生核ヒルベルト空間上に拡張しよう．$(\mathcal{X}, \mathcal{B}_\mathcal{X}), (\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ を可測空間とし，(X, Y) は $\mathcal{X} \times \mathcal{Y}$ に値をとる確率変数とする．$\mathcal{X}$ と $\mathcal{Y}$ 上に，それぞれ可測な正定値カーネルを持つ再生核ヒルベルト空間 $(\mathcal{H}_\mathcal{X}, k_\mathcal{X}), (\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ を与える．ここで，確率変数と正定値カーネルはそれぞれ (A-1) の仮定を満たすとする．平均の場合と同様，$\Phi_\mathcal{X}(x) = k_\mathcal{X}(\cdot, x)$, $\Phi_\mathcal{Y}(y) = k_\mathcal{Y}(\cdot, y)$ を定義し，直積 $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$ 上の確率変数 $G_{XY} = (\Phi_\mathcal{X}(X) - m_X) \otimes (\Phi_\mathcal{Y}(Y) - m_Y)$ を考える．平均の場合と同様にして，条件 (A-1) により $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y} \ni f \otimes g \mapsto E[\langle f \otimes g, G_{XY} \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}}] = \mathrm{Cov}[f(X), g(Y)]$ が連続汎関数となることが示されるので，G_{XY} の平均 m_G は，任意の $f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}$ に対して

$$\langle m_G, f \otimes g \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} = \mathrm{Cov}[f(X), g(Y)]$$

を満たす．直積 $\mathcal{H}_1 \otimes \mathcal{H}_2$ は $\mathcal{H}_1$ から $\mathcal{H}_2$ への線形写像と同一視できるので，結局

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = \mathrm{Cov}[f(X), g(Y)] = E[f(X)g(Y)] - E[f(X)]E[g(Y)] \tag{3.3}$$

を満たす有界線形作用素 $\Sigma_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ が定まる．この作用素 Σ_{YX} を**相互共分散作用素**と呼ぶ[6]．すなわち，再生核ヒルベルト空間内の任意の関

[6]相互共分散作用素の一般的な理論は Baker (1973) に詳しい．

数 f, g によって変換された確率変数 $f(X), g(Y)$ の共分散が，再生核ヒルベルト空間の線形作用素と内積によって与えられる．特に $Y = X$ の場合，Σ_{XX} は自己共役作用素で，これを共分散作用素と呼ぶ．

例として，2 次元確率ベクトル (X, Y) に対し，X, Y それぞれに $\mathbb{R}$ 上の d 次多項式カーネル（ただし $c > 0$）を適用する．その相互共分散作用素 Σ_{YX} は，式 (3.3) で $f(x) = x^l, g(y) = y^r$ とおくことにより，

$$\langle y^\ell, \Sigma_{YX} x^r \rangle_{\mathcal{H}_y} = \mathrm{Cov}[Y^\ell, X^r] \qquad (0 \leq \ell, r \leq d)$$

を満たす．したがって，Σ_{YX} は d 次以下の任意の高次モーメントに関する情報を保持していると解釈できる．

3.2 カーネル法による分布の特徴づけ

前節で見たように，確率変数を再生核ヒルベルト空間に写像すると，その平均はさまざまなモーメントの情報を含んでいる．大雑把に言えば，確率変数に対してすべてのモーメントを考えればその分布は決まるので，十分広いクラスの関数を含むような再生核ヒルベルト空間における平均を考えれば，確率変数を一意的に定めることが期待できる．そこでまず，再生核ヒルベルト空間の豊かさについて定義しよう．

$(\mathcal{X}, \mathcal{B}_\mathcal{X})$ を可測空間，$\mathcal{P}$ をその上の確率測度全体の族とする．$\mathcal{X}$ 上の有界かつ可測な正定値カーネル k が（$(\mathcal{X}, \mathcal{B}_\mathcal{X})$ に関して）**特性的** (characteristic) であるとは，写像

$$\mathcal{P} \to \mathcal{H}_\mathcal{X}, \quad P \mapsto m_P$$

が単写であることをいう．すなわち，k の定める再生核ヒルベルト空間を $\mathcal{H}$ とするとき，任意の $f \in \mathcal{H}$ に対して $\int f dP = \int f dQ$ ならば，$P = Q$ であることをいう．

次の事実は，正定値カーネルが特性的であることを示す際に有用である．

補題 1 上の記法のもと，k が特性的であるための必要十分条件は，任意の確率分布 $P \in \mathcal{P}$ に対して $\mathcal{H} + \mathbb{R}$ が $L^2(P)$ で稠密となることである．ここで，$\mathcal{H}+\mathbb{R}$ は再生核ヒルベルト空間としての直和，すなわち $f+c$（$f \in \mathcal{H}, c \in \mathbb{R}$）の形の関数からなる再生核ヒルベルト空間を意味する．

証明 十分性：$\mathcal{H} + \mathbb{R}$ の稠密性を仮定し，相異なる確率分布 P, Q に対し，$m_P = m_Q$ としよう．$\mathcal{H}+\mathbb{R}$ は $L^2(|P-Q|)$（$|P-Q|$ は $P-Q$ の全変動を表す）

で稠密となるので，任意の $\varepsilon > 0$ と， $\mathcal{X}$ の任意の可測集合 A に対し，ある $f \in \mathcal{H}$ と $c \in \mathbb{R}$ があって $\|f + c - \chi_A\|_{L^1(|P-Q|)} < \varepsilon$ を満たす．ここで χ_A は A の定義関数である．このとき $|(\int f dP - P(A)) - (\int f dQ - Q(A))| < \varepsilon$ であるが，一方 $m_P = m_Q$ により $\int f dP - \int f dQ = 0$ なので， $|P(A) - Q(A)| < \varepsilon$ を得る．$\varepsilon > 0$ は任意なので $P(A) = Q(A)$ となり，$P \neq Q$ に矛盾する．

必要性：k を特性的とし，ある確率分布 P があって，$\mathcal{H} + \mathbb{R}$ が $L^2(P)$ で稠密でないとする．このとき零でない $\varphi \in L^2(P)$ が存在して，任意の $f \in \mathcal{H}$ に対して $\int \varphi g dP = 0$, $\int \varphi dP = 0$ が成り立つ．$\varphi \neq 0$ に注意して，2つの確率変数 Q_1, Q_2 を $Q_1(A) = \int_A |\varphi| dP / \|\varphi\|_{L^1(P)}$, $Q_2(A) = \int_A (|\varphi| - \varphi) dP / \|\varphi\|_{L^1(P)}$ により定義すると，$Q_1 \neq Q_2$ であるが，任意の $f \in \mathcal{H}$ に対して $\int f dQ_1 - \int f dQ_2 = \int \varphi dP / \|\varphi\|_{L^1(P)} = 0$ となり，k が特性的であることに矛盾する． ∎

式 (3.2) からわかるように，特性的な正定値カーネルによる平均は，$\mathbb{R}^m$ 上の確率分布 P に対する特性関数

$$E[e^{\sqrt{-1} u^T X}]$$

（X は P を分布に持つ確率変数）と類似性を持つ．よく知られているように，特性関数は確率分布 P を一意的に定めるが、特性的なカーネルによる平均 $E[k(u, X)]$ もこれと同じ性質を持つ．

付録で示したように，ガウスRBFカーネル $k^G_\sigma(x, y) = \exp\{-\|x-y\|^2/(2\sigma^2)\}$ やラプラスカーネル $k^L_\lambda(x, y) = \exp\left(-\lambda \sum_{i=1}^m |x_i - y_i|\right)$ など，応用上よく使われるカーネルが $\mathbb{R}^m$ 上の特性的なカーネルとなっている．

また，コンパクトな距離空間に対しては，普遍性という概念も再生核ヒルベルト空間の豊かさを表現するのに有用である (Steinwart, 2001)．$\mathcal{X}$ をコンパクト距離空間とし，k をその上の連続な正定値カーネルとする．k が**普遍**（universal）であるとは，k が定める再生核ヒルベルト空間が，$\mathcal{X}$ 上の連続関数全体に $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ によりノルムを入れて定義されるBanach空間の中で稠密であることをいう．ガウスRBFカーネルやラプラスカーネルなどが，$\mathbb{R}^m$ 上の任意のコンパクト集合上で普遍であることが知られている (Steinwart, 2001)．

普遍な正定値カーネルは特性的でもある．

命題 2 コンパクトな距離空間上の普遍な正定値カーネルは特性的である．

証明 P, Q を距離空間上の確率分布とするとき，任意の有界連続関数 f に関して $\int f dP = \int f dQ$ ならば $P = Q$ であることが知られている (Dudley, 2002, , Theorem 9.3.2)．このことから主張は容易に従う．∎

特性的な正定値カーネルによる平均 m_P が分布 P を識別できることを利用して，グレトン (Gretton) らは分布の均一性検定に m_P の推定量を用いる方法を提案している (Gretton et al., 2007)． 以下これを紹介する．

分布の均一性検定とは，２つのサンプル $(X_1, \ldots, X_\ell)$ と $(Y_1, \ldots, Y_n)$ を発生した分布が同じかどうかを判定する問題である．以下では $X_1, \ldots, X_\ell$ と $Y_1, \ldots, Y_n$ はそれぞれ可測空間 $(\mathcal{X}, \mathcal{B})$ 上の確率分布 P および Q に独立に従う i.i.d. サンプルとし，$P = Q$ を帰無仮説，$P \neq Q$ を対立仮説として検定を行うことを考える．

$\mathcal{X}$ 上の特性的なカーネルによる再生核ヒルベルト空間で，P および Q による平均 m_P, m_Q を構成し，

$$M(P, Q) \equiv \|m_P - m_Q\|^2_{\mathcal{H}}$$

と定義すると，$P = Q$ と $M(P, Q) = 0$ は同値である．平均の定義により $\langle m_P, m_Q \rangle_{\mathcal{H}} = E[m_Q(X_i)] = E[\langle k(\cdot, X_i), m_Q \rangle_{\mathcal{H}}] = E[k(Y_i, X_i)]$ であることなどに注意すると，$M(P, Q)$ の値は

$$M(P, Q) = E[k(X, \tilde{X})] - 2E[k(X, Y)] + E[k(Y, \tilde{Y})] \tag{3.4}$$

と計算することができる．ここで X, $\tilde{X}$ は P に，Y, $\tilde{Y}$ は Q に従う確率変数であり，すべて互いに独立であるとする．m_P および m_Q の推定量は，式 (2.5) と同様に

$$\widehat{m}_P = \frac{1}{\ell} \sum_{i=1}^{\ell} k(\cdot, X_i), \qquad \widehat{m}_Q = \frac{1}{n} \sum_{j=1}^{n} k(\cdot, Y_j) \tag{3.5}$$

で与えられることから，検定統計量として

$$\hat{M}_n(P, Q) = \|\widehat{m}_P - \widehat{m}_Q\|^2_{\mathcal{H}}$$

を用いることにする．これを展開すると

$$\hat{M}_n(P, Q) = \frac{1}{\ell n} \sum_{a,b=1}^{\ell} k(X_a, X_b) + \sum_{c,d=1}^{\ell} k(Y_c, Y_d) - 2 \sum_{a=1}^{\ell} \sum_{c=1}^{n} k(X_a, Y_c) \tag{3.6}$$

である．

式 (3.6) の和から $a = b, c = d$ の項を差し引いて不偏化して得られる統計量

$$U_{\ell,n} = \frac{1}{\ell(\ell-1)} \sum_{a=1}^{\ell} \sum_{b \neq a} k(X_a, X_b) + \frac{1}{n(n-1)} \sum_{c=1}^{n} \sum_{d \neq c} k(Y_c, Y_d) - \frac{2}{\ell n} \sum_{a=1}^{\ell} \sum_{c=1}^{n} k(X_a, Y_c)$$

は U 統計量となることがわかるので，その一般論から，$\ell, n \to \infty$ における漸近的性質を知ることができる (例えば van der Vaart, 1998, 12 章)．本稿では詳細は省略するが，帰無仮説 $P = Q$ のもと，$N = \ell + n$ とおいて $\ell/N \to \gamma$, $n/N \to 1 - \gamma$（$0 < \gamma < 1$）を満たすように $\ell, n \to \infty$ とすると，

$$NU_{\ell,n} \quad \Rightarrow \quad \sum_{i=1}^{\infty} \lambda_i \Big(Z_i^2 - \frac{1}{\gamma(1-\gamma)} \Big) \qquad (n \to \infty) \tag{3.7}$$

と法則収束する．ここで，Z_i は平均 0 分散 $1/\gamma(1-\gamma)$ の正規分布に従う独立な確率変数であり，$\{\lambda_i\}$ は

$$\tilde{k}(x, y) = k(x, y) - E[k(x, X)] - E[k(X, y)] + E[k(X, \tilde{X})] \tag{3.8}$$

（$\tilde{X}, X$ は独立に P に従う確率変数）を積分核に持つ $L^2(P)$ 上の積分作用素の非零固有値を重複度だけ並べたもの，すなわち，ある単位ベクトル $\phi_i \in L^2(P)$ に対して

$$\int \tilde{k}(x, y) \phi_i(y) dP(y) = \lambda_i \phi_i(x) \tag{3.9}$$

を満たす非負実数 λ_i を重複度だけ並べたものとなる．また，k が特性的とすると，対立仮説 $P \neq Q$ のもとでは，$M(P, Q) \neq 0$ であり，$\sqrt{N}\big(U_{\ell,n} - M(P, Q)\big)$ は正の分散を持つ正規分布に法則収束することがわかり，この検定は一致性を持つ．

以上により，λ_i が決定できれば帰無仮説のもとでの漸近分布がわかることになる．式 (3.8) の積分核は中心化された正定値カーネルに一致するので，固有値 λ_i の一致推定量が，式 (2.7) で定義される中心化グラム行列の固有値によって推定されることがわかる (Gretton et al., 2009a)．そこで，カイ 2 乗分布に従う独立なサンプルを発生させることによって，検定の棄却域を求めることができる．

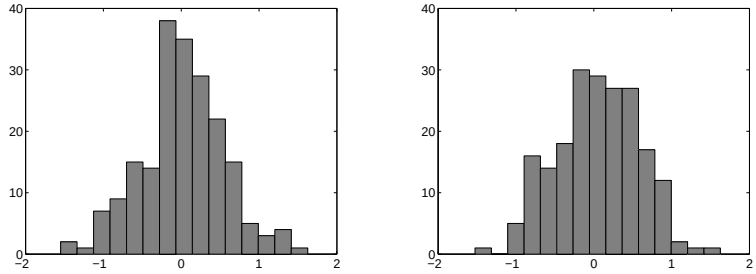


図 1: データ数 200 の場合のヒストグラムの例．左が正規分布 $N(0, 1/3)$，右が $a = 0.5$ による混合．

	$\hat{M}(P,Q)$					Kolmogorov-Smirnov				
$N\backslash a$	1	0.75	0.5	0.25	0	1	0.75	0.5	0.25	0
100	96.94	97.20	94.84	79.64	83.34	97.38	98.98	98.58	84.66	91.44
1000	96.40	98.52	82.96	45.16	0.18	97.52	98.90	87.32	85.54	13.12

表 1: 正定値カーネルによる方法と コルモゴロフ＝スミルノフ検定による均一性検定の結果．有意水準を $\alpha = 5\%$，データ数を $N = 100, 1000$ とし，5000 回の数値実験のうち帰無仮説が受容された割合 (%) を示した．

ここでは計算機実験として，P を正規分布 $N(0, 1/3)$，Q_a を 区間 $[-1, 1]$ 上の一様分布と $N(0, 1/3)$ との混合分布

$$Q_a : \qquad a\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} + (1-a)\frac{1}{2}I_{[-1,1]}(x)$$

とし，a を変化させて，$\hat{M}(P,Q)$ による検定を行った結果を示す．P と Q_a は平均と分散が常に一致するため，２次モーメントまでの情報ではこれらを識別できない．図 3.2 にヒストグラムの例を示した．正定値カーネルはガウス RBF カーネルを用い，標準偏差に相当するパラメータ σ には，データ間の距離 $\|X_i - X_j\|$ の中央値を用いた．表 1 に示された結果をみると，分布の均一性に対するノンパラメトリック検定の標準的方法であるコルモゴロフ＝スミルノフ (Komogorov-Smirnov) 検定と同程度以上の検出力を持っていることがわかる．

Gretton et al. (2007) では，ブートストラップによって α-%点を決める方法や，帰無分布をモーメントによってピアソンカーブにフィットさせる方法などを提案し，実データを用いて，均一性検定の他の方法との比較を行って

いる．

3.3 カーネル法による独立性の特徴づけ

2つの確率変数 X, Y の関係を見るために共分散や相関を調べることは基本的な方法であるが，これは線形な関係しか考慮していない．再生核ヒルベルト空間への写像 $\Phi_{\mathcal{X}}(X), \Phi_{\mathcal{Y}}(Y)$ は高次の情報を含んでいるため，関数空間上で相関や共分散を考えれば，確率変数の独立性や依存性が調べられることが期待できる．以下ではこの考えに基づいて，3.1 節で定義した再生核ヒルベルト空間上の相互共分散作用素を用いて確率変数の独立性や依存性をはかる方法を議論する．以下では X と Y が独立であることを $X \perp\!\!\!\perp Y$ で表す（Dawid の記法）．

まず，相互共分散作用素は以下のように確率変数の独立性を特徴付ける．

定理 3 $(\mathcal{H}_1, k_1)$, $(\mathcal{H}_2, k_2)$ をそれぞれ $\mathcal{X}$, $\mathcal{Y}$ 上の再生核ヒルベルト空間とし，積 $k_{\mathcal{X}} k_{\mathcal{Y}}$ は $\mathcal{X} \times \mathcal{Y}$ 上特性的であるとする．(X, Y) を $\mathcal{X} \times \mathcal{Y}$ 上の確率変数とし，それぞれ (A-1) の条件を仮定するとき，

$$X \perp\!\!\!\perp Y \quad \Longleftrightarrow \quad \Sigma_{XY} = O \tag{3.10}$$

の同値関係が成り立つ．

証明 (X, Y) の同時分布を P_{XY}，また X,Y と同じ周辺分布を持ち，互いに独立な確率分布を $P_X \otimes P_Y$ と書く．3.1 節で述べたように，直積 $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ と線形写像 $\mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ 全体の空間との同一視のもと，Σ_{YX} は，$\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ 上の平均として

$$\Sigma_{YX} = m_{P_{XY}} - m_{P_X \otimes P_Y} \tag{3.11}$$

と表現できるので，定理の同値性は $k_{\mathcal{X}} k_{\mathcal{Y}}$ が特性的なことから従う． ∎

定理 3 は，特性関数を用いた，よく知られた独立性の特徴づけ

$$X \perp\!\!\!\perp Y \quad \Longleftrightarrow \quad E_{XY}[e^{\sqrt{-1}u^T X} e^{\sqrt{-1}v^T Y}] = E_X[e^{\sqrt{-1}u^T X}] E_Y[e^{\sqrt{-1}v^T Y}] \tag{3.12}$$

の一般化とみなすことができる．式（3.10）と式（3.12）の右式はともに，$k(u, X)$, $k(v, Y)$ の形の変換を施した変数の共分散が任意の u, v に対して 0 であることを意味している．

平均の場合と同様に，有限個のサンプル $(X_1, Y_1), \ldots, (X_n, Y_n)$ が与えられたとき，確率分布を経験分布に置き換えることにより標本相互共分散作用素が以下のように与えられる．

$$\widehat{\Sigma}_{YX}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\big(k_{\mathcal{Y}}(\cdot, Y_i) - \widehat{m}_Y\big)\big\langle\big(k_{\mathcal{X}}(\cdot, X_i) - \widehat{m}_X\big), \cdot\big\rangle_{\mathcal{H}_{\mathcal{X}}} \tag{3.13}$$

定理3より，作用素 $\widehat{\Sigma}_{YX}^{(n)}$ の大きさを，X と Y の独立性あるいは依存性の尺度として用いることは自然である．Σ_{YX} に対する式（3.11）の表現から，

$$\left\| m_{P_{XY}} - m_{P_X \otimes P_Y} \right\|^2_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}}$$

を独立性の尺度として用いることが可能である．実は，作用素 Σ_{YX} の言葉で述べると，この値はヒルベルト＝シュミット（Hilbert-Schmidt）ノルムの2乗に一致する．

一般にヒルベルト空間 H_1 から H_2 への作用素 $A : H_1 \to H_2$ が**ヒルベルト＝シュミット**であるとは，H_1 と H_2 の正規直交基底 $\{\phi_i\}_{i=1}^{I}$ と $\{\psi_j\}_{j=1}^{J}$ $(i, j \in \mathbb{N} \cup \{\infty\})$ に対し，

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\langle\psi_j, A\phi_i\rangle^2_{H_2} < \infty$$

が成り立つことをいう．A がヒルベルト＝シュミットであるとき，そのヒルベルト＝シュミットノルム $\|A\|_{HS}$ を

$$\|A\|^2_{HS} = \sum_{i=1}^{I}\sum_{j=1}^{J}\langle\psi_j, A\phi_i\rangle^2_{H_2}$$

により定義する．ヒルベルト＝シュミット作用素およびヒルベルト-シュミットノルムの定義が正規直交基底の取り方によらないことは容易に示されるので，ある一組の正規直交基底に関して考えれば十分である．

また，作用素 $A : H_1 \to H_2$ が直積 $H_1 \otimes H_2$ の要素 ξ によって，

$$\langle g, Af\rangle_{H_2} = \langle\xi, f \otimes g\rangle_{H_1 \otimes H_2} \qquad (\forall f \in H_1, g \in H_2)$$

と表現されているとき，A はヒルベルト＝シュミットで，

$$\|A\|_{HS} = \|\xi\|_{H_1 \otimes H_2}$$

である．これは，H_1, H_2 の正規直交基底 $\{\phi_i\}, \{\psi_j\}$ に対し，$H_1 \otimes H_2$ の正規直交基底が $\{\phi_i \otimes \psi_j\}_{ij}$ によって与えられることから容易にわかる．

この事実を用いると，式 (3.4),(3.6) の特別な場合として，以下の表示が得られる．

$$\|\Sigma_{YX}\|_{HS}^2 = E[k_{\mathcal{X}}(X,\tilde{X})k_{\mathcal{Y}}(Y,\tilde{Y})] - 2E\big[E[k_{\mathcal{X}}(X,\tilde{X})|X]E[k_{\mathcal{Y}}(Y,\tilde{Y})|Y]\big] + E[k_{\mathcal{X}}(X,\tilde{X})]E[k_{\mathcal{Y}}(Y,\tilde{Y})]$$

（ただし，$(\tilde{X},\tilde{Y})$ は (X,Y) と独立で同一の分布に従う．）

$$\begin{aligned}\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2 =& \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\mathcal{X}}(X_i,X_j)k_{\mathcal{Y}}(Y_i,Y_j) \\ &- \frac{2}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n} k_{\mathcal{X}}(X_i,X_j)\sum_{\ell=1}^{n} k_{\mathcal{Y}}(Y_i,Y_\ell) \\ &+ \frac{1}{n^4}\sum_{i,j=1}^{n} k_{\mathcal{X}}(X_i,X_j)\sum_{\ell,r=1}^{n} k_{\mathcal{Y}}(Y_\ell,Y_r)\end{aligned}$$

$\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$ は $\|\Sigma_{YX}\|_{HS}^2$ に収束する．実はさらに強く，$\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\|_{HS} = O_p(n^{-1/2})$ が示されている (Fukumizu et al., 2007)．従って，$\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$ の値を，独立性・依存性をはかる尺度として用いることが可能である．また，均一性検定と同様に独立性検定に用いることが可能であるが，その詳細に関しては本稿では省略する．詳しくは，Gretton et al. (2007, 2009b) を見ていただきたい．ここで定義した尺度は確率密度関数の陽な推定を行わないので，高次元の確率変数に対して有効である．独立性・依存性をはかる尺度としてよく用いられるのは相互情報量であるが，サンプルからの推定を行う際には，確率密度関数の推定を必要とし，高次元連続変数の場合には精度よい計算が容易ではない．

$\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$ が独立性・依存性を捉えていることを見るために，以下のような計算機実験を行った．$X^{(0)}, Y^{(0)}$ を，それぞれ区間 $[-2,2]$ 上の一様分布，および原点対称な 2 つの区間上の一様分布とし，両者の平均と分散が一致するように区間を選んだ．$(X^{(\theta)}, Y^{(\theta)})$ は $(X^{(0)}, Y^{(0)})$ を θ だけ回転させて得られる確率変数とする．サンプルの例を図 2 に示した．$\theta = 0$ の時には $X^{(0)}$ と $Y^{(0)}$ は独立であるが，それ以外では独立ではない．しかし，$(X^{(0)}, Y^{(0)})$ の分散共分散行列がスカラー行列であるため，すべての θ に対して $X^{(\theta)}$ と $Y^{(\theta)}$ は無相関である．正定値カーネルにはガウス RBF カーネルを用い，パラメータ σ はデータの距離の中央値を用いた．データ数 200 個に対する

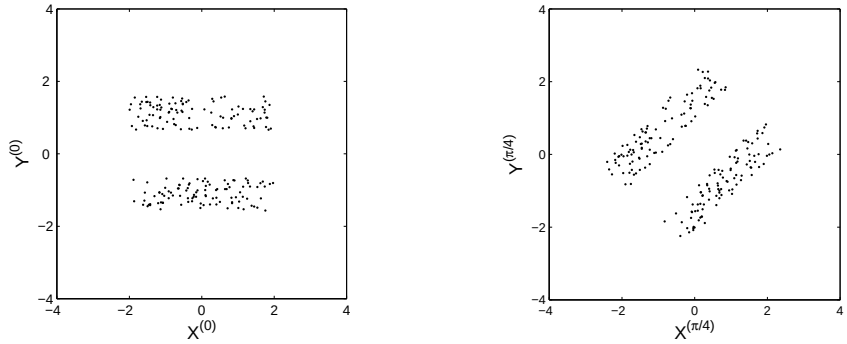


図 2: 独立性・依存性の計算機実験に用いたデータの例．左は独立，右は $\pi/4$ だけ回転したもので，無相関であるが独立ではない．

$\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$ の値を図 3 に示す． ベースラインとして，独立性を帰無仮説とする並べ替え検定の 5% 点の値（$X_1^{(\theta)}, \ldots, X_n^{(\theta)}$ の順序をランダムに並べ替えることにより $Y_j^{(\theta)}$ と独立にしたサンプルを多く発生させ，$X \perp\!\!\!\perp Y$ の場合の $\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$ の値の分布の 5% 点を近似的に求めたもの）も示した．

3.4 カーネル法による条件付独立性の特徴づけ

確率変数の条件付独立性は，グラフィカルモデリングや因果推論をはじめ，多くの統計的推論において重要な役割を果たす概念である．相互共分散作用素を用いると，カーネル法による条件付独立性の特徴付けが可能となる．

まず，有限次元ガウス確率変数の場合を復習しておこう．X, Y, Z を有限次元ガウス確率変数とするとき，Z が与えられたもとでの X と Y の条件付共分散行列は

$$C_{YX|Z} = C_{YX} - C_{YZ}C_{ZZ}^{-1}C_{ZX}$$

によって定義された．ここで C_{YX} などは共分散行列を表し，C_{ZZ} は可逆と仮定する．よく知られているように，ガウス確率変数に対しては，Z が与えられたもとで X と Y が条件付独立であることと $C_{YX|Z} = O$ であることが同値である．この事実は多くの変数の相互関係を調べるグラフィカルモデリングなどで頻繁に用いられている．

この事実を再生核ヒルベルト空間へ拡張すると，一般の確率変数の条件付独立性を特徴付けることが可能となる．以下では Fukumizu et al. (2004,

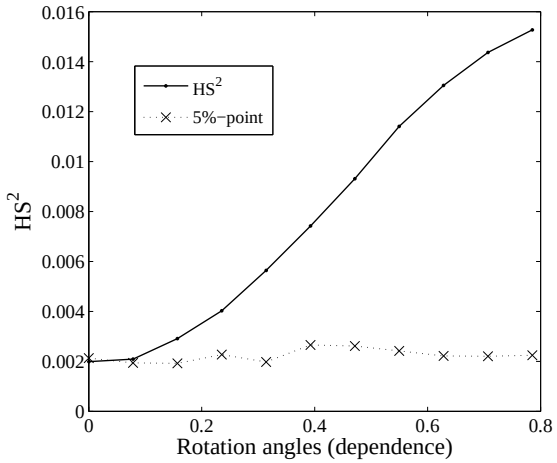


図 3: 独立性尺度の値の例．200 データからなるサンプルに対して，各回転角度における尺度の値と，並べ替え検定における 5% 点を示している．

2009a,b) に従ってその方法を説明する．

以降，$(\mathcal{X}, \mathcal{B}_\mathcal{X}), (\mathcal{Y}, \mathcal{B}_\mathcal{Y}), (\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ を可測空間，(X, Y, Z) を $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ 上の確率変数とし，$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ 上にそれぞれ (A-1) を満たす可測な正定値カーネルを持つ再生核ヒルベルト空間 $(\mathcal{H}_\mathcal{X}, k_\mathcal{X}), (\mathcal{H}_\mathcal{Y}, k_\mathcal{Y}), (\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ が与えられていると仮定する．また，確率変数の分布を P_X, P_Y, P_Z で表す．このとき，**条件付相互共分散作用素** $\Sigma_{YX|Z} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ を

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \tag{3.14}$$

により定義する．Σ_{ZZ} は逆作用素を持つとは限らないが，一般に相互共分散作用素 Σ_{YX} に対し，作用素ノルムが 1 以下の有界作用素 V_{YX} が一意的に存在し

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$$

かつ $\mathcal{R}(V_{YX}) \subset \overline{\mathcal{R}(\Sigma_{YY})}, \mathcal{N}(V_{YX})^\perp \subset \overline{\mathcal{R}(\Sigma_{XX})}$ となることが知られている (Baker, 1973)．したがって，式 (3.14) は正確には

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YY}^{1/2} V_{YZ} V_{ZX} \Sigma_{XX}^{1/2} \tag{3.15}$$

として定義される．条件付相互共分散作用素は，確率変数の条件付共分散と以下の定理のように関係している．

定理 4 正定値カーネルはすべて (A-1) を満たし，$k_{\mathcal{Z}}$ は特性的と仮定する．このとき任意の $f \in \mathcal{H}_{\mathcal{X}}$ と $g \in \mathcal{H}_{\mathcal{Y}}$ に対し

$$\langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = E_Z[\mathrm{Cov}[f(X), g(Y)|Z]]$$

が成立する．

証明 Σ_{ZZ} は自己共役なヒルベルト＝シュミット作用素なので，ある完全正規直交系 $\{\phi_i\}_{i=1}^N$（$N \in \mathbb{N} \cup \{\infty\}$）と $\lambda_i \geq 0$ が存在して $\Sigma_{ZZ}\phi_i = \lambda_i \phi_i$ とできる (例えば Reed and Simon, 1980, Theorem VI.16)．$I_0 = \{i \mid \lambda_i > 0\}$ とし，$i \in I_0$ に対して

$$\tilde{\phi}_i = \frac{1}{\sqrt{\lambda_i}}(\phi_i - E[\phi_i(Z)])$$

とおくと，$\langle \phi_i, \Sigma_{ZZ}\phi_j \rangle_{\mathcal{H}_{\mathcal{Z}}} = \sqrt{\lambda_i \lambda_j} E[\tilde{\phi}_i(Z)\tilde{\phi}_j(Z)]$ により，$\{\tilde{\phi}_i\}_{i \in I_0}$ は $L^2(P_Z)$ の正規直交系である．また，もし $i \notin I_0$ なる i があったとすると，$\mathrm{Var}[\phi_i(Z)] = \langle \Sigma_{ZZ}\phi_i, \phi_i \rangle_{\mathcal{H}_{\mathcal{Z}}} = 0$ により，ϕ_i は確率 1 で定数関数に限られる．補題 1 より $\mathcal{H}_{\mathcal{Z}} + \mathbb{R}$ は $L^2(P_Z)$ で稠密であるので，$\{\tilde{\phi}_i\}_{i \in I_0} \cup \{1\}$ は $L^2(P_Z)$ の完全正規直交系となる．

定理を証明するためには

$$\langle g, \Sigma_{YY}^{1/2} V_{YZ} V_{ZX} \Sigma_{XX}^{1/2} f \rangle_{\mathcal{H}_{\mathcal{X}}} = E\Big[E[(f(X) - E[f(X)])|Z] E[(g(Y) - E[g(Y)])|Z]\Big]$$

を示せばよいが，$\mathcal{R}(V_{ZY})$ および $\mathcal{R}(V_{ZX})$ が $N(\Sigma_{ZZ})$ と直交することに注意すると，パーセバルの等式により上式の左辺は

$$\begin{aligned}
&\sum_{i=1}^{N} \langle V_{ZY} \Sigma_{YY}^{1/2} g, \phi_i \rangle_{\mathcal{H}_{\mathcal{Z}}} \langle \phi_i, V_{ZX} \Sigma_{XX}^{1/2} f \rangle_{\mathcal{H}_{\mathcal{Z}}} \\
&= \sum_{i \in I_0} \langle V_{ZY} \Sigma_{YY}^{1/2} g, \phi_i \rangle_{\mathcal{H}_{\mathcal{Z}}} \langle \phi_i, V_{ZX} \Sigma_{XX}^{1/2} f \rangle_{\mathcal{H}_{\mathcal{Z}}} \\
&= \sum_{i \in I_0} \Big\langle \Sigma_{ZY} g, \frac{\phi_i}{\sqrt{\lambda_i}} \Big\rangle_{\mathcal{H}_{\mathcal{Z}}} \Big\langle \frac{\phi_i}{\sqrt{\lambda_i}}, \Sigma_{ZX} f \Big\rangle_{\mathcal{H}_{\mathcal{Z}}} \\
&= \sum_{i \in I_0} E\Big[\tilde{\phi}_i(Z)(g(Y) - E[g(Y)])\Big] E\Big[\tilde{\phi}_i(Z)(f(X) - E[f(X)])\Big]
\end{aligned}$$

と書き直せる．ここで，$E[f(X) - E[f(X)]|Z]$ および $E[g(Y) - E[g(Y)]|Z]$ が $L^2(P_Z)$ に属することは簡単に示されるので，$\{\tilde{\phi}_i\}_{i \in I_0} \cup \{1\}$ が $L^2(P_Z)$

の完全正規直交系となることを用いると，上式の末行はさらに

$$\Big(E[g(Y)-E[g(Y)]|Z],\, E[f(X)-E[f(X)]|Z]\Big)_{L^2(P_Z)} - \Big(E\big[g(Y)-E[g(Y)]|Z\big],\, 1\Big)_{L^2(P_Z)} \Big(1,\, E\big[f(X)-E[f(X)]|Z\big]\Big)_{L^2(P_Z)}$$

と表せる．ただし $(\ ,\ )_{L^2(P_Z)}$ は $L^2(P_Z)$ の内積を表す．上式第 2 項が 0 になることから，定理が証明される． ∎

以下では Z が与えられたもとでの X と Y の条件付独立性を $X \perp\!\!\!\perp Y \mid Z$ で表す．ガウス確率変数の場合と異なり，再生核ヒルベルト空間の場合には $\Sigma_{YX|Z} = O$ と $X \perp\!\!\!\perp Y \mid Z$ は同値とは限らない．

いま，確率変数 (X, Y, Z) を用いて，$\mathcal{X} \times \mathcal{Y}$ 上の確率分布 $E_Z[P_{X|Z} \otimes P_{Y|Z}]$ を，任意の $A \in \mathcal{B}_{\mathcal{X}}, B \in \mathcal{B}_{\mathcal{Y}}$ に対して

$$E_Z[P_{X|Z} \otimes P_{Y|Z}](A \times B) = E_Z\big[E[\chi_A(X)|Z]\, E[\chi_B(Y)|Z]\big]$$

を満たすように定める．また，(X, Y) の同時確率分布を P_{XY} で表す．このとき，以下の定理が成り立つ．

定理 5 正定値カーネルと確率変数は定理 4 と同じ仮定を満たすとし，さらに積 $k_{\mathcal{X}} k_{\mathcal{Y}}$ は $\mathcal{X} \times \mathcal{Y}$ 上の特性的なカーネルと仮定する．このとき

$$\Sigma_{YX|Z} = O \qquad \Longleftrightarrow \qquad P_{YX} = E_Z[P_{Y|Z} \otimes P_{X|Z}] \tag{3.16}$$

の同値関係が成立する．

証明 定理 4 の表示により左向きの矢印は明らかなので，右向きを示す．簡単のため $Q = E_Z[P_{X|Z} \otimes P_{Y|Z}]$ と書く．$\Sigma_{YX|Z} = O$ のとき，定理 4 より，任意の $f \in \mathcal{H}_{\mathcal{X}},\ g \in \mathcal{H}_{\mathcal{Y}}$ に対して $E_Q[f(X)g(Y)] = E_{P_{XY}}[f(X)g(Y)]$ が成り立つ．さらに，直積 $\mathcal{X} \otimes \mathcal{Y}$ の任意の元が $\sum_{i=1}^n f_i g_i$（$f_i \in \mathcal{H}_{\mathcal{X}}, g_i \in \mathcal{H}_{\mathcal{Y}}$）の形の元の極限として得られることと，定数関数に対してはこの関係が自明に成り立つことから，$(\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}) + \mathbb{R}$ の任意の関数 ϕ に対して $E_Q[\phi(X, Y)] = E_{P_{XY}}[\phi(X, Y)]$ を得る．$k_{\mathcal{X}} k_{\mathcal{Y}}$ は特性的なので $Q = P_{XY}$ が成り立つ． ∎

上の定理は，$\Sigma_{YX|Z} = O$ が $X \perp\!\!\!\perp Y|Z$ よりも弱い条件であることを示している．ガウス確率変数の場合には，$\mathrm{Cov}[Y, X|Z]$ が Z に依存しないという特

別な性質を持つため，条件付共分散行列が条件付独立性を特徴づけた．一方，再生核ヒルベルト空間上の条件付共分散作用素は，定理 3.15 が示すように Z に関する期待値しか表せないため，一般には条件付独立性を特徴づけることができない．しかしながら，X と Y の代わりに (X, Z) と Y を用いると，次のような特徴付けが可能となる．

定理 6 正定値カーネルと確率変数は定理 4 と同じ仮定を満たすとする．$W = (X, Z)$ とし，$\mathcal{X} \times \mathcal{Z}$ 上の正定値カーネルを $k_{\mathcal{W}} = k_{\mathcal{X}} k_{\mathcal{Z}}$ により定めるとき，積 $k_{\mathcal{W}} k_{\mathcal{Y}}$ が $(\mathcal{X} \times \mathcal{Z}) \times \mathcal{Y}$ 上特性的と仮定する．このとき，

$$\Sigma_{YW|Z} = O \qquad \Longleftrightarrow \qquad X \perp\!\!\!\perp Y \,|\, Z \tag{3.17}$$

の同値関係が成立する．

証明 一般に，任意の可測集合 $A \in \mathcal{B}_{\mathcal{X}}, B \in \mathcal{B}_{\mathcal{Y}}, C \in \mathcal{B}_{\mathcal{X}}$ に対し，

$$\begin{aligned}
&E\Big[E[\chi_{A\times C}(X,Z)|Z]\, E[\chi_B(Y)|Z]\Big] - E\Big[\chi_{A\times C}(X,Z)\chi_B(Y)\Big] \\
&= E\Big[E[\chi_A(X)|Z]\chi_C(Z)E[\chi_B(Y)|Z]\Big] - E\Big[E[\chi_A(X)\chi_B(Y)|Z]\chi_C(Z)\Big] \\
&= \int_C \Big\{P_{X|Z}(A|z)P_{Y|Z}(B|z) - P_{XY|Z}(A\times B|z)\Big\} dP_Z(z)
\end{aligned}$$

が成り立つが，定理 5 により，$\Sigma_{YW|Z} = O$ という条件は，上式末行の積分が 0 であること，すなわち，$P_{X|Z}(A|z)P_{Y|Z}(B|z) - P_{XY|Z}(A \times B|z) = 0$ が P_z に関して確率 1 で成り立つことと同値である．これは $X \perp\!\!\!\perp Y \,|\, Z$ を意味する． ∎

式 (3.14) から条件付相互共分散作用素の推定量を得るためには，逆作用素を考える必要があるが，これは一般に存在するとは限らないため，正則化を用い，

$$(\widehat{\Sigma}^{(n)}_{ZZ} + \varepsilon_n I)^{-1}$$

($\varepsilon_n > 0$) を推定量として使う．ここで ε_n は正則化のための正定数で $n \to \infty$ のときに $\varepsilon_n \to 0$ となるように定める．これを用いて，条件付相互共分散作用素の推定量 $\widehat{\Sigma}^{(n)}_{YX|Z}$ を

$$\widehat{\Sigma}^{(n)}_{YX|Z} := \widehat{\Sigma}^{(n)}_{YX} - \widehat{\Sigma}^{(n)}_{YZ}(\widehat{\Sigma}^{(n)}_{ZZ} + \varepsilon_n I)^{-1}\widehat{\Sigma}^{(n)}_{ZX} \tag{3.18}$$

により定める．

$\ddot{X} = (X, Z)$ とするとき，独立性の場合と同様に，$\widehat{\Sigma}^{(n)}_{Y\ddot{X}|Z}$ のヒルベルト＝シュミットノルムを条件付独立性の尺度として用いることが可能である．さらに対称性を重視して，$\ddot{Y} = (Y, Z)$ とおき，$\|\widehat{\Sigma}^{(n)}_{\ddot{X}\ddot{Y}|Z}\|^2_{HS}$ を用いることもできる．これを中心化グラム行列 $G_{\ddot{X}}, G_{\ddot{Y}}, G_Z$ を用いて書き下すと，

$$\|\widehat{\Sigma}^{(n)}_{\ddot{X}\ddot{Y}|Z}\|^2_{HS} = \frac{1}{n^2}\mathrm{Tr}\Big[G_{\ddot{X}}G_{\ddot{Y}} - 2G_{\ddot{X}}G_Z\big(G_Z + n\varepsilon_n I_n\big)^{-1}G_{\ddot{Y}} + G_{\ddot{X}}G_Z\big(G_Z + n\varepsilon_n I_n\big)^{-1}G_{\ddot{Y}}G_Z\big(G_Z + n\varepsilon_n I_n\big)^{-1}\Big]$$

となる．$\varepsilon_n \to 0$ かつ $n\varepsilon_n^3 \to \infty$ のとき $n \to \infty$ において $\|\widehat{\Sigma}^{(n)}_{\ddot{X}\ddot{Y}|Z}\|^2_{HS}$ が $\|\Sigma_{\ddot{X}\ddot{Y}|Z}\|^2_{HS}$ に確率収束することも示される．

Sun et al. (2007) では，ここで述べたヒルベルト＝シュミットノルムによる独立性および条件付独立性の尺度を用いて，変数間の因果関係の推論を行う方法を提案している．

3.5 カーネル次元削減法

カーネル次元削減法 (Fukumizu et al., 2004, 2009b) は，m 次元説明変数 X を用いて従属変数 Y を説明する回帰の問題において，Y に関する情報を保持するような X の低次元部分空間への射影を見つける方法である．近年では，画像，テキスト，遺伝子発現データなど極めて高次元のデータを扱う必要が高く，データの説明や可視化，予測・決定の精度向上のためのノイズ削減，計算量の軽減などさまざまな目的のために次元削減は重要な方法となっている．カーネル次元削減法では，次元削減の問題を条件付独立性によって定式化し，正定値カーネルを用いて条件付独立性をなるべく満たすような部分空間を見つける．

まず，次元削減が有効に働くための問題設定として，$\mathbb{R}^m$ の r 次元部分空間 S が存在して，

$$p_{Y|X}(y|x) = p_{Y|\Pi_S X}(y|\Pi_S x) \tag{3.19}$$

が成り立つと仮定する．ここで Π_S は部分空間 S への直交射影である．式 (3.19) を満たす部分空間 S のことを Li (1991) にならって**有効部分空間**と呼ぶことにする．これは，X に含まれる Y の情報を完全に保持する部分空間である．ここでは，次元削減の問題を，与えられた有限サンプルから有効部分空間 S を推定する問題として定式化する．以降では有効部分空間 S の次元 r は既知として話を進める．

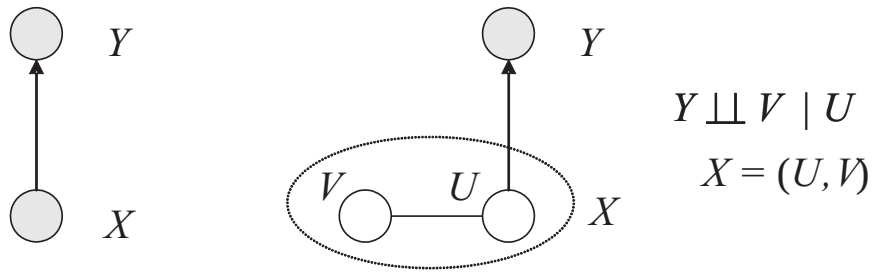


図 4: 回帰問題における次元削減のグラフィカル表現

有効部分空間は，以下のように条件付独立性によって特徴付けられる．S とその直交補空間 $S^{\perp}$ の正規直交基底を並べた行列を，それぞれ B, C とおく．B と C はぞれぞれ $m \times r$, $m \times (m-r)$ 行列であり，(B,C) は m 次元直交行列となる．S と $S^{\perp}$ への X の直交射影をそれぞれ $U = B^T X$, $V = C^T X$ で表すことにすると，(B,C) が直交行列であることから，確率密度関数に関して $p_X(x) = p_{U,V}(u,v)$, $p_{X,Y}(x,y) = p_{U,V,Y}(u,v,y)$ が成り立つ．これにより，式 (3.19) は

$$p_{Y|U,V}(y|u,v) = p_{Y|U}(y|u) \tag{3.20}$$

と同値である．すなわち，S が有効部分空間であることと，$U = \Pi_S X$ が与えられたもとでの Y と $V = \Pi_{S^{\perp}} X$ の条件付独立性とは同値である（図 4）．

条件付独立性の特徴づけには定理 6 を用いることもできるが，今の問題設定には以下の条件付共分散作用素を用いるのがよい．

定理 7 $k_{\mathcal{U}}, k_{\mathcal{V}}, k_{\mathcal{Y}}$ をそれぞれ可測集合 $\mathcal{U}, \mathcal{V}, \mathcal{Y}$ 上の正定値カーネル，(U,V,Y) を $\mathcal{U} \times \mathcal{V} \times \mathcal{Y}$ に値をとる確率変数とし，これらはそれぞれ条件 (A-1) を満たすとする．また，$\mathcal{X} = \mathcal{U} \times \mathcal{V}$ 上の正定値カーネルを積 $k_{\mathcal{X}} = k_{\mathcal{U}} k_{\mathcal{V}}$ によって定め，$k_{\mathcal{X}}$ と $k_{\mathcal{U}}$ が特性的であると仮定する．このとき，自己共役作用素の半順序に関して

$$\Sigma_{YY|U} \geq \Sigma_{YY|X} \tag{3.21}$$

が成立する．さらに $k_{\mathcal{Y}}$ が特性的であるとき

$$\Sigma_{YY|X} = \Sigma_{YY|U} \qquad \Longleftrightarrow \qquad Y \perp\!\!\!\perp X \mid U \tag{3.22}$$

の同値関係が成立する．

線形回帰の平均2乗誤差の場合から類推できるように，条件付分散 $E_X[\mathrm{Var}_{Y|X}[g(Y)|X]]$ は，X を用いて $g(Y)$ を推定したときの推定誤差を表すものと考えることができる．したがって定理4により，式(3.21)が表しているのは，情報が部分的になれば Y の推定誤差が増加するという当然の事実である．また，推定誤差が増加しなければ，X と U は Y に関して同じだけの情報量を持つと解釈できるので，式(3.22)の同値性は自然である．

証明 定理4により，

$$\langle g, (\Sigma_{YY|U} - \Sigma_{YY|X})g\rangle_{\mathcal{H}_\mathcal{Y}} = E_U\big[\mathrm{Var}[g(Y)|U]\big] - E_X\big[\mathrm{Var}[g(Y)|X]\big]$$

が成り立つ．ここで，条件付分散に関するよく知られた関係式

$$\mathrm{Var}[g(Y)|U] = E\big[\mathrm{Var}[g(Y)|U,V]|U\big] + \mathrm{Var}\big[E[g(Y)|U,V]|U\big]$$

の期待値をとると

$$E\big[\mathrm{Var}[g(Y)|U]\big] - E\big[\mathrm{Var}[g(Y)|X]\big] = E\big[\mathrm{Var}[E[g(Y)|X]|U]\big] \geq 0$$

が得られ，式(3.21)が成り立つ．等号成立は，ほとんどすべての X に対して $E_{Y|X}[g(Y)|X] = E_{Y|U}[g(Y)|U]$ となる場合であるが，$k_\mathcal{Y}$ は特性的なので，定理の同値性を得る． ∎

条件付共分散作用素の推定量は，式(3.18)と同様に定めればよい．カーネル次元削減法においては，$U = B^TX$ の形で与えられるので，有効部分空間への射影行列 B の推定関数として

$$\min_{B:B^TB=I_r} \mathrm{Tr}\big[\widehat{\Sigma}^{(n)}_{YY|B^TX}\big]$$

を用いることができる．これを中心化グラム行列を用いて表示し，B に関係する項だけを用いて書くと，

$$\max_{B:B^TB=I_r} \mathrm{Tr}\left[\widehat{G}_Y\big(\widehat{G}_{B^TX} + n\varepsilon_n I_n\big)^{-1}\right] \tag{3.23}$$

となる．この規準により部分空間を求める方法をカーネル次元削減法と呼ぶ．上式で与えられる推定量は $\varepsilon_n \to 0,\ n\varepsilon_n^3 \to \infty$ のもとで一致性を持つことが証明されている (Fukumizu et al., 2009b)．

カーネル次元削減法を実行するためには，上記の推定関数の最適化を行う必要があるが，この関数は非凸であり，勾配法などによる非線形最適化

手法が必要となる．この最適化には $n \times n$ 行列の演算を数多く行う必要があり，サンプル数 n が大きいと計算量が増大する．これに対して，不完全コレスキー分解によって中心化グラム行列を低ランク行列で近似すると演算量を大幅に削減することが可能である (Bach and Jordan, 2002)．

カーネル次元削減法の導出には，周辺分布，条件付分布および可測集合に関する条件をほとんど必要としない点に注意してほしい．したがって，カーネル次元削減法は，離散変数などを含んだ幅広い状況に応用可能である．回帰問題での次元削減に対する従来法としては，Sliced Inverse Regression (SIR, Li, 1991) や Principal Hessian Directions (pHd, Li, 1992) などが有名であるが，これらの手法では，X の分布の楕円性が必要であったり，Y が二値の場合には 1 次元部分空間しか発見できないなどの強い制約が存在する．また，正準相関分析 (CCA, canonical correlation analysis) や PLS 回帰 (partial least square regression) なども用いられることがあるが，これらは線形モデルを仮定している．こういった仮定を置かないカーネル次元削減法は，より広い問題に応用可能である．

3.5.1 カーネル次元削減法の応用例

データ可視化の能力を見る目的で，UCI machine learning repository (Frank and Asuncion, 2010) の Wine データを用いた．このデータは 3 種類のワインに対する 13 次元の連続値属性を 178 サンプル集めたデータである．クラスの情報をなるべく保持するように，各手法で 2 次元部分空間を求めた結果が図 5 である．KDR が 3 クラスを最もよく判別しており，2 次元空間で完全な識別が可能なことがわかる．CCA も 3 クラスを完全に分けているが，境界はそれほど明確ではない．SIR と PLS の結果では判別は不完全である．Fukumizu et al. (2009b) では，さらにさまざまなデータへの応用例が示されている．

4 おわりに

本稿では，機械学習や統計的学習理論の分野で近年研究が盛んとなった「カーネル法」の概要をごく簡単に述べ，著者がたずさわっている，カーネル法による確率変数の依存性解析に関する最近の研究について述べた．サポートベクターマシンの提案以来，さまざまなデータ解析の手法がカーネル法の方法論に基づいて提案されたが，それらの統計理論的解析やそれに

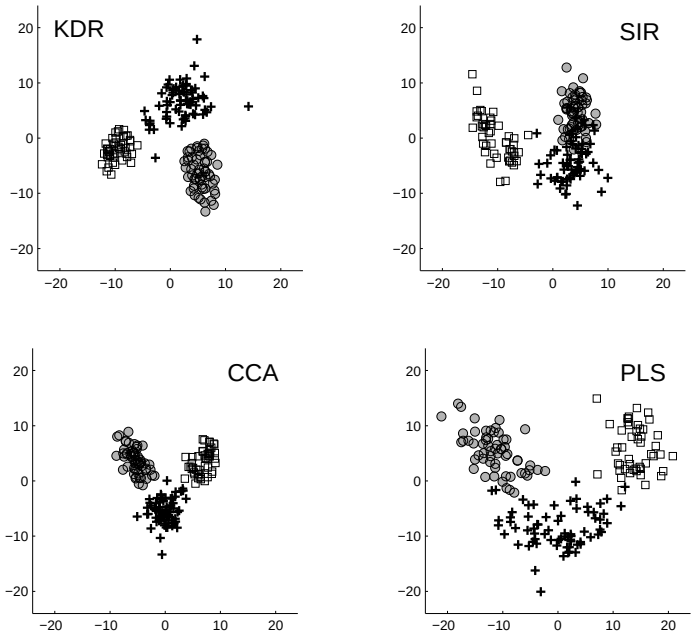


図 5: *Wine* データの 2 次元射影. ”+”, ”●”, ”□”が 3 クラスに対応.

もとづく改良は今後の重要な課題であり，統計科学からのアプローチが重要な分野である．

カーネル法による確率分布の表現やそれを用いた依存性解析は新しい研究分野であり，これから解決すべき課題が多いが，正定値カーネルによる方法は特性関数による方法の自然な拡張であり，ノンパラメトリック推定の有望な方法論を提供していると考える．

最後に，ごく最近の研究の展開に関して触れておく．カーネル法によるノンパラメトリック推定は，ベイズの定理のカーネル表現を実現することによって，ベイズ推論一般に適用可能であることが示されている (Fukumizu et al, 2012)．これは今後の興味深い方向性と考える．また，本稿で紹介したカーネル次元削減法の発展として，数値最適化が不要な，固有値問題により簡単に解が求められる方法も提案され (Fukumizu and Leng, 2011)，非常に高次元のデータにも適用可能であることが示されている．

謝辞

本稿をまとめるにあたり，マックス･プランク研究所の Bernhard Schölkopf 氏と Arthur Gretton 氏には多くの示唆を受けた．また本研究の一部は，フンボルト財団フェローシップ，科学研究費補助金 15700241, 22300098，稲盛財団研究助成，三菱財団自然科学研究助成，および情報・システム研究機構融合研究プロジェクトによる支援を受けた．

付録　特性的な正定値カーネル

$\mathbb{R}^n$ 上の関数 ϕ に対し、$\phi(x-y)$ が正定値カーネルとなるとき、ϕ を $\mathbb{R}^n$ 上の**正値関数**という。$\mathbb{R}^m$ 上定義される正定値カーネルのなかで，連続かつ平行移動不変なもの，すなわち $\mathbb{R}^m$ 上の連続な正値関数 $\phi(z)$ があって

$$k(x,y) = \phi(x-y)$$

と表される正定値カーネルを考える．このような形の正定値カーネルは，フーリエ変換による特徴づけを持つ．

定理 8 (ボホナー (Bochner) の定理) $\mathbb{R}^m$ 上の複素数値連続関数 $\phi(z)$ が正値関数となるための必要十分条件は，$\mathbb{R}^m$ 上の有限な非負ボレル測度 Λ があって，

$$\phi(z) = \int e^{\sqrt{-1}u^T z} d\Lambda(u)$$

と表されることである．また，このような Λ は一意である．

定理の式で与えられる関数が連続な正値関数を与えることは簡単に確認できるが，ボホナーの定理はこの形で尽くされることを主張している．証明は例えばReed and Simon (1980, Theorem IX.9) を見ていただきたい．

ボホナーの定理を用いると，連続で平行移動不変な正定値カーネルが特性的となる条件を容易に述べることができる．この時重要なのは，平行移動不変な正定値カーネル $k(x,y)=\phi(x-y)$ に対し，確率 P の $\mathcal{H}_k$ における平均 m_P^k が

$$m_P^k(x)=\int k(x,y)dP(y)=\int \phi(x-y)dP(y)=(\phi * P)(x),$$

すなわち ϕ と P の畳み込みとして表現できる点である．したがって，特性的であることは，

$$\phi * P=\phi * Q \quad \Longrightarrow \quad P=Q$$

と同値である．ここで畳み込みのフーリエ変換がフーリエ変換の積で与えられることを用いると，厳密性に多少目を瞑れば，上の条件はさらに

$$\widehat{\phi}\widehat{P}=\widehat{\phi}\widehat{Q} \quad \Longrightarrow \quad P=Q$$

と書き直せる．この条件は $\widehat{\phi}$ が全空間で正であれば成立することが予想されるが，実際以下に見るように，上の議論を厳密化することが可能である．

定理 9 ϕ を $\mathbb{R}^n$ 上の連続な複素数値正値関数とし，Λ をボホナーの定理の表示

$$\phi(x)=\int e^{\sqrt{-1}\omega^T x}d\Lambda(\omega)$$

を与える有限非負ボレル測度とする．このとき，$\mathrm{Supp}(\Lambda)=\mathbb{R}^n$ であれば[7]，$\phi(x-y)$ は特性的な正定値カーネルである．

証明 定理の直前の議論により，有限な実測度 μ が $\mu*\phi=0$ を満たすとき，$\mu=0$ を示せばよい．フビニの定理を用いると

$$\begin{aligned}
&\int (\mu*\phi)(x)d\mu(x)=\int\int \phi(x-y)d\mu(y)d\mu(x)\\
&=\int\int\int e^{\sqrt{-1}(x-y)^T\omega}\Lambda(\omega)d\mu(y)d\mu(x)\\
&=\int\int e^{\sqrt{-1}x^T\omega}d\mu(x)\int e^{-\sqrt{-1}y^T\omega}d\mu(y)d\Lambda(\omega)=\int |\widehat{\mu}(\omega)|^2 d\Lambda(\omega)
\end{aligned}$$

[7] $\mathrm{Supp}(\Lambda)$ は Λ の台を表す．非負測度 μ の台は，$\mathrm{Supp}(\mu)=\{x \mid x$ を含む任意の開集合 U に対して $\mu(U)>0\}$ と定義される．

である．ここで，$\mu * \phi = 0$ より

$$\int |\widehat{\mu}(\omega)|^2 d\Lambda(\omega) = 0$$

を得るが，$\widehat{\mu}$ が $\mathbb{R}^n$ 上連続であること[8]と，$\mathrm{Supp}(\Lambda) = \mathbb{R}^n$ であることから，$\widehat{\mu} = 0$ が結論される．フーリエ変換の一意性により $\mu = 0$ を得る． ∎

特に ϕ が実数値の正値関数の場合は，上の条件は必要十分である．

定理 10 ϕ を $\mathbb{R}^n$ 上の連続な実正値関数とし，Λ をボホナーの定理の表示

$$\phi(x) = \int e^{\sqrt{-1}\omega^T x} d\Lambda(\omega)$$

を与える有限非負ボレル測度とする．このとき，$\phi(x-y)$ が特性的な正定値カーネルであるための必要十分条件は $\mathrm{Supp}(\Lambda) = \mathbb{R}^n$ である．

証明 以下では集合 $A \subset \mathbb{R}^n$ に対して $-A = \{-a \in \mathbb{R}^n \mid a \in A\}$, $A - A = \{a - b \in \mathbb{R}^n \mid a, b \in A\}$, $A + b = \{a + b \mid a \in A\}$ と表す．

定理 9 より必要性のみ示せばよい．$k(x,y) = \phi(x-y)$ が特性的であるとき，$\mathrm{Supp}(\Lambda) \neq \mathbb{R}^n$ と仮定して矛盾を導く．そのために，2 つの異なる確率分布のフーリエ変換の差として表される関数 h で，$\mathrm{Supp}(h) \cap \mathrm{Supp}(\Lambda) = \emptyset$ となるものを構成する．

まず ϕ が実関数であることから，任意のボレル集合 E に対して $\Lambda(-E) = \Lambda(E)$ が成り立つ．よって $\mathbb{R}^n \backslash \mathrm{Supp}(\Lambda)$ は原点対称な空でない開集合である．

$\omega_0 \in \mathbb{R}^n \backslash \mathrm{Supp}(\Lambda)$（$\omega_0 \neq 0$）を固定する．このとき，原点の開近傍 W が存在して，$\pm\omega_0 \notin \mathrm{cl}(W - W)$，$(\mathrm{cl}(W-W) + \omega_0) \cap (\mathrm{cl}(W-W) - \omega_0) = \emptyset$, かつ $\mathrm{cl}(W-W) \pm \omega_0 \subset \mathbb{R}^n \backslash \mathrm{Supp}(\Lambda)$ とできる．

上のような W を固定し，$g = \chi_W * \chi_{-W}$ と定める．ここで χ_W は集合 W の定義関数である．関数 g は 0 でない連続関数で，$\mathrm{Supp}(g) \subset \mathrm{cl}(W-W)$, さらに g が $\mathbb{R}^n$ 上の正値関数であることが容易に示される．したがってボホナーの定理により，ある有界な非負ボレル測度 μ があって

$$g(\omega) = \int e^{\sqrt{-1}\omega^T x} d\mu(x)$$

[8] $\widehat{\mu}(\omega) = \int e^{\sqrt{-1}x^T \omega} d\mu(x)$ の連続性は優収束定理からすぐにわかる．

が成り立つ．$h(\omega) = g(\omega - \omega_0) + g(\omega + \omega_0)$ と定めると，W の取り方により h は 0 ではなく，また

$$h(\omega) = \int e^{\sqrt{-1}\omega^T x} 2\cos(\omega_0^T x) d\mu(x)$$

を得る．さらに $\mathrm{Supp}(g) \subset \mathrm{cl}(W - W)$ より，$\mathrm{Supp}(h) \cap \mathrm{Supp}(\Lambda) = \emptyset$ である．ここで，$\pm\omega_0 \notin W - W$ より $h(0) = 0$ であるから，$2\cos(\omega_0^T x)\mu$ を実の符号付測度とみなすと

$$(2\cos(\omega_0^T x)\mu)(\mathbb{R}^n) = 0$$

が成立する．h が 0 でないことから $(2\cos(\omega_0^T x)\mu)$ は零測度ではない．そこで，$c = |2\cos(\omega_0^T x)\mu|(\mathbb{R}^n)$（$|\cdot|$ は全変動）とおき，2 つ異なる確率測度 μ_1, μ_2 を

$$\mu_1 = \frac{1}{c}|2\cos(\omega_0^T x)\mu|, \qquad \mu_2 = \frac{1}{c}\{|2\cos(\omega_0^T x)\mu| - 2\cos(\omega_0^T x)\mu\}$$

により定義する．このときフビニの定理から，

$$\begin{aligned} c((\mu_1 - \mu_2) * \phi)(x) &= \int \phi(x - y) 2\cos(\omega_0^T y) d\mu(y) \\ &= \int 2\cos(\omega_0^T y) \int e^{\sqrt{-1}(x-y)^T\omega} d\Lambda(\omega) d\mu(y) \\ &= \int e^{x^T\omega} \int \overline{\{e^{\sqrt{-1}y^T(\omega-\omega_0)} + e^{\sqrt{-1}y^T(\omega+\omega_0)}\}} d\mu(y) d\Lambda(\omega) \\ &= \int e^{x^T\omega} h(\omega) d\Lambda(\omega) \end{aligned}$$

となるが，$\mathrm{Supp}(h) \cap \mathrm{Supp}(\Lambda) = \emptyset$ により $(\mu_1 - \mu_2) * \phi = 0$ を得る．これは $k(x, y)$ が特性的であることに矛盾する． ∎

定理 9 により，ガウス RBF カーネル $k_\sigma^G(x, y) = \exp\{-\|x-y\|^2/(2\sigma^2)\}$ やラプラスカーネル $k_\lambda^L(x, y) = \exp\left(-\lambda \sum_{i=1}^m |x_i - y_i|\right)$ が $\mathbb{R}^m$ 上の特性的なカーネルであることがわかる．実際，$\exp\{-\|x-y\|^2/(2\sigma^2)\}$ と $\exp\{-\lambda \sum_{i=1}^m |x_i - y_i|\}$ のフーリエ変換は，正の定数倍を除いてそれぞれ $\exp\{-\sigma^2\|x - y\|^2/2\}$ と $\prod_{i=1}^m 1/(u_i^2 + \lambda^2)$ であり，至るところ正である．

参考文献

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 69(3):337–404, 1950.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2nd edition, 2002.

A. Frank and A. Asuncion. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences. 2010.

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

K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8: 361–383, 2007.

K. Fukumizu, B.K. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic Kernels on Groups and Semigroups. Advances in Neural Information Processing Systems 21, 473–480, MIT Press, 2009.

K. Fukumizu, F. R.Bach, and M. I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*. 37(4), 1871–1905, 2009.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' Rule. *Advances in Neural Information Processing Systems 25*, 2012 to appear.

K. Fukumizu and C. Leng. Gradient-based kernel dimension reduction for supervised learning. arXiv:1109.0455v1 [stat.ML], 2011.

F. Girosi. An equivalence between sparse approximation and support vector machine. *Neural Computation*, 10:1455–1480, 1998.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and

T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A Fast, Consistent Kernel Two-Sample Test. *Advances in Neural Information Processing Systems 22*, 673–681. MIT Press, 2009.

A. Gretton, K. Fukumizu and B. Sriperumbudur. Discussion of: Brownian distance covariance. *Annals of Applied Statistics* 3(4), 1285–1294. 2009.

K.-C. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of American Statistical Association*, 86:316–342, 1991.

K.-C. Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of American Statistical Association*, 87:1025–1039, 1992.

M. Reed and B. Simon. *Functional Analysis*. Academic Press, 1980.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

X. Sun, D. Janzing, B. Schölkopf, and K. Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of International Conference on Machine Learning*, pages 855–862, 2007.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

赤穂昭太郎. **カーネル多変量解析**. 岩波書店, 2008.

福水健次. **カーネル法入門**. 朝倉書店, 2010.

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2011 年 10 月

第10章　グラフマイニングとその統計的モデリングへの応用

鷲尾　隆
(大阪大学産業科学研究所)

統計的モデリングを多変数大規模なデータに適用すると，どの程度複雑なモデルを用いるべきか，データの性質を反映する安定したモデル構造はどのようなものであるか，というようなモデル選択の問題に直面する．この問題を乗り越えるために変数間の依存関係を表すグラフ構造探索を行おうとすると，今度は探索の組み合わせ爆発に直面してしまう．筆者は，これまでにデータ中に見て取れるグラフ構造の探索・発掘アルゴリズムを研究してきた．ここではその成果を大規模な遺伝子発現データの統計的モデリングに適用し、モデル選択の困難さ軽減する解析例を示す．

1　はじめに

データマイニングは，膨大なデータからその部分的な特徴を計算機を用いて探索する技術である．最近では，データがHTMLのような入れ子構造を持ったテキスト（半構造テキスト）である場合や，木，記号系列，グラフ，論理関係式など，複雑な構造を持つものにまで，適用が拡大されつつある．その中でも特にグラフは，数学的に基本的な構造であり，かつ統計数理や機械学習においても，グラフィカルモデリングやベイジアンネットワーク，ニューラルネットワークなどのように，頻繁に用いられる構造である．更に生物学や化学，材料化学，社会通信ネットワークなど様々な実分野で，グラフ構造を持つデータが幅広く扱われている．しかし一方で，膨大なグラフデータの中から特徴的な部分構造を見つける問題の多くが，本質的に非常に大きな計算量を必要とすることが知られている．たとえば，ある大きなグラフが別のより小さなグラフを部分グラフとして含むか否かを調べる部分グラフ同型問題は，NP-完全問題という困難な問題であることが知られている (Garey and Johnson (1979))．このような背景から，近年，グラフ構造データを対象として特徴的な部分構造を効率的に発掘するグラフマイニング手法が盛んに研究されるようになった．

グラフマイニング研究の発端は，大規模なグラフから何らかの基準によって特徴的な部分グラフを発見的に探索するアルゴリズムの研究であった．これら代表的研究としては，1990年代半ばのCookとHolderによるSUBDUE (Cook and Holder (1994)) 及び吉田等によるGBI (Yoshida et al. (1994)) が挙げられる．1998年になってDehaspeとToivonenは，多数のグラフの集合に多頻度で現れる部分グラフを，網羅的に探索（完全探索）することを目指すWARMRを発表した (Dehaspe and Toivonen (1998))．2000年には猪口等がAprioriというデータマイニングアルゴリズムをグラフ理論によって拡張し，高速に多頻度部分グラフの完全探索を行うAGMを発表した (Inokuchi et al. (2000))．これらの先駆的研究の後，グラフマイニング研究は急速に盛んになった．

本章では，以下にグラフマイニングを理解する上で重要な基礎概念として，部分グラフ，同型問題，グラフ不変量，マイニング基準を説明する．その後，グラフマイニングで必要とされる探索原理とそれを用いる代表的手法について解説する．最後に，別章で解説しているベイジアンネットワークを用いた遺伝子発現因果関係に関する統計的モデリングに，本グラフマイニングを組み合わせることで，大規模次元データの統計的モデリングの可能性を示唆する．

2　グラフマイニングの基礎

グラフマイニングの背景には，グラフ理論や探索理論に関する豊富な研究が存在している．ここでは，多くのグラフマイニング手法を理解する上で必要となる幾つかの基礎原理を，ラベル付き無向グラフの場合について説明する．グラフがラベル付きであるとは，グラフが複数種類の頂点や辺により構成され，それぞれ種類に応じて識別ラベルが付いていることである．グラフが無向であるとは，グラフの各辺が結ぶ２頂点間に矢印で表される順序が無いことである．尚，説明は省略するが，ここで述べる基礎原理は，有向グラフやラベル無しグラフについても成り立つ.

2.1　一般部分グラフと誘導部分グラフ

１つのグラフは，それを構成する頂点の集合V，同じくそれら頂点のペアを結ぶ辺の集合E，辺による頂点の接続関係を表す関数$f : E \to V \times V$，頂点や辺にラベルを付与する関数ℓの４項組$G(V, E, f, \ell)$で表される．頂点のラベル集合をL_v，辺のラベル集合をL_eとした時，ℓは更に頂点をラベル付けする関数$\ell_v : V \to L_v$と辺をラベル付けする関数$\ell_e : E \to L_e$の２項組$\ell(\ell_v, \ell_e)$で表される．例えば，図1 (a)に示されるグラフでは，$V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$，$E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9\}$ となる．Eに含まれる各辺e_hは，Vに含まれるv_iとv_jを$f(e_h) = (v_i, v_j)$によって関連づける．図の場合には，例えば$f(e_1) = (v_1, v_2)$，$f(e_2) = (v_1, v_2)$，$f(e_4) = (v_1, v_4)$，$f(e_7) = (v_4, v_4)$となる．また，Vに含まれる各頂点v_i，Eに含まれる各辺e_hは，それぞれℓ_vとℓ_eによってラベル$\ell_v(v_i)$と$\ell_e(e_h)$を有する．

グラフ$G(V, E, f, \ell)$の“一般部分グラフ”$G_s(V_s, E_s, f_s, \ell_s)$は，以下の条件を満たすグラフである．

(1) $V_s \subset V$ かつ $E_s \subset E$ である，

(2) すべての$v_i \in V_s$について，$\ell_{sv}(v_i) = \ell_v(v_i)$である，

(3) すべての$e_h \in E_s$について，
$f_s(e_h) = (v_i, v_j)$ かつ $\ell_{se}(e_h) = \ell_e(e_h)$ である $v_i, v_j \in V_s$ が存在する．

条件 (1) は，単にグラフG_sの頂点と辺が元のグラフGの頂点と辺の一部であることを示している．条件 (2) は，G_sの各頂点とそれが対応するGの各頂点は同じラベルを持つこと，同様に条件 (3) は，G_sの各辺とそれが対応

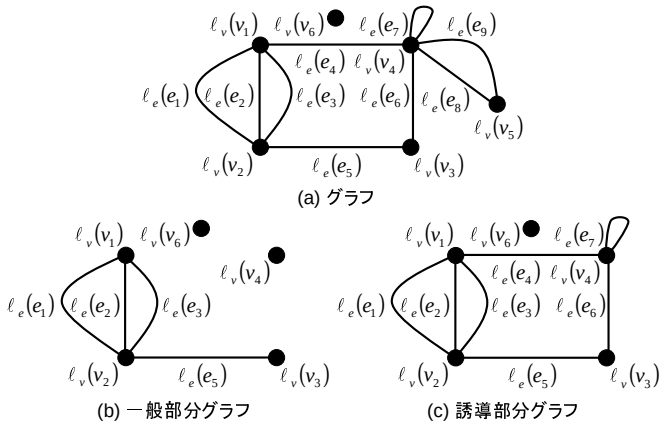


図 1: グラフと部分グラフの例.

する G の各辺が同じラベルを持つこと示している．より直感的に言えば，G の一部の頂点とその間の一部の辺を切り取って得られるグラフが，一般部分グラフ G_s である．図 1 (b) は，元グラフ (a) から頂点 v_1, v_2, v_3, v_4, v_5 及びそれらを結ぶ辺から e_1, e_2, e_3, e_5 のみを切り取って得た一般部分グラフの例である．

もう 1 つの代表的な部分グラフは “誘導部分グラフ”$G_s(V_s, E_s, f_s, \ell_s)$ であり，上記一般部分グラフの条件に加えて以下の条件を満たすものである．

(4) $f(e_h) = (v_i, v_j)$ かつ $v_i, v_j \in V_s$ であるすべての $e_h \in E$ について, $e_h \in E_s$ が存在する.

この条件は，G の辺でその両端の頂点が G_s の頂点に対応するものは，必ず G_s にも含まれることを示している．図 1 (c) は，元グラフ (a) から頂点 v_1, v_2, v_3, v_4, v_5 を選んだ誘導部分グラフの例である．選ばれなかった v_5 に直接繋がる辺 e_8 と e_9 (c) には含まれないが，(b) と異なり元グラフ G の v_1, v_3, v_4 間に存在する辺 e_4, e_6, e_7 は含まれる．G_s が G の一般ないしは誘導部分グラフであることを，ここでは $G_s \subseteq G$ と表すことにする．

2.2 部分グラフ同型問題

グラフ理論では，あるグラフが他のグラフの部分グラフであることを部分グラフ同型という．一方，グラフマイニングは多数のグラフに共通して現

れる部分グラフを探索するので，ここでは多数のグラフに対して部分グラフ同型である部分グラフを求める問題を，“部分グラフ同型問題” と定義する．すなわち，n 枚のグラフからなる集合 $D = \{G_d(V_d, E_d, f_d, \ell_d)|d = 1, ..., n\}$ が与えられた時，すべての $G_d(V_d, E_d, f_d, \ell_d) \in D$ について，一般ないしは誘導部分グラフの意味で $G_s(V_s, E_s, f_s, \ell_s) \subseteq G_d(V_d, E_d, f_d, \ell_d)$ であるグラフ $G_s(V_s, E_s, f_s, \ell_s)$ を見つける問題を，“部分グラフ同型問題” とする．この時，各 $d = 1, ..., n$ に対して，G_s の $e_{sh} \in E_s$ 各々が結ぶ頂点 $f_s(e_{sh}) = (v_{si}, v_{sj})$ を，e_{sh} に対応する G_d の $e_{dh} \in E_d$ が結ぶ頂点 $f_d(e_{dh}) = (v_{di}, v_{dj})$ に対応付ける，すなわち $v_{di} = g_{sd}(v_{si}), v_{dj} = g_{sd}(v_{sj})$ を満たす V_s から V_d $(d = 1, ..., n)$ への一対一写像を g_{sd} とする．例えば，図 1 のグラフ (b) と (c) からなる $D = \{(b), (c)\}$ について，部分頂点集合 $V_s = \{v_{s1}, v_{s2}, v_{s3}\}$ と部分辺集合 $E_s = \{e_{s1}, e_{s5}\}$ からなる部分グラフ $G_s(V_s, E_s, f_s, \ell_s)$ は一般部分グラフ同型である．また，一対一写像 g_{sd} は，$v_{(b)1} = g_{s(b)}(v_{s1}), v_{(b)2} = g_{s(b)}(v_{s2}), v_{(b)3} = g_{s(b)}(v_{s3})$,, $v_{(c)1} = g_{s(c)}(v_{s1}), v_{(c)2} = g_{s(c)}(v_{s2}), v_{(c)3} = g_{s(c)}(v_{s3})$, となる．更に，部分辺集合が $E_s = \{e_{s1}, e_{s2}, e_{s3}, e_{s5}\}$ である場合には，部分グラフ $G_s(V_s, E_s, f_s, \ell_s)$ は $D = \{(b), (c)\}$ について誘導部分グラフ同型である．1 つの小さなグラフが 1 つのより大きなグラフの部分かどうかの判定は，グラフの大きさに対して急激に必要計算量が増大する NP-完全と呼ばれる問題があることが判っている (Garey and Johnson (1979))．従って，複数グラフ間の部分グラフ同型問題も，NP-完全より必要計算量が少ないことはあり得ず，極めて効率的な計算アルゴリズムを用いないと実用時間内に計算処理を終えることが困難である．

2.3 正準ラベルと正準形

2 つのグラフ $G_i(V_i, E_i, f_i, \ell_i)$ と $G_j(V_j, E_j, f_j, \ell_j)$ が，互いに $G_i \subseteq G_j$ かつ $G_i \supseteq G_j$ である時，それらは同型であるという．すなわち，2 つの同型なグラフは，両者の間で頂点同士かつ辺同士が互いに漏れなく一対一対応して同じラベルを有している．同型なグラフ同士ならば，頂点数や各頂点に接続する辺の数（線度），閉路の数などは等しい．このように同型なグラフを特徴付ける量をグラフ不変量という。しかし，一般には不変量値が等しくても同型なグラフとは限らない．これに対して，グラフ構造を正確に反映する特殊な不変量として，“正準ラベル” がある．同型なグラフは等しい正準ラベルを持ち，正準ラベルが等しければ同型なグラフとなる．様々な正準ラベルの定義が可能であるが，ここではグラフの隣接行列に基づく定

義を説明する．

グラフ G の i-番目の頂点 v_i を i-番目の行と列に対応させ，要素によって頂点間の辺の接続関係を表した行列を “隣接行列” という (Inokuchi et al. (2000))．隣接行列の i,j-要素は，辺のラベル集合 $E_{i,j} = \{\ell_e(e_h)|f(e_h) = (v_i, v_j)$ であるすべての $e_h \in E\}$ で表される．この隣接行列は，厳密には要素が数ではないので行列ではないが，ここでは都合上行列と呼ぶ．頂点 v_i と v_j 間に辺が存在しない場合，要素は 0 とする．以下は図 1 (a) の隣接行列の 1 例である．

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 0 & \{\ell_e(e_2),\ell_e(e_3),\ell_e(e_4)\} & 0 & \{\ell_e(e_1)\} & 0 & 0 \\ \{\ell_e(e_2),\ell_e(e_3),\ell_e(e_4)\} & 0 & \{\ell_e(e_5)\} & 0 & 0 & 0 \\ 0 & \{\ell_e(e_5)\} & 0 & \{\ell_e(e_6)\} & 0 & 0 \\ \{\ell_e(e_1)\} & 0 & \{\ell_e(e_6)\} & \{\ell_e(e_7)\} & \{\ell_e(e_8),\ell_e(e_9)\} & 0 \\ 0 & 0 & 0 & \{\ell_e(e_8),\ell_e(e_9)\} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{2.1}$$

隣接行列の行同士や列同士を入れ替えても，行及び列に対応する頂点も一緒に入れ替えて，行や列と頂点の対応を変えなければ同じグラフを表す．即ち，G を表す隣接行列は多数存在する．

以上の隣接行列の各行及び列には，関数 $\ell_v : V \to L_v$ によって頂点ラベルが付与されているが，ここで互いに異なる頂点ラベルを異なる整数で表す．すなわち，i 番目の頂点 v_i に対応する行及び列のラベル $\ell_v(v_i)$ にある整数 ℓ_i を，$\ell_v(v_i) \to \ell_i$ と対応付けることにする．また，隣接行列の各要素に示される辺には，関数 $\ell_e : E \to L_e$ によって辺ラベルが付与されているが，同じく互いに異なる辺ラベルを異なる整数で表すことにする．これによって，各 i,j -要素の辺ラベル集合 $E_{i,j}$ にある整数 $x_{i,j}$ を，$E_{i,j} \to x_{i,j}$ と対応付けることにする．例えば上の隣接行列の場合，v_1, v_3 同士，v_2, v_4 同士，e_1, e_5 同士のラベルが等しく，頂点ラベルに適当に $\ell_v(v_6) \to 1$, $\ell_v(v_1) = \ell_v(v_3) \to 2$, $\ell_v(v_2) = \ell_v(v_4) \to 3$, $\ell_v(v_5) \to 4$ と整数を割り当て，辺ラベルからなる集合である各要素にも適当に $\{\ell_e(e_1)\} = \{\ell_e(e_5)\} \to 1$, $\{\ell_e(e_6)\} \to 6$, $\{\ell_e(e_7)\} \to 7$, $\{\ell_e(e_8),\ell_e(e_9)\} \to 89$, $\{\ell_e(e_2),\ell_e(e_3),\ell_e(e_4)\} \to 234$ と整数を割り当てると，以下の隣接行列に変換される．

$$\begin{array}{c} \\ 2 \\ 3 \\ 2 \\ 3 \\ 4 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccccc} 2 & 3 & 2 & 3 & 4 & 1 \end{array} \\ \begin{pmatrix} 0 & 234 & 0 & 1 & 0 & 0 \\ 234 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 6 & 0 & 0 \\ 1 & 0 & 6 & 7 & 89 & 0 \\ 0 & 0 & 0 & 89 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

そして，G の $n \times n$ 隣接行列の各行または列の頂点ラベル $\ell_v(v_i)$ の整数 ℓ_i と各 i,j-要素の整数 $x_{i,j}$ から，以下のようなコードを定義する．

$$code(G) = x_{1,1}x_{1,2}x_{2,2}x_{1,3}x_{2,3}x_{3,3}\cdots x_{1,n}\cdots x_{n-1,n}x_{n,n}$$

$$CODE(G) = \ell_1 \cdots \ell_n code(G)$$

$code(G)$ の部分は，無向グラフの隣接行列の対角対称性より，上三角部分の要素のみで表される．上記の例では，

$$CODE(G) = 2323410\{234\}00101067000\{89\}0000000$$

となる．ここで，頂点ラベル及び辺ラベル集合の整数の最大値 N を基数として，上記コードを N 進法の数字と見なすことにする．上述のようにグラフ G を表す隣接行列やそれに対応するコードは多数あるが，その中で N 進法の数字が最小（あるいは最大）のコードを，グラフ G の正準ラベルという．そして，そのコードに対応する隣接行列を “正準形” と呼ぶ．図 1 (a) のグラフ G の正準ラベルとその正準形は以下のようになる．

$$CODE(G) = 12233400000001670\{234\}100000\{89\}00$$

$$\begin{array}{c} \\ v_6 \\ v_1 \\ v_3 \\ v_4 \\ v_2 \\ v_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} v_6 & v_1 & v_3 & v_4 & v_2 & v_5 \end{array} \\ \left(\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \{\ell_e(e_1)\} & \{\ell_e(e_2),\ell_e(e_3),\ell_e(e_4)\} & 0 \\ 0 & 0 & 0 & \{\ell_e(e_6)\} & \{\ell_e(e_5)\} & 0 \\ 0 & \{\ell_e(e_1)\} & \{\ell_e(e_6)\} & \{\ell_e(e_7)\} & 0 & \{\ell_e(e_8),\ell_e(e_9)\} \\ 0 & \{\ell_e(e_2),\ell_e(e_3),\ell_e(e_4)\} & \{\ell_e(e_5)\} & 0 & 0 & 0 \\ 0 & 0 & 0 & \{\ell_e(e_8),\ell_e(e_9)\} & 0 & 0 \end{array} \right) \end{array}. \tag{2.2}$$

式 (2.1)，(2.2) の隣接行列は，同じグラフを表している．正準ラベルは最小（あるいは最大）のコードなので，G について一意でありグラフ不変量である．従って，隣接行列の正準形も G について一意に与えられる．正準ラベルと正準形によって，グラフ表現の多様性や部分グラフ同型問題の探索空間は著しく削減される．

2.4 マイニングの基準

データマイニングでは，ある基準を満たすデータ部分に着目して部分的特徴を発掘する．どのような部分的特徴に着目するかによって，用いられる基準は様々である．代表的基準として，ある商品（アイテム）集合 a が

スーパーマーケットの各顧客の購入商品（アイテム）集合 t からなるデータ D に現れる出現頻度（支持度）

$$sup(a) = \frac{|\{t|t \in D, a \subseteq t\}|}{|D|}$$

が，ある閾値（最小支持度）$minsup$ 以上であること

$$sup(a) \geq minsup$$

が挙げられる．この条件を満たす a を多頻度アイテム集合という．a が多頻度なら，その任意の部分集合 $a'(\subseteq a)$ も多頻度である．代表的なデータマイニング手法であるバスケット分析は，この基準を満たすすべての a を発掘する (Agrwal and Srikant (1994))．

グラフマイニングにおいても，ラベル付きの頂点及び辺からなる多数のグラフの集まりであるデータ $D = \{G_d(V_d, E_d, f_d, \ell_d)|d = 1, ..., n\}$ が与えられた時，D におけるある部分グラフ G_s の出現頻度（支持度）を以下のように定義する．

$$sup(G_s) = \frac{|D_s|}{|D|}$$

ここで，D_s は D の中で G_s が一般ないし誘導部分グラフ同型であるグラフの集合 $D_s = \{G_d|G_d \in D, G_s \subseteq G_d\}$ である．スーパーマーケットの例と同様に，支持度の最小閾値（最小支持度）を $minsup$ とした時，G_s が多頻度である，すなわち

$$sup(G_s) \geq minsup$$

を満たす全ての G_s を発掘する．この条件を満たす G_s を多頻度部分グラフという．D において G_s が多頻度なら，その任意の部分グラフ $G'_s(\subseteq G_s)$ もやはり多頻度である．

3　グラフマイニングの探索原理

前節で述べたように，ここで説明するグラフマイニングでは，多数のグラフの集まりであるデータ $D = \{G_d(V_d, E_d, f_d, \ell_d)|d = 1, ..., n\}$ から，多頻度な一般ないし誘導部分グラフ G_s をすべて探索する．このような G_s を探索する単純な方法は，D のグラフが含みうる可能なすべての部分グラフについて，多頻度か否かを調べることである．しかし，例えば頂点数が８つの部分グラフに限っても，ラベルの付け方を無視したとしても，頂点間の辺

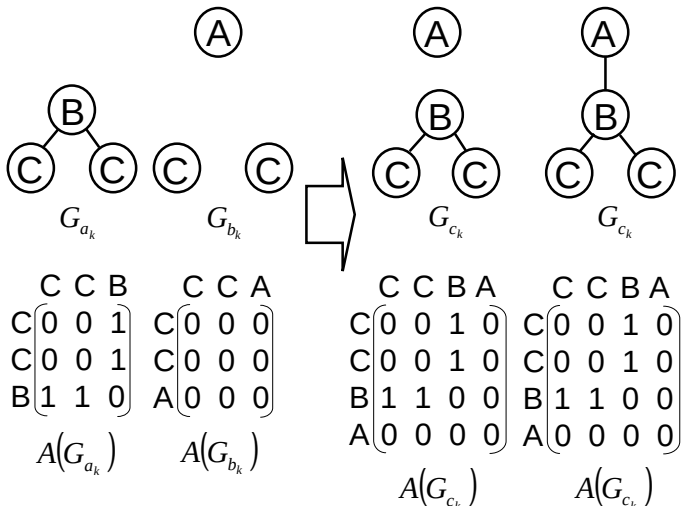


図 2: ２つの部分グラフの結合例.

の有無の組み合わせは $2^{8C_2}(=2^{28})$ 通りも存在する．しかも，既に述べたように個々の部分グラフの同型性判定は，NP-完全もしくはそれ以上に困難な問題である．このような探索上の組み合わせ爆発を回避するためには，効率的なアルゴリズムを用いなければならない.

本節では効率的なアルゴリズムの 1 つとして，データマイニングの代表的手法であるバスケット分析の Apriori アルゴリズムをグラフに拡張したものを説明する．このアルゴリズムは，多頻度グラフの任意の部分グラフも多頻度である性質を効果的に利用する．ある部分グラフ G_{c_k} が多頻度ならば，それから頂点を 1 つ除いて得られる部分グラフ G_{a_k} も多頻度である．同じく，G_{c_k} から別の頂点を 1 つ除いて得られる部分グラフ G_{b_k} も多頻度である．それならば例えば，図 2 の左側に示すような各々 1 つの頂点を除いて共通な 2 つの部分グラフ G_{a_k}, G_{b_k} がそれぞれ多頻度である時，それらの共通部分を重ねて結合した右側の 1 頂点多い部分グラフ G_{c_k} も多頻度である可能性が高い．ただし，G_{c_k} には左側で共通でなかった頂点 A と B を，辺で結ぶ場合と結ばない場合の２通りが存在する．これらより，右側の部分グラフを候補として，多頻度部分グラフであるか否かを調べる．このように

探索候補を絞り込むことで，単純しらみ潰しに候補を生成して調べるよりも，遥かに高速に多頻度部分グラフを完全探索可能となる．

以上を多頻度誘導部分グラフを導出する場合について，より詳細に説明する．今，１つの頂点を除いて共通な，大きさが k の２つの多頻度誘導部分グラフ G_{a_k}, G_{b_k} が知られているとする．そして，これらを表す隣接行列 $A(G_{a_k}), A(G_{b_k})$ は，$(k-1)\times(k-1)$ 左上部分行列が両者の共通部分を表しており同じであるとする．例えば，図２の左側の２つの隣接行列がこれに相当する．そこで，G_{a_k} と G_{b_k} の共通部分を重ねて結合して，多頻度誘導部分グラフ候補 G_{c_k} を得ることを考える．この時，計算機内では直接に隣接行列を操作するのではなく，前述のコード表現で以下のように結合することで高速かつメモリーを節約した処理が可能になる．

$$\begin{aligned} CODE(G_{a_k}) &= \ell_1\cdots\ell_{k-1}\ell_k x_{1,1}x_{1,2}x_{2,2}x_{1,3}x_{2,3}x_{3,3}\cdots x_{1,k}\cdots x_{k-1,k}x_{k,k} \\ CODE(G_{b_k}) &= \ell_1\cdots\ell_{k-1}\ell'_k x_{1,1}x_{1,2}x_{2,2}x_{1,3}x_{2,3}x_{3,3}\cdots x'_{1,k}\cdots x'_{k-1,k}x'_{k,k} \end{aligned} \tag{3.3}$$

$$\begin{aligned} CODE(G_{c_{k+1}}) &= CODE(G_{a_k})\cup CODE(G_{b_k}) \\ &= \ell_1\cdots\ell_{k-1}\ell_k\ell'_k \\ &\quad x_{1,1}x_{1,2}x_{2,2}\cdots x_{1,k}\cdots x_{k-1,k}x_{k,k}x'_{1,k}\cdots x'_{k-1,k}z_{k,k+1}x'_{k,k} \end{aligned} \tag{3.4}$$

但し $CODE(G_{a_k})$ は正準ラベルであり，$CODE(G_{a_k}) \le CODE(G_{b_k})$ とする．これは同一のグラフを表すコード間の結合や同じコード組の異なる順序での結合といった冗長な結合を避けるためである．ここで，$CODE(G_{c_{k+1}})$ には，最後の $x'_{k,k}$ の前に新たな要素 $z_{k,k+1}$ が挿入されている．この要素は G_{a_k} と G_{b_k} の k 番目の頂点間の辺ラベルを表す．各コードに対応する隣接行列 $A(G_{a_k}), A(G_{b_k}), A(G_{c_{k+1}})$ は以下のようになる．

$$A(G_{a_k}) = \begin{pmatrix} X_{k-1} & \boldsymbol{x}_1 \\ \boldsymbol{x}_2^T & x_{k,k} \end{pmatrix},\quad A(G_{b_k}) = \begin{pmatrix} X_{k-1} & \boldsymbol{x}'_1 \\ \boldsymbol{x}_2'^T & x'_{k,k} \end{pmatrix},$$

$$A(G_{c_{k+1}}) = \begin{pmatrix} X_{k-1} & \boldsymbol{x}_1 & \boldsymbol{x}'_1 \\ \boldsymbol{x}_2^T & x_{k,k} & z_{k,k+1} \\ \boldsymbol{x}_2'^T & z_{k+1,k} & x'_{k,k} \end{pmatrix}.$$

ここで X_{k-1} は G_{a_k} と G_{b_k} に共通する大きさ $k-1$ のグラフを表す隣接行列であり，$\boldsymbol{x}_i$ と $\boldsymbol{x}'_i (i=1,2)$ は $(k-1)\times 1$ の列ベクトルである．$z_{k,k+1}$ と $z_{k+1,k}$ の値は無向グラフの場合には対称性より同一であるが，元の $A(G_{a_k})$，$A(G_{b_k})$

からは決まらず２つの場合が考えられる．１つは結合して得られるグラフ $G_{c_{k+1}}$ の k 番目と $k+1$ 番目の頂点の間にラベル $\{\ell_e(e_i)|f_{c_{k+1}}(e_i)=(v_k,v_{k+1})\}$ を持つ辺を付加する場合，もう１つはそれらの頂点間に辺を付加しない場合である．これによって $z_{k,k+1}$ と $z_{k+1,k}$ が “$\{\ell_e(e_i)|f_{c_{k+1}}(e_i)=(v_k,v_{k+1})\}$” か “0” である複数通りの隣接行列が，多頻度誘導部分グラフ候補として生成される．図２の例では，右側の２つの多頻度誘導部分グラフが候補である．このようにして得られた $CODE(G_{c_{k+1}})$ に対応する隣接行列を，グラフ $G_{c_{k+1}}$ の “正規形” 表現という．

以上の結合操作により，逐次的に頂点数の多い多頻度誘導部分グラフを効率的に完全探索することが可能になる．図３に逐次処理による多頻度誘導部分グラフのマイニングアルゴリズムを示す．初期化のステップに示す頂点数１の多頻度誘導部分グラフ（孤立した１個の頂点のみからなるグラフ）のコード集合 $FCODE(1)$ からはじめて，ステップ１に示す上記２つの多頻度誘導部分グラフの結合によって，ボトムアップ的に頂点数 k の多頻度誘導部分グラフから頂点数 $k+1$ の多頻度誘導部分グラフの候補を表すコードの集合 $CFCODE(k+1)$ を作り出す．仮に $G_{c_{k+1}}$ が多頻度誘導部分グラフであるなら，その部分である G_{a_k} や G_{b_k} も多頻度誘導部分グラフである．頂点数 k の多頻度誘導部分グラフが既に全て見つかっているならば，G_{a_k} と G_{b_k} も必ず既に見つかっている．即ち，今見つかっている頂点数 k の多頻度誘導部分グラフの内，式 (3.3) を満たすグラフの全ての組み合わせについて式 (3.4) の結合をとれば，漏れなく多頻度誘導部分グラフ候補を得ることができる．

ただし，上記の結合のみでは，不要な多頻度誘導部分グラフ候補が生成されてしまう．$G_{c_{k+1}}$ から１つの頂点を除去した G_{a_k} や G_{b_k} は多頻度であるが，他の頂点を除去して得られる G_{a_k}, G_{b_k} 以外の頂点数 k の部分グラフの中に多頻度ではないものが存在すれば，$G_{c_{k+1}}$ は明らかに多頻度ではない．その場合には，$G_{c_{k+1}}$ が実際に D において多頻度か否かをチェックする必要はない．そこで，このような多頻度ではない部分グラフが存在するかを図３のステップ２において確認し，存在すれば $G_{c_{k+1}}$ を候補としない．具体的には $A(G_{c_{k+1}})$ について，それから i 行 i 列 $(i=1,\cdots,k-1)$ を除去した $k\times k$ の各部分行列が，全て既に探索された多頻度誘導部分グラフを表す場合にのみ，$G_{c_{k+1}}$ を多頻度誘導部分グラフ候補として残す．こうして得た最終候補 $G_{c_{k+1}}$ のコード集合についてのみ，ステップ３において D にアクセスし

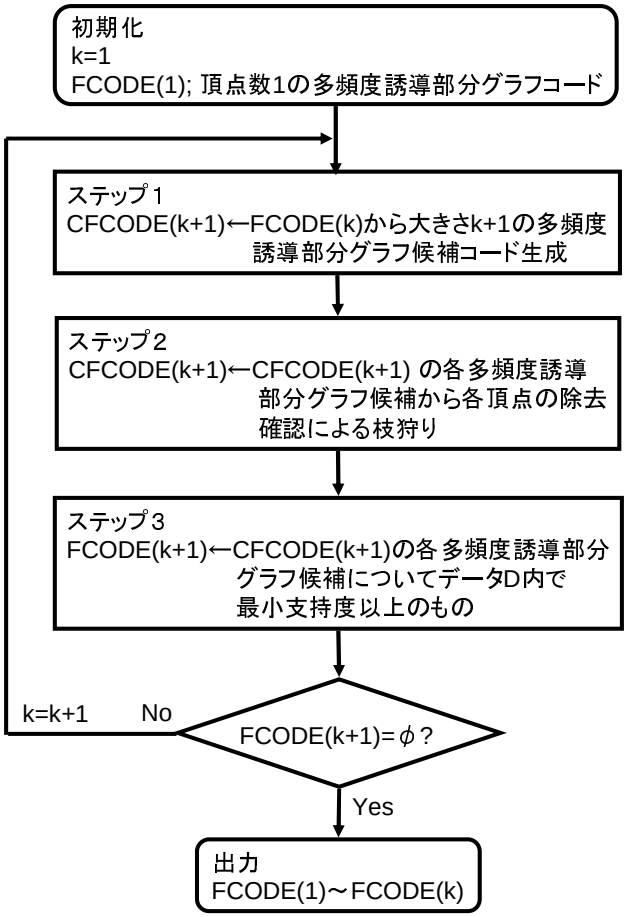


図 3: 多頻度部分グラフマイニングのアルゴリズム.

て多頻度か否かをチェックする．このようにして頂点数 $k+1$ の多頻度誘導部分グラフの候補全てについて 1 回のデータ D のスキャンで支持度を計算し，最小支持度を越えるものを多頻度誘導部分グラフとする．更に k を更新して上記を繰り返す．より多くの頂点からなる部分グラフの支持度は減少するので，$minsup$ 以上の多頻度誘導部分グラフは存在しなくなり，図 3 の最後に示すようにこのアルゴリズムは停止する．データに存在する最も大きな多頻度誘導部分グラフの頂点数を k_{max} とすると，高々 $k_{max}+1$ 回のデータ D のスキャンで，全ての多頻度誘導部分グラフが得られる．

4　統計的モデリングへの応用

多数の観測変数の関係をモデル化する統計的方法には，ベイジアンネットワークに代表されるようにグラフ構造を有するモデルを用いるものが多い．このような統計的モデル化手法にグラフマイニングを組み合わせることで，様々な解析ができる可能性がある．ここではその例として，別章で説明しているマイクロアレイ遺伝子発現プロフィールデータから遺伝子発現程度の関係をベイジアンネットワークで同定した結果に，更にグラフマイニングを適用して各遺伝子発現の因果関係を調べる解析を紹介する．これはベイジアンネットワークによるモデル化の後処理としてグラフマイニングを適用し，対象とする多変数間の因果関係をより明確に把握する試みである．

4.1　遺伝子発現データとベイジアンネットワークモデリング

細胞内の DNA 鎖には多数の遺伝子 (gene) がコードされているが，一般にある遺伝子の発現程度は，他の遺伝子の発現程度によって刺激ないし抑制の影響を受けることが判っている．そこで，細胞内の各遺伝子の発現程度を実験的に測定し，その間の発現の因果関係を統計的にベイジアンネットワークによってモデル化し，各遺伝子の機能や遺伝子集団として働きを明らかにする研究が進んでいる．詳細は別章の説明に譲るが，ここではその概要を簡単に紹介する．

生体の様々な状態にある細胞を複数採取し，その細胞核内物質を DNA マイクロアレイと呼ばれる測定器具で分析することで，各遺伝子が生成する固有のタンパク質の濃度を調べることができる．基準となる細胞に含まれる各固有タンパク質濃度と測定対象とする細胞の同タンパク質濃度の対数比を求めることで，そのタンパク質に対応する遺伝子の基準状態に対する相対的な発現程度を知ることができる．このようなデータをマイクロアレイデータという．今，図 4 に示すように，各マイクロアレーが p 個の遺伝子の発現程度を測定し，そのようなマイクロアレイデータが n 個あるものとする．$i(=1,\cdots,n)$ 番目のマイクロアレイデータの $j(=1,\cdots,p)$ 番目の遺伝子 $gene_j$ に対応する発現程度の測定値を x_{ij} とすると，i 番目のマイクロアレイデータは p 次元ベクトル $\boldsymbol{x}_i = (x_{i1},\cdots,x_{ip})$ で表され，n 個のマイクロアレイデータ全体の集合は $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ で表される．

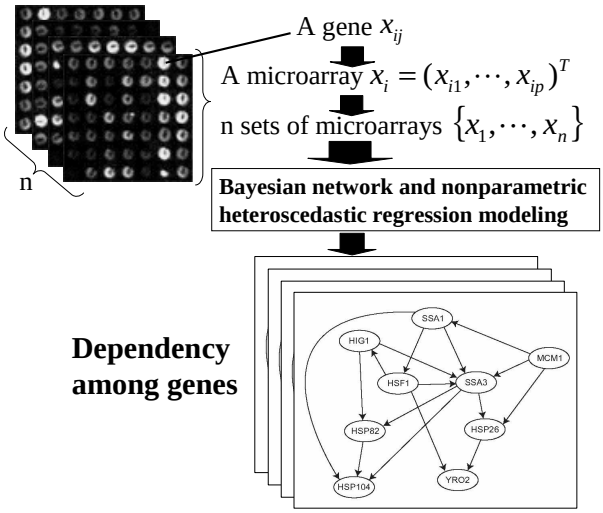


図 4: マイクロアレイデータと遺伝子依存関係モデリング.

このデータから，ある遺伝子の発現程度に対する他の遺伝子の発現程度の影響を表すモデルを導く方法として，ベイジアンネットワークノンパラメトリック加法回帰モデルを用いる (Hastie and Tibshirani (1990), Imoto et al. (2002a))．これは i 番目のマイクロアレイデータにおいて，ある遺伝子 $gene_j$ の発現程度 x_{ij} に直接影響する親遺伝子 $gene_1^{(j)}, \cdots, gene_{q_j}^{(j)}$ の発現程度を $p_{ik}^{(j)} (= 1, \cdots, q_j)$ とした時，それらの関係を

$$x_{ij} = m_{j1}(p_{i1}^{(j)}) + \cdots + m_{jq_j}(p_{iq_j}^{(j)}) + \epsilon_{ij}$$

という回帰式で近似するモデルである．誤差 ϵ_{ij} は平均 0，分散 σ_j を持つ正規分布に独立に従う．また，$m_{jk}(\cdot)$ は，親遺伝子の発現程度と対象とする遺伝子の発現程度の非線形な関係を表す平滑化関数と呼ばれるものである．詳細は別章の説明に譲るが，想定される様々な非線形関係を表すため，$m_{jk}(\cdot)$ は B-スプラインを用いる基底関数展開法によって構成される (Eilers and Marx (1996))．この時，ϵ_{ij} が正規分布に従うことより，各 x_{ij} の確率密度関数 $f_j(x_{ij}|\boldsymbol{p}_{ij}; \boldsymbol{\Theta}_j)$ は，$\boldsymbol{p}_{ij} = [p_{i1}^{(j)}, \cdots, p_{iq_j}^{(j)}]^t$ が与えられた時の平均 $m_{j1}(p_{i1}^{(j)}) + \cdots + m_{jq_j}(p_{iq_j}^{(j)})$，分散 σ_j を持つ条件付正規分布となる．なお，$\boldsymbol{\Theta}_j$ は各 m_{jk}, σ_{ij} を含むパラメータベクトルである．ただし，親遺伝子が存在しない $gene_j$ には，$i = 1, \cdots, n$ に亘る x_{ij} の平均 μ_j，分散 σ_j^2 を用いる．これから $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$ 全体の確率密度関数を，各 x_{ij} の分布の独立性を仮定して以下

で表す.

$$f(\boldsymbol{x}_i;\boldsymbol{\Theta}_G)=\prod_{j=1}^{p}f_j(x_{ij}|\boldsymbol{p}_{ij};\boldsymbol{\Theta}_j)$$

ここで，各遺伝子 $gene_j$ へのその親遺伝子 $gene_k$ からの影響 $m_{jk}(p_{ik}^{(j)})$ が存在する場合に $gene_k$ から $gene_j$ へ有向辺を付与し，影響 $m_{jk}(p_{ik}^{(j)})$ が存在しないないし無視しえる場合に辺を付与しないという規則によって，全遺伝子 p 個間の発現程度の因果関係をあるグラフ G によって表すものとする．n 個のマイクロアレイデータが与えられた時，ある G で表される遺伝子間の因果関係を仮定した場合の $\boldsymbol{x}_i$ の事後確率分布は

$$\pi(G)\int\prod_{i=1}^{n}f(x_i;\boldsymbol{\Theta}_G)\pi(\boldsymbol{\Theta}_G|\boldsymbol{\lambda})d\boldsymbol{\Theta}_G$$

で与えられる．ここで $\pi(G)$ は生物学的な背景知識から決める因果関係グラフ G の事前確率分布，$\boldsymbol{\Theta}_G$ は G を表すパラメータベクトル $\boldsymbol{\Theta}_j$ の集合であり，$\pi(\boldsymbol{\Theta}_G|\boldsymbol{\lambda})$ は $\boldsymbol{\Theta}_G$ の事前確率分布である．この分布関数としては多次元正規分布が用いられ，ハイパーパラメータ $\boldsymbol{\lambda}$ は生物学的な背景知識から決められる．

n 個のマイクロアレイデータが与えらえた際に，原理的には上記 $\boldsymbol{x}_i$ の事後確率分布が最大となるモデル（MAP 解）を求めればよい．しかし，因果関係グラフ G は様々なものが考えられ，また各 G に関する $\boldsymbol{\Theta}_G$ が高次元であるため積分計算も容易ではない．そこで，詳細は省略するが積分計算については，ラプラス近似に基づく $BNRC$(Bayesian network and Nonparametric Regression Criterion) の計算によって，モデルの対数事後確率を評価する方法を用いる (Imoto et al. (2002b))．また，因果関係グラフ G に関する MAP 解を完全探索することは，グラフ構造の組み合わせ爆発により計算量的に困難なため，最良優先探索 (Greedy 探索) を用いる (Imoto et al. (2004))．ある因果関係グラフ G 候補とそれに対応する $\pi(G)$ の下で，それを初期グラフとして各遺伝子間の有向辺の付加，除去，方向の反転を逐次適用して，より事後確率が大きいモデルを探索していく．探索の袋小路に至るとバックトラックして更に事後確率の大きいモデルを探し続け，規定の r 個のモデルを探し終えて停止する．従って，探索中途を含め規定個数の多数のモデルが得られる．各 $gene_k$ から $gene_j$ への影響の強さは，その有向辺を除去した場合の $BNRC_{\overline{kj}}$ としない場合の $BNRC_{kj}$ の差 $\Delta BNRC^{kj}=BNRC_{kj}-BNRC_{\overline{kj}}$ の大きさによって評価される．この差が大きいほど，影響の大きな有向辺であると考えられる．

4.2 グラフマイニングによる主要因果関係の抽出

探索によって得られる多数のモデルは，それぞれ遺伝子の発現程度に関する異なる因果関係グラフの候補を表す．各グラフにおいて $\Delta BNRC^{kj}$ の大きい有向辺は，実際の遺伝子発現程度の因果関係を説明するために必要である可能性が高いが，最良優先探索の過程において，たまたまいくつかの有向辺の $\Delta BNRC^{kj}$ が大きく評価されてしまう可能性もある．従って，本当に必要な可能性の高い有向辺及びそれらが繋がった因果関係グラフは，$\Delta BNRC^{kj}$ の大きさに加えて探索途中の多くの因果関係グラフに安定して見られる構造であると考えられる．そこで，探索過程で導かれる因果関係グラフモデルの集合 $\{G_1, \cdots, G_r\}$ から，ある最小支持度以上頻出する因果関係の主要な部分グラフを抽出することを考える．以下では，具体的データへの適用解析を通じてこの抽出過程を示す．

あるマイクロアレイデータ $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ に最良優先探索を適用して，$BNRC$ が極大な $r = 5000$ 個のベイジアンネットワークノンパラメトリック回帰モデル $\{G_1, \cdots, G_{5000}\}$ を得た．各モデルは平均 184 個の頂点（遺伝子）とその間に平均 115 個の有向辺を有する．ここでは，各因果関係グラフ G_h 中の個々の遺伝子間の因果関係ではなく，遺伝子作用のプロセス間の因果関係を分析することにする．各遺伝子の機能は一般に Gene Ontology (GO) Term と呼ばれる記述子によって簡潔に表される (Gene Ontology Consortium (2000,2005))．GO では 1 つの遺伝子に対して Process, Function, Component の 3 つの側面から GO Term と呼ばれる記述を割り当てている．ここでは，遺伝子が如何なるプロセスで作用するかを表す 33 種類の Process GO Term で，データ中の各遺伝子固有名を置き換えた．そして，3 節で述べた原理を基に多頻度連結誘導部分グラフを完全探索する AcGM 手法 (Inokuchi et al. (2002)) を適用した．

ベイジアンネットワークノンパラメトリック回帰モデルは，遺伝子の発現程度間の因果関係の方向性を表す有向グラフで表される．ただし因果の方向性が異なっても，データを同様に説明可能な等価なモデルが複数存在する場合が多く，影響の有無に関るモデルの構造に比較すれば，その方向性に関するモデル構造の信憑性は低い．そこで，ここでは因果関係グラフの各辺の方向性を無視し，無向グラフの多頻度連結誘導部分グラフを導出した．データ中の 2/3 以上のネットワークに共通して発見された 3 遺伝子以上からなる主要因果関係を表す部分グラフを図 5 に示す．これは大半のモデルに共通して見られる最大の大きさの主要な部分グラフであると考えられる．個々のモデルが多数の頂点から成る大規模な因果関係グラフであ

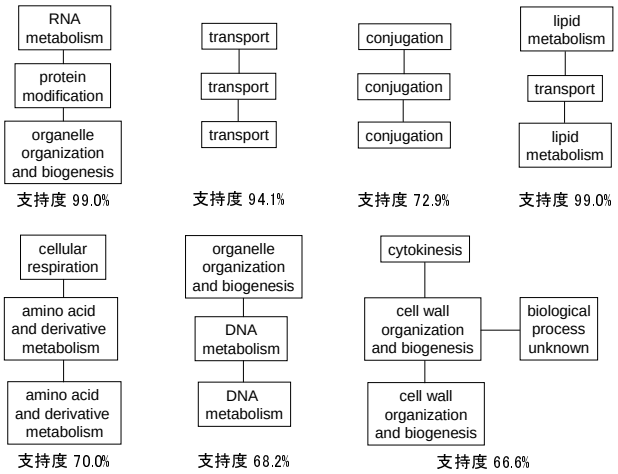


図 5: 2/3 のモデルに現れる遺伝子作用プロセス間依存性の主要部分ネットワーク.

るにもかかわらず，遺伝子作用プロセス間の強い因果関係からなる部分グラフは非常に小規模なものに限られることが分かる．図 6 は 1/3 以上のモデルに共通する主要因果関係部分グラフ，図 7 は全体の 10%以上に共通する主要因果関係部分グラフである．このように，より大きな部分グラフの中には一部の因果関係グラフに特徴的に現れるものが存在するが，各因果関係グラフ全体の大きさから見ると，特徴的部分の大きさは限られていることが判る．このことから，遺伝子の作用プロセス間には，安定した大きな因果関係の構造は見られないことが判る．しかしながら，DNA metabolism 同士の関係や DNA metabolism と organelle organization and biogenesis の関係，cell cycle と protein modification の関係などには多くに共通性が見られ，この解析によって結びつきの強い遺伝子作用プロセスを把握できると考えられる．

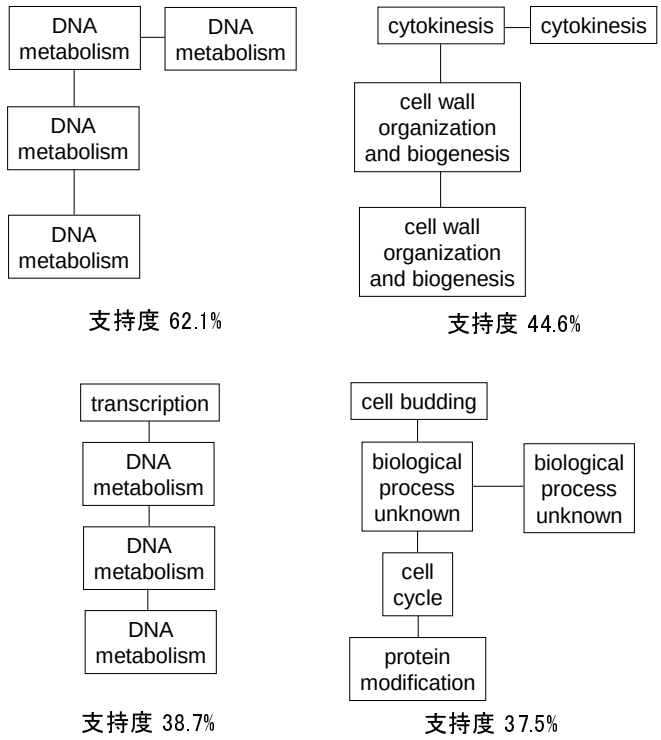


図 6: 1/3 のモデルに現れる遺伝子作用プロセス間依存性の主要部分ネットワーク.

5　グラフマイニング関連研究

最初の節で早期の代表的グラフマイニング手法について概説したが，グラフマイニングに興味を持つ読者の参考として，最後に最近の研究を紹介する．より効率的に多頻度かつ連結な部分グラフを完全探索する手法としては，グラフ不変量で部分グラフ同型判定を行う FSG (Kuramochi and Karypis (2001))，部分グラフ同型判定を深さ優先探索で効率的に実現する gSpan (Yan and Han (2002))，連結部分グラフのみを多頻度部分グラフ候補として探索する AcGM (Inokuchi et al. (2002))，頂点同士を結ぶ辺が少ない，いわゆる疎グラフデータから，非常に高速に多頻度部分グラフを発掘する Gaston (Nijssen and Kok (2004)) などが提案されている．一方，一般的な多頻度連結部分グラフではなく，より限定された部分グラフを発掘する手法も多く研究されている．例えば，帰納推論データベースの枠組みを用いて，グラフデータ中で与えられた条件を満足する部分経路を完全探索する MolFea (de

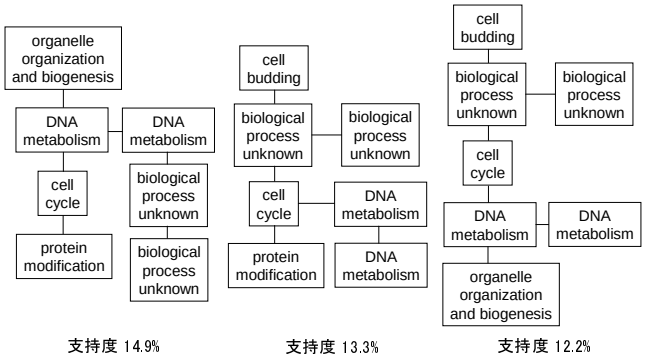


図 7: 10%以上のモデルに現れる遺伝子作用プロセス間依存性の特徴的部分ネットワーク.

Raedt and Kramer (2001))，ある頻度を有する極大な部分グラフを探索する CloseGraph (Yan and Han (2003)) などが挙げられる．更に最近では，与えられた条件を満足する部分自由木を完全探索する FreeTreeMiner (Ruckert and Kramer (2004))，1 枚の大きな疎グラフから互いに重ならない多頻度連結部分グラフを発掘する SiGraM (Kuramochi and Karypis (2004))，グラフデータ中から極大な多頻度連結部分グラフを発掘する SPIN (Huan et al. (2004))，階層的 (Taxonomy) なラベル付けを有するグラフデータ中の多頻度連結部分グラフを探索する Generalized AcGM (Inokuchi (2004))，部分グラフ同型探索に様々な制約を導入してグラフに限らずデータ中に埋め込まれた多頻度の部分経路や部分木の発掘を可能にした B-AGM (Inokuchi et al. (2005)) など，様々なものが提案されている．これらの内，比較的時期の早い手法については文献 (Washio and Motoda (2003)) が詳しい．この他にもグラフデータから種々の部分グラフを発掘する手法が提案されており，ライデン大学のホームページ “Homepage for Mining Structured Data” で最新の手法を含めた紹介や比較を見ることができる (Nijssen (2005))．

6 おわりに

本章では，近年，複雑な構造を持つ膨大なデータが増大していることを背景に発展しているグラフマイニング手法研究の概観と，その重要な基礎概念，代表的手法について述べた．そして，遺伝子発現の因果関係解析への応用を通じて，グラフマイニング手法と統計統計的モデリングの結合，融

合の可能性を示した．このような融合によって，これまでの統計的モデリングでは取り扱いが困難であった膨大な変数を含む大規模な対象の解析が，種々の側面から可能になっていく可能性がある．この分野の研究はまだ緒に就いたばかりあり，今後の発展が待たれるところである．

謝辞

本研究は，部分的に情報・システム研究機構，新領域融合研究センター，機能と帰納プロジェクトの研究費補助を受けた．また本研究は，東京大学医科学研究所ヒトゲノム解析センター・スーパーコンピュータシステムの計算機を利用して行った．

参考文献

[1] Agrwal, R. and Srikant, R. (1994). First algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, 487–499.

[2] Cook, J. and Holder, L. (1994). Substructure discovery using minimum description length and background knowledge, *Journal of Artificial Intelligence Research*, **1**, 231–255.

[Dehaspe and Toivonen (1999)]

[3] Dehaspe, L. and Toivonen, H. (1999). Discovery of frequent datalog patterns, *Data Mining and Knowledge Discovery*, **3**, No.1, 7–36.

[4] de Raedt, L. and Kramer, S. (2001). The levelwise version space algorithm and its application to molecular fragment finding, *Proceedings of IJCAI01: Seventeenth International Joint Conference on Artificial Intelligence*, Vol.2, 853–859.

[5] Eilers, P.H.C. and Marx, B. (1996) Flexible smoothing with B-splines and penalties (with discussion), *Statistical Science*, **11**, 89–121.

[6] Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology, *Nature Genetics*, **25**, 25–29.

[7] Gene Ontology Consortium (2005). http://www.yeastgenome.org/GOContents.shtml

[8] Garey, M. and Johnson, D. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, New York.

[9] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

[10] Huan, L., Wang, W. and Prins, J. (2004). SPIN: Mining Maximal Frequent Subgraphs from Graph Databases, *Proceedings of the 2004 Conference on Knowledge Discovery and Data Mining (SIGKDD2004)*, 581–586.

[11] Imoto, S., Goto, T. and Miyano, S. (2002a). Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression, *Proceedings of Pacific Symposium on Biocomputing*, No.7, 175–186.

[12] Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. and Miyano, S. (2002b). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network, *Proceedings of 1st IEEE Computer Society Bioinformatics Conference*, 219–227.

[13] Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *Journal of Bioinformatics and Computational Biology*, **2**, No.1, 77–98.

[14] Inokuchi, A. (2004). Mining Generalized Substructures from a Set of Labeled Graphs, *Proceedings of Fourth IEEE International Conference on Data Mining (ICDM2004)*, 415–418.

[15] Inokuchi, A., Washio, T. and Motoda, H. (2000). An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, *Proceedings of PKDD2000: Principles of Data Mining and Knowledge Discovery, 4th Europian Conference, Lecture notes in Artificial Intelligence 1910*, Jan Zytkow Eds., Springer, 13–23.

[16] Inokuchi, A., Washio, T. and Motoda, H. (2003). Complete mining of frequent patterns from graphs: Mining graph data, *Machine Learning*, **50**, 321–354.

[17] Inokuchi, A., Washio, T. and Motoda, H. (2005). A General Framework for Mining Frequent Subgraphs from Labeled Graphs, *Journal of Fundamenta Informatiae, Special issue on Advances in Mining Graphs, Trees and Sequence*, **66**, No.1-2, 53–82.

[18] Inokuchi, A., Washio, T. Nishimura, K. and Motoda, H. (2002). A Fast Algorithm for Mining Frequent Connected Subgraphs, IBM Technical Research Report:RT0448, IBM Tokyo Research Laboratory.

[19] Kuramochi, M. and Karypis, G. (2001). Frequent subgraph discovery, *Proceedings of ICDM'01: 1st IEEE International Conference on Data Mining*, 313–320.

[20] Kuramochi, M. and Karypis, G. (2004). Finding Frequent Patterns in a Large Sparse Graph, *Proceedings of the 2004 SIAM Data Mining Conference* (Web. Proceedings).

[21] Nijssen, S. (2005). Homepage for Mining Structured Data, http://hms.liacs.nl/index.html

[22] Nijssen, S. and Kok, J. N. (2004). A Quickstart in Frequent Structure Mining can make a Difference, LIACS Technical Report, Version April 2004.

[23] Ruckert, U. and Kramer, S. (2004). Frequent Free Tree Discovery in Graph Data, *Proceedings of ACM Symposium on Applied Computing (SAC2004), Special Track on Data Mining*, 564–570.

[24] Washio, T. and Motoda, M. (2003). State of the Art of Graph-based Data Mining, *ACM, SIGKDD Explorations*, **5**, No.1, 59–68.

[25] Yan, X. and Han, J. (2002). gspan: Graph-based substructure pattern mining, *Proceedings of ICDM'02: 2nd IEEE International Conference on Data Mining*, 721–724.

[26] Yan, X. and Han, J. (2003). CloseGraph: Mining Closed Frequent Graph Patterns. *Proceedings of the 2003 Conference on Knowledge Discovery and Data Mining (SIGKDD2003)*, 286–295.

[27] Yoshida, H., Motoda, K. and Indurkhya, N. (1994). Graph-based induction as a unified learning framework, *Journal of Applied Intelligence*, **4**, 297–328.

「21 世紀の統計科学」第 II 巻
日本統計学会 HP 版, 2011 年 10 月

第11章 QTL解析の統計モデルと検定の多重性調整

栗木哲[1]

(情報・システム研究機構 統計数理研究所 教授)

個体の形質を規定する遺伝子 (量的形質遺伝子座，QTL) を探索するための統計的手法を QTL 解析という．QTL 解析においては，ロッドスコアとよばれる尤度関数の極大点を探索することによって，QTL の位置が推測される．本稿では，実験交配生物に対する QTL 解析について，その統計モデルと，QTL 探索においてしばしば問題となる検定の多重性調整について解説する．1 節では，QTL 解析が，多重比較・多重検定の一種であることを説明し，問題の枠組みを与える．2 節では QTL 解析のための代表的な統計モデルを，背景となる遺伝知識と併せて可能な限り簡潔に説明する．また各モデルに対して定まるロッドスコアの確率構造を導く．3 節では 2 つの数学的手法 (非線形再生理論およびオイラー標数法) によって，QTL の有無判定のためのロッドスコアの閾値が合理的に設定されることを見る．

[1]kuriki@ism.ac.jp

1　はじめに

1.1　QTL 解析と変化点問題

個体のある形質 (形態や性質) が遺伝的な効果によってもたらされると考えられる場合，その原因となる遺伝子を探し出すことは，遺伝学研究の重要な目的の一つとなる．その形質が主として連続量で記述され，また一般には複数の遺伝子と環境要因によって規定されるものである場合，量的形質とよばれる．たとえばマウスの脂肪体重比 (肥満度) は，典型的な量的形質である．量的形質の原因となる遺伝子が，QTL (量的形質遺伝子座，quantitative trait loci) である．

QTL を探索するための統計手法を QTL 解析という．QTL 解析は連鎖とよばれる遺伝現象を積極的に利用した，代表的な連鎖解析である．QTL 解析においては，連鎖と形質発現の双方を確率的な現象と捉えて統計モデルを設定することによって，目的遺伝子の探索が行われる (鵜飼 2000, Wu *et al.* 2007)．

最初にマウスの肥満の原因となる遺伝子を探索するためのデータ解析の例を示す．ここでは肥満度の代用特性である血中アディポネクチン濃度 (単位 $\log_{10}[\mathrm{ng/m}\ell]$) に着目し，それを量的形質としている．また解析対象のマウスは，標準的マウス近交系である B6 と，日本産亜種由来の MSM 系統の雑種 206 個体である．この 2 系統は形質が多くの点で対照的であるため，QTL 解析に適したものである．解析結果は，図 1 のロッドスコア (LOD score) に要約されている．図の横軸はマウスの 20 対の染色体における遺伝子座の位置であり，その点に位置する遺伝子の遺伝子型と，血中アディポネクチン濃度とのある種の連関の尺度 (ロッド) がプロットされている．この図によると，第 3 および第 16 染色体上に QTL が存在することが示唆される．

ところでこの解析例のように，観測値として与えられている系列や関数のデータがある時点において変化を示すと考えられるときに，その変化時点を統計的に推測する問題を変化点問題という．一般に変化点問題では (i) 変化点の有無の判断のための統計量の閾値の設定，(ii) 変化点の位置の区間推定，(iii) 複数の変化点が存在する可能性があるときはその個数の推測，などが問題となる．本稿では QTL 解析および関連する連鎖解析において，比較的研究が進んでいる (i) の閾値の設定という問題に焦点を絞り，現時点で知られていることやその背景となる数理について概観する．残念ながら (ii), (iii) については現時点では不十分な結果しか知られておらず，本稿ではほとんど触れることはしない．たとえば汎用的手法とされるブートスト

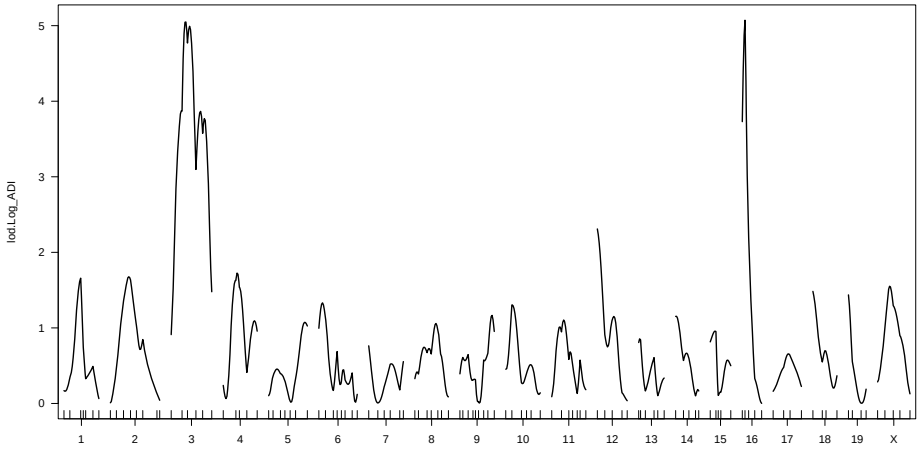


図 1: ロッドスコア

ラップもこれらの解決の役には立たない (Manichaikul *et al.* 2006)．変化点問題は，より広い立場では特異モデルとよばれるクラスの統計モデルであるが，特異モデルにおいては，正則なモデルで成り立つ種々の漸近的性質が成り立たない (福水ほか 2004)．変化点問題において，ブートストラップやモデル選択規準を，少なくとも正則なモデルと同じ形で用いることはできないのは，このことによる．

ところで，連鎖解析，QTL 解析は，ヒトを対象にするものと実験交配が可能な生物を対象にするものに大別され，その両者では解析の対象とするデータの形が大きく異なる．本稿では後者に対するものを想定する．

1.2 多重性調整 (有意水準の調整)

図 1 のようにロッドスコアが明確なピークを持つ場合，その位置付近に QTL が存在すると推測される．しかしながら連鎖や量的形質の発現は確率的な事象であり，それゆえロッドスコアもランダムなグラフである．現れたピークがランダムなゆらぎによる見せかけのものでないかどうかを判定するためには，そのための基準，すなわち閾値を合理的に決める必要がある．

ロッドスコアの定義域を Γ とし，ロッドスコアを $\mathrm{LOD}(\gamma)$, $\gamma \in \Gamma$ で表すとする．$\mathrm{LOD}(\gamma)$ は，“QTL が γ 付近に存在しない” という帰無仮説に対する検定統計量であり，その値がある閾値 c を越えたときに位置 γ の付近に

QTL が存在すると判定することができる．ロッドスコアのピークを探索し，その付近に QTL が存在するかどうかを判定することは，全ての $\gamma \in \Gamma$ について仮説を同時に検定していると考えられる．そのことから多重検定の考え方によって，閾値 c を決めることができる．帰無仮説を

$$H_0 : \text{QTL がどの位置にも存在しない}$$

とし，その仮説が真であるにもかかわらずロッドが閾値 c を越え QTL がどこかで発見される (正確には，発見されたと判断される) 事象を偽陽性 (false positive) と定義する．偽陽性確率を α (0.05 あるいは 0.01 など) 以下に調整するためには，閾値 c を

$$P\Big(\exists\gamma,\ \mathrm{LOD}(\gamma) \geq c_\alpha \,|\, H_0\Big) = \alpha \tag{1.1}$$

となる $c = c_\alpha$ とすればよい．この偽陽性確率は，弱い意味での FWER (family-wise error rate in the weak sense) ともよばれる (Hochberg and Tamhane 1987)．上式は

$$P\Big(\max_{\gamma\in\Gamma} \mathrm{LOD}(\gamma) \geq c_\alpha \,|\, H_0\Big) = \alpha \tag{1.2}$$

と書き換えられる．すなわち閾値 c_α は，確率過程 $\mathrm{LOD}(\cdot)$ の最大値の上側 α 点である．

ところで γ を固定すれば，統計量 $\mathrm{LOD}(\gamma)$ に対する有意点として

$$P\Big(\mathrm{LOD}(\gamma) \geq \tilde{c}_\alpha(\gamma) \,|\, H_0\Big) = \alpha$$

をみたす点 $\tilde{c}_\alpha(\gamma)$ が定義される．$\mathrm{LOD}(\cdot)$ が確率 1 で一定値をとることがない限り，
$\max_{\gamma\in\Gamma} \mathrm{LOD}(\gamma)$ は $\mathrm{LOD}(\gamma)$ よりも確率的に大きな値をとり，$c_\alpha > \tilde{c}_\alpha(\gamma)$ $(0 < \alpha < 1)$ である．逆に，多重検定であることを考慮しないでロッドスコアの棄却点として $\tilde{c}_\alpha(\gamma)$ を用いると，偽陽性確率は α をこえてしまう．この現象を検定の多重性という．また水準 α の有意点として，$\tilde{c}_\alpha(\gamma)$ の代わりにより値の大きな閾値である c_α を用いることを，多重性調整，あるいは有意水準の調整という．

多重検定の多重性調整の方法としてボンフェロニ法がよく知られている．これは，検定の回数 (Γ の要素数) を $|\Gamma|$ とおくとき，全ての $\gamma \in \Gamma$ について $\mathrm{LOD}(\gamma)$ を水準 $\alpha/|\Gamma|$ で検定する，すなわち棄却点として $\tilde{c}_{\alpha/|\Gamma|}(\gamma)$ を用いる

方法である．このとき

$$P\Big(\exists\gamma,\ \mathrm{LOD}(\gamma) \geq \tilde{c}_{\alpha/|\Gamma|}(\gamma)\,|\,H_0\Big) \ \leq\ \sum_{\gamma\in\Gamma} P\Big(\mathrm{LOD}(\gamma) \geq \tilde{c}_{\alpha/|\Gamma|}(\gamma)\,|\,H_0\Big)$$
$$= \ \sum_{\gamma\in\Gamma} \alpha/|\Gamma| = \alpha \qquad (1.3)$$

であるので，偽陽性確率は α 以下に調整される．しかし後で詳しく見るように，LOD(·) は強い相関を持った確率過程であるため，(1.3) の左辺の偽陽性確率は α よりも非常に小さな値となり，それにともなって QTL の検出確率 (検出力) も小さなものとなる．とくにマーカー数が多いとき，あるいはエピスタシスとよばれる複数の QTL による交互作用を検定するときには検定の回数が莫大となりこの傾向が顕著となる．また，2.6 項で説明する区間マッピング法では，ロッドスコアはマーカー間で連続的に補間されるため，$|\Gamma| = \infty$, $\tilde{c}_{\alpha/|\Gamma|}(\gamma) = \infty$ となり，ボンフェロニ法は意味をなさなくなる．以上の理由から，QTL 解析においてはボンフェロニ法を用いることはできない．

次の 2 節では，実験交配における QTL 解析と関連する連鎖解析の統計モデルをいくつか紹介する．さらにそれらのモデルにおいて現れるロッドスコア LOD(·) の確率過程としての構造を調べる．

3 節では，2 節で与えたロッドスコア LOD(·) の構造から，その最大値 $\max_{\gamma\in\Gamma}\mathrm{LOD}(\gamma)$ の上側 α 点 c_α を求める方法を説明する．経験則やシミュレーションに基づく方法に触れた後，理論的な近似法について解説する．確率過程，確率場の最大値の分布については長い研究の歴史がある一方で，近年においても本質的な進展が見られている (Siegmund 1985, Piterbarg 1996, Adler and Taylor 2007, 栗木・竹村 2008)．本稿では，逐次解析，非線形再生理論を用いる方法と，オイラー標数法とよばれる積分幾何的な手法の 2 通りによって，確率過程としてのロッドスコアの最大値分布の近似を与え，そのことを通して多重性調整が可能であることを見る．

2　QTL 解析の統計モデルとロッドスコア

2.1　データの形

ここでは QTL 解析が対象とするデータの形と，その背後に想定される確率構造を説明する．交配の実験計画として，BC (戻し交配, backcross) と F_2 の 2 種類を考える．(これらの実験交配については，次項で説明する．)

個体数を n とする．またマーカー遺伝子座の数を m とする．マーカー遺伝子座 (しばしばマーカーと略す) とは，何らかの方法でその遺伝子型が観測できる遺伝子座をいう．個体のそれぞれ $t = 1, \ldots, n$ について，着目している量的形質 (表現型) の測定値 $y^{(t)}$ (スカラ) と m 個のマーカー遺伝子に対する遺伝子型のベクトル $z^{(t)} = (z_1^{(t)}, \ldots, z_m^{(t)})$ が得られている．遺伝子型 $z_i^{(t)}$ は戻し交配の場合は 2 値，F_2 の場合は 3 値をとる．取り扱いの容易さから，

$$z_i^{(t)} = 1, -1 \ \text{(戻し交配の場合)}, \quad 1, 0, -1 \ \text{(F_2 の場合)}$$

と表すことにする (表 1).

表 1: 個体データ

個体番号	表現型	遺伝子型
1	$y^{(1)}$	$z^{(1)} = (z_1^{(1)}, \ldots, z_m^{(1)})$
$\vdots$	$\vdots$	$\vdots$
n	$y^{(n)}$	$z^{(n)} = (z_1^{(n)}, \ldots, z_m^{(n)})$

これらの個体データとは別に，m 個のマーカー遺伝子 $i = 1, \ldots, m$ のそれぞれについて，それが属する染色体の番号 c_i と，その染色体上での位置 d_i の情報が与えられている．d_i は基準となる点からの遺伝的距離 (単位はモルガン M，またはセンチモルガン cM，これらの意味は次項で説明する) で記述され，同一染色体の中では昇順 ($i < j$ ならば $d_i < d_j$) とする (表 2).

表 2: マーカーデータ

マーカー番号	1	2	$\cdots$	m
マーカー名	$*$	$*$	$\cdots$	$*$
染色体番号	1	1	$\cdots$	c
座の位置 (M)	d_1	d_2	$\cdots$	d_m

QTL 解析では，個体データである遺伝子型と表現型の組 $(z^{(t)}, y^{(t)})$ を確率変数と考え，モデル化を行う．ただし通常の多変量解析と同様に，個体間では独立と考える．

以下では，m 次元の遺伝子型ベクトルデータ $z^{(t)}$ の周辺分布について説明する．これは連鎖によって引き起こされるものである．遺伝子型 $z^{(t)}$ が与えられたときの表現型 $y^{(t)}$ の分布を表現するための統計モデルについては後の項で説明する．マーカーの位置 d_i の単位はモルガンとする．

マーカー遺伝子の添字 $i = 1, \ldots, m$ に対応させる形で，± 1 に値をとる確率変数 ϵ_i $(i = 1, \ldots, m)$ を考える．ただしこの列はマルコフ系列で，

$$
\begin{aligned}
&P(\epsilon_1 = \pm 1) = \frac{1}{2}, \\
&P(\epsilon_{i+1} = \pm\epsilon_i \,|\, \epsilon_i) = \begin{cases} \frac{1}{2}\left(1 \pm e^{-2(d_{i+1}-d_i)}\right) & (i, i+1 \text{ は同じ染色体上}), \\ \frac{1}{2} & (i, i+1 \text{ は異なる染色体上}) \end{cases}
\end{aligned}
$$

で定義されるものとする．同時確率分布は

$$
P(\epsilon_1, \ldots, \epsilon_m) = \frac{1}{2^m} \prod_{i=1}^{m-1} \left(1 + \epsilon_i \epsilon_{i+1} e^{-2(d_{i+1}-d_i)}\right) \tag{2.1}
$$

(ただし座 i と座 $i+1$ が同じ染色体上にないならば $d_{i+1} - d_i = \infty$ とおく) である．任意の i, j について

$$
\begin{aligned}
&P(\epsilon_i = \pm 1) = \frac{1}{2}, \\
&P(\epsilon_j = \pm\epsilon_i \,|\, \epsilon_i) = \begin{cases} \frac{1}{2}\left(1 \pm e^{-2|d_j-d_i|}\right) & (i, j \text{ は同じ染色体上}), \\ \frac{1}{2} & (i, j \text{ は異なる染色体上}) \end{cases}
\end{aligned}
$$

となることに注意する．次に $(\delta_1, \ldots, \delta_m) \in \{-1, 1\}^m$ を $(\epsilon_1, \ldots, \epsilon_m)$ と独立に同じ分布に従うランダムベクトルとする．このとき，遺伝子型 $z^{(t)} = (z_1^{(t)}, \ldots, z_m^{(t)})$ に仮定される確率モデルで最も基本的なものは，

$$
\begin{aligned}
(z_1^{(t)}, \ldots, z_m^{(t)}) &\stackrel{d}{=} (\epsilon_1, \ldots, \epsilon_m) && (\text{戻し交配の場合}), \\
&\stackrel{d}{=} \frac{1}{2}(\epsilon_1 + \delta_1, \ldots, \epsilon_m + \delta_m) && (\mathrm{F}_2 \text{ の場合}), \\
&\quad (\text{各 } t \text{ について独立に}) &&
\end{aligned} \tag{2.2}
$$

と表わされる．ここで $\stackrel{d}{=}$ は両辺の分布が等しいことを意味する．

遺伝子型のこのような確率構造は，連鎖により引き起こされるものである．次項ではそのことについて説明する．

2.2 実験交配と連鎖

一対の染色体 (相同染色体) の一本は母親由来，一本は父親由来である．各個体の遺伝子型を，記法 $A_1B_1\cdots/A_2B_2\cdots$ によって表す．これは，一方の親に由来する染色体の遺伝子型が $A_1B_1\cdots$，もう一方の親に由来する染色体の遺伝子型が $A_2B_2\cdots$ であることを意味するものとする．ここで A_1 と A_2 のペア，あるいは B_1 と B_2 のペアは同じ座に位置する一対の遺伝子 (対立遺伝子) である．全ての遺伝子座についてその遺伝子型がホモ (すなわち $A_1 = A_2, B_1 = B_2, \ldots$) であるとき，近交系という．

近交系は植物の場合は自殖，動物の場合は兄妹 (けいまい) 交配を繰り返すことによって容易に作り出すことができる．実際，自殖あるいは兄妹交配を，遺伝子型を状態とするマルコフチェインによってモデル化すると，その吸収状態が近交系に対応することが確認できる．(遷移行列の固有値の絶対値で 1 でないものの最大は，自殖あるいは兄妹交配のそれぞれについて 0.5, 0.809 であり，前者の方が収束速度が速い．)

ある近交系の個体 P_1 と別の近交系の個体 P_2 を掛け合わせた雑種第 1 代を F_1 世代という．F_1 個体とその親 P_1 (P_2) との掛け合わせを戻し交配 BC_1 (BC_2)，また F_1 個体同士の自殖，あるいは兄妹交配によって得られる雑種第 2 代を F_2 世代という．

以下では 2 つの遺伝子座に着目する．近交系においては，どの遺伝子座においても遺伝子型はホモであるので，異なる近交系の個体の遺伝子型は

$$P_1 = AB/AB, \qquad P_2 = ab/ab$$

と書くことができる．また P_1 と P_2 の配偶子 (生殖細胞，すなわち卵，精子) の遺伝子型はそれぞれ AB, ab であるので，雑種第 1 代の遺伝子型は

$$F_1 = AB/ab$$

となる．

F_1 個体の配偶子の遺伝子型としては，次の 4 種類

$$\begin{aligned} F_1\text{の配偶子} \ &= AB,\ ab \ \ (\text{それぞれ確率 } \tfrac{1-r}{2} \text{ で}), \\ &\quad\ Ab,\ aB \ \ (\text{それぞれ確率 } \tfrac{r}{2} \text{ で}) \end{aligned} \tag{2.3}$$

が現れる．その理由を述べるために，図 2 に沿って配偶子の生成過程 (減数分裂) を説明しよう．

減数分裂では最初に相同染色体の対が分離し，4 本の染色分体となる．さらに互いに別の親に由来する 2 本の染色分体 (図中で色の異なるもの) は，図のように交差を起こす場合がある．(これは確率的な事象である．) その後 4 本の染色分体は互いに分かれて F_1 個体の配偶子が生成される．いま着目している 2 座の間で，奇数回の交差が起きたとすると，生成される生殖細胞の遺伝子は Ab または aB となる．この現象を組換えという．この結果として，F_1 個体の配偶子の遺伝子型は 4 種類 AB, ab, Ab, aB でその頻度は (2.3) の通りとなる．なお r は 2 座間で組換えが起きる確率であり，組換え価とよばれる．P_1, P_2 の各個体においても同じ過程で配偶子が生成されるが，4 つの染色分体の遺伝子型は同じであるので結果として組換えは観察されないことに注意する．

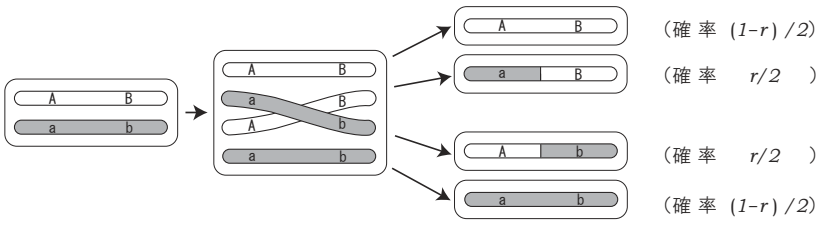


図 2: 減数分裂と交差

さらに $P_1 = AB/AB$ と $F_1 = AB/ab$ の戻し交配 BC_1 を考えると，その遺伝子型は，

$$\begin{aligned} BC_1 \ &= AB/AB,\ AB/ab \ \ (\text{それぞれ確率 } \tfrac{1-r}{2} \text{ で}), \\ &\quad\ AB/Ab,\ AB/aB \ \ (\text{それぞれ確率 } \tfrac{r}{2} \text{ で}) \end{aligned}$$

の 4 通りとなる．1 つの遺伝子座に着目した場合，遺伝子型は A/A または A/a の 2 通りであり，一般性を失うことなく 1, −1 で表すことができる．

また F_1 個体同士の掛け合わせにより得られる F_2 個体では，$4 \times 4 = 16$ 通りの遺伝子型が現れる．ただし通常の方法では，2 種類のヘテロ A/a と a/A は識別されない．A/A, A/a, a/a の 3 通りについては，そのマーカー遺伝子が共優性であれば識別することができる．それらを一般性を失うことなく 1, 0, −1 で表すことにする．2 つの座に着目した場合は，識別できる遺伝子型は $3 \times 3 = 9$ 通りとなる．

交差の確率モデルとして最も基本的なものは，交差の生起をポアソン事象と考えるものである (Haldane 1919)．交差をポアソン事象と考えること

により，染色体上の 2 点間でおこる交差の平均値をもって，その 2 点間の距離と定義することができる．この距離が遺伝的距離で，その単位はモルガン (M) である．この遺伝的距離を遺伝子座の距離と定義することによって，交差の生起事象は強度関数が 1 の定常ポアソン点過程と考えることができる．x (M) 離れた 2 点で，i 回交差がおきる確率は $x^i e^{-x}/i!$ であるので，2 点間の組換え価 $r = r(x)$ は

$$r(x) = P(\text{奇数回交差が起きる}) = \sum_{i:\,\text{odd}} \frac{x^i}{i!} e^{-x} = \frac{1}{2}(1 - e^{-2x}) \tag{2.4}$$

となる．関数 $r(x)$ をホールデンの地図関数という．

F_1 個体の配偶子の遺伝子型の確率構造を考える．マーカー遺伝子座 i の遺伝子型が P_1 由来であるとき $\epsilon_i = 1$，P_2 由来であるとき $\epsilon_i = -1$ とする．ポアソン性の仮定の下で，$(\epsilon_1, \ldots, \epsilon_i)$ と $\epsilon_{i+1} - \epsilon_i$ は独立であるので $\epsilon_1, \ldots, \epsilon_m$ はマルコフ性を持ち，また

$$\begin{aligned} P(\epsilon_{i+1} = \epsilon_i \,|\, \epsilon_i) &= P(\text{座 } i, i+1 \text{ 間で組換えは起らない}) \\ &= 1 - r(|d_{i+1} - d_i|), \end{aligned}$$

すなわち $\epsilon_1, \ldots, \epsilon_m$ は確率分布 (2.1) に従う ± 1 列である．

遺伝子型 z_i の定義の仕方より，戻し交配 BC においては F_1 個体の配偶子の $\epsilon_i = 1, -1$ の値が，BC の遺伝子型に一致する．また F_2 個体においては，$\frac{1}{2}(\epsilon_i + \delta_i) = 1, 0, -1$ の値が遺伝子型となる．以上で (2.2) が導出された．

組換え価 $r(x)$ は，$0 \le r(x) \le 1/2$ に値をとる単調増加関数である．2 座が同じ遺伝子座 $(x = 0)$ のとき組換えは起らず $(r = 0)$，また 2 つの遺伝子座が非常に離れている，あるいは別の染色体上にあるとき $(x = \infty)$ は確率 $r = 1/2$ で組換えが起きる．2 つの遺伝子座が近くにあり，組換え価が小さな値をとっている状態を，2 座が連鎖するという．染色体の長さは典型的には 100cM (=1M) 程度であり，交差は平均 1 回しか起きない．そのため実験交配の個体の遺伝子型は強い正の相関を持ち，多重性の調整においてはそのことの配慮が必要となる．

注 2.1 ここでは交差をポアソン点過程によりモデル化しているが，より一般には再生過程を用いてモデル化することができる (Karlin and Liberman 1983).

2.3 単一マーカー分析

本項では QTL 解析の統計モデルとして基本となる単一マーカー分析 (single marker analysis) を説明する．とくに断らない限りは F_2 集団を扱う．

QTL 解析が対象とするデータは，表 1，表 2 の形であった．単一マーカー分析とは，マーカー遺伝子座 i の遺伝子型で集団を 3 群に分け，その 3 群で形質 (表現型) の平均が等しいという仮説に対する分散分析統計量，あるいは尤度比検定統計量を $T(i)$ とおき，その統計量を最も大きくする遺伝子座 $\hat{i} = \operatorname{argmax} T(i)$ の付近に QTL が存在すると判断するという手順である．

この手順に対応する統計モデルは次のようなものである．遺伝子型 $z_i^{(t)} = 1, 0, -1$ に対応した 3 つの群の形質の平均を 3 つのパラメータによって

$$\mu + \alpha + \delta, \quad \mu - \delta, \quad \mu - \alpha + \delta$$

と表現する．QTL は 1 つだけどこかの座 (i とおく) に存在すると仮定すると，モデルは以下のようになる．

1 つの $i \in \{1, \ldots, m\}$ が存在し，

$$y^{(t)} = \mu + \alpha z_i^{(t)} + \delta w(z_i^{(t)}) + \varepsilon^{(t)}, \quad \varepsilon^{(t)} \sim N(0, \sigma^2) \quad (t = 1, \ldots, n). \tag{2.5}$$

ただし

$$w(z) = \begin{cases} 1 & (z = \pm 1), \\ -1 & (z = 0) \end{cases}$$

とおいた．

このモデル (2.5) に含まれる未知パラメータは $(i, \mu, \alpha, \delta, \sigma^2)$ である．パラメータ α, δ はそれぞれ QTL の加法効果，優性効果と解釈されている．QTL の効果がない ($\alpha = \delta = 0$) ときは，QTL の位置パラメータ i は推測不能な量となる．この例のようにパラメータが特別な値をとるときにモデルの識別性が崩れるモデルを特異モデルという (福水ほか 2004)．

QTL の位置 i が既知であるという仮定をおくと，モデルは正則となる．その仮定の下で，パラメータ $(\mu, \alpha, \delta, \sigma^2)$ の最尤推定量を $(\hat{\mu}(i), \hat{\alpha}(i), \hat{\delta}(i), \hat{\sigma}^2(i))$，また QTL が存在しない (QTL の効果がない) という帰無仮説 $\alpha = 0, \delta = 0$ の下での最尤推定量を $(\tilde{\mu}, 0, 0, \tilde{\sigma}^2)$ とおく．サンプルサイズ n を明示する形で，尤度関数を L_n と書くとき，座 i が QTL であるという仮定の下でのロッドスコア $\mathrm{LOD}_n(i)$ および尤度比検定統計量 $\mathrm{LRT}_n(i)$ は

$$\mathrm{LOD}_n(i) = \log_{10} \frac{L_n(\hat{\mu}(i), \hat{\alpha}(i), \hat{\delta}(i), \hat{\sigma}^2(i))}{L_n(\tilde{\mu}, 0, 0, \tilde{\sigma}^2)} = 0.217\, \mathrm{LRT}_n(i)$$

である．(ロッドスコアは尤度比の常用対数として定義される．すなわち尤度比検定統計量の $(2\log 10)^{-1} = 0.217$ 倍である．) QTL の位置 i も未知とするモデル (2.5) の下で，最尤推定量は $(\hat{\mu}(\hat{i}), \hat{\alpha}(\hat{i}), \hat{\delta}(\hat{i}), \hat{\sigma}^2(\hat{i}))$，ただし

$$\hat{i} = \operatorname{argmax} \mathrm{LOD}_n(i)$$

であり，単一マーカー分析における QTL の推測手順がモデル (2.5) の下での i の最尤推定量を与えることが分かる．

以降では，多重性調整のための $\max_{1\le i\le m} \mathrm{LRT}_n(i)$ の分布計算の準備として，QTL が存在しないという帰無仮説の下での $\mathrm{LRT}_n(i)$ $(i = 1, \ldots, m)$ の同時漸近分布を与える．尤度比検定の一般論より，帰無仮説の下では個体数 n についての漸近的性質として，各 i に対して $\mathrm{LRT}_n(i)$ は漸近的に自由度 2 のカイ 2 乗分布に従う．しかしそれらは独立ではない．カイ 2 乗確率変数の間の相関構造は以下のように表される．

命題 2.1 座間 i, j の組換え価を $\frac{1}{2}(1 - \rho_{ij})$ とおく．QTL が存在しないという帰無仮説 H_0 の下で，$i = 1, \ldots, m$ の同時分布の意味で分布収束

$$\mathrm{LRT}_n(i) \;\Rightarrow\; T_i = U_i^2 + V_i^2 \quad (n \to \infty) \tag{2.6}$$

が成り立つ．ただし $(U_1, V_1, \ldots, U_m, V_m)$ は平均 0 の $2m$ 次元正規分布ベクトルで

$$\mathrm{Cov}(U_i, U_j) = \rho_{ij}, \quad \mathrm{Cov}(V_i, V_j) = \rho_{ij}^2, \quad \mathrm{Cov}(U_i, V_j) = 0$$

をみたすものである．とくに各 i について T_i は自由度 2 のカイ 2 乗分布に従う．

証明 しばしば個体を識別する添字 (t) を省略する．$\epsilon_i^{(t)} = \epsilon_i = \pm 1$ を母由来の相同染色体の第 i 座の遺伝子型，$\delta_i^{(t)} = \delta_i = \pm 1$ を父由来のそれとする．このとき

$$z_i^{(t)} = z_i = \frac{1}{2}(\epsilon_i + \delta_i), \quad w(z_i^{(t)}) = w_i = \epsilon_i \delta_i$$

と表すことができる．

2 つの m 次元ベクトル $(\epsilon_1, \ldots, \epsilon_m)$, $(\delta_1, \ldots, \delta_m) \in \{-1, 1\}^m$ は独立で，それぞれの各成分は平均 0，分散 1，また組換えに由来する相関構造

$$\begin{aligned} E[\epsilon_i \epsilon_j] &= E[\delta_i \delta_j] \\ &= 1 \times P(\epsilon_i = \epsilon_j) + (-1) \times P(\epsilon_i \neq \epsilon_j) \\ &= \frac{1}{2}(1 + \rho_{ij}) - \frac{1}{2}(1 - \rho_{ij}) = \rho_{ij} \end{aligned}$$

を持っていた．z_i, w_i の 1, 2 次モーメントは $E[z_i]=0$, $E[w_i]=0$,

$$\begin{aligned}
&\mathrm{Cov}(z_i,z_j) = (E[\epsilon_i\epsilon_j]+E[\delta_i\delta_j])/4 = \rho_{ij}/2,\\
&\mathrm{Cov}(w_i,w_j) = E[\epsilon_i\delta_i\epsilon_j\delta_j] = E[\epsilon_i\epsilon_j]E[\delta_i\delta_j] = \rho_{ij}^2,\\
&\mathrm{Cov}(z_i,w_j) = (E[\epsilon_i\epsilon_j\delta_j]+E[\delta_i\epsilon_j\delta_j])/2 = 0
\end{aligned}$$

である．

一般性を失うことなくパラメータの真値を $\mu=0$, $\sigma^2=1$ とおく．

$$y = \begin{pmatrix} \vdots \\ y^{(t)} \\ \vdots \end{pmatrix}_{1\le t\le n} \qquad X_i = \begin{pmatrix} \vdots & \vdots \\ z_i^{(t)} & w(z_i^{(t)}) \\ \vdots & \vdots \end{pmatrix}_{1\le t\le n}$$

とおく．テイラー展開より

$$\begin{aligned}
&\mathrm{LRT}_n(i) \;\approx\; y^TQX_i^T(X_i^TQX_i)^{-1}X_i^TQy,\\
&Q = I_n - \mathbf{1}_n\mathbf{1}_n^T/n,\quad \mathbf{1}_n=(1,\dots,1)^T.
\end{aligned}$$

ここで $\approx$ は両辺の差が $o_p(1)$ であること意味する．さらに

$$\frac{1}{n}X_i^TQX_i = \frac{1}{n}\sum_{t=1}^n \begin{pmatrix} z_i^{(t)}-\bar z_i \\ w_i^{(t)}-\bar w_i \end{pmatrix}(z_i^{(t)}-\bar z_i, w_i^{(t)}-\bar w_i) \approx \begin{pmatrix} \frac12 & 0 \\ 0 & 1 \end{pmatrix}$$

に注意すると

$$\mathrm{LRT}_n(i) \;\approx\; \frac{2}{n}\Big\{\sum_{t=1}^n (y^{(t)}-\bar y)z_i^{(t)}\Big\}^2 + \frac{1}{n}\Big\{\sum_{t=1}^n (y^{(t)}-\bar y)w_i^{(t)}\Big\}^2 \;\approx\; u_i^2+v_i^2,$$

ただし

$$u_i = \sqrt{\frac{2}{n}}\sum_{t=1}^n y^{(t)}z_i^{(t)}\ \Big(\approx \sqrt{\frac{n}{2}}\,\hat\alpha(i)\Big), \qquad v_i = \frac{1}{\sqrt{n}}\sum_{t=1}^n y^{(t)}w_i^{(t)}\ \Big(\approx \sqrt{n}\,\hat\delta(i)\Big)$$

である．ここで u_i, v_i の平均は 0, 共分散関数は

$$\begin{aligned}
&\mathrm{Cov}(u_i,u_j) = 2\,\mathrm{Var}(y)\,\mathrm{Cov}(z_i,z_j) = \rho_{ij},\\
&\mathrm{Cov}(v_i,v_j) = \mathrm{Var}(y)\,\mathrm{Cov}(w_i,w_j) = \rho_{ij}^2,\\
&\mathrm{Cov}(u_i,v_j) = \sqrt{2}\,\mathrm{Var}(y)\,\mathrm{Cov}(z_i,w_j) = 0.
\end{aligned}$$

あとは中心極限定理による． ∎

注 2.2 戻し交配の場合は，$i = 1, \ldots, m$ の同時分布の意味で

$$\mathrm{LRT}_n(i) \ \Rightarrow \ U_i^2 \quad (n \to \infty).$$

注 2.3 命題 2.1 は，組換え価がホールデンの地図関数 (2.4) であることは仮定していない．(2.4) の仮定の下では，i, j 座間の遺伝的距離を r_{ij} とするとき，$\rho_{ij} = e^{-2r_{ij}}$ であるので

$$\rho_{ij} = \rho_{i,i+1}\rho_{i+1,i+2}\cdots\rho_{j-1,j}.$$

すなわち $(U_i)_{i\geq 1}$ および $(V_i)_{i\geq 1}$ はマルコフ性を持った正規分布変数列となる．

2.4 分離比の検定

ここで述べる分離比の検定は，QTL 解析ではないが，数理的には前項の単一マーカー分析と非常に似た構造を持つ連鎖解析である．

F_2 個体の 1 つの遺伝子座に着目する．対立遺伝子を A, a とする．(2.2) によると，それが共優性である限り，遺伝子型は分離比の期待比率

$$\mathrm{A/A} : \mathrm{A/a} : \mathrm{a/a} = 1 : 2 : 1$$

を持つはずである．(A/a と a/A は区別されていない．) しかし実際の観測値は，理論値である 1:2:1 の比率を持った母集団からのサンプルとはみなされないことがある．このような現象は分離のゆがみとよばれる．そのような現象が起こる理由として，その近くに致死遺伝子 (生殖隔離障壁) が存在することが考えられる．生殖隔離障壁とは，それが特定の遺伝子型をとったときに生殖率，稔性が下がるような遺伝子をいう．Harushima *et al.* (2001) は，イネの F_2 集団を用いて，そのような分離のゆがみを伴う生殖隔離障壁を検出している．

いま，分離のゆがみの検出のために，表 1，表 2 のデータが利用可能であるとする．ただし表 1 の表現型のデータ $y^{(t)}$ はここでは不要である．

分離比が理論値に従っているかどうかを検定するためには，多項分布のカイ 2 乗適合度検定を用いることができる．各遺伝子型の個体数の観測度数を $n_{\mathrm{A/A}}, n_{\mathrm{A/a}}, n_{\mathrm{a/a}}$ とおく．分離比の検定統計量

$$T_n = \frac{(n_{\mathrm{A/A}} - n/4)^2}{n/4} + \frac{(n_{\mathrm{A/a}} - n/2)^2}{n/2} + \frac{(n_{\mathrm{a/a}} - n/4)^2}{n/4}$$
$$(n = n_{\mathrm{A/A}} + n_{\mathrm{A/a}} + n_{\mathrm{a/a}})$$

は，分離比が 1 : 2 : 1 であるという帰無仮説の下で，自由度 2 のカイ 2 乗分布を漸近分布に持つ．

分離のゆがみを検出するためには，m 個のマーカー遺伝子 $(i = 1, \ldots, m)$ について，このカイ 2 乗検定を同時に行う必要がある．そのために，ここでも多重性の調整が必要となる．第 i 座における分離比の検定統計量を $T_{n,i}$ とする．最大値 $\max_{1\le i\le m} T_{n,i}$ の分布を近似する準備として，$T_{n,i}$ $(i = 1, \ldots, m)$ の同時漸近分布を求めよう．

次のカイ 2 乗統計量の分解に注意する．

$$T_n = \frac{(n_{\mathrm{A/A}} - n/4)^2}{n/4} + \frac{(n_{\mathrm{A/a}} - n/2)^2}{n/2} + \frac{(n_{\mathrm{a/a}} - n/4)^2}{n/4} = U_n^2 + V_n^2,$$

ただし

$$U_n = \sqrt{\frac{2}{n}}(n_{\mathrm{A/A}} - n_{\mathrm{a/a}}), \qquad V_n = \frac{1}{\sqrt{n}}(n_{\mathrm{A/A}} + n_{\mathrm{a/a}} - n_{\mathrm{A/a}}).$$

ここで，もし A/a と a/A が識別可能 (相が既知) とすると，頻度のデータは表 3 の形に集計される．また $V_n = \frac{1}{\sqrt{n}}(n_{\mathrm{A/A}} + n_{\mathrm{a/a}} - n_{\mathrm{A/a}} - n_{\mathrm{a/A}})$ である．

表 3: 相が既知の場合

	A	a	
A	$n_{\mathrm{A/A}}$	$n_{\mathrm{A/a}}$	
a	$n_{\mathrm{a/A}}$	$n_{\mathrm{a/a}}$	
			n

1 つの個体 (t 番目とする) の，この表に対する寄与を考える．4 つのセルのうちの 1 か所で 1 回カウントされているはずである．相同染色体の母方染色体の第 i 座の遺伝子型 $\epsilon_i^{(t)} = \pm 1$ と，父方の対応する遺伝子型 $\delta_i^{(t)} = \pm 1$ に立ち返ると，この個体の，表 3 への寄与は表 4 となる．表 4 において，1 つのセルは 1，他の 3 つのセルは 0 である．

表 4: 個体 t の表 3 への寄与

	A	a	
A	$\frac{1}{4}(1+\epsilon_i^{(t)})(1+\delta_i^{(t)})$	$\frac{1}{4}(1+\epsilon_i^{(t)})(1-\delta_i^{(t)})$	$\frac{1}{2}(1+\epsilon_i^{(t)})$
a	$\frac{1}{4}(1-\epsilon_i^{(t)})(1+\delta_i^{(t)})$	$\frac{1}{4}(1-\epsilon_i^{(t)})(1-\delta_i^{(t)})$	$\frac{1}{2}(1-\epsilon_i^{(t)})$
	$\frac{1}{2}(1+\delta_i^{(t)})$	$\frac{1}{2}(1-\delta_i^{(t)})$	1

このことから，第 i 座の分離比の統計量は $T_{n,i} = U_{n,i}^2 + V_{n,i}^2$，ただし

$$\begin{aligned}
U_{n,i} =& \sqrt{\frac{2}{n}}\Big\{\sum_{t=1}^{n}\frac{1}{4}(1+\epsilon_i^{(t)})(1+\delta_i^{(t)}) - \sum_{t=1}^{n}\frac{1}{4}(1-\epsilon_i^{(t)})(1-\delta_i^{(t)})\Big\} \\
=& \frac{1}{\sqrt{n}}\sum_{t=1}^{n} u_i^{(t)}, \qquad u_i^{(t)} = \frac{1}{\sqrt{2}}(\epsilon_i^{(t)} + \delta_i^{(t)}), \\
V_{n,i} =& \frac{1}{\sqrt{n}}\Big\{\sum_{t=1}^{n}\frac{1}{4}(1+\epsilon_i^{(t)})(1+\delta_i^{(t)}) + \sum_{t=1}^{n}\frac{1}{4}(1-\epsilon_i^{(t)})(1-\delta_i^{(t)}) \\
& - \sum_{t=1}^{n}\frac{1}{4}(1+\epsilon_i^{(t)})(1-\delta_i^{(t)}) - \sum_{t=1}^{n}\frac{1}{4}(1-\epsilon_i^{(t)})(1+\delta_i^{(t)})\Big\} \\
=& \frac{1}{\sqrt{n}}\sum_{t=1}^{n} v_i^{(t)}, \qquad v_i^{(t)} = \epsilon_i^{(t)}\delta_i^{(t)}
\end{aligned}$$

と分解できる．

2 つの座 i, j に着目する．座間の組換え価を $\frac{1}{2}(1-\rho_{ij})$ とすると，前項ですでに計算したように，$E[\epsilon_i^{(t)}\epsilon_j^{(t)}] = E[\delta_i^{(t)}\delta_j^{(t)}] = \rho_{ij}$. したがって，$E[u_i^{(t)}] = 0$, $E[v_i^{(t)}] = 0$,

$$\mathrm{Cov}(u_i^{(t)}, u_j^{(t)}) = \rho_{ij}, \qquad \mathrm{Cov}(v_i^{(t)}, v_j^{(t)}) = \rho_{ij}^2, \qquad \mathrm{Cov}(u_i^{(t)}, v_j^{(t)}) = 0$$

が成り立つ．以上から中心極限定理によって，次が従う．

命題 2.2 座間 i, j の組換え価を $\frac{1}{2}(1-\rho_{ij})$ とおく．$i = 1, \ldots, m$ の同時分布の意味で，分布収束

$$T_{n,i} \;\Rightarrow\; T_i \quad (n \to \infty)$$

が成り立つ．ただし T_i は命題 2.1 の (2.6) で定義したものである．

つまり単一マーカー分析におけるロッドスコアの同時漸近分布と全く同じものが現れる．

2.5　エピスタシス，遺伝子座相互作用の検出

今までは，QTL はたかだか 1 つ存在するというモデルを扱ってきた．しかし量的形質は，複数の QTL によって引き起こされるものと考えられているため，そのようなモデルでは不十分である．とくにエピスタシスとよばれる QTL 間の交互作用を検出するためには，複数の QTL の存在を仮定した統計モデルを使う必要がある．

たとえば 2 つの QTL の存在を仮定した場合，単一マーカー分析のモデル (2.5) に対応するものとして，次のモデルが考えられる．

$i, j \in \{1, \dots, m\}$ が存在し，

$$\begin{aligned} y^{(t)} = \mu &+ \alpha_1 z_i^{(t)} + \delta_1 w_i^{(t)} + \alpha_2 z_j^{(t)} + \delta_2 w_j^{(t)} \\ &+ \beta_1 z_i^{(t)} z_j^{(t)} + \beta_2 z_i^{(t)} w_j^{(t)} + \beta_3 w_i^{(t)} z_j^{(t)} + \beta_4 w_i^{(t)} w_j^{(t)} \\ &+ \varepsilon^{(t)}, \quad \varepsilon^{(t)} \sim N(0, \sigma^2) \quad (t = 1, \dots, n), \end{aligned}$$

ただし

$$w_i^{(t)} = w(z_i^{(t)}) = \begin{cases} 1 & (z_i^{(t)} = \pm 1), \\ -1 & (z_i^{(t)} = 0). \end{cases}$$

座 i と座 j が QTL であるという仮定の下で，エピスタシスが存在しないという帰無仮説 $\beta_1 = \cdots = \beta_4 = 0$ の尤度比検定を考える．サンプルサイズ n に対するロッドスコア (尤度比検定統計量) を，$\mathrm{LRT}_n(i, j)$ と書く．多重性調整のため，エピスタシスが存在しないという帰無仮説の下での $\mathrm{LRT}_n(i, j)$ $(i, j = 1, \dots, m)$ の同時漸近分布を与えたい．尤度比検定の一般論より，帰無仮説の下では個体数 n についての漸近的性質として，各 i, j に対して $\mathrm{LRT}_n(i, j)$ は自由度 4 のカイ 2 乗分布に従う．しかしそれらは独立ではない．その相関構造は複雑なものとなるが，2 座 i, j が同じ染色体の上にない場合は，次のような直積型構造であることが示される．

命題 2.3 座間 i, j の組換え価を $\frac{1}{2}(1 - \rho_{ij})$ とおく．座 i と座 j が同じ染色体上にある (ない) ことを $i \sim j$ $(i \nsim j)$ と書く．エピスタシスが存在しないという帰無仮説の下で，全ての

$$(i, j) \in \{(i, j) \mid 1 \le i, j \le m,\ i \nsim j\}$$

についての同時分布の意味で，分布収束

$$\mathrm{LRT}_n(i, j) \ \Rightarrow\ T_{ij} = U_{1,ij}^2 + \cdots + U_{4,ij}^2 \quad (n \to \infty)$$

が成り立つ．ただし $(U_{k,ij})$ は k が異なると互いに独立な平均 0 の正規分布の配列で

$$\begin{aligned}\mathrm{Cov}(U_{k,ij}, U_{k,i'j'}) =& \rho_{ii'}\rho_{jj'}\ (k=1),\ \rho_{ii'}^2\rho_{jj'}\ (k=2),\\ & \rho_{ii'}\rho_{jj'}^2\ (k=3),\ \rho_{ii'}^2\rho_{jj'}^2\ (k=4)\end{aligned}$$

をみたすものである．

Mizuta *et al.* (2010) は，遺伝子座相互作用によって引き起こされる生殖隔離障壁を検出するために，2 座 i, j の遺伝子型の組合せとして得られる 3×3 表の独立性検定を行い，相互作用を独立性の乖離として検出することを試みた．その検定統計量の同時分布は帰無仮説の下で命題 2.3 と同じであることを示すことができる (Kuriki *et al.* 2010)．

2.6　区間マッピング法と Haley-Knott の回帰分析

単一マーカー分析では，各マーカー遺伝子座についてロッドスコアが計算された．ここで説明する 2 つの方法は，ロッドスコアを補完によってマーカー遺伝子座の位置以外でも定義するものであり，マーカー間隔が密でない場合に有効である．

ここでも F_2 集団で考える．QTL がある位置に 1 つ存在してそれが形質に影響を与えるというモデルを考える．このような仮想 QTL のことを putative QTL という．個体 t の QTL の遺伝子型を $z_*^{(t)}$ とおく．これは $1, 0, -1$ の値をとる潜在変数である．さらにこれは，相同染色体上の母方，父方の遺伝子型 $\epsilon_*^{(t)}, \delta_*^{(t)}$ $(=\pm 1)$ によって，$z_*^{(t)} = \frac{1}{2}(\epsilon_*^{(t)} + \delta_*^{(t)})$ と表される．

以下では混乱のない限り個体の添字 (t) を省略する．仮想 QTL の位置を γ とする．いま $d_i \le \gamma \le d_{i+1}$，つまり QTL はマーカー遺伝子座 i と $i+1$ の間に存在するとする．$\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ と ϵ_* の同時分布，$\delta = (\delta_1, \ldots, \delta_m)$ と δ_* の同時分布は，(2.1) と同様にマルコフ性によって γ の関数として陽に書き下すことができる．このことから，

$$z = (z_1, \ldots, z_m) = \frac{1}{2}(\epsilon + \delta) = \frac{1}{2}(\epsilon_1 + \delta_1, \ldots, \epsilon_m + \delta_m)$$

が与えられたときの，QTL の遺伝子型 z_* の条件付き分布が以下のように与えられる．

$$P(z_* \,|\, z\,; \gamma) = \frac{P(z, z_*\,; \gamma)}{P(z)}, \tag{2.7}$$

ただし

$$P(z,z_*\,;\gamma) = \sum_{z=(\epsilon+\delta)/2,\ z_*=(\epsilon_*+\delta_*)/2} \frac{1}{2^{2(m+1)}} \prod_{j=1,\ j\neq i}^{m-1} \Big(1+\epsilon_j\epsilon_{j+1}e^{-2(d_{j+1}-d_j)}\Big)\Big(1+\delta_j\delta_{j+1}e^{-2(d_{j+1}-d_j)}\Big) \times \Big(1+\epsilon_i\epsilon_* e^{-2(\gamma-d_i)}\Big)\Big(1+\epsilon_*\epsilon_{i+1}e^{-2(d_{i+1}-\gamma)}\Big) \times \Big(1+\delta_i\delta_* e^{-2(\gamma-d_i)}\Big)\Big(1+\delta_*\delta_{i+1}e^{-2(d_{i+1}-\gamma)}\Big), \qquad d_i \le \gamma \le d_{i+1},$$

$$P(z) = \sum_{z=(\epsilon+\delta)/2} \frac{1}{2^{2m}} \prod_{j=1}^{m-1} \Big(1+\epsilon_j\epsilon_{j+1}e^{-2(d_{j+1}-d_j)}\Big)\Big(1+\delta_j\delta_{j+1}e^{-2(d_{j+1}-d_j)}\Big). \tag{2.8}$$

(2.7) の $P(z_* \,|\, z\,;\gamma)$ は，$\gamma \to d_i$ または d_{i+1} のとき z_i または z_{i+1} に確率 1 で値をとる一点分布となる．$P(z_* \,|\, z\,;\gamma)$ は γ の連続関数であるが，マーカー点で滑らかではない．

仮想 QTL の遺伝子型の情報は，m 個のマーカーのなかで，とくに QTL に隣接するマーカー (flanking marker) i と $i+1$ が多く持っていると考えられる．そのため QTL の遺伝子型の予測のために (2.7) の代わりに $P(z_* \,|\, z_i,\ z_{i+1}\,;\gamma)$ を考えることもできる．これは簡便法であるが，戻し交配の場合のように，マーカー遺伝子型自体にマルコフ性が成り立つばあいには，(2.7) と正確に一致する．

Lander and Botstein (1989) の区間マッピング法 (interval mapping) とは，次の統計モデルを仮定した解析法である．

$$z_*^{(t)} \ \sim\ P(z_*^{(t)} \,|\, z^{(t)};\gamma), \tag{2.9}$$

$$y^{(t)} \ =\ \mu + \alpha z_*^{(t)} + \delta w(z_*^{(t)}) + \varepsilon^{(t)}, \quad \varepsilon^{(t)} \sim N(0,\sigma^2) \tag{2.10}$$

$(t=1,\ldots,n)$，ただし

$$w(k) = \begin{cases} 1 & (k=\pm 1), \\ -1 & (k=0). \end{cases}$$

また Haley and Knott (1992) は，区間マッピング法の簡便法として，潜在変数をその期待値で置きかえた次の回帰分析を提案した．

$$y^{(t)} = \mu + \alpha E[z_*^{(t)} \,|\, z^{(t)}\,;\gamma] + \delta E[w(z_*^{(t)}) \,|\, z^{(t)}\,;\gamma] + \varepsilon^{(t)}, \quad \varepsilon^{(t)} \sim N(0,\sigma^2) \quad (t=1,\ldots,n). \tag{2.11}$$

両者において，仮想 QTL の位置 γ の関数として尤度関数が定義されるので，連続な曲線としてロッドスコアが定義される．

また両者ともに論文で提案されたオリジナルの形は，QTL の遺伝子型の分布として隣接マーカーの情報だけを用いた $P(z_* \,|\, z_i, z_{i+1}\,; \gamma)$ を仮定するものであるが，本稿では全マーカーの情報を用いた $P(z_* \,|\, z\,; \gamma)$ で考えることにする．

以下では，区間マッピング法の尤度関数の形を書き下し，ロッドスコアの確率過程としての構造を調べていくことにする．Haley-Knott の回帰分析の場合は最後に触れる．

マーカーの遺伝子型 $z^{(t)}$ と QTL の位置 γ が与えられたとき $z_*^{(t)}$ は 3 値の離散分布に従う．記法を簡単にするために，その確率を (2.7) を使って

$$\pi_k^{(t)}(\gamma) = P(z_*^{(t)} = k \,|\, z^{(t)}\,; \gamma), \quad k = -1, 0, 1$$

とおくと，$z^{(t)}$ が所与のときの $y^{(t)}$ の分布は，コンポーネント数が 3 の正規分布の有限混合分布

$$\sum_{k=-1}^{1} \pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)})$$

となる．ただし正規分布 $N(\mu + \alpha k + \delta w(k), \sigma^2)$ の密度関数を $f_{k,\theta}(\cdot)$, $\theta = (\alpha, \delta, \mu, \sigma^2)$ とおいた．したがって，$(z^t, y^{(t)})$ $(t = 1, \ldots, n)$ の同時密度関数は，(2.8) を用いて

$$\prod_{t=1}^{n} \Big\{ \sum_{k=-1}^{1} \pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)}) P(z^{(t)}) \Big\}$$

と書ける．

ロッドスコアを描くためには，各 γ についてこの尤度を θ について最大化する必要がある．そのために EM アルゴリズム (Wu (1983) など) を用いることができる．ここで扱うモデルは，混合確率は個体 t に依存すること，また (γ を所与とすると) 混合確率には推測対象の未知パラメータが含まれないこと，の 2 点で通常の有限混合モデルとはやや異なっている．

$z_*^{(t)}$ と一対一に対応するダミー変数ベクトル

$$e^{(t)} = (e_{-1}^{(t)}, e_0^{(t)}, e_1^{(t)}), \quad e_k^{(t)} = \begin{cases} 1 & (z^{(t)} = k), \\ 0 & (z^{(t)} \neq k) \end{cases}$$

を導入する．マーカー遺伝子の遺伝子型，表現型，ならびに潜在変数 $(z^{(t)}, y^{(t)}, e^{(t)})$ $(t = 1, \ldots, n)$ の同時分布は

$$\prod_{t=1}^{n}\Big[\prod_{k=-1}^{1}\Big\{\pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)})\Big\}^{e_k^{(t)}} P(z^{(t)})\Big]$$

と書けるので

$$E_\theta\Big[e_k^{(t)} \mid (z^{(t)}, y^{(t)})_{t=1,\ldots,n}\Big] = \frac{\pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)})}{\sum_{k=-1}^{1} \pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)})}$$

となる．これを用いて，EM アルゴリズムは次のようにまとめられる．

1. 初期値
$$\hat{e}_k^{(t)} := \pi_k^{(t)}(\gamma) \quad (k = -1, 0, 1;\ t = 1, \ldots, n).$$

2. 以下を $\hat{\theta}$ と $\hat{e}_k^{(t)}$ が収束するまで繰りかえす．

$$\hat{\theta} := \operatorname*{argmax}_{\theta} \prod_{t=1}^{n}\prod_{k=-1}^{1}\Big\{\pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)})\Big\}^{\hat{e}_k^{(t)}},$$

$$\hat{e}_k^{(t)} := \frac{\pi_k^{(t)}(\gamma) f_{k,\hat{\theta}}(y^{(t)})}{\sum_{k=-1}^{1} \pi_k^{(t)}(\gamma) f_{k,\hat{\theta}}(y^{(t)})} \quad (k = -1, 0, 1;\ t = 1, \ldots, n).$$

ステップ 2 において $\hat{\theta}$ は，

$$Q_n(\theta\,;\,\gamma) = \sum_{t=1}^{n}\sum_{k=-1}^{1} \hat{e}_k^{(t)}\Big\{\frac{1}{\sigma^2}(y^{(t)} - \mu - \alpha k - \delta w(k))^2 + \log\sigma^2 + (\text{定数})\Big\}$$

を $\theta = (\alpha, \delta, \mu, \sigma^2)$ について最小化する操作 (重み付き最小 2 乗法) で陽に求めることができる．

1 節の図 1 は区間マッピング法により描いたものである．同じものを第 3 染色体について拡大して描いたものが図 3 である．ロッドスコア曲線が，マーカー間を補完していることが分かる．

注 2.4 ここで区間マッピング法におけるモデリングを振りかえってみる．統計モデルは，連鎖を記述する部分 (2.9) と，形質発現を記述する部分 (2.10) で構成されていた．後者の形質発現モデルを工夫することによって，QTL が複数ある場合や，表現型が非正規分布や離散分布，あるいは多変量分布に従う場合などを扱うことができる．このような統一的な扱いは，Sen and Churchill (2001) により提案され，R/qtl (Broman *et al.* 2003) として実装されている．

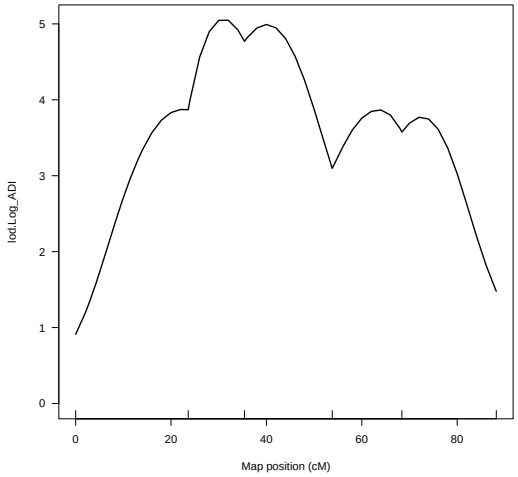


図 3: 区間マッピング法によるロッドスコア (第 3 染色体)

2.7 区間マッピング法のロッドスコア

区間マッピング法のロッドスコア (の定数倍である) $\mathrm{LRT}_n(\gamma)$ は，各 γ について尤度比検定統計量であり，QTL が存在しないという帰無仮説の下で漸近的に自由度 2 のカイ 2 乗分布に従う．ここでは，QTL が存在しないという帰無仮説の下で，$\mathrm{LRT}_n(\cdot)$ を確率過程 (漸近的カイ 2 乗確率過程) とみなしたときの相関構造を確定する．これは多重性調整のために必要である．

対数尤度は

$$L_n^{(\gamma)}(\theta) = \sum_{t=1}^{n} \log \sum_{k=-1}^{1} \pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)}) + (\theta \text{ を含まない項}),$$

$\theta = (\alpha, \delta, \mu, \sigma^2)$，であった．

パラメータ θ を $\theta = (\theta_1, \theta_2)$，ただし $\theta_1 = (\alpha, \delta)$, $\theta_2 = (\mu, \sigma^2)$ と分割表現する．帰無仮説は $H_0 : \theta_1 = 0$ である．一般性を失わずに，真値を $\theta_0 = (\theta_{10}, \theta_{20}) = (0, 0, 0, 1)$ とする．

γ を固定し，θ に関するフィッシャー情報行列を

$$I(\theta_0; \gamma) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} = \mathrm{Var}_{\theta_0}\Big(\frac{\partial L_n^{(\gamma)}}{\partial \theta}\Big|_{\theta_0} \Big) \tag{2.12}$$

とおく．尤度比検定の一般論から，真値 θ_0 の下で

$$\mathrm{LRT}_n(\gamma) \approx \left\{ \left(\frac{\partial L_n^{(\gamma)}}{\partial \theta_1}, \frac{\partial L_n^{(\gamma)}}{\partial \theta_2} \right) \begin{pmatrix} I \\ -I_{22}^{-1} I_{21} \end{pmatrix} I_{11\cdot 2}^{-1} \left(I, -I_{12} I_{22}^{-1} \right) \begin{pmatrix} \frac{\partial L_n^{(\gamma)}}{\partial \theta_1} \\ \frac{\partial L_n^{(\gamma)}}{\partial \theta_2} \end{pmatrix} \right\}_{(\theta_{10}, \theta_{20})}$$

ただし $I_{11\cdot 2} = I_{11} - I_{12} I_{22}^{-1} I_{21}$ である．

帰無仮説 $\alpha = \delta = 0$ のとき，$f_{k,\theta}$ は k によらない．このことから，スコア関数を真値で評価すると

$$\begin{aligned} \frac{\partial L_n^{(\gamma)}}{\partial \theta}\Big|_{\theta_0} &= \sum_{t=1}^{n} \frac{\sum_k \pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)}) \frac{\partial}{\partial \theta} \log f_{k,\theta}(y^{(t)})}{\sum_k \pi_k^{(t)}(\gamma) f_{k,\theta}(y^{(t)})}\Big|_{\theta_0} \\ &= \sum_{t=1}^{n} \sum_{k=-1}^{1} \pi_k^{(t)}(\gamma) \frac{\partial}{\partial \theta} \log f_{k,\theta}(y^{(t)})\Big|_{\theta_0} \\ &= \sum_{t=1}^{n} \begin{pmatrix} y^{(t)} \sum_k k \pi_k^{(t)}(\gamma) \\ y^{(t)} \sum_k w(k) \pi_k^{(t)}(\gamma) \\ y^{(t)} \\ (y^{(t)2} - 1)/2 \end{pmatrix} \end{aligned} \tag{2.13}$$

である．$L_n^{(\gamma)}(\theta_0)$ は γ に依存しない (特異モデルであるため) が，$\partial L_n^{(\gamma)}/\partial\theta|_{\theta_0}$ は γ に依存することに注意する．

スコアベクトルの分散共分散関数を

$$R(\gamma, \tilde{\gamma}) = \frac{1}{n} \mathrm{Cov}_{\theta_0}\Big(\frac{\partial L_n^{(\gamma)}}{\partial \theta}\Big|_{\theta_0}, \frac{\partial L_n^{(\tilde{\gamma})}}{\partial \theta}\Big|_{\theta_0} \Big)$$

とおく．

$$\sum_{k=-1}^{1} k \pi_k^{(t)}(\gamma) = E\big[z_*^{(t)} \,|\, z^{(t)}; \gamma \big],$$
$$\sum_{k=-1}^{1} w(k) \pi_k^{(t)}(\gamma) = E\big[w(z_*^{(t)}) \,|\, z^{(t)}; \gamma \big]$$

に注意すると，簡単な計算により

$$R(\gamma, \tilde{\gamma}) = \begin{pmatrix} R_{11}(\gamma, \tilde{\gamma}) & & O \\ & 1 & 0 \\ O & 0 & 1/2 \end{pmatrix}_{4\times 4}$$

ただし

$$R_{11}(\gamma,\tilde{\gamma}) = E\left[\begin{pmatrix}\sum_k k\pi_k^{(t)}(\gamma)\\ \sum_k w(k)\pi_k^{(t)}(\gamma)\end{pmatrix}\left(\sum_k k\pi_k^{(t)}(\tilde{\gamma}), \sum_k w(k)\pi_k^{(t)}(\tilde{\gamma})\right)\right]$$
$$= E\left[\begin{pmatrix}E[z_* \,|\, z\,;\gamma]\\ E[w(z_*) \,|\, z\,;\gamma]\end{pmatrix}\left(E[z_* \,|\, z\,;\tilde{\gamma}], E[w(z_*) \,|\, z\,;\tilde{\gamma}]\right)\right] \quad (2.14)$$

が分かる．ここで外側の期待値はマーカーの遺伝子型 z についてとる．フィッシャー情報行列 (2.12) は $I(\theta_0;\gamma) = nR(\gamma,\gamma)$ であり，ブロック対角行列 $(I_{12} = I_{21}^T = 0)$ となる．

$C(\gamma)$ を各成分が γ について滑らかな 2×2 行列で $C(\gamma)R_{11}(\gamma,\gamma)C(\gamma)^T = I_2$ をみたすものとする．中心極限定理より以下が示される．

命題 2.4 有限個の γ の，有限次元周辺分布の分布収束の意味で

$$\mathrm{LRT}_n(\gamma) \;\Rightarrow\; T(\gamma) = U(\gamma)^2 + V(\gamma)^2 \quad (n\to\infty).$$

ただし，$U(\cdot)$, $V(\cdot)$ は平均 0 の正規過程で，その共分散関数は (2.14) の $R_{11}(\gamma,\tilde{\gamma})$ を用いて

$$R^{UV}(\gamma,\tilde{\gamma}) = \mathrm{Cov}\left(\begin{pmatrix}U(\gamma)\\ V(\gamma)\end{pmatrix}, \begin{pmatrix}U(\tilde{\gamma})\\ V(\tilde{\gamma})\end{pmatrix}\right) = C(\gamma)R_{11}(\gamma,\tilde{\gamma})C(\tilde{\gamma})^T \quad (2.15)$$

として記述される．とくに固定した γ について $\mathrm{LRT}_n(\gamma)$ は漸近的に自由度 2 のカイ 2 乗分布に従う．

注 2.5 上の命題は，確率過程としての収束は述べていない．確率過程としての収束は，尤度比確率場の弱収束に関する Ibragimov and Has'minskii (1981) の方法で示すことができる．(Yoshida (2011), Introduction を参照．)

注 2.6 Haley-Knott の回帰分析 (2.11) のスコア関数を真値 θ_0 で評価すると，(2.13) と同じ形となる．ロッドスコアは区間マッピング法と同じ極限分布を持つ．

3　多重性調整のための方法

3.1　経験的方法とシミュレーション

前節ではいろいろな QTL 解析のモデルについて，QTL が存在しないという帰無仮説のもとで，ロッドスコアの確率構造を確定した．その結果を

出発点として，本節では多重性調整のために必要な，ロッドスコアの最大値の分布の近似法について説明する．そのために利用可能な数学的な方法として，非線形再生理論やオイラー標数法がある．それらについては，後の 3.2, 3.3 項で説明することとし，本項では QTL 解析の多重性調整のために行われている経験則とシミュレーションによる方法を紹介する．

Lander and Kruglyak (1995) は，ロッドスコアの多重性調整の目安として，多重性未調整の p 値を表 5 に従って解釈することを提唱している．しかしゲノムワイドの多重性調整は，染色体長のみならずマーカーの密度に大きく依存する．(3.2 項，表 4 の数値実験を参照．) そのため，機械的にこの表を用いることは行われてはいない (石川 (2006) など).

表 5: 多重性未調整 p 値の解釈

実験交配の手法	suggestive	significant
BC (1 d.f.)	3.4×10^{-3}	1.0×10^{-4}
F_2 (1 d.f., 加法効果)	3.4×10^{-3}	1.0×10^{-4}
F_2 (1 d.f., 優性効果)	2.4×10^{-3}	7.2×10^{-4}
F_2 (2 d.f)	1.6×10^{-3}	5.2×10^{-5}

多重性調整の方法として現在広く用いられている方法は，Churchill and Doerge (1994) の提案による並べ替え検定である．その手順は次のようなものである．

1. 個体数を n とし，集合 $\{1, \ldots, n\}$ の置換の全体を Π_n とおく．
2. N を十分大きな数とし，以下の手順を N 回繰りかえす．その繰り返しを $k = 1, \ldots, N$ とする．
   (i) ランダムに $\pi \in \Pi_n$ を選ぶ．
   (ii) 表現型 $(y^{(t)})_{t=1,\ldots,n}$ を，置換 π によって入れ替えたデータセット (表 6) について QTL 解析を行い，ロッドスコアの最大値を数値的に探索する．それを MaxLOD_k とおく．
3. $\{\mathrm{MaxLOD}_k\}_{k=1,\ldots,N}$ の経験分布をロッドスコアの最大値の経験分布とみなして p 値の推定値を

$$\widehat{p\text{値}} = \frac{\#\{k \mid \mathrm{MaxLOD}_k \geq \mathrm{MaxLOD}\ (\text{実現値})\}}{N}$$

と計算する．

表 6: 並べ替え検定のためのデータセット

個体番号	表現型	遺伝子型
1	$y^{(\pi(1))}$	$z^{(1)} = (z_1^{(1)}, \ldots, z_m^{(1)})$
$\vdots$	$\vdots$	$\vdots$
n	$y^{(\pi(n))}$	$z^{(n)} = (z_1^{(n)}, \ldots, z_m^{(n)})$

この並べ替え検定は直感的に分かりやすい方法であり，R/qtl などの多くのプログラムに実装されている．しかし次のような問題点がある．

最初の問題点は，並べ替え検定の一般論に関わる問題である．並べ替え検定が生成する $\{\mathrm{MaxLOD}_k\}_{k=1,\ldots,N}$ の経験分布は，帰無仮説が成り立つ場合とそうでない場合とでは当然異なるものである．そのために，帰無仮説が正しい場合には閾値を正しく推定できたとしても，帰無仮説が正しくない場合には閾値を過大評価する可能性がある．そのときは，ピークの検出確率 (検出力) の低下を招くことになる．

もうひとつの問題点は，計算量の問題である．説明した並べ替え検定の手順では，ロッドスコアの最大値を数値的に探索する必要があるため，計算時間がかかる．QTL を 1 つしか想定しない場合は問題ないが，エピスタシスの検定などで複数の QTL を想定する場合は，探索すべき組合せ数が莫大となり，計算が実行可能でなくなる場合がある．

並べ替え検定は汎用的でしばしば用いられる手法であるが，これらの理由のためその解釈や利用に注意を払う必要がある．可能な限り，他の多重性調整の方法の結果と併用するのがよいと思われる．問題を単一マーカー分析あるいは分離比の検定における多重性調整に限定すると，AR モデルを用いたモンテカルロシミュレーションも利用可能である．

1. $\epsilon_1, \ldots, \epsilon_m,\ \delta_1, \ldots, \delta_m$ を標準正規分布 $N(0,1)$ に従う i.i.d. 列とする．$U_1 = \epsilon_1, V_1 = \delta_1$ とおく．

2. $i = 2, \ldots, m$ について以下を計算する．

$$U_i = \alpha_i U_{i-1} + \sqrt{1-\alpha_i^2}\epsilon_i \quad (\alpha_i = e^{-2(d_i - d_{i-1})}),$$
$$V_i = \beta_i V_{i-1} + \sqrt{1-\beta_i^2}\delta_i \quad (\beta_i = e^{-4(d_i - d_{i-1})}),$$
$$T_i = U_i^2 + V_i^2.$$

3. $\max_{1\le i\le m} T_i$ を計算する．

上の手続きを十分な回数繰り返すことによって，命題 2.1, 2.2 の $\max_{1\le i\le m} T_i$ の経験分布を求めることができる．しかしながら，並べ替え検定のときと同様，複数の QTL を想定した場合には計算量が膨大となる．

3.2 非線形再生理論による近似

本項以降では，ロッドスコアの最大値の分布を理論的に求める (近似する) ための方法を 2 つ紹介する．最初に述べる方法は，逐次解析，非線形再生理論を用いるものであり，単一マーカー分析や分離のゆがみの検定において，マーカーが密で間隔が等間隔に近い場合 ($|d_{i+1} - d_i| \approx \Delta \ll 1$) に有効な方法である (Dupuis and Siegmund 1999)．

まず一般的な形で問題設定を行う．$Z_k(t),\ t \in I \subset \mathbb{R}$ $(k = 1, 2)$ を互いに独立な正規過程で，平均 0，共分散関数が

$$\mathrm{Cov}(Z_k(t), Z_k(\tilde{t})) = e^{-\rho_k|t-\tilde{t}|} \quad (\rho_k > 0)$$

であるもの (Ornstein-Uhlenbeck 過程) とする．自由度 2 のカイ 2 乗確率過程の平方根を

$$Y(t) = \sqrt{Z_1(t)^2 + Z_2(t)^2} \quad (t \in I)$$

で定義する．

命題 3.1 格子間隔を $\Delta > 0$ とし，$J = \{j \in \mathbb{Z} \mid j\Delta \in I\}$ とおく．$b \to \infty$, $\Delta \to 0$, $b\sqrt{\Delta} \to c\ (> 0)$ のとき

$$P\Big(\max_{j\in J} Y(j\Delta) \ge b\Big) \sim |I|\, b^2 e^{-b^2/2} \int_0^{2\pi} \frac{d\theta}{2\pi}\, \bar{\rho}(\theta)\, \nu(c\sqrt{2\bar{\rho}(\theta)}). \tag{3.1}$$

ここで $|\cdot|$ は集合のルベーグ測度，$\bar\rho(\theta)=\rho_1\cos^2\theta+\rho_2\sin^2\theta$．また $\Phi(\cdot)$ を標準正規分布の分布関数とするとき

$$\nu(x)=\begin{cases}2x^{-2}\exp\Big\{-2\displaystyle\sum_{n=1}^{\infty}n^{-1}\Phi\Big(-\frac{1}{2}x\sqrt{n}\Big)\Big\} & (x>0),\\ 1 & (x=0).\end{cases}$$

注 3.1 $x\le 2$ くらいでは $\nu(x)\approx\exp(-0.583x)$ と近似できる．実用的にはこの範囲で足りることが多い．

注 3.2 形式的に $\Delta=0$, $c=0$ とおいて得られる式

$$P\Big(\sup_{t\in I}Y(t)\ge b\Big)\ \sim\ |I|Kb^2e^{-b^2/2},\quad K=\int_0^{2\pi}\frac{d\theta}{2\pi}\bar\rho(\theta)=\frac{1}{2}(\rho_1+\rho_2)\quad(b\to\infty)$$

も成り立つ (Piterbarg (1996), Corollary 7.1 より)．

$I=[0,1]$, $\Delta=0.01$ (染色体長 1M, マーカー数 $m=100$ に相当) の場合の数値例を図 4 に示す．一点鎖線 $(-\cdot-)$ と実線 (—) はそれぞれ格子点上の最大値の上側確率 $P\big(\max_{j\in J}Y(j\Delta)\ge b\big)$ の命題 3.1 による近似値とシミュレーションによる推定値を表している．すなわち図の縦軸は p 値に対応する．ただしここでは最大値分布の近似として (3.1) の右辺をそのまま用いるのではなく，それと漸近的に同等な $1-\exp\{-(3.1)$ 右辺 $\}$ を用いている．この両者は p 値が 0.5 程度より小さい範囲で非常に精度良く一致していることが見て取れる．

また破線 $(--)$ は $\Delta=0$，すなわち連続集合 I 上の最大値分布 $P\big(\sup_{t\in I}Y(t)\ge b\big)$ である．Lander and Botstein (1989) は区間マッピング法を提案するとともに，Ornstein-Uhlenbeck 過程の最大値の分布を用いてゲノムワイドな多重性調整を行うことを提唱している．それがこの場合に対応するが，マーカー間隔 10cM ($\Delta=0.01$) という密な場合であってもその近似はあまり良くない．

点線 $(\cdots)$ は自由度 2 のカイ 2 乗分布の上側確率である．これは多重調整を行わない場合の p 値に対応する．

命題 3.1 の証明は，節末で与える．なおここで与えた近似は，マーカー間距離が常に一定値 Δ であるというモデルに基づくものである．実際問題への適用の際には，Δ にマーカー間距離の平均値を代入して用いることになる．Dupuis and Siegmund (1999) は類似の問題設定において，そのような場合であってもよい近似を与えることを数値計算によって確認している．

また命題 2.3 で与えた 2 次元格子上のカイ 2 乗確率場の最大値についても，同様の近似を与えることができる (Kuriki *et al.* 2010)．

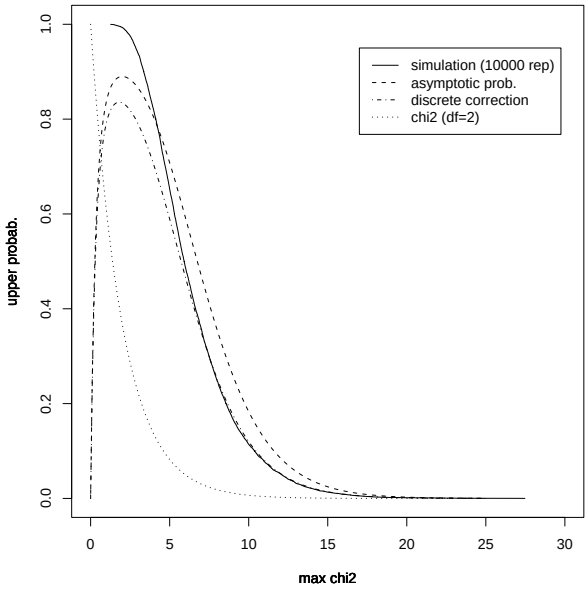


図 4: 格子点上の最大値の分布

(–·–：命題 3.1 による近似，—：シミュレーション，– –：連続近似，· · ·：χ^2_2 分布)

3.3 ランダム関数の零点の個数の期待値

次に，滑らかなサンプルパスを持つ確率場の最大値の上側裾確率を近似するためのオイラー標数法 (Euler characteristic heuristic) を，添字集合が 1 次元という特殊な場合について説明する．この方法は，信号処理の分野でライスの公式 (Rice's formula) として知られているものと同等である．区間マッピング法のロッドスコアは，漸近的には区分的に滑らかなサンプルパスを持つカイ 2 乗確率過程であったので，オイラー標数法によってその最大値の分布を近似することができる．オイラー標数法の一般論については，栗木・竹村 (2008) を参照のこと．

$Z(t), t \in I \subset \mathbb{R}$ は，実数に値をとるランダムな C^1 関数で，各 t について $(Z(t), \dot{Z}(t))$ $(\dot{Z}(t) = dZ(t)/dt)$ が縮退しない分布を持つとする．$Z(t) = u$ となる t (方程式 $Z(t) - u = 0$ の零点) の個数を

$$N_u = \#\{t \in I \mid Z(t) = u\}$$

とおく．

ところで，任意の連続関数 f に対して

$$\int |\dot{Z}(t)|\, f(Z(t))\, dt = \int N_u f(u)\, du \tag{3.2}$$

が成り立つことが容易に分かる (図 5)．

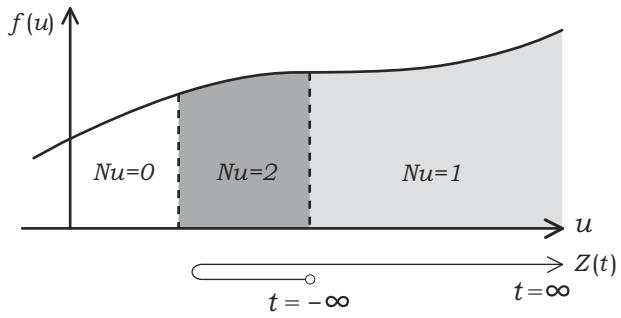


図 5: (3.2) の図による説明

(3.2) の両辺について，$Z(\cdot)$ の期待値をとると

$$\begin{aligned}\int E[N_u] f(u) du &= \int E\big[|\dot{Z}(t)|\, f(Z(t))\big]\, dt \\ &= \int\int E\big[|\dot{Z}(t)| \mid Z(t) = u\big] f(u)\, p_{Z(t)}(u)\, du\, dt\end{aligned}$$

を得る．最右辺において，$p_{Z(t)}$ は $Z(t)$ の周辺密度である．これが任意の連続関数 f で成り立つので，

$$E[N_u] = \int E\big[|\dot{Z}(t)| \mid Z(t) = u\big]\, p_{Z(t)}(u)\, dt \quad \text{a.s.} \tag{3.3}$$

である (Azaïs and Wschebor 2005)．

いま閾値 u が大きいとする．このとき，関数 $Z(t)$ は閾値 u を超えることは稀であり，もし一度超えてもまたすぐに閾値を下回ることが予想される．つまり N_u が 0 または 2 以外の値をとる確率が小さいことが期待される．そのような状況では

$$\begin{aligned}\frac{1}{2}E[N_u] &\approx P(N_u = 2) \\ &\approx P\big(\exists t \in I,\ Z(t) \geq u\big) \ = \ P\Big(\sup_{t\in I} Z(t) \geq u\Big)\end{aligned}$$

である．オイラー標数近似とは，最大値分布の近似式として

$$P\Big(\sup_{t\in I} Z(t) \geq u\Big) \approx \frac{1}{2}E[N_u] \quad (u \text{ が大きいとき}) \tag{3.4}$$

とおき，(3.4) の右辺として，(3.3) を用いるというものである．ここではこのような直感的な議論にとどめるが，正規確率場，カイ 2 乗確率場に対するオイラー標数法は，正則条件の下で非常に良い近似を与えることが知られている．(栗木・竹村 (2008) の参考文献を参照．)

このオイラー標数法によって，F_2 集団の区間マッピング法によるロッドスコアの最大値の分布を近似しよう．仮想 QTL の位置を γ とする．区間マッピング法で補間されたロッドスコア (尤度比検定統計量) $\mathrm{LRT}_n(\gamma)$, $\gamma \in \Gamma$ の，QTL が存在しないという帰無仮説のもとでの漸近分布は，連続かつ隣り合うマーカーの間では滑らかなパスを持つカイ 2 乗確率過程 $T(\gamma)$, $\gamma \in \Gamma$ であった．全ての隣り合うマーカー間について $E[N_u]/2$ を計算し，それを全て足しあわせることによって最大値分布の近似が可能である (Rebaï *et al.* 1994).

命題 3.2 $R^{UV}(\gamma, \tilde{\gamma})$ を，命題 2.4 で与えた共分散関数 (2.15) とする．

$$A(\gamma) = \frac{\partial}{\partial\tilde{\gamma}} R^{UV}(\gamma,\tilde{\gamma})\Big|_{\tilde{\gamma}=\gamma}, \qquad B(\gamma) = \frac{\partial^2}{\partial\gamma\partial\tilde{\gamma}} R^{UV}(\gamma,\tilde{\gamma})\Big|_{\tilde{\gamma}=\gamma}$$

とおく．$h = (h_1, h_2)^T$ を円周上の一様分布 $\mathrm{Unif}(\mathbb{S}^1)$ とする．ロッドスコアの極限過程 $T(\cdot)$ の，$\Gamma = [d_1, d_m]$ 上最大値のオイラー標数近似は

$$\begin{aligned} P\Big(\sup_{\gamma\in\Gamma} T(\gamma) \geq u\Big) \approx{}& \frac{1}{\sqrt{2\pi}} u^{1/2} e^{-u/2} \\ &\times \sum_{i=1}^{m-1} \int_{d_i}^{d_{i+1}} E\Big[\sqrt{h^T(B(\gamma) - A(\gamma)^T A(\gamma))h}\Big]\, d\gamma \end{aligned} \tag{3.5}$$

で与えられる．

この命題の証明も，本節の節末で与える．その導出は Davies (1987), Theorem A.1 と本質的に同じである．

3.4 命題の証明

3.4.1 命題 3.1 の証明

ここで与える証明は，Kim and Siegmund (1989) と Siegmund (1992) を組合せたものである．t を固定し，$j \in \mathbb{Z}$ を添字とみなした確率変数列

$$\widetilde{Y}_j = b(Y(t+j\Delta) - y)\Big|_{(Z_1(t),Z_2(t))=(y\cos\theta, y\sin\theta)}$$

を定義する．$Y(t) = \sqrt{Z_1(t)^2 + Z_2(t)^2} = y$ に注意する．$b, y \to \infty$, $b \sim y$, $\Delta \to 0$, $b\sqrt{\Delta} \to c$ (> 0) の極限を考える．$Y(t+j\Delta)$ を t のまわりでテイラー展開し，また正規分布の条件付分布の公式より，$(Y_j)_{j\in\mathbb{Z}}$ の有限次元の周辺分布は正規分布で，その平均，分散は

$$\begin{aligned} E[\widetilde{Y}_j] &= -\bar{\rho}(\theta)c^2|j|, \\ \mathrm{Cov}(\widetilde{Y}_j, \widetilde{Y}_{\tilde{j}}) &= \bar{\rho}(\theta)c^2(|j| + |\tilde{j}| - |j-\tilde{j}|) \\ &= \begin{cases} 2\bar{\rho}(\theta)c^2 \min(|j|, |\tilde{j}|) & (j \text{ と } \tilde{j} \text{ は同符号}), \\ 0 & (\text{異符号}) \end{cases} \end{aligned}$$

であることを確認できる．これは X_i $(i \in \mathbb{Z})$ を独立な正規分布 $N(-\bar{\rho}(\theta)c^2, 2\bar{\rho}(\theta)c^2)$ の列としたときの，両側ランダムウォーク

$$S_j = \begin{cases} X_1 + \cdots + X_j & (j > 0), \\ 0 & (j = 0), \\ X_{-1} + \cdots + X_{-|j|} & (j < 0) \end{cases}$$

の分布に等しい．すなわち

$$(\ldots, \widetilde{Y}_{-1}, \widetilde{Y}_0, \widetilde{Y}_1, \ldots) \quad \Rightarrow \quad (\ldots, S_{-1}, S_0, S_1, \ldots) \tag{3.6}$$

である．

条件 $(Z_1(j^0\Delta), Z_2(j^0\Delta)) = (y\cos\theta, y\sin\theta)$ を与えた条件付確率測度を $\widetilde{P}$ で表す．$-x = b(b-y)$ (すなわち $y = b + x/b$) とおく．まず (3.6) より

$$\begin{aligned} \widetilde{P}\Big(\max_{j>j^0} Y(j\Delta) < b\Big) &= \widetilde{P}\Big(\max_{j>j^0} b(Y(j\Delta) - y) < -x\Big) \\ &= \widetilde{P}\Big(\max_{j>0} \widetilde{Y}_j < -x\Big) \ \to \ P\Big(\max_{j>0} S_j < -x\Big) \end{aligned}$$

に注意する．

これからが証明の本体である．$Y(j\Delta) \geq b$ をみたすような点 j の最大値を j^0 とおく．事象 $\{\max_{j\in J} Y(j\Delta) \geq b\}$ は j^0 の値によって排反に分割される．

$$P\Big(\max_{j\in J} Y(j\Delta) \geq b\Big) = \sum_{j^0\in J} P\Big(\max_{j>j^0} Y(j\Delta) < b,\ Y(j^0\Delta) \geq b\Big). \qquad (3.7)$$

さらに事象を $\tan^{-1}\frac{Z_2(j^0\Delta)}{Z_1(j^0\Delta)}$ の値によって排反に分割する．

$$\begin{aligned}(3.7)\text{ 右辺} &= \int_0^{2\pi} \sum_{j^0\in J} P\Big(\max_{j>j^0} Y(j\Delta) < b,\ Y(j^0\Delta) \geq b,\ \tan^{-1}\frac{Z_2(j^0\Delta)}{Z_1(j^0\Delta)} \in d\theta\Big) \\ &= \int_{y\geq b} \int_0^{2\pi} \sum_{j^0\in J} P\Big(\max_{j>j^0} Y(j\Delta) < b, \\ &\qquad\qquad Y(j^0\Delta) \in dy,\ \tan^{-1}\frac{Z_2(j^0\Delta)}{Z_1(j^0\Delta)} \in d\theta\Big),\end{aligned}$$

ただし $d\theta = (\theta, \theta + d\theta)$, $dy = (y, y + dy)$ である．

$$Y(j^0\Delta) = y, \quad \tan^{-1}\frac{Z_2(j^0\Delta)}{Z_1(j^0\Delta)} = \theta$$

と値を与えることは $(Z_1(j^0\Delta), Z_2(j^0\Delta)) = (y\cos\theta, y\sin\theta)$ と値を与えることと等価なので，

$$\begin{aligned}&P\Big(\max_{j>j^0} Y(j\Delta) < b,\ Y(j^0\Delta) \in dy,\ \tan^{-1}\frac{Z_2(j^0\Delta)}{Z_1(j^0\Delta)} \in d\theta\Big) \\ &\quad = \widetilde{P}\Big(\max_{j>j^0} Y(j\Delta) < b\Big) \times P\Big(Y(j^0\Delta) \in dy,\ \tan^{-1}\frac{Z_2(j^0\Delta)}{Z_1(j^0\Delta)} \in d\theta\Big) \\ &\quad \to P\Big(\max_{j>0} S_j < -x\Big) \times P(Y(j^0\Delta) \in dy) \times \frac{d\theta}{2\pi}\end{aligned}$$

である．ここで $y = b + x/b$ と変数変換する．$\int_{y>b} = \int_{x>0}$ ならびに

$$P(Y(j^0\Delta) \in dy) = ye^{-y^2/2}dy \ \sim\ e^{-b^2/2}e^{-x}dx \quad (b \to \infty)$$

に注意する．

さらに逐次解析で知られた関係式

$$\int_0^\infty e^{-x} P\Big(\max_{j>0} S_j < -x\Big)dx = \bar{\rho}(\theta)c^2\nu(c\sqrt{2\bar{\rho}(\theta)})$$

および

$$\Delta^{-1}c^2 \sim b^2, \quad \Delta \sum_{j^0 \in J} \sim \int_I dt \quad (\Delta \to 0)$$

を組合せると (3.1) を得る． ∎

3.4.2 命題 3.2 の証明

仮想 QTL の位置を $\gamma \in [d_i, d_{i+1}]$ とする．命題 2.4 の共分散関数 $R^{UV}(\gamma, \tilde{\gamma})$ (2.15) は引数について十分に滑らかである．とくに

$$\Big(\frac{\partial}{\partial\gamma}\Big)^3 \Big(\frac{\partial}{\partial\tilde{\gamma}}\Big)^3 R^{UV}(\gamma, \tilde{\gamma})$$

が $\gamma = \tilde{\gamma}$ の近傍で存在し有界なので，$U(\cdot), V(\cdot)$ は確率 1 で連続微分可能で，また微分過程 $\dot{U}(\cdot), \dot{V}(\cdot)$ も連続な正規過程となる (Adler and Taylor (2007), Theorem 1.4.2)．

微分過程 $\dot{U}(\cdot), \dot{V}(\cdot)$ も平均 0 で，その共分散関数は

$$\mathrm{Cov} \begin{pmatrix} U(\gamma) \\ V(\gamma) \\ \dot{U}(\gamma) \\ \dot{V}(\gamma) \end{pmatrix} = \begin{pmatrix} I_2 & A(\gamma) \\ A(\gamma)^T & B(\gamma) \end{pmatrix}$$

である．ここで

$$0 = \frac{d}{d\gamma}\mathrm{Cov}(U(\gamma), U(\gamma)) = 2\mathrm{Cov}(\dot{U}(\gamma), U(\gamma)),$$
$$0 = \frac{d}{d\gamma}\mathrm{Cov}(U(\gamma), V(\gamma)) = \mathrm{Cov}(\dot{U}(\gamma), V(\gamma)) + \mathrm{Cov}(U(\gamma), \dot{V}(\gamma))$$

であるので $A(\gamma)^T = -A(\gamma)$ (交代行列) である．

以下では引数の γ は省略する．$(\dot{U}, \dot{V})$ の，(U, V) を与えた条件付き分布は

$$\begin{pmatrix} \dot{U} \\ \dot{V} \end{pmatrix}\Big|_{(U,V)} \sim N\Big(A^T \begin{pmatrix} U \\ V \end{pmatrix}, B - A^T A\Big)$$

であり，また

$$T = (U, V)\begin{pmatrix} U \\ V \end{pmatrix}, \qquad \dot{T} = 2(U, V)\begin{pmatrix} \dot{U} \\ \dot{V} \end{pmatrix}$$

より，

$$\begin{aligned}\dot{T}\Big|_{(U,V)} &\sim N\Big(2(U,V)A^T\begin{pmatrix}U\\V\end{pmatrix}, 4(U,V)(B-A^TA)\begin{pmatrix}U\\V\end{pmatrix}\Big)\\ &= N\Big(0, 4(U,V)(B-A^TA)\begin{pmatrix}U\\V\end{pmatrix}\Big)\end{aligned}$$

である．したがって

$$\frac{\dot{T}}{\sqrt{4(U,V)(B-A^TA)\binom{U}{V}}}\Big|_{(U,V)} = \xi \sim N(0,1)$$

となり，ξ の条件付き分布は条件 (U,V) によらないことが分かるので，

$$\dot{T} = \xi\sqrt{4(U,V)(B-A^TA)\begin{pmatrix}U\\V\end{pmatrix}} \qquad (\xi \sim N(0,1) \text{ は } (U,V) \text{ と独立})$$

と表すことができる．

さらに

$$h = \begin{pmatrix}h_1\\h_2\end{pmatrix} = \frac{1}{\sqrt{T}}\begin{pmatrix}U\\V\end{pmatrix}$$

とおくと，$\dot{T} = 2\xi\sqrt{T}\sqrt{h^T(B-A^TA)h}$ であり，h は ξ, T と独立に，円周上の一様分布 $\mathrm{Unif}(\mathbb{S}^1)$ に従う．

以上より

$$\begin{aligned}E\big[|\dot{T}| \mid T=u\big]\, p_T(u) &= E\big[|\dot{T}|\, 1_{\{T\in(u,u+du)\}}\big]/du\\ &= 2E[|\xi|]\, E\big[\sqrt{T}\, 1_{\{T\in(u,u+du)\}}\big]/du \times E\Big[\sqrt{h^T(B-A^TA)h}\Big]\\ &= 2\sqrt{\frac{2}{\pi}}u^{1/2}\frac{1}{2}e^{-u/2}E\Big[\sqrt{h^T(B-A^TA)h}\Big].\end{aligned}$$

これを γ で積分して，

$$\frac{1}{2}E[N_u] = \frac{1}{\sqrt{2\pi}}u^{1/2}e^{-u/2}\int_{d_i}^{d_{i+1}} E\Big[\sqrt{h^T(B(\gamma)-A(\gamma)^TA(\gamma))h}\Big]\, d\gamma,$$

ただし右辺の期待値は $h \sim \mathrm{Unif}(\mathbb{S}^1)$ についてとる．

これを全ての隣り合うマーカーについて足しあわせることによって (3.5) が得られる． ∎

謝辞

1 節で紹介したマウスの QTL 解析例 (図 1, 3) は，国立遺伝学研究所の城石研究室 (前野哲輝氏，城石俊彦氏) によるものです．また同研究所の春島嘉章氏，倉田のり氏からは，本稿の内容について有益なコメントをいただきました．本稿は，数理科学総合セミナー II (2006 年度前期，東京大学数理科学研究科) の講義録に加筆したものです．同大学の吉田朋広氏には，注 2.5 をご指摘いただいたきました．これらの方々に感謝いたします．

参考文献

[1] Adler, R. J. and Taylor, J. E. (2007). *Random Fields and their Geometry*, Springer.

[2] Azaïs, J-M. and Wschebor, M. (2005). On the distribution of the maximum of a Gaussian field with d parameters, *Ann. Appl. Probab.*, **15** (1A), 254–278.

[3] Broman, K. W., Wu, H., Sen, Ś. and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses, *Bioinformatics*, **19** (7), 889–890.

[4] Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative triat mapping, *Genetics*, **138** (3), 963–971.

[5] Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternatives, *Biometrika*, **74** (1), 33–43.

[6] Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers, *Genetics*, **151** (1), 373–386.

[7] 福水健次, 栗木哲, 竹内啓, 赤平昌文 (2004). 『特異モデルの統計学 — 未解決問題への新しい視点』, 統計科学のフロンティア 7, 岩波書店.

[8] Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors, *J. Genetics*, **8**, 299-309.

[9] Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers, *Heredity*, **69**, 315–324.

[10] Harushima, Y., Nakagahra, M., Yano, M., Sasaki, T. and Kurata, N. (2001). A genome-wide survey of reproductive barriers in an intraspecific hybrid, *Genetics*, **159** (2), 883–892.

[11] Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*, Wiley.

[12] Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*, Springer.

[13] 石川明 (2006). 動物モデルによる多因子性疾患の QTL 解析：その基礎的理論と解析方法 (改訂版), NAGOYA Repository, `http://hdl.handle.net/2237/6779`

[14] Karlin, S. and Liberman, U. (1983). Measuring interference in the chiasma renewal formation process, *Adv. Appl. Probab.*, **15** (3), 471–487.

[15] Kim, H.-J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression, *Biometrika*, **76** (3), 409–423.

[16] Kuriki, S., Harushima, Y., Fujisawa, H. and Kurata, N. (2010). Approximate tail probabilities of the maximum of a chi-square field on multidimensional lattice points and their applications to detection of loci interactions, arXiv:1012.4921, `http://arxiv.org/abs/1012.4921`

[17] 栗木哲, 竹村彰通 (2008). チューブの体積と正規確率場の最大値の分布, 数学, **60** (2), 134–155.

[18] Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics*, **121** (1), 185–199.

[19] Lander, E. S. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics*, **11** (3), 241–247.

[20] Manichaikul, A., Dupuis, J., Sen, S. and Broman, K. W. (2006). Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus, *Genetics*, **174** (1), 481–489.

[21] Mizuta, Y., Harushima, Y. and Kurata, N. (2010). Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes, *Proc. Natl. Acad. Sci. USA*, **101** (47), 20417–20422.

[22] Piterbarg, V. I. (1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields*, Translations of Mathematical Monographs, 148, AMS.

[23] Rebaï, A., Goffinet, B. and Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection, *Genetics*, **138** (1), 235–240.

[24] Sen, S. and Churchill, G. A. (2001). A statistical framework for quantitative trait mapping, *Genetics*, **159** (1), 371–387.

[25] Siegmund, D. (1985). *Sequential Analysis*, Springer.

[26] Siegmund, D. O. (1992). Tail approximations for maxima of random fields, in *Probability Theory*, L. H. Y. Chen, K. P. Choi, K. Hu and J-H. Lou (eds.), Walter de Gruyter, 147–158.

[27] 鵜飼保雄 (2000). 『ゲノムレベルの遺伝解析 — MAP と QTL』, 東京大学出版会.

[28] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm, *Ann. Statist.*, **11** (1), 95–103.

[29] Wu, R., Ma, C.-X., and Casella, G. (2007). *Statistical Genetics of Quantitative Traits*, Springer.

[30] Yoshida, N. (2011). Polynomial type large deviation inequalities and convergence of statistical random fields, *Ann. Inst. Statist. Math.*, **63** (3), 431–479.