統計数理 第66巻 第2号

（通巻128号）

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目　次

特集「サービス科学の今」

2018年12月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）

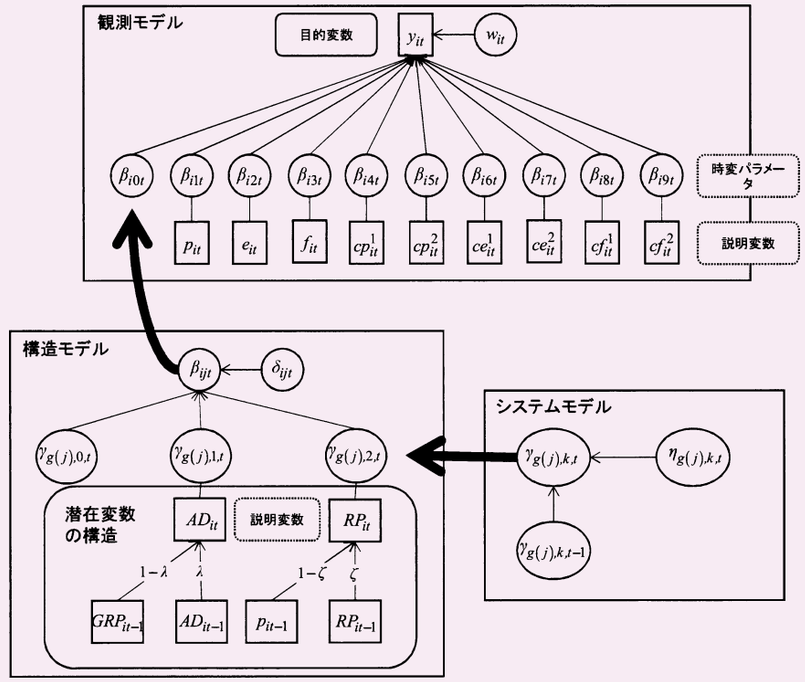
本号の内容はすべて https://www.ism.ac.jp/editsec/toukei/ からダウンロードできます

ISSN 0912-6112

統計数理　PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS　第66巻 第2号 2018

統計数理 Vol. 66, No.2

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS



統計数理研究所

統計数理

（年 2 回発行）

「統計数理」は，統計数理研究所における研究成果を掲載する統計数理研究所「彙報」として 1953 年に歴史を始め，1985 年に誌名を変更し今の形となりました．現在は，統計数理研究所の研究活動に限らず，広く統計科学に関する投稿論文を掲載し，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指しています．

投稿を受け付けるのは，次の 6 種です．

a. 原著論文　b. 総合報告　c. 研究ノート
d. 研究詳解　e. 統計ソフトウェア　f. 研究資料

投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定します．投稿規程，執筆要項は，本誌最終頁をご参照ください．

また，上記以外にも統計科学に関して編集委員会が重要と認める内容について，編集委員会が原稿作成を依頼することがあります．

その他，「統計数理」に関するお問い合わせは，各編集委員にお願いします．

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
〒 190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）
https://www.ism.ac.jp/

印刷：笹氣出版印刷株式会社

表紙の図は本誌 258 ページを参照

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 66, No. 2

Contents

December, 2018
Research Organization of Information and Systems
The Institute of Statistical Mathematics
10–3 Midori-cho, Tachikawa, Tokyo 190–8562, JAPAN

統計数理(2018)
第 66 巻 第 2 号 189–191

「特集 サービス科学の今」について

丸山 宏[1]・中野 純司[2](オーガナイザー)

統計数理研究所では，その時期において重要と思われる課題に対して研究所内外の研究者を集めて研究センターを開設し，集中的に研究教育活動を行っている．サービス科学研究センターは平成 24–28 年度に活動を行った研究センターであり，センター長としては丸山宏(当時統計数理研究所，現 Preferred Networks)が着任し，その後，中野純司(統計数理研究所)が引き継いだ．本特集はその活動の一部を論文としてまとめたものである．

サービス科学研究センターの開設に際しては，以下のように考えた．

世界の産業構造は急速に変化している．我が国の就業人口の 8 割以上が，今では医療，流通，金融などのサービス産業に従事していると言われている．しかしながら，サービスビジネスを設計し，運用するための科学的方法論は未だに確立されておらず，経営における意思決定の多くは勘と経験にもとづいている．これは，機械工学，化学工学，電子工学などのディシプリンを持つ製造業とは対照的である．サービス科学とは，サービス産業や公共サービス事業，それにその周辺の産業エコシステムに対して，科学的な方法論全体を持ち込む試みを指す言葉で，2006 年ころから使われるようになってきている．一方，新しい科学の方法論として，第 4 の科学と呼ばれるデータ中心の考え方が注目を浴びている．理論的なモデルを立てて，そこからの演繹によって自然界を説明できるかどうかを問うのが今までの科学における主流の方法論であった．しかし，情報技術の進展にともなってあらゆる分野で大量のデータが得られるようになって，モデルを予め立てるのではなく，これらのデータを元に帰納的に世の中を説明しようというものがデータ中心科学である．そして，データ中心科学を実施するのに中心的な役割を果たすのが，情報技術と統計科学である．われわれは，統計数理研究所が伝統的に持つ統計科学の深い知見と，最新の情報技術を組み合わせ，それをサービス科学に適用することで，データ中心のサービス科学を推進していく．サービスとは何か，については多くの議論がある．サービス業と対比して語られるのは製造業であるが，今日では製造業とサービス業の境界が曖昧になってきている．われわれは，サービスを「顧客に価値を提供する事業」として広く捉えることにする．そして，価値を生み出すプロセスと，そこにおけるデータに基づく意思決定のメカニズムに焦点を当てる．サービスにおける顧客価値は，多分に主観的なものであり，そのサービスが提供される文脈によって価値が大きく変動する．マーケティングは，このように大きく変動する顧客価値を定量化し予測する試みの分野と言える．同時に，事業の価値は，より大きな顧客価値をより小さな資源投入で達成することによって最大化される．このように，顧客価値の最大化とコスト最小化は，どちらもサービスにおける「価値を生み出すプロセス」でありサービス科学の主要な課題である．われわれは，サービス科学 NOE(Network Of Excellence)を軸に，国内外の多くの研究者とともに，特にデータ中心科学の観点からこれらの課題に挑戦していく．また，データに基づく意思決定のためには，データの収集・キュレーション・分析に関して確立された方法論と，データの流通に関する多様な利害関係者間の共通理解が必要で

[1] 株式会社 Preferred Networks：〒 100–0004 東京都千代田区大手町 1–6–1 大手町ビル 2F
[2] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

あり，その分野においても，積極的に貢献していきたいと考える．

そして，次のようなプロジェクトを開始した．

1)製品・サービスの質保証・信頼性研究プロジェクト

本プロジェクトの目的は質保証・信頼性に資する統計的方法の開発と産業界への展開を推進し，品質・サービスの質確保と安全の実現に寄与することである．ロバストパラメータ設計は，創始者である田口玄一博士が半世紀かけて体系化した品質を向上させるための技術方法論で，使用環境条件などの誤差因子に対してロバストになるように制御因子を設計することにより特性や機能性のばらつきを低減する方法である．これらの方法論は制御因子の水準変更のみでばらつきの低減を図れるという，経済的かつ効果的であるため，我が国の「ものづくり」の設計開発の現場を中心に利用されてきたものである．

2)ビッグデータ対応型ベイズモデル開発・研究プロジェクト

本プロジェクトは，情報化社会の副産物であるビッグデータを活用して，サービスの個別対応を実現するビッグデータ対応型のベイズモデル開発・研究を行うものである．情報化の進展に伴う知識社会においては，従来の大量生産・大量消費と異なり，一人ひとりの個性や置かれている状況要因の違いによる異質なニーズに適応するサービスの提供が求められる．ビッグデータに対して，人文社会科学の知見と急速に高度化しているベイズモデリング技術を適用することにより，サービスの供給者と受給者相互の価値を生み出しながら，生産性向上と新たなサービス創造する可能性がある．供給側の視点による大量生産消費社会から受給者の理解と知識獲得にもとづく生活者満足度実現社会への転換，さらに両者の価値の共創による個人対応サービス社会の実現に寄与したい．

3)レジリエント社会システム研究プロジェクト

社会は自然災害や経済・産業の変化などに柔軟に対応して行かなければならない．このプロジェクトでは，社会の数理的モデル化を通して，想定外の事象にも柔軟に対応できる社会とは何かを追求する．レジリエンスとは，環境の大きな変換に対して，一時的に機能を失ったとしても柔軟に回復できる能力を指す言葉で，生態学等ではよく知られた概念である．情報・システム研究機構では，多様な分野におけるレジリエンスを調べることによって，レジリエントなシステムを構築・運用するための共通な知識体系を構築すべく，領域横断型研究プロジェクト「システムズ・レジリエンス」を立ち上げた．サービス科学研究センターでは，このプロジェクトと連携しながら，社会や企業・組織がレジリエントであるための方法論を研究する．

4)人間社会のコミュニティモデルに関する研究プロジェクト

近年，実社会生活において人間関係が希薄になる中，安心安全で文化的な社会の実現には，人間関係が密な有機的コミュニティが必要不可欠である．本プロジェクトでは，人間社会のコミュニティを統計的にモデル化し，そのモデルを解析することにより，コミュニティを密な関係に深化させるための知見や長く持続させるための知見を見出すことを目的とする．インターネット上のソーシャルネットワーク(SN)は，現情報化社会に即応して形成されるコミュニティとして捉えることができる．本プロジェクトでは，この SN に注目し，SN の発生から消滅に至る生存時間をモデル化する技術，および発展や衰退，分岐や合流などを経ながら動的に変化する SN の生存状態をモデル化する技術を開発する．それらのモデルを解析して，SN を密な関係に深化させ，持続させるための知見を見出し，安心安全で文化的にも豊かな社会の実現に貢献する．

5)サービス産業のためのシンボリックデータ解析手法開発プロジェクト

サービス産業における大量の複雑なデータをシンボリックデータとしてとらえ，その構造を明らかにするための手法を研究することにより，サービス科学において有用かつ解りやすい手法の構築に寄与する．サービス産業においては，大規模かつ多様なデータが日々大量に蓄積さ

れている．そのような状況において，オリジナルデータそのものではなく，自然な，あるいは意味のあるグループについての情報に興味がある場合が多く，そのようなグループを新しいデータ（シンボリックデータ）と考えて統計解析を行うために，その可視化や解析に着目する．シンボリックデータはこれまで個体のグループを表現するために周辺分布の情報だけを利用することが多かったが，われわれはより多くの情報を用いる集約的シンボリックデータに対して必要な解析手法を研究することにより，サービス科学における有用かつ解りやすい統計分析手法の構築を行う．

6）データ・キュレーションプロジェクト

統計科学の知識をもとに，データ処理に関する技術，方法論，ポリシーを統合した知識の体系を確立し，サイバーフィジカルシステム（CPS）に対応したデータ分析手法を構築する．近年の IT の進歩により，情報が氾濫する世の中にわれわれは暮らしている．将来的には，これら IT が作るサイバースペースと呼ばれる仮想空間と現実の空間とが，多くのつながりを持った CPS と呼ばれる空間が構築される．CPS では，大量のデータが瞬時に集まる．これらのデータの有効利用は多くの利益をもたらすと期待されるが，情報としてどのような価値を持つか不明確なデータも多く，有効な利用にはこれまでと違った手法が求められている．そこで，この分析手法の構築を進める．

データ科学研究センターでは 5 年間にわたり，これらの目標に向かって共同研究，研究集会などを通して研究教育活動を進めてきた．もちろん，当初掲げた目標をすべて達成したとは言えないが，サービス科学の研究に対していくらかの新しい貢献をすることができたと考えている．本特集が，われわれの考える「サービス科学の今」をお伝えすることができれば幸いである．

統計数理(2018)
第 66 巻 第 2 号 193–212

特集「サービス科学の今」
[総合報告]

アジャイルな社会に向けて

丸山 宏†

(受付 2018 年 2 月 6 日；改訂 5 月 31 日；採択 6 月 1 日)

要　　旨

筆者は 2011 年 4 月より 2016 年 3 月まで，統計数理研究所サービス科学研究センターのセンター長として勤務した．その中でいくつか一見関連のない研究活動を行ったが，それらを俯瞰してみると「変化にどのように向き合うか」という一貫したモチーフが見えてくる．我々はますます変化が激しくなる社会に直面していて，それに対して我々自身もアジャイルに変化していかねばならない．本稿では，サービス科学，情報技術の変遷，レジリエントなシステム，個人のキャリアとスキル，という 4 つの観点から，我々が日々直面する変化にどのように対応していくか，を議論する．

キーワード：Service Science, Statistical Machine Learning, Resilience, Data Scientists.

1. はじめに—変化にどのように向き合うか

我々は急速に変化する社会に生きている．我が国では主要先進国でも類を見ない速さで少子・高齢化が進んでいる．一方で，情報技術や生命科学などの技術革新が進み，それに応じて社会のインフラや仕組みも変化しつつある．グローバル化によって国の境界を超えて産業のエコシステムが複雑に絡みあうと同時に，特に 21 世紀に入ってからはテロなどの宗教的・民族的な問題が世界を引き裂こうとしている．このような変化に対して我々はどのように向き合っていけばよいのだろうか．

筆者は 2011 年 4 月より 2016 年 3 月まで，統計数理研究所サービス科学研究センターのセンター長として勤務した．その中でいくつか一見関連のない研究活動を行ったが，それらを俯瞰してみると「変化にどのように向き合うか」という一貫したモチーフが見えてくる．本稿では，サービス科学，情報技術の変遷，レジリエントなシステム，個人のキャリアとスキル，という 4 つの観点から，我々が日々直面する変化にどのように対応してくか，を議論したい．

第 2 章では，産業構造の変化に対する取り組みとして，サービス科学とは何かを議論し，特にサービス科学研究センターで主要なテーマとした「データに基づく意思決定」について述べる．

製造業からサービス業への業態の変革を遂げた業種の 1 つが情報産業である．第 3 章では，情報産業が，いかに変化に適応していったかを概括し，さらに，統計的機械学習に基づく帰納的なシステム開発によって今後の開発がより適応的になっていくことを説明する．

そもそも変化に強いシステムとは何だろうか．サービス科学研究センターは 2011 年，東日本大震災の直後に立ち上がったので，災害にどのように対応するかをセンターの 1 つのテーマとすることは自然な発想だった．自然災害のような大きな擾乱が起きても，何らかの形で回復して機能を取り戻すシステムをレジリエントなシステムと呼ぶ．我々はレジリエンスとは何

† 株式会社 Preferred Networks：〒 100–0004 東京都千代田区大手町 1–6–1 大手町ビル 2F

か，レジリエンスを実現するにはどのような方略があるかを，多分野のエキスパートを集めて 4 年間研究した．そこで得られた知見について第 4 章で議論する．

サービス科学研究センターではまた，2013 年度から 3 年間「データサイエンティスト育成ネットワークの形成」という文部科学省委託事業を行った．少子・高齢化が進む社会では，個人が長い期間労働することが求められる．変化の激しい社会で長期間働くためには，我々のスキルもそれに合わせて変化していかなければならない．第 5 章では，この委託事業の知見を振り返りつつ，変化する社会で私たち個人が身につけなければならないスキルとは何かについて考える．

2. ビジネスの変化—サービス科学

18 世紀後半に始まった産業革命から，20 世紀後半になるまで，世界の産業・経済の仕組みは製造業を中心に構築されてきた．工場に大きな投資をし，工場労働者が製品を低コストで大量生産し販売する，という考え方である．そこでは，同じ製品が多くの顧客に，また長期間に渡って売れるだろうという「ものづくり」の前提があった．

消費者の観点からは，20 世紀までは「ものを所有することが価値であった時代」と言えるかもしれない．高度経済成長期やそれに続くバブル経済においては，自家用車を所有することはステータスであり，所有することに価値があった．しかし，人々が物質的に豊かになるにつれ，所有することの価値が相対的に小さくなってきている．Rifkin (2014)は，生産性の向上が進んだことによって製造物の生産にかかる限界費用が限りなくゼロに近づいていく社会を描いている．限界費用がゼロであれば，製品を所有すること自体には価値がなくなっていくだろう．もし，所有すること自体が目的でないのであれば，自動車は人々を運ぶ，という機能を提供すればよいことになる．そのことに注目して新たなサービス業を始めたのが，例えば自動車の配車アプリで急成長した Uber である．将来，産業の長い歴史を振り返ったとき，製造業を中心とした産業構造の時代は，産業革命の 18 世紀後半から 20 世紀後半までのたった 200 年だと記憶されることだろう．21 世紀からは，サービス産業の時代である．

サービス科学は，このような産業構造の変化に対する必要から生まれてきた．現在，我が国の労働人口約 5,800 万人の 7 割以上がサービス産業に従事している．サービス科学を提唱した IBM は 1980 年代までは大型計算機を製造・販売する製造業であったが，1990 年代に入ってサービス産業に大きく転換した．その転換の過程で，製造業に見られる工学的な知識体系をサービス産業にも導入しようとして提案したのがサービス科学と呼ばれる学問領域である．サービス科学は，情報科学，経営科学，組織論など複数の領域にまたがった学際領域と考えることもできる．

2.1 サービスの特質—文脈依存性

製造業における機械工学，材料工学などの知識体系に対比してサービス産業における知識の体系化を考えるとき，サービス産業には製造業と対比してどのような特質があるのかを考えてみよう．Parasuraman et al. (1985)は，サービスの品質を議論する上でサービスと製造物の主要な違いが，次の 4 つであると指摘している．

- **無形性(intangibility)**—製品と違って，サービスで提供されるものは通常手に触れることができない．提供されるものは，サービス提供者の活動の結果としての効果・効能である．提供されるものが目に見えたり手で触れたりすることができないため，その価値を客観的に見極めるのが難しく，価格決定の困難さにつながる．
- **異質性(heterogeneity)**—同じサービスでも，提供する人，提供される場所，利用者の置

かれている環境や心理状態により，サービスの効果や利用者の受け止め方が異なる．同じサービスでも利用者によって受け取る価値が異なるので，これも価格決定の難しさにつながる．

- **同時性(inseparability)**―製造物では，製品が製造され，それが顧客に販売されて所有権が移転してから，製造物が利用される．一方，サービス業では生産と消費が双方向的に，時間的・空間的に同時に起こる．このため，提供物の品質の事前チェックが難しい．
- **消滅性(perishability)**―製造物と違って，サービスは予め作って在庫しておくことができない．このため，需要の変動の影響を，製造業より強くうけることになる．

また，Vargo et al. (2008)は，サービスとは顧客との価値共創と述べている．これらの特質はいずれも，サービスの価値がその文脈に強く依存することを示している．文脈は変化するものであり，サービス提供者はその文脈に合わせてサービスの形態を変えたり，価格を変えたりしなければならない．

一方で，多くのサービスでは，製造業と異なり，構築に時間とコストのかかる生産設備を要しない．このため，文脈に合わせて身軽にサービスの内容を変化させていくことができる．特に，近年のサービスの多くは情報技術を使ってネット上の Web サービスとして提供されていて，それらはクラウドを使うことによって，変動する需要に弾力的に対応したり，サービスの内容をダイナミックに変化させていくことが容易である．

2.2 データに基づく意思決定

サービス科学には，経営，マーケティング，オペレーションズ・リサーチ，品質管理，イノベーション，組織論など多くの領域があり，全体を網羅的に研究することには無理がある．そこで，統計数理研究所サービス科学研究センターではそのテーマを「データに基づく意思決定」とした．

21 世紀のサービス業は，Web やスマートフォンアプリで提供されたり，IoT(Internet of Things)によって大量のセンサーデータを集めるなど，情報技術に強く結びついている．折しもビッグデータという考えが注目を浴びてきていて，情報技術によって得られるデータをどのように経営の意思決定に結びつけるかをセンターの中心テーマに据えることは時流にもかなったものであった．

データ分析と一口で言っても，様々なタイプの分析手法があり，異なるビジネス局面では，異なるデータ分析が必要となる．どのような状況において，どの種のデータ分析が適しているのか，それが何を与えてくれるのか，そのためにはどのような前提が必要なのか，意思決定者は理解しておかなければならない．

我々は，Evans and Lindner (2012)に基づき，データ分析を目的によって Descriptive(説明的)，Predictive(予測的)，Prescriptive(指示的)の 3 つに分類して考える．

説明的データ分析(Descriptive Data Analysis)

データ分析が教えてくれることの第一は，「何が起きたか」という事実に関するものである．ある会社で，中南米市場向けの在庫が急速に不足するようになった．経済の発展によって市場が拡大しているのだろうか．自社の生産や流通に問題があるのだろうか．それとも，何か特別な理由があって末端の小売店が自分の在庫を確保するために注文を一時的に増やしているのだろうか．何が起きたかを知るには，様々なデータを眺める必要がある．社内の生産・在庫・売上などのデータに加えて，中南米の経済状況，政治状況，流行している商品，人々の消費動向など，問題の在庫状況に影響を与える可能性のある要因はいくらでもある．このような様々なデータ・ソースを結びつけて，何らかの説明をつけなければならない．

データ・マイニングは Descriptive なデータ分析のための強力なツールである．データ・マイニングは，データから何か「面白い」兆候を見つけてくれる．有名なのはビールとオムツの例である．スーパーマーケットの POS 端末データに基づいて，何と何が同時に売れる可能性が高いか，を調べると，もちろん，パンとミルク，パスタとパスタソースなどが同時に売れるのは，すぐにわかる．だが，データ・マイニングを用いると，ビールとオムツが同時に売れる傾向にあることもわかる．子供のいる若い男性がスーパーマーケットに来るときに，それらを同時に買うことが多いということなのかもしれない．

このようなデータ分析は自動化するのは難しい．何が「面白い」発見であるかを，予め決めておくことは，その性質上困難だからである．面白さを決めるのは多くの場合人間であり，そのため，Descriptive なデータ分析は通常は人間と機械の協調作業となる．統計データ分析パッケージで各社がデータの視覚化に力をいれているのはそのためである．

逆に，これらのツールを使いこなす人間の側に要求されるのは，データの裏にあるストーリーを描き出してみせる想像力である．例えば攻撃を受けた Web サイトの調査の場合，大量のログデータを分析し，攻撃者がいつどのような方法で攻撃を行い，どれだけの機密情報を盗み出したかを再構成するのは，想像力を働かせて仮説を立て，それをデータで検証する，という作業の繰り返しとなる．

予測的データ分析（Predictive Data Analysis）

「ある日にビールを買った人が 400 人，オムツを買った人が 350 人いて，そのうちの 300 人はオムツとビールの両方を買った」という面白い事実を教えてくれるのはデータ・マイニングである．しかし，別の日にビールを買う人が 250 人になったら，どうだろうか？ 1000 人になったら？ それらが 30 代の男性だったら？ ビールを買う人に関するパラメターに対して，それらのうちの何人がオムツを買うか，という予測をする数式を立てることは，予測的データ分析である．

データ・マイニングは確かに何か面白い事象をデータの中から探してきてくれるが，そのような事象が常に起きるかどうかについては，語ってくれない．あくまでも過去に起きたことを見せてくれるだけである．将来に何が起きるか，このまま行くと中南米の市場シェアは下がるのか，あるいは生産力を増強すれば利益があがるのか，そういうことを意思決定者は知りたいだろう．このためには，過去のデータから法則性を導き出し，それを使ってまだ見ぬデータを推測するしかない．この「データから法則性を導き出す」ことは統計モデリングに他ならない．

統計モデリングは，将来に対する予測を与えてくれると同時に，現在や過去のデータの欠損値を埋めてくれる（オムツのデータが今日に限って，システム故障のために得られなかった場合，など）．さらに，モデルがあれば，様々な戦略をシミュレーションしてみることができる．ある販促を行って，ビールの売上が 20% 増えるとすれば，オムツの売上も増えるだろうか？統計的機械学習は，統計モデリングとそれに基づく予測を行うための重要なツールである．深層学習など統計的機械学習の技術は近年急速に進歩していて，金融，製造，流通，行政などあらゆる分野での応用が始まっている．統計的機械学習による予測には，大量のデータ処理が必要なため，今までは高度なプログラミングのできるスキルを持つ組織などに限られていたが，現在では多くのツールがあり，手軽に利用できるようになっている．

もちろん，統計的機械学習の技術がいくら進歩しても，通常完璧なモデルはあり得ない．統計学者の George Box は，「本質的にすべてのモデルは正しくない．ただし，役に立つモデルはある」と言っている（Box, 1976）．世の中のすべての機序を書き下すことはできないのだから，モデルは常に世の中の近似値にすぎない．だから，モデルによる予測も外れることがある．しかし，過去のデータに基づくモデルがあれば，より合理的な予測ができることは間違いない．

指示的データ分析(Prescriptive Analytics)

データ分析を行う究極の目的はよりよい意思決定につなげるためである．ビールとオムツが同時に売れる傾向にあることがわかっても，売上増や顧客満足度向上につながる意思決定ができなければ役に立たない．モデルがあればシミュレーションができる．だが，シミュレーションにどのような仮説を入れるか，すなわちどのようなシナリオを作りパラメターを設定するかがわからなければ，そもそもシミュレーションは成り立たない．

最適化はそのようなシナリオとパラメターの設定の可能性，つまりビジネスにおける「次の一手」の中で，最も良いものを選んでくれる技術である．可能な指し手が数個であれば，それらを次々に予測モデルに入れて，それらを評価することで，最もよい指し手を見つけることができる．もし，可能な指し手の数が大きければ，その中からベストなものを見つけるのは容易でない．最適化の手法は，単純な数え上げの不可能な，非常に大きいパラメター空間の中で最適な設定値を探してくれる．

もう一つ，強力な最適化の手法として，「実際にやってみる」というものがある．最近 Web マーケティングの世界でよく使われる **A/B テスティング**というのはその一つである．A/B テスティングとは，たとえば Web ページの広告を A パターン，B パターンと分けてそれをランダムに表示し，どちらのほうがよりクリックされやすいか，を調べる．それによって自動的に Web ページのデザインを最適化するのである．より一般的には，**強化学習**と呼ばれる一連の手法があり，特に深層学習の普及によって広く使われるようになってきている．

もちろん，「何が最適か」は目的によって違う．クリック率を最大化したいのか，それとも売上を最大化したいのか，顧客満足度を上げたいのか，あるいは利益を上げたいのか，はたまたそれらの組み合わせかもしれない．「望ましさ」の尺度となる量を得る関数を，**目的関数**と呼ぶ(**効用関数**あるいは**報酬関数**と呼ばれることもある)．最適化アルゴリズムはこの目的関数を最大化するようにパラメター空間の中を探索する．ただし，複雑な条件が絡み合う問題設定の中では，得られた解が必ずしも意思決定者にとってのぞましいものではないかもしれない．Dewey (2014)は強化学習における報酬関数の決め方の重要性と難しさを議論している．

ビジネスの状況に応じて，意思決定に必要なデータ分析は descriptive なものか，predictive なものか，それとも prescriptive なものであるかを見極めなければならない．データ分析は意思決定のために行うのであり，「どのような結果が得られれば意思決定できるか」を常に問い続けなければならない．統計数理研究所サービス科学研究センターでは，いくつもの企業との共同研究を行ったが，「何がわかれば意思決定できますか」という質問は，データ分析の手法を考える上で常に有効であった．

もちろん，同じ課題設定でも，場面によって部分問題として descriptive, predictive, prescriptive な分析の必要性が現れてくる．しかし，最終的に欲しいものは何か，それが説明なのか，予測なのか，最適化なのかを常に意識しておくことが重要であることがわかった．

本章の始めに述べた様に，サービスは変化する文脈に依存する概念であり，よりよいサービスを提供するにはその文脈の変化を理解し，予測し，それに合わせてサービスを最適化していかなければならない．深層学習など最新のデータ分析技術を利用することにより，今後もサービス産業はどんどん高度化していくことだろう．

3. 情報産業は環境変化にどのように対応するか

3.1 情報産業の変化

製造業からサービス産業への大きな転換を果たした産業分野の 1 つが情報産業である．情報技術が社会の様相を変化させ始めてからおよそ 50 年になる．最初は情報技術の利用は，主に既

存の社会の仕組みを効率化することに重点が置かれていた．昭和 30 年代の情報技術の普及していないオフィスの様子を想像してみるとよい．給与計算や経理の仕事がそろばんと手作業で行われていたビジネス現場に，情報技術が導入されることによって，飛躍的な生産性向上が図られたことは想像に難くない．同様に，金融・製造・物流など多くの産業，また企業会計・サプライチェーンマネジメント・顧客管理・オフィスの効率化など多くの職種において，情報技術が既存のビジネスを効率化させてきた．この効率化を主眼にした情報技術の利用(**Systems of Record** と呼ばれることがある)は，そろそろ社会の隅々まで行き渡って，これまでのような急速な生産性向上は望めなくなってきた．

Uber など新たなサービス産業で現在起きているイノベーションは，既存ビジネス・プロセスの効率化ではなく，情報技術で顧客体験をダイレクトに作り出すことによって起きている(このようなシステムを **Systems of Engagement** と呼ぶことがある)．

Systems of Engagement は，第 1 世代の情報技術すなわち Systems of Record とどのように違うのだろうか．Systems of Record の開発においては，ものづくりの考えの延長で，事前に要件定義をきっちり行う，いわゆる「ウォーターフォール型」の開発が行われてきた．給与計算や銀行の勘定系システムのように，要件が長期にわたって安定しているシステムでは，このような考え方で問題なかった．一方，サービス・イノベーションを牽引する Systems of Engagement の開発では，ビジネス環境が刻々変化するため，短いサイクルで要件のバックログを見直すアジャイル開発[1)] や，開発時と運用時の切り分けを行わない DevOps (Hüttermann, 2012) という運用が行われていて，こららの方法論がソフトウェア工学の知識として体系化されている．

このように，情報システムの開発手法は，環境の変化に追随するために変化してきた．これからますます変化が加速する社会において，今後のシステム開発はどうなっていくのだろうか．そのための切り札の 1 つが，統計的機械学習による帰納的開発である．

3.2 統計的機械学習による帰納的開発

システム開発の一例として，摂氏を華氏に変換するプログラムを考えてみよう．通常のプログラム開発では，まず「摂氏を入力として取り，それに対応する華氏を出力する」という要求仕様を定義し，その計算方法を我々が持つ先験的な知識(ここでは，$F = 1.8 \times C + 32$ という変換式)に基づいてモデル化する．このモデルに基づいて設計を段階的に詳細化していき，実装を得る．これを，**演繹的プログラミング**(あるいは**モデルベース開発**)と呼ぶ．

一方，**帰納的プログラミング**(あるいは**モデルフリー開発**)においては，入出力の例を作ることから開発が始まる．例えば，摂氏と華氏の 2 つの温度計を調達して，時々それらの値を同時に読むことで訓練データセットを得る．訓練データセットに対して，統計的機械学習アルゴリズムを適用し，モデルを帰納的に求める．このモデルを用いて入出力の変換を行う．

これは統計モデリングに他ならない．ただし，パラメトリックな確率分布を仮定する伝統的な統計モデリングでは，モデル選択が正しく行われていることを前提としていて，このモデル選択が難しいことが知られている．一方，近年注目を浴びている深層学習においては，モデル選択がそれほどシビアでない．大量のパラメタを持つ深層ニューラルネットワークにおいても，実際上それほど過学習しないことが知られているからである．

深層学習の表現力はどうだろうか．任意の計算可能関数について，それを十分な精度で近似できるニューラルネットワークが存在することが知られている(Cybenko, 1989)．このため，深層学習は擬似的にチューリング完全と考えることができる．この汎用計算機構は，今までのプログラミングとは異なり，入出力の例示により帰納的にプログラミングすることが可能である．

システム開発の方法論としてみたとき，統計的機械学習に基づく帰納的開発は，仕様を訓練データセットの形で表現し，実装は訓練(training)によって半自動的に行われる．もともと，

システム開発がウォーターフォール型で行われてきた背景には，一度システムを実装してしまうと，仕様など上流工程に手戻りすることが非常にコスト高である，という「モノづくり」と共通する制約があった．アジャイル開発では，要件が変化することを見越して小さなサイクルで開発を回すことで，手戻りのコストを最小化する．しかし，統計的機械学習に基づく帰納的開発では，新しい要件を表現する訓練データセットが低コストで用意できる限り，手戻りのコストは(訓練にかかる計算コストを除けば)ほとんどかからない，というメリットがある．このため，環境の変化に柔軟に対応できるシステムを開発できるという可能性がある．

このような新しいスタイルのプログラミングにおいて，効率的に品質の良いソフトウェアを開発するにはどうしたらよいのだろうか．統計的機械学習を取り入れたシステム(本稿では**機械学習応用システム**と呼ぶ)の開発はまだ発展途上にあり，このような方法論(我々はソフトウェア工学に習って**機械学習工学**と呼んでいる)の議論は新たにスタートしていて，今後の発展が望まれる(Maruyama and Kido, 2017)．

4. レジリエンス—そもそも変化に強いシステムとは何か

サービス科学や帰納的システム開発は，それぞれ特定の分野で変化に対応するための手法だと言える．分野に関わらず，そもそも変化に強いシステム(我々はレジリエントなシステム，と呼ぶ)というものは考えられるだろうか．あるとしたら，それらの共通戦略はなんだろうか．我々は，サービス科学研究センターの活動の一環として，システムのレジリエンスを科学的に解明するため，情報・システム研究機構の領域横断的研究プロジェクト「システムズ・レジリエンス」を 2012 年度から 2015 年度にかけて実施した(システムズ・レジリエンス・プロジェクト, 2016)．本章では，その主要な成果を概括する．

4.1 レジリエンスのタクソノミ

多くのレジリエンスに関する先行研究を調査した結果，我々はレジリエンスを，少なくも 1)擾乱のタイプ，2)対象とするシステム，3)回復のタイプの 3 つの軸に整理した(Maruyama et al., 2014)．

1) 擾乱のタイプ
意図の有無：意図を持たない擾乱(例：自然災害)，意図的な攻撃(例：戦争)
頻度：高頻度な擾乱(例：交通事故)，極めて稀な事象(巨大隕石の衝突)
予測可能性：かなりの精度で予測できるもの(例：台風の進路)，事前に正確に予測することが不可能なもの(例：巨大地震)
継続時間：発生から終了までが極めて短時間な擾乱(例：落雷)，継続時間が極めて長い擾乱(例：地球温暖化)
内部性：システム外部からの擾乱(例：自然災害)，システムの内部から発生する脅威(例：複雑さで自己崩壊するシステム(Per Bak and Wiesenfeld, 1987))
2) 対象システム
対象領域：生態学，生物学，金融，社会コミュニティ，組織など
粒度：個別システム(例：個人)を対象とするのか，集合体(例：社会全体)か
能動性：擾乱に対する回復のメカニズムを内在的に持つ(例：生態系)か，その維持に人間の知的作業による能動的な介入が必要か(例：組織)
機能：目的関数が明確(例：営利組織)か，多数のステークホルダのためシステム全体の目的関数が不明確(例：コミュニティ)か

3) 回復のタイプ

工学的レジリエンス：システムが擾乱の前と全く同じ構造に戻る場合(例：制御工学によるフィードバック制御，故障時の部品交換など)

機能的レジリエンス：異なる構造で，同等以上の機能に回復する場合(例：製造業からサービス業に転換した IBM．企業として利益を上げるという機能は同じだが，内部構造が異なる)

適応的レジリエンス：別の機能・目的を持った新たなシステムとして生まれ変わる場合(例：戦前・戦後の日本．全体主義から民主主義へと価値観は転換したが，民族・文化などのアイデンティティーは保たれた)

4.2 レジリエンス戦略

システムをレジリエントにする戦略は様々なものが考えられる．我々は，それらの戦略を，図 1 に示すレジリエンス・サイクルに沿って整理する．

設計時のレジリエンス戦略

冗長性：マージンの増大(例：より高い津波に耐える防潮堤を作る)，多重化(例：データセンターにおいて電源やネットワークを多重化)，相互運用性(バックアップ資源間の相互運用性により，少ないバックアップ資源で多重化の効果を出す)，など

多様性：多様性指標管理(例：マイノリティ従業員の割合を管理)，ポートフォリオ(例：金融の世界でリスク分散)など．なお，システムに多様性を導入するための戦略として，収穫逓減則の効果があることが知られている(Akashi et al., 2012)．

資源・管理の分散化：資源や管理を 1 点に集中しておくと，そこが被害を受けた時にシステム全体が止まってしまう．このため，資源や管理を分散化しておくことが望ましい(例：インターネット)．

リスク転移：リスクを他者に転移することはリスク管理でよく知られた手法である(例：保険をかける)．

運用時のレジリエンス戦略

訓練：訓練には，事前に通告して行うスケジュールされた訓練と，抜き打ちで行う訓練がある．Google などのデータセンターにおいては，“Game Day” と称して抜き打ちで意図的に障害を注入する訓練も行っている(Limoncelli et al., 2012)．

マネジメントサイクル：環境の経時変化に対して，システムの本来設計された機能を維持するためには，PDCA サイクルを回す．

抑止：意図的な攻撃に対しては，防護手段や報復手段を見せることで抑止する．

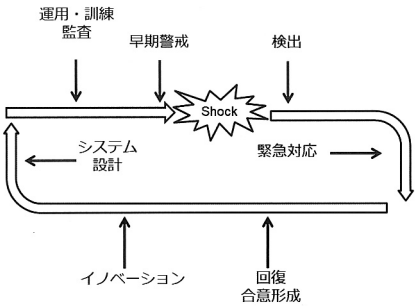


図 1．レジリエンス・サイクル．

早期警戒に関するレジリエンス戦略

擾乱の予測：擾乱が来ることが事前に予測できればそれに応じた対応を取ることができる（例：台風の上陸予測）．

早期対策：擾乱が来ることが確実でなくても，擾乱の可能性が高まった場合には，早期対策が有効である（例：テロに対する警備）．

緊急時のレジリエンス戦略

擾乱の検出・損害の評価：まずは発生した擾乱を検出しなければならない．擾乱の検出には，予測と同様にデータの収集とその分析が必須である．

ダメージコントロール：被害拡大防止（例：ウィルス汚染された PC の切断）

ポリシーの切替：緊急時には，通常時と異なるポリシーを適用する必要があるかもしれない（例：緊急時に個人情報保護よりも人命救助に必要な情報アクセスを優先する（Maruyama et al., 2013））．

現場への権限委譲：現場で危機対応を行う第一応答者（first responder）にかなりの自由裁量を与える（例：危機対応の国際規格である ISO22320 では，「意思決定は，可能な限り低い階層で行うことを許可し，連携及び支援は必要とされる中で最も高い階層から提供することを許容することがのぞましい」としている）．

回復時のレジリエンス戦略

資源割当の最適化：システムの回復のためには資源を投入する必要がある．しかし，地震などの広域災害の場合，資源の配分が必ずしも最適に行われるわけではない．このため，資源の割当を最適化することが重要である．

利他主義：自分の利益を後回しにして他人を助けることで，コミュニティ全体の復興を促進する（例えば 2011 年に行われた慶応大学の調査によれば，震災後に日本人の利他性が高まったとしている[2)]）．

境界拡大：対象システムが救えない時に，スコープを拡げて上位のシステムの回復を目指す（例：ある企業のビジネス継続が難しくなったときに，問題をその会社のレジリエンスではなく親会社のレジリエンスの問題と捉え直し，親会社に吸収することで事業の一部や従業員の雇用を救済する）．

イノベーション時のレジリエンス戦略

システムに擾乱があったとき，それはリスクでもあるが，同時にシステムを再構成し，より良いシステムに発展させるチャンスでもある．我々はこれを，通常の回復によるレジリエンスを表現する “bounce back” という言葉に対して，“bounce forward” という言葉で表現することがある（Yamagata and Maruyama, 2016）．“Bounce forward” のための戦略もいくつか考えることができる．

事後調査：擾乱の事後調査を行い，再発を防ぐ（例：サイバー攻撃が起きた場合に侵入経路など原因の調査を行い，対策する）．

合意形成：“bounce forward” させるために，どのようなシステムに回復させるか，合意形成を行わなければならない．

研究開発投資：擾乱をきっかけにシステムをより良くするために，長期的にイノベーションを起こすために研究開発に投資する．

以上，様々なレジリエンス戦略を概括したが，これらの戦略は常に有効であるとは限らない．それぞれの文脈に応じて，効果のある戦略もあれば，そうでない戦略もある．このため，我々は戦略選択の支援ツールとして，レジリエンスの文脈と戦略との対応表を作成した（図 2）．

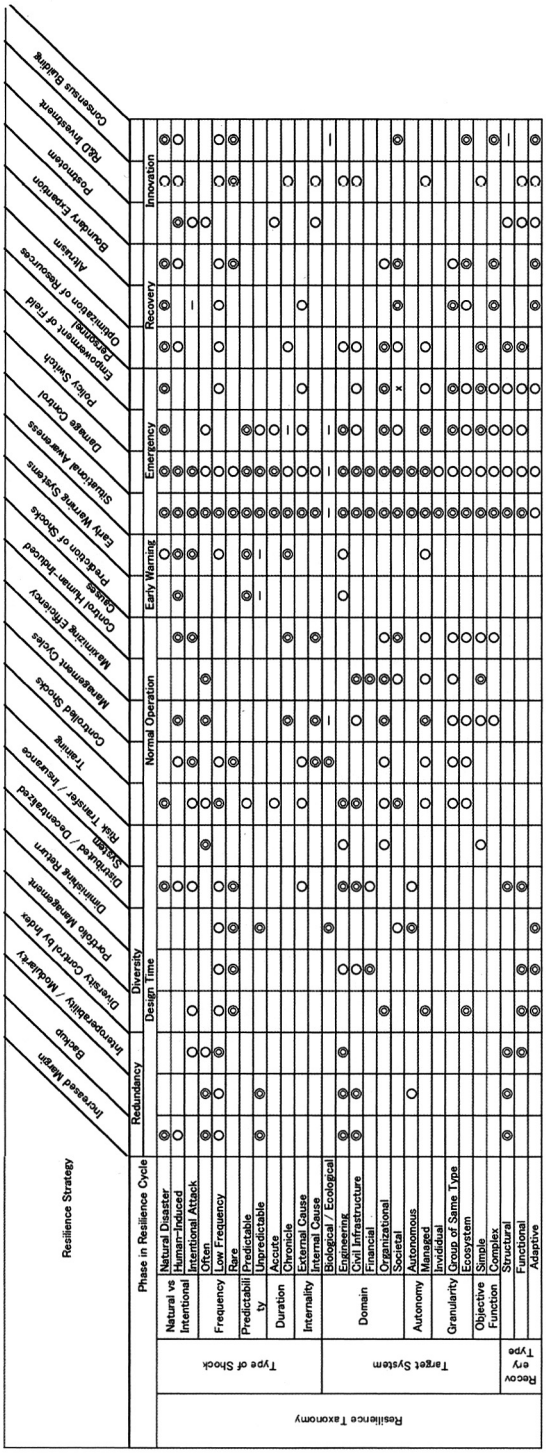


図 2. レジリエンスのメタ戦略.

アジャイルな社会に向けて

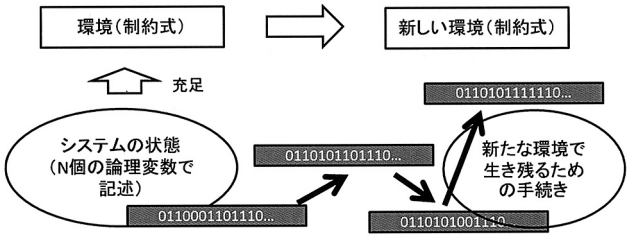


図 3. SR Model.

4.3 レジリエンスの数理モデル

複雑かつ大規模なシステムのレジリエンス性を評価するためには，汎用的な数理モデル上で様々な動的特性を解明する計算論的手法が不可欠である．我々は，計算機科学の知見を元に，レジリエンスの原始的な数理モデルを提案し，そこからいくつかの性質を論理的に導びいた．

4.3.1 SR Model

我々は図 3 に示すようにシステムの状態が，有限の記述で表現できると仮定する．このことは，システムの状態を長さ n のビット列で表すことができる，と言い換えることができる．システムは，2^n 通りの異なる状態を持つことができる．

システムはある時点で，環境に適応している．環境への適応性は，これらのビット列に対する制約という形で表現することにする．環境(制約)C は，システムの状態空間の部分集合として表す．一般にシステムの状態 s の環境 C に対する適応度はシステムの状態空間における適応度関数で定義できる．ここでは説明を簡略化するため，状態 s が環境 C に適応($s \in C$)しているかまたは不適応($s \notin C$)であるかの 2 値を考慮する適応度関数を考えることにする．システムの状態 s が環境 C に適応しているとは，$s \in C$ が成り立つと定義する．

さて，環境 C が変化して，C' になったとしよう．もし，今のシステム状態 s が新たな環境に適応していない，すなわち $s \notin C'$ とする．その場合，システムは新しい環境に適応すべく，システム状態を変化させなければならない．例えば，一度に 1 ビットを反転させて $s \in C'$ になるにはどれだけのステップがかかるか，という問題を考えよう．もし，環境の変化に対して常に k ステップ以内で $s \in C'$ とすることができれば，このシステムは k-レジリエントと呼ぶことにする．

これは非常にシンプルな数理モデルであるが，我々はこの考えを拡張してレジリエンスの数理モデル SR-Model を構築した(Schwind et al., 2013)．このモデルでは，レジリエンスは，外界の擾乱とシステム管理者の対応が交互に行われる，2 プレイヤーのゲームの軌跡(System State Trajectory, SST)として解釈される(図 4)．この軌跡のコスト(あるいはコストの移動平均)がある閾値を超えない場合，システムはこの軌跡に関してレジリエントだったと定義する．

4.3.2 レジリエンスの指標

このモデルから得られる帰結の一つは，レジリエンスの指標に関するものである．レジリエンスの指標としてよく知られているのは，Bruneau et al. (2003)による，Resilience Triangle というものである(図 5)．これは，起きた事象に対して，システムのパフォーマンスを計測し，本来あるべきパフォーマンスとの差を積分したもの(図でいえば三角形の部分の面積)をレジリエンスの指標とするものである．これを，事後レジリエンス指標(**performance metric**)と呼

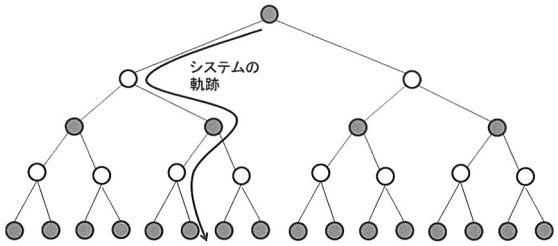


図 4. System State Trajectory.

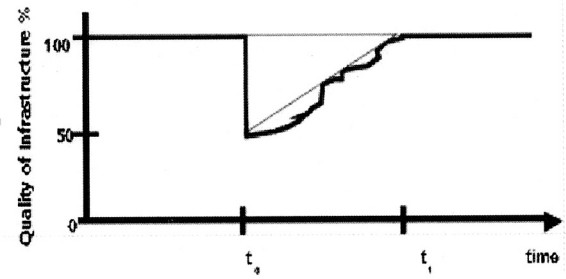


図 5. Resilience Triangle.

ぶ．図 4 における軌跡に対してコスト関数を割り当てることに相当する．

事後レジリエンス指標は，起きてしまった事象に対してシステムのレジリエンスを評価するのには使えるが，将来起きうる事象に対して，システムがレジリエントにふるまうかについての知見を与えてくれるわけではない．我々が欲しいのは多くの場合，将来に起きうる事象に対して，システムがレジリエントであるかどうかの指標である．これを我々は**事前レジリエンス指標**(**competency metric**)と呼ぶ．事前レジリエンス指標を客観的に求めることは難しい．将来にどのような事象が生起するかわからないからである．このため，事前レジリエンス指標は多くの場合，ドメインに依存した主観的な指標となる．例えば都市におけるレジリエンスの評価においては，市民一人あたりの GDP，教育水準，失業率などの指標を組み合わせて，どの都市が将来の擾乱に対してよりレジリエントであるかを判断する，ということが行われている．このため，事後レジリエンス指標と事前レジリエンス指標を明確に結びつけることができない．

我々の SR-Model を用いると，事前レジリエンス指標は，「起きうるすべての将来に対しての事後レジリエンス指標の最大値」のような形で，事後レジリエンス指標と事前レジリエンス指標の間の関連付けを行うことができる．SR-Model において，「今後起きうるすべての将来事象を数え上げられる」ことを仮定しよう．図 6 において，現在から将来に向かってすべての軌跡を数え上げることができれば，それらの軌跡の代表値(コストの最大値，平均値など)を求めることができる．この値をシステムの現時点での事前レジリエンス指標と考えれば，事後レジリエンス指標と事前レジリエンス指標を結びつけることができる．

4.3.3 レジリエンスの時間地平線

このモデルから理論的に得られる一つの帰結は，レジリエンスを語るには，有限の時間地平線を設ける必要がある，ということである．図 4 において，外界の擾乱は 1/2 の確率でシステムのコストを +1 し，それ以外はコストを変えないとしよう．同様にシステム回復の試みは

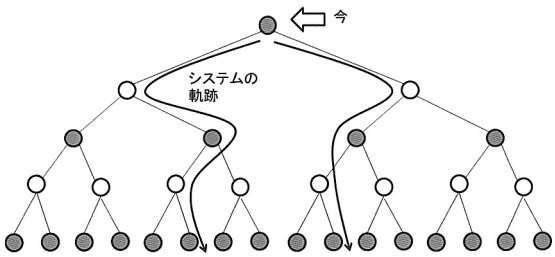


図 6. Set of Possible Trajectories.

1/2 の確率でシステムコストを −1 すると仮定する．そうすると，システムのコストの軌跡はランダムウォークとなる．ランダムウォークの結果，コストが最終的に取る値は，このランダムウォークを無限回試行すれば初期コストを中心とした正規分布になることが知られている．

このことは，どんな有限の閾値を設定しても，時間軸を無限に取る限り，ゼロでない確率でシステムのコストが閾値を超えることを意味する．したがって，時間軸を無限に取る限り，レジリエンスを保証するシステムは実現できないことが理論的に示される．

レジリエンスを実際に議論する時には，時間地平線について意識することは少ないと思われるが，我々が狙うレジリエンスとは，次の 10 年なのか，100 年なのか，1,000 年なのか，というタイムスケールを意識してレジリエンスを議論することが重要である，というのが我々の数理モデルから言えることの一つである．

4.3.4 SR-Model の限界

以上のように，SR-Model はレジリエンスに関する原始的な数理モデルとして用いることができる．しかし，SR-Model には限界もある．1 つには，上記の議論は静的なコスト関数を仮定していることである．システムが “Bounce Back” する工学的レジリエンスや機能的レジリエンスの場合には，コスト関数は変化しないが，擾乱をきっかけに全く新しい価値観を持つシステムに生まれ変わるようなシステムについては，SR-Model はうまくモデル化できない．

もう 1 つの限界は，本質的に閉世界を仮定していることである．回復の戦略の中には，システムの境界そのものを見直して，より上位のシステムを救う，というものもある．このようなレジリエンスは，SR-Model では扱えない．

4.4 終わりのない対話

瀬名・鈴木 (2009) は，パンデミックのような重大な擾乱に向き合った時，我々は「真実へ至る対話」「合意へ至る対話」「終わりのない対話」という 3 つの対話を繰り返していかなければならないのだ，と述べている．「真実へ至る対話」とは，擾乱の際に「何が起きたのか」に関する理解であり，この理解が得られれば “bounce back” するような対策を考えることができる．「合意に至る対話」とは，SR-Model におけるコスト関数を決めるものであり，私たちがどのように多様なステークホルダの間で，優先順位をつけて回復していくかを導いてくれる．そして「終わりのない対話」こそ，私たちが “bounce forward” するレジリエンス，すなわち人類社会の価値をどのように再定義していくか，という価値観の議論に他ならない．これらの対話に真摯に取り組むことがレジリエントな社会を作っていく本質なのではないだろうか．

5. 個人のキャリア—変化の時代を生き抜くスキル

変化が激しい社会においては，個人もその変化に対応してスキルを磨いていかねばならな

い．2 章，3 章で見てきたように，サービス産業，またその根幹を支える情報技術の変化は，特にデータ分析の領域において顕著である．このような変化に対して，サービス産業，情報産業に従事する者はデータ分析のスキルを身につける必要があるだろう．本章では，まず，このような変化を先導するべき専門家としてのデータサイエンティスト育成の試みについて述べる．その上で，情報技術やデータ分析が広く行きわたった社会で，一般の市民がどのようにリテラシーを身につけるべきかについて考える．

5.1 データサイエンティストの育成

サービス科学研究センターで「データに基づく意思決定」の研究活動を行う中で，実際にビジネス環境の中でデータ分析ができる人材をどのように育成したらよいかが大きな課題として浮き上がってきた．このために，文部科学省からの委託事業として「データサイエンティスト育成ネットワークの形成」を 2013 年度から 2015 年度までにかけて実施した（丸山, 2014）．

この事業が始まってから，我が国のデータサイエンティストを取り巻く環境は大きく変化した．データサイエンティストという言葉が広く浸透し，データサイエンティストに関する多くの書籍が刊行され，データサイエンティストを育成する多くの教材が現れた．滋賀大学においては，本邦初のデータサイエンス学部が平成 29 年度より設置され，進学を考える高校生と，人材を求める企業の双方から注目されている．新しい職種であるデータサイエンティストの育成と業界の健全な発展を目指す民間団体であるデータサイエンティスト協会は 2018 年 1 月現在 50 を超える法人会員と，数千名の個人会員を擁するようになった．データサイエンティストとしての実務を経験する，データサイエンティスト向けのインターンシップ・プログラムも，民間の営利事業として軌道に乗り始めている．産官学の有識者からなる「ビッグデータの利活用に係る専門人材育成に向けた産学官懇談会」は，我が国におけるビッグデータ利活用人材育成の青写真とも言える提言を 2015 年 7 月に公開し[3)]，その提言に沿った施策の検討が行われた．これらの動きの多くには，直接・間接に本事業で形成されたネットワークが関わっていて，その意味で本事業は一定の成果を挙げてきたと考える．

本事業期間全体を通して得られた主要な知見は以下の 3 点にまとめることができる．

(1) データサイエンティスト像の多様性

データサイエンティストという概念は米国で発生したものであるが，「ビッグデータをビジネス上の価値に変えることのできるプロフェッショナル」という観点で見ると，共通に求められる資質はあるものの，その専門性のレベル（見習い，独り立ち，棟梁クラス，業界代表レベル，など），そもそものバックグラウンド（自然科学，情報科学，統計学，経済学，経営学など）や働き方（サービスプロフェッショナル，部署内での専門家，フリーランスなど），キャリアの形成には多くのバリエーションがあることがわかった（Maruyama et al., 2015）．これらの多様性をサポートする育成を考える必要がある．

(2) 現場体験の重要性

データサイエンティストには，統計学や統計的機械学習を中心とするデータサイエンス力，情報科学やソフトウェア工学を中心とするデータエンジニアリング力，さらにはビジネスを理解し推進するビジネス力という分野横断型のスキルが求められる．特に後者の 2 つは，現場での経験から学ぶ割合の大きい分野であり，このため PBL など現場体験型の育成が不可欠になる．本事業でも現場体験を推進するため「人材ローテーション」を柱の 1 つに掲げてインターンシップ・プログラムなどを実施したが，そのフィードバックからも，現場体験の重要さは確認することができた．また，PBL としては，データ分析ハッカソンを行った（丸山 他, 2016）．

(3) データサイエンティスト利用側のリテラシー・洞察

初年度におけるデータサイエンティスト現状調査や，その後に実施した利活用ベストプラクティス調査などを含め，事業全体を通して繰り返し感じられたのは，データサイエンティストを育成するだけでなく，データサイエンティストを雇用し利用する側のリテラシーの重要性である．そのためには，社会全体のデータ・リテラシーも向上させる必要がある．

3 番目の，利用側のリテラシーは社会全体の課題であり，情報技術やデータ分析の専門家だけの問題ではない．この問題について，次の節でより詳しく議論する．

5.2 リベラルアーツ

現代社会は，個人の自由と基本的人権を普遍的な価値と認める，という理念にもとづいている．個人の自由とは，様々な事柄について自分で決められる，ということだ．それを実践するスキルを「リベラルアーツ」という．私たちは社会の中で生きているので，「自分で決める」ということは社会の中で合意形成するということに他ならない．多人数間での合意形成は時として非常に難しいものになるが，科学における方法論や民主主義など，長い人類の歴史の中で私たちは「社会的動物の生きる知恵」としての合意形成の仕組みを作り上げてきた．一方で，情報技術，特にデータ分析の急速な発展が，それら人類の叡智とも言える合意形成の仕組みの前提を揺るがしている．人工知能の時代になっても，「自由人」であるためのアーツ (技芸) とは何か，を考える必要がある．

5.2.1 合意に関する人類の叡智

人類の長い歴史の中で築き上げられた「合意形成の叡智」として，科学と民主主義について考えてみよう．

科学において，何かが真であるということが認められるのはどういうときだろうか．最も古い科学の方法論の 1 つが，実験や観察を行い，その中から共通の原理を帰納的に見つけていくやり方であり**実験科学**と呼ぶ．19 世紀の終わりになって提案された統計的仮説検定によって定量的な尺度を与えられるようになった．統計的仮説検定は帰無仮説を仮定すると実験結果が確率的にありそうもない時に，帰無仮説を棄却することで対立仮説を示す．この際，「確率的にありそうもない」値として，5% あるいは 1% などの有意水準が使われる．これは極めて強力な推論ツールであり，これによって帰納的な実験科学は初めて，客観的でかつ定量的な裏付けを得ることができた．現在も，多くの科学の分野において，統計的仮説検定に基づく推論が行われている．統計が「科学の文法」(Pearson, 1900) と言われる所以である．

科学の方法論が「X は真である」という形の命題についての合意であるのに対して，集団の意思決定の合意についての人類の知恵の一つが民主主義である．民主主義とは，市民による統治であり，独裁者や一部のエリート階級による支配ではない．民主主義においては，すべての構成員が平等に政策決定に参画する権利を持つ．ある政策が特定の個人やグループには有利だが，別の個人やグループに不利に働くとき，エリート政治家やエリート官僚がそれを決めるのではなく，民主主義においては全員参加で議論を尽くし，最終的には投票で決定する．この合意形成の方略は常にうまく働くわけでは無いし，民主主義が唯一無二の政治形態というわけでもない．しかし，人類の長い歴史の中で今のところもっともうまく機能している政治形態と言って良いだろう．

5.2.2 情報技術による「前提の崩壊」

科学や民主主義は，合意形成に関する人類の叡智と言えるが，その根底にある前提はしかし，情報技術の急速な発展によってくずれつつある．

まず，科学のプロトコルである実験科学について考えてみよう．情報技術のある世界では，統計的仮説検定は，Excel 上でマウスクリック一つで行うことができる．だから，1 回の実験に対していくらでも複雑な仮説を生成して，試してみることができる．そういう世界で，前出の統計的有意水準を 5% としてたくさんの仮説を試してみるとどういうことになるだろうか．p 値が 5% ということは，仮説が成り立たないにも関わらず間違って統計的仮説検定を通ってしまうことが，20 回の実験で 1 回くらいありそうだ，ということである．互いに関係の無さそうな(独立な)仮説を 20 個試せば，その仮説が成り立っていないにも関わらず一つくらいの仮説は，よく認められた「科学の文法」によって，正しいものと認められてしまうのである．

英エコノミスト誌は 2013 年 10 月 19 日号の「なぜ科学が間違うのか」という記事[4] の中で，がん研究で「画期的な成果」とされた研究論文のうち，たったの 11% が追実験によって再現可能だったと報じている．この理由の一つは，上述の有意水準から説明することができる．(半)自動化された実験・観測装置から大量のデータが得られ，それらに対してまた大量の異なるモデルを適用し，組み合わせ的に統計的仮説検定が行われる世界では，このような擬陽性も数多く生産されることを我々は知っていなければならない．このような問題を受けて，2016 年に米国統計学会は p 値の利用に関して注意を呼びかける声明を出している[5]．情報技術を前提とした科学においては，19 世紀の終わりに確立された文法では不十分であるのは明らかだ．

情報技術がもう一つの人類の叡智，民主主義に及ぼす影響について考えてみよう．民主主義においては，集団の構成員がすべて政治的に平等に意思決定に参画できなければならず，そのことは「全員が質問，議論，熟考をすることによって集団にとっての問題点を知る機会を十分に持っているという仮定」に基づく(Dahl, 2008)．Web などの情報技術によって，市民一人ひとりが最新の情報にアクセスし，Blog などを通して自分の考えを世に問うことができるようになり，その結果民主主義がより理想に近いものになるという期待がある．そこには，各人がそれぞれ普遍的な「理性」を持ち，他人に強制されるのではなく，自分自身の合理的な推論によって判断ができる，という啓蒙思想に基づいた前提がある．

しかし，私達は常に合理的な判断が下せるかというとそうでもないようだ．そこにも情報技術，特に統計的機械学習を始めとするデータ分析技術の進化の影響がある．データ分析は強力な技術だが，悪用すれば人々の心に内在する認知バイアスを操作することができる．人々の行動は合理的ではなく，様々な認知バイアスに晒されていることが知られている．Kahneman (2011)は，人間の合理的な判断を司る「システム 2 思考」は意識の集中を必要とするために，人々は努力しなくて済む直感的な思考「システム 1 思考」に頼りがちだという．だがシステム 1 思考には様々な認知バイアスがあることが今では知られている．

Twitter や Facebook などのソーシャルメディアでは，同じ意見を持つ人どうしがコミュニティを作ることが容易で，これによってすべての人が同じ意見を持つと錯覚してしまう「エコーチャンバー効果」が知られている．これには人々の認知バイアスの一つである確証バイアスが影響している．また，コマーシャルに人気のある俳優を使うことも認知バイアスを利用したものである．このような，人々が持つ認知バイアスをビジネスに積極的に利用することは，主に広告業界でよく知られたテクニックである．

特に近年では統計的機械学習に代表される人工知能技術が進化していることもあり，「この商品を買った人はこれも買っています」のようなリコメンデーションなどが広く利用されている．人々の購買行動に影響を与えるだけならまだ良いが，このような認知バイアスを政治の世界でも利用しようとする動きがあり，Gore (2017) や Heath (2014)などが，警鐘を鳴らしている．ビッグデータ・統計的機械学習の進歩によって，啓蒙思想に対する新たな脅威が訪れていることを看過すべきでない．

5.2.3 ビッグデータ時代を生きるリベラルアーツ

このように，情報技術が意思決定の前提をあちこちで崩しつつある社会にあって，それでは私達が自由人として行動していくためにはどのようなスキルが必要なのだろうか．

まず私達は，今の世の中がなぜそうなっているかを知っておかねばならない．科学には科学の方法論，プロトコルがある．政治にもプロトコルがある．今の時代には合わずに，欠陥だらけのものに見えることもあるだろう．だが多くの場合それらの仕組みは，長い歴史の中で必然として現れてきたものだ．民主主義は確かに効率の悪い政治形態かもしれないが，圧政やそれに起因する戦争・飢餓を避けるための人類の叡智と見ることもできる．歴史を理解し，今の世の中の仕組みが何を乗り越えようとしてそうなったかを理解する，それが自由人として必要な知識の第 1 である．

もし情報技術が私達の自由意思を阻害しているのだとすれば，情報技術で何ができるのか，できないのかを知らねばならない．自由を脅かす敵を知る，それが自由人として必要な知識の第 2 である．「人工知能」という言葉が独り歩きしているが，マスコミ等で喧伝される擬人化された「人工知能」はまだ技術的には遠く，我々の社会に影響を与えるとしてもだいぶ将来のことになるはずである．一方で，デジタルゲリマンダーと呼ばれる，人間の認知バイアスを操作する技術は急速に発展している．その裏にある技術は統計的機械学習であり，基本的には大量のデータから，その確率分布を推定する統計の理論を元にしている．統計は，データから人が気が付かなかったパターンを見出すことが得意であるが，一方で，常に過去のデータに基づいて未来を予測するので，過去と未来が基本的に変わらない（統計的には同じ確率分布からサンプルされる）場合にしかうまく働かない（Silver, 2012）．「人工知能」という言葉に惑わされず，ビッグデータの本質が統計であることを知れば，情報技術に過剰に期待したりむやみに恐れたりする必要はない．

同時に私達は，自分たちの「弱さ」も知っておかねばならない．特に自分の認知バイアスに気づくことが重要である．ここに難しさがある．認知バイアスとは，自分が気づいていないことだからである．ただし，どうしたらより自分のバイアスに気づけるか，についてのテクニックがいくつかある．それらについては，知識として理解しておくとともに，普段から実践することが重要ではないか．例えば Heath（2014）は，合理的思考であるシステム 2 には時間がかかるので，重要な意思決定は瞬間に判断せず，わずかでもよいから時間をかけるのが良いとしている．

6. おわりに

5 年間にわたる統計数理研究所サービス科学研究センターの活動を振り返った．そこには「変化にどのように向き合うか」という共通のテーマがあった．

「変化に対応する知恵」としてサービス科学研究センターの活動を振り返った時，そこには常に統計の考え方があった．もちろん，統計数理研究所の組織であったので統計が主要な道具立てであることは当然なのだが，それだけではなさそうだ．統計とは畢竟，観測される現象に起きる小さな変化（ノイズ）を捨象して，その裏にある本質を捉える手法と考えることができる．その意味で，統計が「変化に対応する」ための手法として普遍的に現れるのは自然なことに思える．

注．

1) http://agilemanifesto.org/

2) https://www.keio.ac.jp/ja/press_release/2011/kr7a430000094z75-att/120215_1.pdf

[3] http://www.rois.ac.jp/open/pdf/bd_houkokusho.pdf
[4] https://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong
[5] http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

謝　辞

サービス科学研究センターでの研究の機会を与えてくださった北川源四郎前情報・システム研究機構機構長，樋口知之統計数理研究所所長に感謝します．また，本稿では取り上げることができなかった多くの研究を推進してくださったサービス科学研究センターメンバーの方々に感謝いたします．

参 考 文 献

Akashi, H., Osada, N. and Ohta, T. (2012). Weak selection and protein evolution, *Genetics*, **192**(1), 15–31.

Box, G. E. P. (1976). Science and statistics, *Journal of the American Statistical Association*, **71**(356), 791–799.

Bruneau, M., Chang, S. E., Eguchi, R. T., Lee, G. C., O'Rourke, T. D., Reinhorn, A. M., Shinozuka, M., Tierney, K., Wallace, W. A. and Von Winterfeldt, D. (2003). A framework to quantitatively assess and enhance the seismic resilience of communities, *Earthquake Spectra*, **19**(4), 733–752.

Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions, *Mathematics of Control, Signals, and Systems*, **2**(4), 303–314.

Dahl, R. A. (2008). *On Democracy*, Yale University Press, New Haven & London.

Dewey, D. (2014). Reinforcement learning and the reward engineering principle, *2014 AAAI Spring Symposium Series*, AAAI Publications, Palo Alto, California.

Evans, J. R. and Lindner, C. H. (2012). Business analytics: The next frontier for decision sciences, *Decision Line*, **43**(2), 4–6.

Gore, A. (2017). *The Assault on Reason*: *Our Information Ecosystem, from the Age of Print to the Era of Trump*, Bloomsbury Publishing, London.

Heath, J. (2014). *Enlightenment 2.0*, Harper Collins, Tronto.

Hüttermann, M. (2012). *DevOps for Developers*, Apress, New York.

Kahneman, D. (2011). *Thinking, Fast and Slow*, Penguin Books, London.

Limoncelli, T., Robbins, J., Krishnan, K. and Allspaw, J. (2012). Resilience engineering: Learning to embrace failure, *Communications of the ACM*, **55**(11), 40–47.

丸山宏 (2014). 我が国におけるデータ分析人材の育成と活用(特集 ビッグデータへのアプローチ), Estrela, No.244, 8–13.

Maruyama, H. and Kido, T. (2017). Machine learning engineering and reuse of AI work products, *The First International Workshop on Sharing and Reuse of AI Work Products*.

Maruyama, H., Watanabe, K., Yoshihama, S., Uramoto, N., Takehora, Y. and Minami, K. (2013). ICHIGAN security-a security architecture that enables situation-based policy switching, *2013 Eighth International Conference on Availability, Reliability and Security* (*ARES*), 525–529, IEEE, Regensburg, Germany.

Maruyama, H., Legaspi, R., Minami, K. and Yamagata, Y. (2014). General resilience: Taxonomy and strategies, *2014 International Conference and Utility Exhibition on Green Energy for Sustainable Development* (*ICUE*), 1–8, IEEE, Pattaya City, Thailand.

Maruyama, H., Kamiya, N., Higuchi, T. and Takemura, A. (2015). Developing data analytics skills in Japan: Status and challenge, 日本経営工学会論文誌, **65**(4E), 334–339.

丸山宏, 神谷直樹, 宮園法明 (2016). クラウド環境を利用したデータ分析ハッカソンの計画と実施, Estrela, No.272, 30–37.

Parasuraman, A., Zeithaml, V. A. and Berry, L. L. (1985). A conceptual model of service quality and its implications for future research, *the Journal of Marketing*, **49**(4), 41–50.

Pearson, K. (1900). *The Grammar of Science*, Adam and Charles Black, London.

Per Bak, T. and Wiesenfeld, K. (1987). Self-organized criticality: An explanation of 1/f noise, *Physical Review Letters*, **59**, 381–384.

Rifkin, J. (2014). *The Zero Marginal Cost Society*: *The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism*, St. Martin's Press, New York.

Schwind, N., Okimoto, T., Inoue, K., Chan, H., Ribeiro, T., Minami, K. and Maruyama, H. (2013). Systems resilience: A challenge problem for dynamic constraint-based agent systems, *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, 785–788, International Foundation for Autonomous Agents and Multiagent Systems, Saint Paul, Minnesota.

瀬名秀明, 鈴木康夫 (2009).『インフルエンザ 21 世紀』, 文藝春秋, 東京.

Silver, N. (2012). *The Signal and the Noise*: *Why So Many Predictions Fail—but Some Don't*, Penguin, London.

システムズ・レジリエンス・プロジェクト (2016).『システムのレジリエンス—さまざまな擾乱からの回復力』, 近代科学社.

Vargo, S. L., Maglio, P. P. and Akaka, M. A. (2008). On value and value co-creation: A service systems and service logic perspective, *European Management Journal*, **26**(3), 145–152.

Yamagata, Y. and Maruyama, H. (2016). *Urban Resilience*, Springer, Switzerland.

212 Proceedings of the Institute of Statistical Mathematics Vol. 66, No. 2, 193–212 (2018)

Towards Agile Society

Hiroshi Maruyama

Preferred Networks, Inc.

The Institute of Statistical Mathematics operated the Service Science Research center from April 2011 to March 2016. We conducted several seemingly unrelated projects during this period, but in retrospect there has been a common theme, which is how to deal with changes. In this manuscript we discuss our insights on how we make ourselves more agile in order to survive changes. These insights were acquired from these projects, namely, service science, new programming paradigm, systems resilience, and individual career development.

Key words: Service science, statistical machine learning, sustainability, data scientists.

統計数理(2018)
第 66 巻 第 2 号 213–224

特集「サービス科学の今」
[総合報告]

ビッグデータを活用する確率モデリング技術

―社会実装の取り組みと課題―

本村 陽一†

(受付 2017 年 11 月 17 日;改訂 2018 年 3 月 15 日;採択 4 月 11 日)

要　　旨

ビッグデータを活用した機械学習により現在，人工知能技術の実用化が劇的に進んでおり，それによる産業構造変革や Society5.0 と呼ばれるスマート社会の実現も期待されている．本稿では実社会のビッグデータとして利用者の ID が付いたサービス利用履歴データ(ID=POS データや ID 付きアンケート，ID 付き操作履歴など)から確率潜在意味解析(PLSA)，ベイジアンネットを用いて確率モデルを構築し，利用者の行動や嗜好性を予測する確率モデリング技術について概説する．またそれによりサービスの価値や生産性向上の実現に寄与する人工知能技術としての応用例や社会実装を進めるための取り組みについても紹介する．

キーワード：サービス工学，人工知能技術，確率モデリング，ベイジアンネット，確率的潜在意味解析，ビッグデータ．

1. はじめに

物質的価値だけでなく生活の質(QoL)や経験価値が重視される時代になり，生活やサービス(コト)を直接の対象にして価値を生み出すことが期待されている．それにともない従来の技術開発の延長ではない，新たなコトづくりのための技術開発が必要になっている．技術競争の激化が進み機能や品質の均質化が起こると，どの技術や製品を選んでも大差がないコモデティ化(汎用化)が進む．製品は独自の付加価値を失うために製品選択の基準が価格にしかない状態となるために価格競争にさらされる．こうしたことから，これまで技術力を高めることで競争力を誇っていた日本の製造業が苦戦するようになってきたと言われている．製品の機能や技術自体では差別化することが難しくなると，製品が使われる背景であるニーズや市場を知り，市場に投入するタイミングをはかり，製品の魅力を高める外装デザインや使い勝手など機能以外の付加価値も合わせて考えることに競争力の原泉がシフトする．技術が優位であっても，それが価値として十分に伝わらなければその技術は使われない．使われないので技術開発の優先順位が下がり，せっかく優れた技術であるのに開発が継続できず最悪の場合には技術が不当に安く放出されたり，朽ち果てることすらある．競争力を高めるために全く新しいジャンルの製品開発によりイノベーションを起こすという考え方もあるが，利用者のニーズやマーケットの動向やトレンド，投入するタイミングを考えることなしに全く新しい製品の価値を理解してもらうことはさらに難しくなる．これまでに成功した新製品では技術それ自体の良さだけでなく，利用者側に潜在的なニーズと適切なタイミング，普及のための条件などが整っているところに投

† 産業技術総合研究所 人工知能研究センター：〒 135–0064 東京都江東区青海 2–4–7

入されていたことも多い．つまり，利用者の側，ユーザーサイドの状況にももっと目を向けることが成功の確率を高めることにつながる．イノベーションとは技術革新のみではなく，革新的技術が受容されることを通じた社会革新であり，製造業がもっと利用者側の状況や潜在的ニーズを考慮して利用者にとって価値の高い製品を持続的に開発できるような構造変革を目指すことが重要になっている．

一方，国民総生産(GDP)や労働人口の 7 割近くがサービス産業に関連している．サービス産業の実態は幅広いが，顧客と直接接する，顧客接点の多いサービスでは，いわゆる「おもてなし」として日本のサービス全般に質が高いと言われている．しかし，一般にそのサービス価値が高いと言っても，それは定量的な指標で比較できるものになってはいない．顧客満足度という概念はあっても，それが定量化されていないためにサービスの質を正当な価格に反映することができず，わかる人にはわかるが大きな利益を生むことはできないということになる．品質のよいサービスを提供している従業員がその結果に対する貢献度を示すことが難しく，待遇の改善や仕事へのモチベーションを上げることが難しい問題がある．また品質の高いよいサービスを提供する技術があくまで個々の従業員の経験により支えられ再現性が低いために，大きな波及効果を上げることは難しい．これまでの製造業の成長を支えてきた産業技術が高品質な製品を大量に生産，提供するバリューチェーンとして再現性の高い仕組みを構築することで成長してきたように，サービス業でもレバレッジ(乗数効果)の大きな効果や生産性を上げるための仕組みづくりが重要である．サービスは提供されるものであると同時に，そのサービスを受け取る利用者側の反応でその品質が評価されるものである．つまりサービスでは本質的に利用者の側，ユーザーサイドの反応に目を向けることがとても重要になる．

こうした背景のもと，製品や製品の機能も含めた価値の提供であるサービスと，さらにその結果生じる利用者側に起こる反応も含めた価値共創の仕組みを明らかにし，それを工学的な枠組みとして確立するサービス工学の試みが行われてきた(吉川, 2008)．2008 年経済産業省の研究拠点整備事業が開始されると同時に，産業技術総合研究所においてサービス工学研究センターが設立され(本村 他, 2008)，2012 年にはサービス学会も誕生した．モノと違って，サービスは複数の人の相互作用(コト)として伝搬するので，モノの機能評価と違い，人の心理や行動，状況に基づく相互作用や経験を計算する方法が必要である．この価値共創の主体が人であることが，明示的な表現体系がすでに確立されてきたモノを対象に発展してきた工学の体系だけでは取り扱うことができない本質的な問題である．つまり，社会システム，つまり人の集団活動をいかに工学的，体系的に取り扱うことができるか，がサービス工学の大きな課題であり，そのためにはサービスのシステム観が重要になる(図 1)．

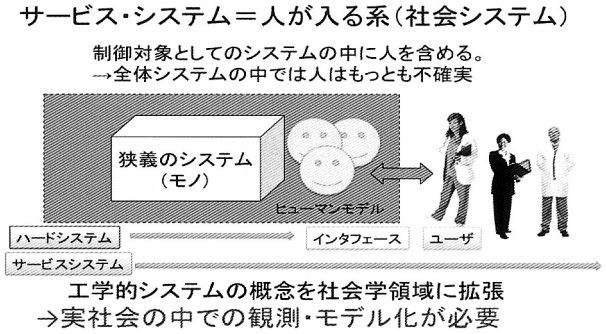


図 1．社会システムも含めたサービスのシステム観．

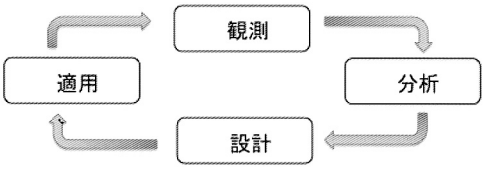


図 2. サービスの最適設計ループ.

サービスの特性はサービス提供と消費が同時に行われ，その品質はサービス利用者や状況に依存し，保存できないというもので，これらは同時性・異質性・消滅性というように整理されている．簡単に言えば，モノと違って，サービスは複数の人の相互作用(コト)として伝搬しており，人の心理や行動，状況に基づく相互作用や経験を計算する方法が必要というわけである．これは製品それ自体の価値ではなく，利用する時の価値についても成り立つ性質でもある．この価値の主体が人である，人起点であることが，これまでのモノを対象に発展してきた工学の体系だけでは取り扱うことができない本質的な問題である．サービスは無形であるが，サービス現場における人の相互作用の結果である行動履歴を大規模データとして客観的に観測することはできる．サービスの現場で起こる人の行動の結果を客観的に観測し，得られたデータを分析して得られる計算論的なモデルに基づいて，あるべきサービスを設計し，それを現場に適用するという「最適設計ループ」によって，価値を連続的に改良するアプローチを実践している(本村 他, 2008); (産業技術総合研究所, 2014)(図 2)．

このアプローチを実現するために，サービス中に生成されるデータを使って工学的に扱うことのできる計算モデルを活用してサービスを設計する方法がある．計算機の中で表し，実際に変数を計算したりすることで，あたかも実在するシステムを近似できるようなものを計算モデルという．この計算モデルによって，サービスシステムをとりあえず近似して説明できれば，その動きを模倣(シミュレート)することができる．すると，その動き方をいろいろと変えてみることで，より良い改善を考える．サービスの結果として起こる現象を説明する背後の関係を全部書き下すことは難しい．また，全ての人について成立する決定的なモデルを天下り的に見つけることも難しい．そこで，その日の行動を(例えば天気や曜日，利用者のライフスタイルなどから)確率的に予測し，その予測精度が実用上十分であるような不確実性を考慮した計算モデルを実際のデータから構築することで，サービスの現象を実用十分な範囲で近似できるモデルとして活用する．データからこうした計算モデルを構築する技術は機械学習(Machine Learning)と呼ばれる人工知能技術である．これらの技術はここ十数年の間，インターネットのデータについて使われ大きな成功を収め，今や日常の中でなくてはならないものになりつつある amazon.com や Google などの実際の IT サービスを支える基盤技術であり，さらに様々なモノがネットで接続される IoT(Internet of Things)時代においても注目されている技術である．現実の空間でのサービスの場合にはインターネットと違って機械学習やデータマイニングのために大量のデータを効率よく集める仕組みがまだ十分整っていない．しかし，電子マネーにより買い物をしたり，メンバーズカードに共通ポイントを貯めたりする利用者が増えており，さらに，今後行動履歴などの大量のデータが簡単に収集できるようになることが予想され，実社会物理空間(フィジカル)の活動と計算モデルの空間(サイバー)の活動が統合したサイバーフィジカルシステムと呼ばれるものとして社会活動をとらえる時期が来ている．サービス現場で得られるビッグデータを活用して実際の店舗やレストラン，病院などで起きている実際のサービス現象の計算モデルを作り，それを利用した支援技術を現場に提供できれば，実際のサービスの改善に役立つ．同様に製品の使われ方や利用した時の利用者が感じる価値を定量的に評価す

るモデルをデータから構築して，設計や製品の制御に活用すれば，それは製品開発や製品の制御にも役立つ仕組みでもある．本稿ではサービス現場で生成されるビッグデータとして利用者の ID をひもづけた購買履歴データ（ID-POS データ）や共通ポイントデータ，行動履歴データなどを活用して確率的モデルを構築する事例を紹介し，これが人工知能技術として社会実装されると，産業構造変革にどのような役割を果たすかを議論する．

2. ビッグデータを活用する人工知能技術：確率モデリング

人の行動は毎回同じとは限らず決まった通りに動くものではない．したがって，不確実性を考慮したモデルが必須になる．また実生活場面の現象を説明するモデル化においては記述量・計算量の点から，扱う対象自体を完全に記述することは無理であるので，現象を確率的・統計的なものとして扱うことにする．人の行動が起こる確率を考えて，その行動が起こる条件として典型的な（相互情報量が高い）状況をみつけると，条件付確率 P（行動 | 状況）という形で不確実性を含めて表すことができる．さらに人のタイプごとにとる行動が異なる場合には，さらにこれを条件部に加えて P（行動 | 状況，人のタイプ）とすればよい．この人のタイプは利用者の「異質性」とも呼ばれる．この人の異質性はサービスにおける基本的な特性であり，これをいかに取り扱うかはサービス工学における重要な課題である．この条件付確率の条件部に入る変数を加えていくことで，来店行動やある商品を買う購買行動の確率を精度良く予測できれば，あるタイプの顧客がお店に来る人数やその顧客が買いそうな商品の数も推定できるので，適切な人員配置や商品の準備をすることで人員不足や品切れを防ぐことができる．つまりサービスの最適化がはかれる．また，日常業務の中でその日の状況や顧客のデータをさらに大量に持続的に集め，確率モデルも新たなデータによって更新することができれば，予測精度をさらに高めることができる．これを実現するために条件付確率のモデルをデータから構築する技術がベイジアンネットである（本村・岩崎, 2006）．

利用者の異質性については，行動が似ている人を集めてセグメントを作ることが行動の予測を行う場合には適している．実サービス中に集積されている購買行動のデータであれば，行動は購入した製品の ID として見分けられる．会員カードなどの顧客 ID と製品 ID が記録された ID-POS データに対して確率的潜在意味解析（PLSA）を活用して利用者の異質性を潜在クラスとして発見する事例（本村 他, 2012, 石垣 他, 2010; 2011a, Ishigaki et al., 2010, 石垣 他, 2011b）がある．大規模な数年分の ID-POS データを使って，数千人の顧客を比較的少数のセグメントに分類し，セグメントごとに商品選択確率や来店行動などの予測の精度が向上することなどが確認されている．個人の ID 付データの場合，個人の行動やプライバシーの保護が問題になるのに対して，適切なセグメントにより異質性を表す方法は情報量を失わずにプライバシーも保護できる（山下・本村, 2014）．こうした適切なセグメントを探索するアルゴリズムとして確率的潜在意味解析が応用できる．

3. 確率的潜在意味解析

確率的潜在意味解析（Probabilistic Latent Semantic Analysis: PLSA）は，顧客毎の商品の買い方，商品から見ると買われ方が似ているものを同一クラスタに併合する操作を繰り返すことによって ID 付き POS データから顧客と商品を同時分類しカテゴリ生成を行う方法である．具体的にはまず，顧客ごとに購入した商品の数を集計した共起行列 を 作成する．また，顧客と商品が所属する潜在クラスの数を決める．その上で，顧客 x_i，商品 y_j，潜在クラス z_k の関係を次式，

$$P(x_i, y_j) = \sum_k P(x_i|z_k)P(y_j|z_k)P(z_k) \tag{3.1}$$

としてモデル化し，次式の対数尤度

$$L = \sum_i \sum_j n(i,j) \log P(x_i, y_j) \tag{3.2}$$

を最大化する方法が確率的潜在意味解析である．

本来は顧客ごとに購入した商品の数をベクトルとして考えて，類似した顧客をセグメント化したいが，商品の種類が数千点以上の規模である場合，それらの商品全てを購入することはほとんどない．したがってたいていの顧客はそのうちほとんどの商品の購買数がゼロとなり，顧客毎の商品の数をベクトルにした類似度計算がうまくいかない．そこで，次元を圧縮するために潜在クラスを導入して，適当な初期値で割り当てた商品分類を顧客の特徴ベクトルとして顧客を分類する．次に今度は商品の類似性を求めるために，先に得られた顧客分類を特徴ベクトルとして使って，新たに商品分類を更新する．この操作を繰り返すことで対数尤度は増大し，やがて収束する．最適なカテゴリの数は事前に決めることは難しいため，情報量規準(AIC)に基づいて最適なクラスタ数を探索する．このアルゴリズムによって，数千から数万の膨大な数の顧客と商品を比較的少数の潜在クラスへ分類する確率ベクトル $P(x|z)$，$P(y|z)$ が得られる．

ここで得られたセグメントは商品選択を対象に情報量が大きくなるように顧客を分割する場合の解である．本質的にどのように分類することが望ましいかどうかはそのセグメントをどのような目的に用いるかに依存する．商品の購入を促すためのマーケティング施策として，提供する商品や商品情報を選ぶ商品選択に対する情報量が大きいセグメントに顧客を分類することが最適である．一方，ある一日や場所ごとに実施する施策を最適化したいのであれば，商品の代わりに，時間や場所を対象にした PLSA を実行して得られた各セグメントに対して最適な施策を決定すべきであろう．このようにして PLSA の結果得られるセグメントは顧客セグメントというだけではなく，商品や時間や場所のセグメントとして活用することもできる．この PLSA のセグメントがどういう意味を持つのかを説明するために，ベイジアンネットを用いてセグメント間の構造をモデル化する(本村, 2016b)．これまでサービス現場での実証研究(吉田・本村, 2013)や応用，企業による実用化も多数行われている．

4. ベイジアンネットワーク

日常生活中における現象をモデル化するためには人の行動や心理などからくる様々な不確実性に対処することが必要である．そのため確率モデルを使って対象をモデル化することで，知りたい変数の確率分布を推定し，おこりえる各状態の確率(確信度)を評価する枠組みが有効である．複数の確率変数の間の定性的な依存関係をグラフ構造によって表し，個々の変数の間の定量的な関係を条件付確率表で表したモデル，ベイジアンネットワークがある(本村・岩崎, 2006)．

ベイジアンネットワークの中の一つの子ノードに注目した依存関係，つまり一つの目的変数(従属変数：Y)と，それに対する説明変数(独立変数：X)の間の依存関係について着目すると X-Y 空間を条件付確率表にしたがって量子化し，個々の確率値を割り当てたものになっており，これにより任意の非線形性，非正規性を表現できる．また複数の親ノードによる交互作用を表せること，つまり複数の親を持つ場合，非線形を持つ交互作用も表すことができることが現実の社会で起こる多様な事象をモデル化するために有用な特長になっている．日常生活場面では個人差，状況依存性などを反映する必要があり，この点で交互作用や非線形性，非正規性

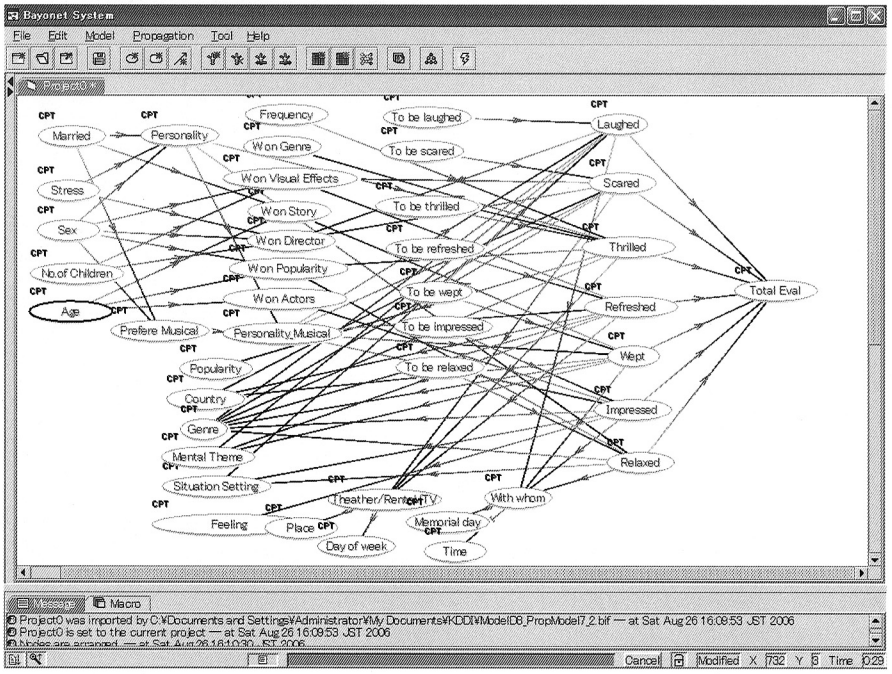


図 3. 映画推薦のためのベイジアンネット.

を含むモデル化が果たす役割が非常に大きい．また構築したベイジアンネットワークの一部の変数に観測値を代入して確率伝搬と呼ばれる計算を実行することで，それ以外の変数の確率分布が高速に求められるアルゴリズムが存在することも，様々な情報サービスに応用する上で重要な特長である．

ベイジアンネットワークと，その上で確率推論を実行することで，利用者の選択行動などを推定することができる．そこでユーザが選択する可能性の高いコンテンツを推定することで情報推薦に応用できる．とくに新たに追加されたコンテンツにも対応できるようにコンテンツ属性を変数として用い，さらにユーザ属性や状況を表す変数もベイジアンネットワークのノードとしてモデルに組み込むことで，状況やユーザの傾向に応じた推薦が可能になる．筆者らと KDDI 研究所のグループによる，携帯電話サービスのためにベイジアンネットワークを用いた映画コンテンツを推薦する事例(本村・岩崎, 2006)では約 1600 名の被験者に対して映画コンテンツを提示するアンケート調査により収集したユーザ属性，コンテンツ属性，コンテンツ評価履歴からベイジアンネットワークモデルを構築した(図 3)．

このモデルを用いて状況とユーザの嗜好性に応じて映画を推薦する携帯情報システムのプロトタイプを開発した．システムはデータベースから登録済みのユーザ属性情報と状況情報を使って確率推論を実行し，その結果選択される確率が高いと判断されたコンテンツを上位から推薦する．こうしてユーザの状況と嗜好性に応じて映画を推薦するサービスを実際に運用することで，持続的にデータの収集とモデルの改善が可能になる．こうして得られる各種の大規模データからベイジアンネットを構築すれば，人間の生活行動を推定する再利用可能な計算モデルとして様々な情報サービスにも活用できる(図 4)．

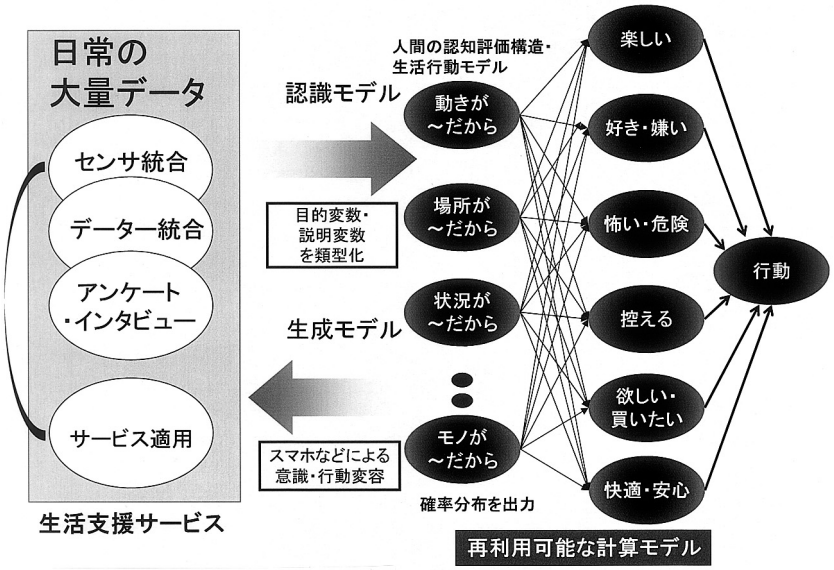


図 4. ビッグデータから構築される生活行動モデル.

5. 確率潜在意味構造モデリングと利用者のモデル化手法としての応用

確率的潜在意味解析とベイジアンネットを組み合わせることで，ID-POS データや共通ポイントカードの使用履歴データなどのサービス現場で大量に集積されているビッグデータから顧客の異質性を表し購買行動や嗜好性，アンケート回答の推定を行うことができる確率モデル，利用者モデルが構築できる(図 5).

この利用者モデルを使って，顧客それぞれに対して対応を個別に最適化することや，ある時間やエリア，ある商品に対して主たる利用者を推定して利用者の集団の特性を推定することでサービスを最適化する方法などが実現されはじめている．前者は会員カードやスマートホンなどと連携したレコメンド(情報推薦)やナビゲーション，後者は利用者セグメントごとの施策やサービスの最適化という形で実行されることになる(図 6)．またこれは従来，マーケティングやマネジメントとして実行されている業務のインテリジェント化としても位置付けることもでき，実際の産業応用としてはサービス分野の IT 化，あるいは人工知能技術の普及を進めるものとしても考えられる．今後，「サービスのシステム化」「社会基盤技術のインテリジェント

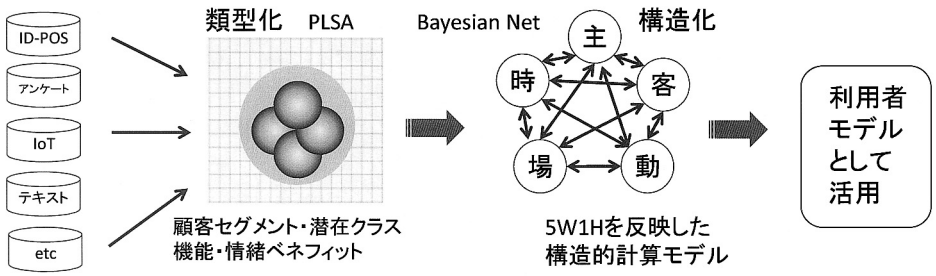


図 5. 確率的潜在意味解析(PLSA)とベイジアンネット.

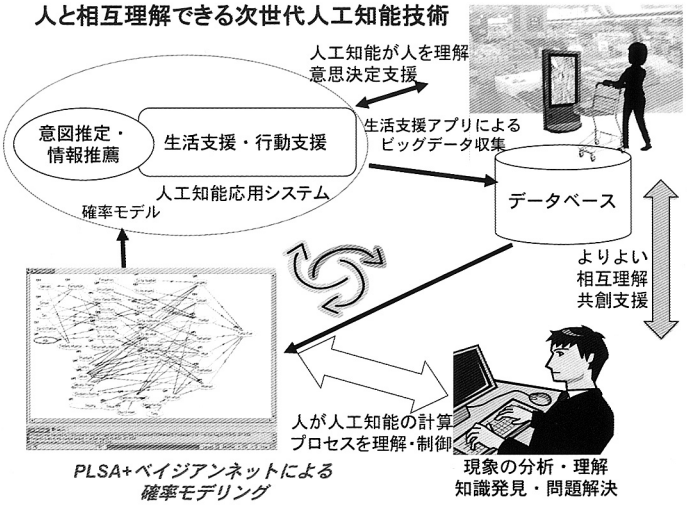


図 6. 人と相互理解できる AI による生活支援技術.

化」というものを考えた時，ビッグデータを計算モデル化して利用者の行動を確率的に予測することのできる人工知能技術をさらに多様なデータと組みあわせることによって，付加価値が高く，効率よく実行できる新たなサービスが生まれることになるだろう．

6. 産業構造変革のための人工知能技術としての役割

本稿で述べた確率的潜在意味解析(PLSA)とベイジアンネットを用いて確率モデルを自動的に構築する技術を情報サービスにも適用することが可能なソフトウェア(PLASMA)として実装した．このソフトウェアを使ったマーケティングサービスや，データを複数の企業が共有，連携して活用する新規事業などの立ち上げも開始され，こうした活動を支援するために後で述べる「人工知能技術コンソーシアム」が産総研内に設置され，本稿で述べた人工知能技術を活用することで，飛躍的な生産性の向上と，既存の価値創出の拡大を第一段階とし，さらにデータと計算モデルを共有することで，有機的に連携できる企業ネットワークを第二段階とした産業構造変革の実現に向けて活動している．

機械学習に基づく人工知能技術を社会の中で活用するためには，持続的なデータ収集を行う仕組みが不可欠であり，AI 技術の研究開発はウォーターフォール型よりは，スパイラル型の研究推進，つまり生活やサービスの現場でビッグデータの収集と技術活用を併行して行う必要性がある．そのため，産業応用プロジェクトの中では社会的なニーズの高い問題設定と，それに関与する多くのステークホルダーとの連携が不可欠になる．インターネットの発展とともに成長した巨大 IT 産業が，自身のサービスやビジネスを展開しながら最新の人工知能技術の研究開発を進めているのと同様，次世代人工知能技術の研究開発においても，現実的な場面における社会実装と技術検証，つまりユーザにとっての有用性や安全性，信頼性を初期の段階で示しながら性能を高度化するといった方法論(growth hack)が有効であると考えられる(本村, 2016a)．そのため，単に技術を研究開発するだけでなく，実際にどのようなデータから機械学習が実行され，社会の中でどんなサービスを実現するかを研究開発と合わせて考える工夫も重要である(本村, 2015)．

Internet of Things(IoT)が爆発的に進み，実空間における様々な現象がビッグデータとして記

録され，それらが計算機空間でモデル化，シミュレーション可能になる Cyber Physical System（CPS）の概念がある．IoT デバイスの普及とそこから生成されるビッグデータを活用することで，社会の現象を計算モデルとして構築することで新たな現象が計算可能になる．本稿で述べた確率的潜在意味解析（PLSA）とベイジアンネットを統合した手法はビッグデータを扱う統計手法であると同時に，これが社会実装されると人工知能技術として次のような役割を果たす．実空間の活動が IoT デバイスにより時間，空間情報とともにデータ化され，これが PLSA により情報量の高い潜在クラスに確率的に分類，すなわち確率ベクトルにコード化される．簡単のために所属確率が高いベクトルを 1 に，それ以外を 0 のように二値化されると考えると，実空間の現象の統計的な共起関係が PLSA の尤度極大化により，できるだけ情報量が高くなるようにデジタル化したことに相当する．さらにこのデジタル化された現象間の関係がベイジアンネットによって構造化されることで，時間や場所，人の特性，行動，対象物などがいわゆる「いつ」「どこで」「誰が」「何を」「どうした」という形で関係性や交互作用を確率的に表現できるように計算モデル化される．この計算モデルの入出力が実社会のサービスとして実行され，新たな実空間の現象を生じさせるならば，これは，計算モデルの空間（サイバー）上で，実空間（フィジカル）の現象が密に連携したサイバーフィジカルシステムの一つの実装方法となっている．

サイバー空間で最適化された計算結果をスマホのアプリやサービスを通じて人々に提供し，意思決定や行動を支援することで，良い現象の発生確率を上げ，事故などの良くない現象の発生確率を下げるという意味での物理世界の制御，マネジメントが可能になる．こうした IoT デバイスと AI 技術により構成される Cyber Physical System（CPS）を活用することで，産業構造変革を進め，生産性向上や付加価値の向上に寄与することが期待され，実際に 2018 年度から多数の政府系プロジェクトや企業活動が開始している．人工知能技術を活用し実社会の産業構造変革に貢献するためには，現実の社会構造や生活と乖離することなく，人々にとって扱いやすい形でその技術とサービスが提供され，制度や文化の進化とも歩調を合わせて社会実装を円滑に進めていく必要がある．

7. おわりに

機械学習やビッグデータを活用する人工知能技術により，サイバーフィジカルシステムを構築することで複雑な社会問題の解決も期待されている．ただし，計算モデルが高度で複雑なものになるにつれ，学習のために必要なデータ量が増大する．表層的に観測可能なセンサデータなどは比較的容易に取得できるが，人間行動の内部的状態は心理的なものであるため，被験者を用いたアンケート調査も必須になりコストが大きい．またデータを取得する上で，プライバシーの問題や，単に研究目的のためには協力が得られにくいという現実的な問題もある．またたとえ外部的な要因で観測容易な事象だとしても，実際に使う場面において，状況依存性の高い説明変数を網羅的に収集するためには，データを観測する環境が日常的な利用環境とできるだけ合致するように統制しておく必要がある．

そこで，こうした問題に対して実サービスと調査・研究を一体化すべきであるとする「サービスとしての調査・研究（Research as a service）」という概念が提唱されている（本村, 2009）．調査・モデル化の段階とそのモデルを用いた応用を切り離すことなく，情報サービスを社会の中で実行しながら，そこで得られる観測や評価アンケート，利用者のフィードバック（心理的調査）の結果を網羅的に収集する．これは古くはサイバネティクス，また信頼性工学ではデミングサイクルとして知られる PDCA（Plan, Do, Check, Action）サイクルを，実問題を通じて回し続けることで，モデルを常に修正していくというものである．

不確実性に対する本質的な解決のためには対象を実データによりモデル化し，そのモデルを

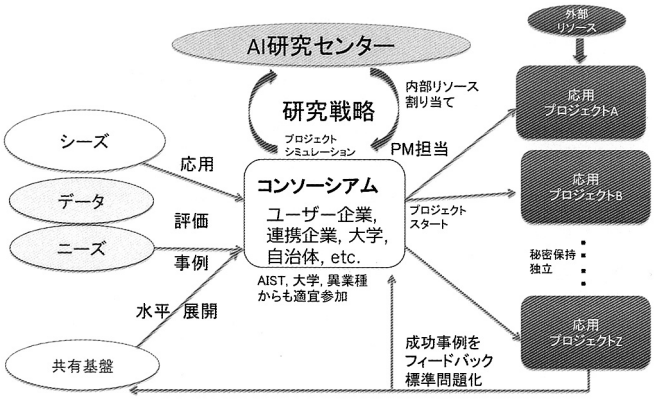


図 7. 産総研人工知能技術コンソーシアム.

用いて制御しながらさらにデータを収集する，というサイクルを永続的に続けるアプローチが重要になる．大規模データと人工知能技術を活用した情報システムを導入した実サービスの開発と応用を通じて，多様な生活者の特性を計算モデル化し，有用な知識モジュールとして社会全体で活用できるビッグデータ活用技術の仕組みを確立することが期待できるようになっている．こうした技術を産業基盤として社会の実サービスの中で大規模データの収集とモデルの構築，活用を持続することでさらに多くの情報サービスの実現が容易になる．こうしたビッグデータと人工知能技術により産業生産性を向上し，生活の質も向上することを目指し異業種連携を加速するための場づくりとして，産総研内に「人工知能技術コンソーシアム」が設立され，2018 年度 180 社以上の企業が 10 を越えるワーキンググループ，関西，東海，九州の地域支部により活動し，ユーザ参加型の社会実装，技術評価を進めている (図 7).

共同で行動を行うコミュニティは実践コミュニティ (Community of practice) と呼ばれる．この実践コミュニティの形成や運用を人工知能技術によって支援し，誰もが自由に出入りして参加したりできるような「場」ができることによって，柔軟にサービスシステムが共創的に価値を創出できるようになることも期待できる．こうした価値共創を持続的に再現，発展できる仕組みによって社会や暮らしを今よりもよいものにするという方向性を強く打ち出すことが，人工知能技術により新たな産業や社会構造を変革するための社会実装のために有効であると考えられる.

参 考 文 献

石垣司, 竹中毅, 本村陽一 (2010). 確率的潜在意味解析を用いた大規模 ID-POS と顧客アンケートの統合利用による顧客-商品の同時カテゴリ分類, 電子情報通信学会技術研究報告. NC, ニューロコンピューティング, **109**(461), 425–430.

Ishigaki, T., Takenaka, T. and Motomura, Y. (2010). Category mining by heterogeneous data fusion using PdLSI model in a retail service, *Proceeding on IEEE International Conference on Data Mining* (ICDM), 857–862.

石垣司, 竹中毅, 本村陽一 (2011a). 日常購買行動に関する大規模データの融合による顧客行動予測システム：実サービス支援のためのカテゴリマイニング技術, 人工知能学会論文誌, **26**(6), 670–681.

石垣司, 竹中毅, 本村陽一 (2011b). 百貨店 ID 付き POS データからのカテゴリ別状況依存的変数間関係

の自動抽出法, オペレーションズ・リサーチ, **56**(2), 77–83.
本村陽一 (2009). 大規模データからの日常生活行動予測モデリング, シンセシオロジー, **2**(1), 1–11.
本村陽一 (2015). サービス工学におけるビッグデータ活用技術, 日本機械学会誌, **118**(1163), 628–631.
本村陽一 (2016a). 次世代人工知能技術, 情報処理学会誌, **57**(5), 466–469.
本村陽一 (2016b). 第 9 章『ベイジアンネットワークと確率的潜在意味解析による確率的行動モデリング』,『確率的グラフィカルモデル』(鈴木譲 他), 共立出版, 東京.
本村陽一, 岩崎弘利 (2006).『ベイジアンネットワーク技術〜ユーザ・顧客のモデル化と不確実性推論』, 東京電機大学出版局, 東京.
本村陽一, 西田佳史, 持丸正明, 赤松幹之, 内藤耕, 橋田浩一 (2008). サービスイノベーションのための大規模データの観測・モデリング・サービス設計・適用のループ, 人工知能学会誌, **23**(6), 736–742.
本村陽一, 竹中毅, 石垣司 (2012).『サービス工学の技術〜ビッグデータの活用と実践〜』, 東京電機大学出版局, 東京.
産業技術総合研究所 (2014).『社会の中で社会のためのサービス工学〜モノ・コト・ヒトづくりのための研究最前線』, カナリア書房, 東京.
山下真一郎, 本村陽一 (2014). 確率的潜在意味解析を用いた集団匿名化法における来店店舗予測精度の評価, 人工知能学会全国大会 1L2-OS-17a-3.
吉川弘之 (2008). サービス工学序説—サービスを理論的に取り扱うための枠組み—, シンセシオロジー, **1**(2), 111–122.
吉田真, 本村陽一 (2013). ベイジアンネットワークによるセグメント説明モデルと映画推薦への応用, 第三回 Web インテリジェンスとインタラクション研究会.

Probabilistic Modeling Technology Using Big Data: Activity for Social Implementation

Yoichi Motomura

Artificial Intelligence Research Center, National Institute of Advance Industrial Science and Technology

Currently, the practical application of artificial intelligence is being dramatically advanced by machine learning using big data. These efforts are also expected to help realize industrial structural reform and the smart society ("Society 5.0"). In this paper, we introduce probabilistic modeling using probabilistic latent semantic analysis and Bayesian networks. To realize the value of service and improvementin productivity, the user's behavior and preference are predicted by probabilistic models constructed from service history data (ID-POS data, questionnaire with ID, operation history with ID). Examples of real applications and efforts at social implementation are also discussed.

Key words: Service engineering, artificial intelligence technology, probabilistic modeling, Bayesian networks, probabilistic latent semantic analysis, big data.

統計数理(2018)
第 66 巻 第 2 号 225–233

特集「サービス科学の今」
[総合報告]

位置情報軌跡の統計的プライバシー保護

南 和宏†

(受付 2017 年 12 月 5 日；改訂 2018 年 3 月 27 日；採択 4 月 6 日)

要　　旨

スマートフォンの普及に伴い，我々の位置情報の取得が容易になり，多くのユーザーの移動履歴は，交通情報の提供，都市設計といった社会サービス，また商圏分析等の企業活動にも活用されている．一方，位置情報から個人の興味に関するプライバシーに関する情報が漏洩する危険性が懸念されている．位置情報の時系列データは，既存の匿名化手法の適用が困難な多次元データであり，個人の行動習慣，移動経路の制約等を反映した時空間の相関性を利用した統計的推論攻撃に対する防護策が必要となる．本記事では，位置情報軌跡を安全に分割する動的仮名更新手法，および，時空間の相関性による情報漏洩リスクを考慮した状態空間モデルに基づく匿名化データの安全性評価手法を紹介する．

キーワード：位置情報，匿名化，仮名化，マルコフ過程，状態空間モデル．

1. はじめに

近年，スマートフォンによる GPS 座標の位置情報の取得に加え，携帯電話の基地局，WiFi のアクセスポイント，IC カードによる電車の乗車履歴等，様々な種類の位置情報を入手することが可能になった．それら位置情報を統合することで多数のユーザーの長期間，広域での移動履歴が把握でき，首都圏でのリアルタイムの人口統計の提供（寺田 他, 2012），また商圏分析（清嶋, 2012）等の企業活動にも活用されている．さらには，運転操作データのような他の IoT データと位置情報を組み合わせることで，交通事故の発生する可能性の高い地点を表示するヒアリハット地図の作成（中野・豊田, 2015）に活用されている．

その一方，位置情報から，個人の習慣，興味，行動，交際範囲等，プライバシーに関する情報が明らかになる危険性がある．例えば，病院への定期的な訪問は重大な病気が推測され，カフェなど同一の場所での複数人の集まりは秘密の会議の開催を示唆するかもしれない．また位置情報は，ストーカーや空き巣のような犯罪に利用される可能性もある．よって位置情報の安全な 2 次利用には匿名化と呼ばれる個人の識別情報を取り除くデータ加工が不可欠である．

通常，匿名データを作成する際，氏名等の個人の識別子を削除するだけでは不十分である．なぜなら「年齢」，「性別」といった個人を断片的に識別する準識別子と呼ばれる属性情報が存在し，それらの組み合わせで個人の特定が可能になるからである．したがって，一般的には人々の属性に関する準識別子の情報を用いて k 未満のユーザーに絞り込むことを防ぐ k-匿名化処理（Sweeney, 2002）を行う．k-匿名化では元データを類似する k 個以上のレコードを含むグループに分割し，他のデータセットと照合されても特定のレコードの客体が再識別されることを防止するため，同一グループ内のレコードが同じ値を取るように一般化する．位置情報の

† 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

場合，情報の粒度を粗くすることが一般化処理に対応する．

しかし通常の k-匿名化の手法を位置情報軌跡に適用する場合，2 つの課題が存在する．1 つは，位置情報軌跡のような各時刻の位置情報を含む多次元データの場合，k-匿名化を実施するとデータの有用性が著しく劣化してしまう問題である．位置情報は個人のユニークな行動パターンを反映しており，長期的な位置情報軌跡を匿名化する場合，互いに類似する軌跡のグループを見つけることは困難である．そのような軌跡群をグループ化して一般化処理すると情報の損失が大きくなり，有益なデータ分析に堪えない．

2 つめは，位置情報軌跡のデータ間に時空間の相関性が存在し，匿名化した位置情報から統計的推論により元の軌跡情報が復元される問題である．位置情報軌跡には，人の移動に関する物理的制約が反映し，短時間での移動範囲は局所的であり，車，電車といった交通手段により移動経路は限定される．また長期的な移動軌跡には通勤，病院への通院といった個人の生活習慣を反映した特徴的なパターンが現れる．そのような移動パターンに関する外部知識を用いると一般化された匿名化データから元に位置情報が復元される危険性がある．

本記事では，この 2 つの課題を解決するための 2 つの手法を中心に位置情報の匿名化技術を紹介する．1 つは位置情報軌跡を複数のセグメントに分割する動的仮名交換手法（Tanjo et al., 2014）であり，ミックスゾーンと呼ばれる複数ユーザーの集積点でのランダムな仮名の再割当により移動先の不確定性を確保する手法である．もう 1 つは，状態空間モデルに基づく匿名化データの安全性評価手法である．ユーザーの移動パターンをマルコフチェーンでモデル化し，隠れマルコフモデルにおける内部状態の推定問題として匿名化データの安全性の評価を行う．

2. 攻撃者モデル

図 1 は位置情報データの流通形態を示す．位置情報サーバーは各ユーザーから定期的に時刻でタグ付けされた位置情報を受け取り，図 2 に示す表データに集計する．この表の各行はユーザーの位置情報を保持し，各列はある時刻の位置情報の値を示す．ただし，説明を簡略化するため，本記事では位置情報は座標値ではなく，座標値から変換された位置情報の領域に対応するグリッド ID を示す．位置情報サーバーは集計した表データから「氏名」等の識別子情報を削除する等，匿名化処理を施した匿名化データをデータ分析者に提供する．

匿名化技術における攻撃者は公開された匿名化データから標的とするユーザーの元データにおける位置情報軌跡を復元しようとする．この攻撃者は匿名化データ以外に標的とするユーザーの住所，勤務先，その他の目撃情報等の外部知識を利用する．このような攻撃者を想定すると単純な「氏名」等の識別子を仮名に置換する単純な匿名化処理では不十分である．例えば，図 3 は 3 人のモバイルユーザーの仮名化された軌跡を示す．もし攻撃者が「鈴木さん」の住所情

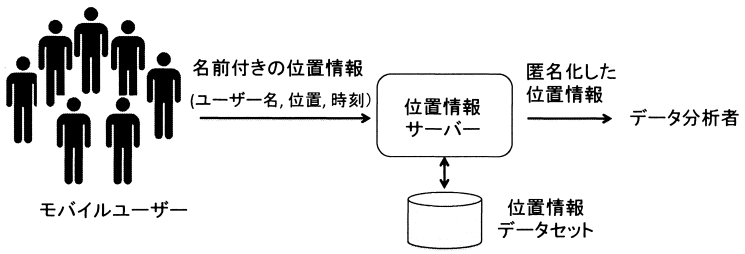


図 1．位置情報サービスの流通形態．データ分析者への位置情報の提供は一般公開の形態をとるため，元データから個人を識別する情報を除いた匿名化データを提供する．

2017年11月8日

氏名	8:00	8:30	9:00	9:30	10:00	10:30	11:00	
伊藤	1	5	4	8	12	15	9	
加藤	10	15	24	14	21	20	19	.....
鈴木	3	8	6	6	7	10	15	
高橋	23	24	19	11	9	4	5	

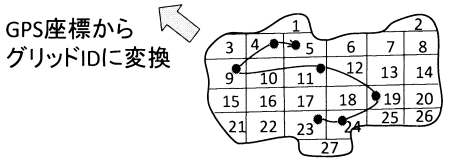


図 2. 位置情報軌跡の表データ.

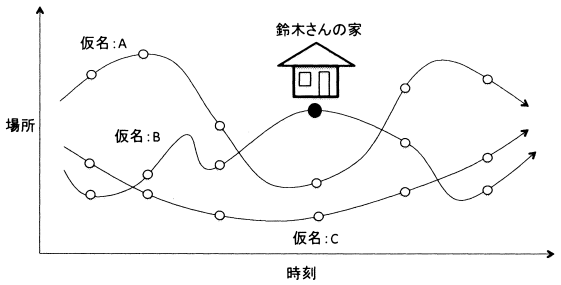


図 3. 仮名化された移動軌跡の概念図. 3 人のモバイルユーザーの実名を仮名 A，B，C に置き換える匿名化処理が行われている.

報を入手しており，仮名 B の軌跡がその場所を経由していることが分かると，仮名 B の軌跡は識別されてしまう.

実際これに類似する問題が 2012 年，JR 東日本が IC カード乗車券「Suica」の仮名化した乗車履歴を日立製作所に提供した時に起きている．Kikuchi ら（Kikuchi and Takahashi, 2015）は，駅の平均乗降数がジップの法則（Zipf's law）に従うと仮定してこの乗車履歴の安全性を評価し，個人が普段使う駅が 3 つ分かるだけで膨大な数の乗車履歴レコードから元の個人名が再識別されると分析している.

3. ミックスゾーンにおける動的仮名割当

2 章で示したように，個人の行動パターンが顕著に現れる位置情報軌跡の場合，その中の幾つかの点に過ぎない外部知識を用いて個人を識別することが可能性である．またいったん移動軌跡が識別されるとその軌跡全体の情報が開示されることとなり，位置情報軌跡の情報漏えいリスクは非常に高い.

本章では位置情報軌跡に紐付けられる仮名を動的に更新し長期間の軌跡データを複数の軌跡セグメントに分割することで情報漏えいリスクの局所化を実現する方式（Tanjo et al., 2014）を紹介する．この仮名の更新は複数のユーザーが同一の場所に存在する「ミックスゾーン」と呼ばれる場所で仮名交換の形式で実施され，ミックスゾーンを経由することで軌跡セグメント間の

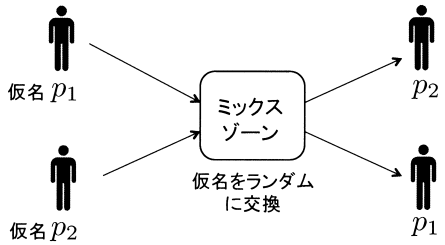


図 4. ミックスゾーンにおけるランダムな仮名再割当.

関連性を分断する.

3.1 ミックスゾーン

図 4 にミックスゾーンの概念を示す．各ユーザーの位置情報は実 ID を置き換えた仮名と紐付けられており，複数のユーザーが同時刻に集まるミックスゾーンで仮名をランダムに再割当が行われる．図 4 では仮名 p_1 と p_2 をもつ 2 人のユーザーがミックスゾーンに入り，そこで 2 つの仮名がランダムに再割当される．再割当の可能性は仮名が両者で交換される場合と交換されずに同じ仮名を持ち続ける場合の 2 通りである．ミックスゾーンを経由したとき，どちらの再割当が実施されたか匿名化データからは判別できないため，最初に仮名 p_1 のユーザーがミックスゾーン経由後，引き続き仮名 p_1 の経路をとる場合と p_2 の経路に移る場合の 2 つの可能性が共存する．このようにミックスゾーンを経由すると個人の移動軌跡の可能経路が分岐し，全体の位置情報軌跡の不確定性を増加させることができる.

3.2 安全性評価と排他的辺素パス問題

このミックスゾーンにおける仮名交換方式では，ユーザー u の時刻 t におけるプライバシー指標は到達可能な位置の数として定式化でき，到達可能な位置の数が多いほど高いプライバシーが保証できる．ただし一般のユーザーは自宅を始点として出発して最後はやはり終点である家に戻るという攻撃者が容易に知りうる拘束条件を持つため，全てのミックスゾーンの分岐経路が利用できるとは限らない．さらに，あるユーザーが特定の経路を使うと別のユーザーの始点から終点への経路が存在しなくなるという問題も生じる．つまりあるユーザーの代替経路を列挙する場合，他の全てのユーザーについても妥当な経路が存在することを保証する必要がある．もしデータ・セットに n 人の位置情報軌跡が含まれるとすると，この問題は図 5 に示すようなミックスゾーンをノードとするグラフ上で n 個の(始点，終点)の組が与えられたときに排他的辺素パスを列挙する問題に相当する．ユーザー u_2 が始点から終点に到達する経路は単独では $\langle 1 \to 3 \rangle$, $\langle 1 \to 2 \to 3 \rangle$ の 2 通りの順序でミックスゾーンを通過する経路が存在するが，後者の経路を選択するとユーザー u_1 が終点に向かう経路を分断してしまう．したがってこの例では排他的辺素パスは一組しか存在しない.

ユーザー数が入力として指定される場合，排他的辺素パス問題は NP 完全問題であることが知られている（Karp, 1975）．また攻撃者が始点，終点以外の中間の地点の情報(例えば，勤務先)を外部情報として持つ可能性もある．したがって仮名更新による安全性を評価するには排他的辺素パス問題をさらに一般化する必要がある．Tanjo et al., 2014 では一般化された排他的辺素パス問題を制約充足問題（Rossi et al., 2006)に変換し，仮名の変数の全ての異なる解の数を求めることで安全性の評価を行うシステムを開発した.

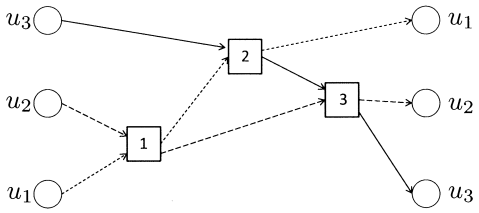


図 5. ミックスゾーンにおける排他的辺素パス問題．丸ノードは各ユーザーの始点，終点，四角のノードはミックスゾーンを表す．

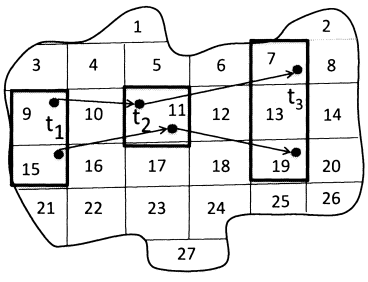
4. 状態空間モデルに基づく匿名化データの安全性評価

3 章で紹介した仮名更新による位置情報軌跡の分割は位置情報の識別リスクを局所化する手法である．しかしある位置情報が識別されると同じ軌跡セグメント内の位置情報は依然として漏洩してしまう．したがって分割して次元を削減した軌跡セグメント単位に対して k-匿名化を実施することが望ましい．ただし，位置情報軌跡には時空間の相関性が存在するので，通常の k-匿名化では不十分な場合が多い．本章では，統計的推論攻撃のリスクに対処するための状態空間モデルに基づく安全性評価の手法を紹介する．

4.1 k-匿名化の課題

k-匿名化は，標的ユーザーの軌跡の k 個未満への絞り込みを防ぐ実用的なプライバシー指標である．図 6 の例では，2 人のユーザーの位置情報を領域区分の ID で示しており，2 人の移動軌跡を同一にする 2-匿名化を実現するためには，粒度を粗くした太枠の区分に位置情報を変換する必要がある．

しかし位置情報の時系列データには，個人の行動習慣，移動経路の制約等を反映した時空間の相関性が存在し，k-匿名化した位置情報から元の軌跡の復元が可能な場合がある．例えば，図 7 は，2 人の車を運転するユーザーの移動軌跡を 2-匿名化した例である．太枠の区画に位置情報が一般化されているものの，破線で示す道路の経路情報が与えられれば，2 人のドライバーの位置を詳細に推測することは容易である．つまり k-匿名化は匿名化したデータの形式（シンタックス）のみを要件とするプライバシー指標であるため，軌跡データ間の相関性による情報漏えいのリスクが考慮できていない．



元データ

ユーザー	時刻 t_1	時刻 t_2	時刻 t_3
田中	{9}	{11}	{7}
鈴木	{15}	{11}	{19}

一般化処理

2-匿名化データ

仮名	時刻 t_1	時刻 t_2	時刻 t_3
A	{9, 15}	{11}	{7, 13, 19}
B	{9, 15}	{11}	{7, 13, 19}

図 6. 位置情報の 2-匿名化の例．位置情報はグリッドの集合として表現される．

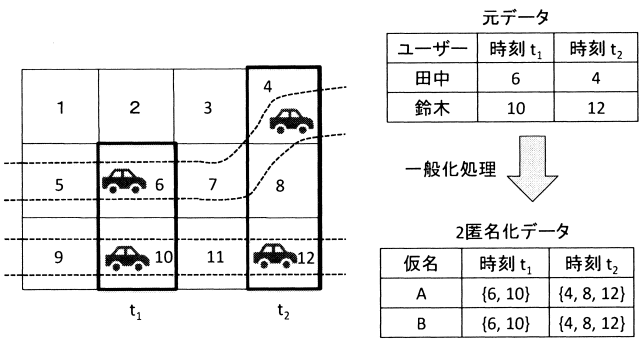


図 7. 道路情報を用いた位置情報の復元．破線は道路の形状を表す．

4.2 ユーザーの移動モデル

人々が移動する経路は，道路や建物の形状，通勤，通学等の生活習慣など様々な要因の影響を受ける．このような移動経路の特徴を統一的に捉えるには，時系列データの確率モデルであるマルコフ過程が適している．なぜなら人の移動を場所間の状態遷移と考える場合，移動経路に影響を与える要因は移動時の遷移確率に間接的に反映されるからである．例えば，歩行者が急に遠方の場所に移動できない事実や，侵入不可の建物による移動の制限は，それらの場所への遷移確率がないと記述すればよい．つまり移動に関する個別の要因をそれぞれ明示的に記述する必要がない．

このようなマルコフ行列は，各ユーザーの過去の移動軌跡を学習して作成することができる．ユーザー u の移動範囲が N 個の離散的な場所とすると，その状態遷移は $N \times N$ のマルコフ行列 P^u で記述される．マルコフ行列の各行 i が現在位置のグリッド i，各列 j が次の移動先のグリッド j に相当する．そしてグリッド i から j へ移動する確率はマルコフ行列の要素 $P^u_{i,j}$ で示される．

4.3 隠れマルコフモデルによる匿名化処理のモデル化

ユーザーの移動パターンをマルコフチェーンでモデル化し，匿名化技術の安全性評価を隠れマルコフモデルにおける観測情報から内部状態の推定問題として定式化する（Minami, 2014; Shokri et al., 2011）．図 8 のモデルの観測情報は匿名化データ，内部状態遷移は秘匿すべき元の位置情報に相当する．そして，匿名化処理は，内部状態から観測情報への確率的な変換を定義する記号出力行列として表現される．このモデル化の主な利点は，攻撃者が標的とする客体の移動パターンに関する知識を保持する場合，匿名化データの安全性を真の内部状態を推定する条件付き確率として定量的に評価できる点にある．

位置情報の匿名化手法は，4.1 章で述べた位置データの粒度を変える一般化処理以外にも幾つか存在する．例えば，図 8 において位置 l_1 は粒度が粗い l'_1 に一般化されているのに対し，位置 l_2 は省略を意味する空の文字($\perp$)に置換されており，また末端の l_T にはノイズが付加され，真の位置とは異なる l'_T に変換されている．

但し，図 8 は説明の簡略化のために一人のユーザー u_i の情報のみを表示しているが，実際のモデルは n 人のユーザーの情報を表現する必要がある．つまり内部状態は時刻 t における n 人の位置情報をもつベクトルとして表現し，それに応じて状態遷移を表すマルコフ行列，匿名化処理を記述する記号出力行列も拡張することになる．

例えば，図 9 は，2 人のユーザー u_1 と u_2 が 4 つのグリッド領域 $\{1, 2, 3, 4\}$ を移動する状況

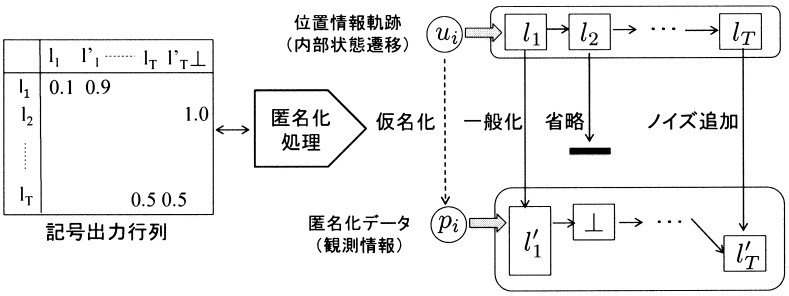


図 8. 隠れマルコフモデルによる匿名化技術のモデル化．元の位置情報軌跡は記号出力行列で定義された匿名化処理が適用され，匿名化データに変換される．省略された位置データは ⊥ で示す．

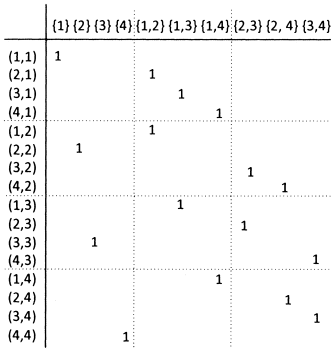


図 9. 2-匿名化関数を表現する記号出力行列．グリッド集合 $\mathcal{G} = \{1, 2, 3, 4\}$ とする．2 人のユーザーがそれぞれグリッド i, j に位置する状態を (i, j) と表す．2-匿名化関数は入力 (i, j) に対して決定的に $(\{i, j\}, \{i, j\})$ を出力する．2 人の匿名化された移動軌跡が同一なので，記号出力行列の各列の見出しは $(\{i, j\}, \{i, j\})$ を $\{i, j\}$ と簡潔化している．

で 2-匿名化を実施する場合の記号出力行列を示す．2 人のユーザーが時刻 t にそれぞれグリッド k, l に位置する場合，内部状態は (k, l) である．この 2-匿名化関数は単純な決定的関数であり，$f(k, l) = (\{k, l\}, \{k, l\})$ である．もし 2 人のユーザーが同じグリッド k に位置する場合，一般化加工は行われず出力は $f((k, k)) = (\{k\}, \{k\})$ となる．

図 9 は決定的で単純な匿名化アルゴリズムを記号出力行列として記述することが可能であることを示した．しかし，記号出力行列の形式で一般の匿名化アルゴリズムを記述するのは困難と予想され，隠れマルコフモデルの適用可能な匿名化アルゴリズムに一定の制限があることは明らかである．したがって，モデル化の適用範囲を明らかにし，匿名化アルゴリズムの汎用的な記述手法を確立することが今後の研究課題である．

5. まとめ

本記事では，一般に広く普及している k-匿名化手法を位置情報軌跡に適用する場合の 2 つの課題に対する解決策を提示した．1 つは，多次元データの匿名化における「次元の呪い」の問題 (Aggarwal, 2005) を回避するための位置情報軌跡の仮名更新による軌跡分割手法である．仮名

化された位置情報の安全性評価は排他的辺素パス問題の解列挙の問題に帰着され，制約充足問題ソルバーを利用した効率的な実施手段を実装した．

もう 1 つは位置情報の時空間の相関性を用いた推論攻撃に対する対策である．k-匿名化は標的となるユーザーと匿名化データの紐付け防止には一定の効果があるものの，移動に関する統計情報を用い，匿名化データから詳細の元データを復元する攻撃の危険性を見過ごしている．匿名化処理を隠れマルコフモデルでモデル化し，匿名化データの安全性を観測情報から内部状態への推定問題として定式化できることを示した．ただし，モデル化の対象に任意の匿名化アルゴリズムが含まれることになり，計算論的アルゴリズムと統計モデルの融合は今後の長期的課題となると予想される．

参 考 文 献

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality, *Proceedings of the 31st International Conference on Very Large Data Bases*, 901–909.

Karp, R. M. (1975). On the computational complexity of combinatorial problems, *Networks*, **5**, 45–68.

Kikuchi, H. and Takahashi, K. (2015). Zipf distribution model for quantifying risk of re-identification from trajectory data, *2015 13th Annual Conference on Privacy, Security and Trust* (*PST*), 14–21.

清嶋直樹 (2012). 電通 Draffic, http://itpro.nikkeibp.co.jp/article/JIREI/20121005/427881/.

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

中野美由紀, 豊田正史 (2015). ビッグデータがもたらす超情報社会—すべてを視る情報処理技術：基盤から応用まで．ビッグデータ時代を生きる, 情報処理, **56**(10), 958–961.

Rossi, F., van Beek, P. and Walsh, T. (2006). *Handbook of Constraint Programming*, Elsevier Science Inc., New York.

Shokri, R., Theodorakopoulos, G., Boudec, J. Y. L. and Hubaux, J. P. (2011). Quantifying location privacy, *2011 IEEE Symposium on Security and Privacy*, 247–262.

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

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

寺田雅之, 永田智大, 小林基成 (2012). モバイル空間統計における人口推計技術（社会・産業の発展を支える「モバイル空間統計」：統計情報に基づく人口推計技術とその活用）, NTT DoCoMo テクニカル・ジャーナル, **20**(3), 11–16.

Statistical Privacy Protection of Location Trajectories

Kazuhiro Minami

Institute of the Statistical Mathemetics

Nowadays, trajectory location data, which is collected from peoples' smart phones, can be used for various analytic purposes, such as traffic monitoring, urban city planning. However, due to significant concern about location privacy, location data must be anonymized properly before making it available for secondary usage. Unfortunately, trajectory location data is inherently difficult to anonymize due to its high-dimensionality. Furthermore, we need to take additional measures to prevent inference attacks exploiting strong temporal and spatial correlations among data points. In this article, we present a technique of dynamically pseudonyms that devides a location trace into multiple segments and describe a state-space model to evaluate the safety of anonymized location data.

Key words: Location privacy, anonymization, Markov chain, state space model.

統計数理(2018)
第 66 巻 第 2 号 235–247

特集「サービス科学の今」
[研究詳解]

大規模集計POSデータの高次元スパースモデリング

李 銀星[1]・照井 伸彦[2]

(受付 2017 年 11 月 16 日；改訂 2018 年 5 月 17 日；採択 6 月 7 日)

要　　旨

多様な消費者ニーズをきめ細かく捉えて顧客を獲得して維持するための効果的マーケティングのために，主体(消費者)の異質性の統計モデリングが他の分野に先駆けて開発された．他方，実店舗において，個別対応は必ずしも容易ではないのも現実である．本稿では集計 POS データに対して機械学習などの新しい統計分析による高度情報処理を適用することにより，多くの実店舗で活用できる大規模データを活用したマーケティングモデルの可能性を展望する．高次元データについては，2 種類の次元圧縮法，すなわち，トピックモデルによる次元圧縮と購買文脈による部分市場分解，階層因子回帰モデルによる次元圧縮とパラメータの高次元空間への還元が議論される．

全商品データを分析に取り入れることで，目的変数を説明する意外な変数の発見のみならずその量的関係が弾力性の形で測定可能となり，実店舗のきめ細かいマーケティング戦略に有用な情報が提供可能となることを展望する．

キーワード：集計 POS データ，購買状況の異質性，トピックモデル，高次元スパースデータ，階層因子回帰．

1. はじめに

顧客データベースの整備を背景にして，多様な消費者ニーズをきめ細かく捉えて顧客を獲得して維持するための効果的マーケティングのために，主体(消費者)の異質性の統計モデリングが他の分野に先駆けて開発された．例えば，顧客データベースから消費者選択を階層ベイズモデルでモデル化して顧客ごとの市場反応を推定する Rossi et al. (1996)は，個別化モデリングの先駆けである．価格感度の違いに応じて個別化したクーポンの発行による個別対応は Rossi et al. (1996)や Terui and Dahana (2006)などにおいて提案されている．これらの包括的な説明は照井 (2018)でなされている．その後，個別対応のマーケティングは，機械学習の手法も巻き込んで発展し，E ビジネスの世界で日常的に行われている．

実店舗においてはマーケティングの個別対応は必ずしも容易ではないのも現実である．POS システムはほとんどの実店舗で導入されており，集計データは無自覚的に日々蓄積されている．これら集計データのマーケティング分析は古くから回帰や時系列モデルにより行われてきた．Hanssen et al. (2001)では，集計データのマーケティングモデルについて包括的に説明してい

[1] 東北大学大学院 経済学研究科：〒 980–8576 宮城県仙台市青葉区川内；dgod1028@gmail.com
[2] 東北大学大学院 経済学研究科：〒 980–8576 宮城県仙台市青葉区川内；terui@tohoku.ac.jp

るが，これらは分析対象を特定カテゴリーに限定した低次元空間上での少数変数間のモデリングである．他方，アソシエーションルールによるマーケットバスケット分析などカテゴリーの枠を超えた高次元変数の分析も行われているが，マーケティング変数と市場構造の関係などマネジメントに必要なきめ細かい情報は抽出できない．また POS データは実務レベルで十分活用されているとは言えない状況にある．

本稿では，多くの実店舗が保有する集計 POS データに対して，機械学習や高次元データのモデリングなどの新しい統計分析による高度情報処理によって，多くの店舗で活用できるマーケティングの市場反応モデルを視野において，その活用の可能性を展望する．具体的には，(i)自然言語処理分野で提案されたトピックモデルの大規模集計 POS データへの適用による次元圧縮と市場細分化，(ii)大規模高次元スパースデータの市場反応モデルについて詳解して展望する．

2. 大規模集計 POS データのトピックモデルによる次元圧縮と市場細分化

2.1 トピックモデルのマーケティングへの展開

自然言語処理分野で潜在的話題—トピック—を抽出するために開発されたトピックモデル(Blei et al., 2003, 2012; Blei and McAuliffe, 2007)は，モデルの汎用性が高く拡張しやすい特徴をもつため，マーケティングにおいても広く使われるようになってきた．まず，元来の目的であるテキスト情報を直接活用するものとしては，ソーシャルメディアから収集したテキストデータから抽出したトピック(話題)をビジネスやマーケティングの問題に活用した研究がある．Si et al.（2013)は Twitter のテキストデータに対して，トピックモデルの一種の潜在ディリクレ配分(LDA: Latent Dirichlet Allocation)モデルを拡張したディリクレ過程混合 LDA モデルによりトピックを抽出し，それらの動きが株価予測に有効であることを示した．Wang et al.(2016)は，Amazon など商品に対するコメントのテキスト解析により，従来のトピックモデルのような全体的話題ではなく，スクリーン，バッテリーなど商品の特徴をターゲットにした話題を抽出するモデルを提案し，抽出された商品の潜在的特徴が売上に有効な情報をもつことを示した．また Morimoto and Kawasaki（2016)では，時系列テキスト分析であるダイナミックトピックモデル(Blei and Lafferty, 2006)を用いて，ファイナンス市場におけるボラティリティを予測する研究が行われている．

これらに対し，トピックモデルを購買や売上など数量データに適用した研究もある．ID-POS データと呼ばれる顧客ごとの非集計データの日々の記録をテキスト解析での一つの文章に対応させ，購買に関するトピックを抽出してマーケティングに活用する研究として，例えば，Ishigakiet al.(2011)がある．Christidis(2010)の研究では，ネット通販での顧客の購買履歴データをトピックモデルで分析し，潜在的マーケットバスケットの発見や個別顧客に対し商品を推薦できるリコメンデーションシステムが可能であることを示した．Iwata et al.（2009)は，提案したトピック・トラッキング・モデルを映画やアニメの購買データに適用し，顧客の趣味・嗜好を潜在トピックにより追跡できることを示した．さらに，Iwata and Sawada（2013)では，ID-POS データに付随する価格情報も考慮した LDA モデル分析を行っている．以上の研究は集計 POS データおよび非集計 ID-POS データを利用するものの，購買行動の背後にある潜在的トピックの情報により顧客を分類することに留まっている．したがって企業が駆使するマーケティング変数による最適化を可能とする制御モデルとは必ずしもなっていない．これに対し，Ishigaki et al.(2017)は，上記を拡張して，顧客の異質性をモデルに取りこみ，潜在購買トピックを割り出すと同時に顧客ごと/商品ごとの価格反応，プロモーション反応を個別に推定するモデルを展開している．

2.2 集計 POS 売上データのトピック分解による市場細分化

集計 POS データは，日々の SKU 単位の商品の売上とその価格およびその商品に対してプロモーションを行ったか否かに関する情報の記録であり，非集計の ID-POS データやレシートデータと異なり顧客の個別の購買状況が見えない．同じ商品の購買であっても購入目的や購買状況の違い，すなわち購買文脈によって，商品の評価やプロモーションへの反応などマーケティング戦略の効果は異なると考えるのが自然である．

Terui and Li (2017)では，集計データに埋没した購買文脈の異質性を潜在変数とし，トピックモデルにより分析に取り入れる．具体的には，集計データをその日に購買された商品の売上情報を用いて複数のショッピングバスケット(ここではトピック)に振り分けて市場細分化を行い，バスケットごとの市場反応を測定して実店舗のきめ細かいマーケティング戦略のための情報提供を可能とするものである．データは高次元でスパースな性質を持っており，これらのための統計モデルを提案している．

自然言語処理の手法として通常使われるトピックモデル(Blei et al., 2003; Blei, 2012)では，単語 v が文書 d に現れる確率は，潜在的トピック k の存在のもとでは，次式の有限混合モデルで表現できると仮定する．

$$p(v|d) = \sum_{k=1}^{K} p(v|k)\, p(k|d) = \sum_{k=1}^{K} \phi_{v|k}\theta_{k|d}. \tag{2.1}$$

Terui and Li (2017)では，店舗の集計売上数量データを同時購買情報のもとで各トピックに分類する．商品 j はテキスト分析における単語 v，日 t は文章 d に対応させる．t 日における商品 j の売上数を Y_{jt} とするとき，$\boldsymbol{Y}_t = (Y_{1t}, Y_{2t}, \ldots, Y_{n_t})'$ は同じ日のすべての商品 j の売上数量ベクトルである．この場合，t 時点で商品 j が購入される確率は $p(j|t) = \sum_{k=1}^{K} p(j|k)p(k|t) = \sum_{k=1}^{K} \phi_{j,k}\theta_{k,t}$ である．さらに，$\phi_{j,k}\theta_{k,t}$ に基づいて商品 j の売上数 Y_{jt} を各トピックごとに分解し，トピック k に入った売上数を $Y_{jt}^{(k)} = Y_{jt} \times E(\phi_{j,k}\theta_{k,t})$ で表す．そのとき t 日における商品 j の売上数は $Y_{jt} = \sum_{k=1}^{K} Y_{jt}^{(k)}$ と表現できる．トピックへの割当て確率に比例する $(\phi_{j,k}\theta_{k,t})$ は均一ではなく，少数のトピックに集中する場合が多いことから，トピックモデルにより次元圧縮が可能である．

2.3 実証分析(1)

実証分析として，Terui and Li (2017)での結果を紹介する．まずデータは，一店舗の 2002 年 5 月 6 日-2003 年 5 月 7 日間の日次集計 POS データを利用する．このデータセットは 363 日(1 年中 3 日休み)で，7912 個の商品種類の総計 3,720,419 回の購買記録が含まれている．POS データは各商品の毎日の売上数量と売上金額，また三種類のプロモーション(実施の有無のバイナリーデータ)の情報が含まれ，購買されていない商品のマーケティング変量の情報は含まれていない．

トピックの数 $k = 10$ と設定し，トピック分布と単語分布はハイパーパラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ によるディレクレー分布を事前分布とし，崩壊型ギッブス・サンプリングで推定を行う．

$$\theta_d \sim Dir(\alpha)\ (d = 1, \ldots, M); \phi_k \sim Dir(\beta)\ (k = 1, \ldots, K)$$

ここでは Griffiths and Steyvers (2004)の研究に従って，$\alpha = 50/k$，$\beta = 0.1$ と設定した(トピックの構成については紙面の都合上 Terui and Li (2017)を参照)．

後段で売上を価格や各種プロモーションなどマーケティング変数の関数として規定する市場反応モデルを定義することを念頭にして，いまターゲットとする商品として特定ブランドの牛乳(JANcode:4902705065161)を取り上げる．その売上数量 Y は 344 日で総計 21,482 個の売上

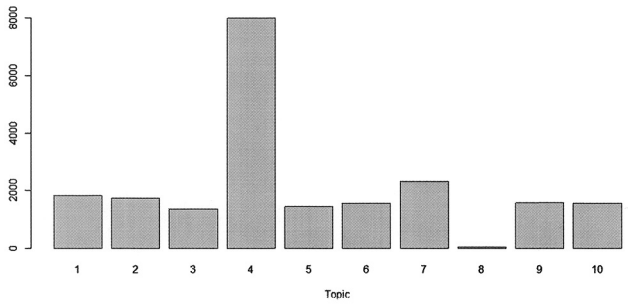


図 1．平均トピック分布．

数である．

図 1 は，7912 個の商品の集計 POS データにトピックモデルを適用したときの平均トピック分布

$$\hat{\phi}_{j,k}\hat{\theta}_{k|\bullet} = \frac{1}{N}\sum_{t=1}^{N}\hat{\phi}_{j,k}\hat{\theta}_{k|t}, k = 1, \ldots, K, \tag{2.2}$$

を表している．ここで，$N = 344$，$K = 10$ であり，一番頻度が高いトピック $k = 4$，と一番頻度が低いトピック $k = 8$ 以外は，ほぼ均等に配分されていることがわかる．図 2 は，各トピックの確率 $\hat{\phi}_{j,k}\hat{\theta}_{k|t}$ でターゲット商品である特定ブランドの牛乳の集計売上数を按分し，トピック分解した $Y_{jt}^{(k)}$ の時系列データである．

他と比べて極端に数量の少ないノイズ的意味をもつトピック 8 を除くと，各トピックは 2 − 4 か月ごとの季節性のトピックを表している．

3. 大規模集計 POS データの市場反応モデル

3.1 高次元スパースデータの統計モデル

高次元スパースデータに関する統計モデルはこれまで多くの研究がある．まず回帰の枠組みでは，Meinshausen and Yu (2009)が Tibshirani (1996)による LASSO 回帰の高次元スパースデータでの変数選択問題の理論的研究を行い，説明変数間に相関が強い場合には問題が生じることを示した．LASSO 回帰は変数間の相関が高い POS データでは有効とは言えない．Chen and Ishwaran (2012)は，バギング (Breiman, 1996)というアルゴリズムを用いる回帰木のランダムフォーレスト (Breiman, 2001)が高次元スパースデータにおいて，チューニングの難しさの問題も指摘しながらも高い予測精度を持つことを示した．ただし，変数間の構造の推定ができない限界も有している．

つぎに次元圧縮手法として，主成分分析，因子分析，正準相関分析を用いる方法がある．まず主成分分析による研究として，Zou et al. (2006)によるスパース主成分モデル，Tipping and Bishop (1999)による確率的主成分分析をスパースデータに拡張する Zeng et al. (2017)によるスパース確率的主成分がある．因子モデルを用いる Lopes and West (2004)は，ベイズ因子分析モデルが計算時間的に通常の因子モデルより優位性を持つため，高次元データに適していると主張した．また，West (2003)は因子分析による説明変数空間の次元圧縮によるスパースベイズ因子回帰モデルを提案し，高い予測精度を持つことを示したが，高次元空間での変数間の構造を推定する議論とはなっていない．

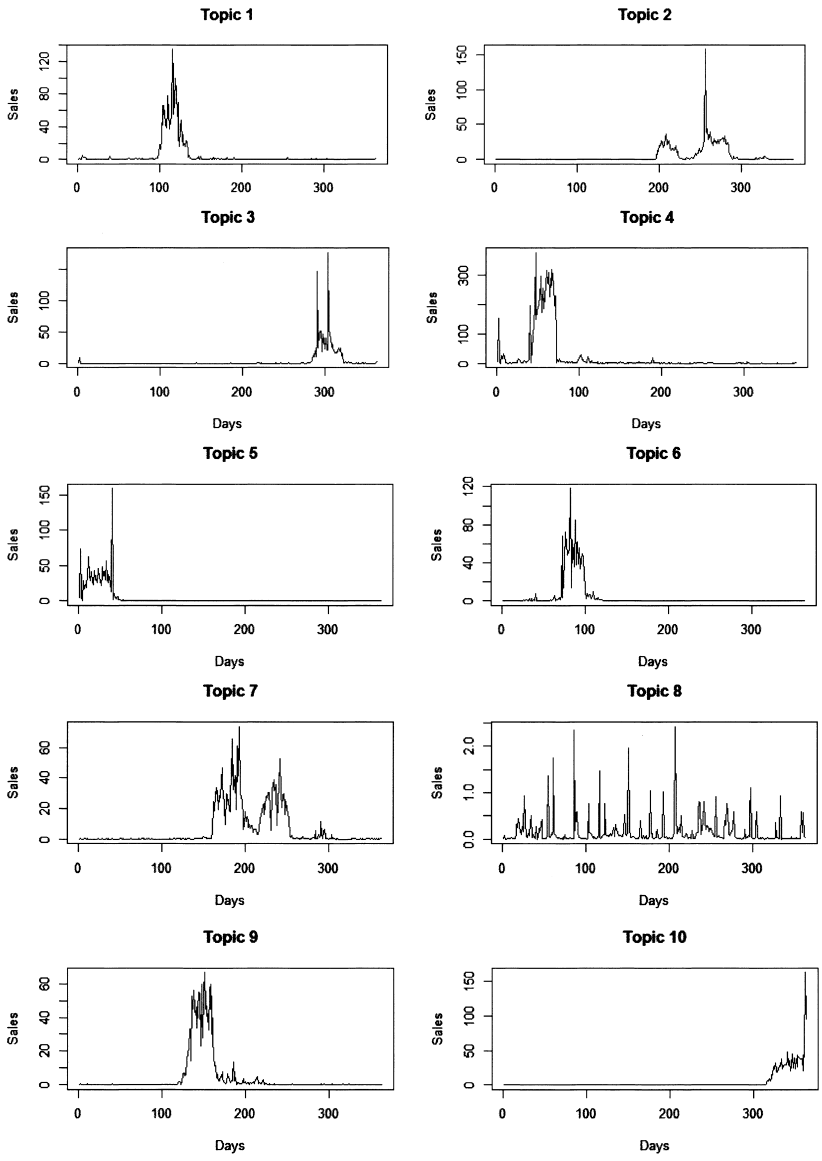


図 2. 売上のトピック分解.

多変数間の関係をモデル化する正準相関分析を用いる最近の研究として，Klami et al.（2013）によるベイズ正準相関分析があり，高次元データ解析での有効性を示した．また Brynjarsdottir and Berliner（2014）は，2 つの高次元時空間大規模データに対し，正準相関分析を基にして変数をそれぞれ圧縮させ，低次元変数間の関係を階層回帰モデルとして定義し，密度の濃い低次元空間で構造を利用して予測モデルを構築している．そのほか，Çiğdem（2009）は正準相関分析をマーケティング問題に応用と評価はしているものの，これらの研究は高次元空間での構造推定は行っていない．マーケティングにおいては，解釈不能な低次元での推定にもとづく予測

だけが目的とはなり得ず，日々のマーケティング意思決定のために，ある特定の商品の売上と価格の弾力性など元の高次元空間での変数間の関係の構造推定が必要となる．

3.2 高次元スパースデータに対する市場反応測定モデル

2 節で行われたショッピングバスケット(トピック)を前提にして，トピックごとの市場反応関数を測定する問題を考える．まずトピック k における商品 j の売り上げ $Y_{jt}^{(k)}$ を説明する市場反応関数を下記で定義する．

$$Y_{jt}^{(k)} = \alpha_{0j}^{(k)} + \boldsymbol{\alpha}'^{(k)}_{j} \boldsymbol{X}_{jt} + \sum_{m \neq j} \beta_m^{(k)} Y_{mt}^{(k)} + \sum_{m \neq j} \boldsymbol{\gamma}'^{(k)}_{m} \boldsymbol{X}_{mt}^{(k)} + \varepsilon_{jt}^{(k)}, \quad t = 1, \ldots, N \tag{3.1}$$

ここで $\boldsymbol{X}_{jt}$ は商品 j のマーケティング変数ベクトル，$Y_{mt}^{(k)}$ はトピック k に入っている目的商品以外($m \neq j$)の商品の売上数，$\boldsymbol{X}_{mt}^{(k)}$ はその商品のマーケティング変数である．ターゲットとする商品の予測は各トピックの予測値の合計 $\hat{Y}_{jt} = \hat{Y}_{jt}^{(1)} + \hat{Y}_{jt}^{(2)} + \cdots + \hat{Y}_{jt}^{(K)}$ で計算できる．

つぎに Terui and Li (2017)で提案した高次元スパースデータの市場反応構造推定のための回帰モデルを紹介する．(3.1)式の回帰モデルの説明変数を改めて $\boldsymbol{X}^{(k)}$ と定義し，同じトピックに入る商品の売上を目的変数 $\boldsymbol{Y}^{(k)}$ とし，各トピックにおける高次元の $\boldsymbol{Y}^{(k)}$ と $\boldsymbol{X}^{(k)}$ の回帰問題について考える．読みやすくするため，ここからはトピックの表記(k)を省略する．

まず P_y 次元の $\boldsymbol{Y}_t$ と P_x 次元の $\boldsymbol{X}_t$ の多変量回帰式を下記で定義する．

$$\boldsymbol{Y}_t = \boldsymbol{F}\boldsymbol{X}_t + \boldsymbol{e}_t, \tag{3.2}$$

ここで，係数行列 $\boldsymbol{F}$ は変数がデータの数を多く超えている高次元であり，いわゆる NP 問題のため直接に推定はできない．そこで，まず $\boldsymbol{Y}_t$ と $\boldsymbol{X}_t$ の周辺分布は，因子構造

$$\boldsymbol{Y}_t = \boldsymbol{U}\boldsymbol{a}_t + \boldsymbol{\eta}_{yt}; \boldsymbol{X}_t = \boldsymbol{V}\boldsymbol{b}_t + \boldsymbol{\eta}_{xt}, \quad t = 1, \ldots, N, \tag{3.3}$$

をもつと仮定する．ここで $\boldsymbol{a}_t$ は $f_y(< P_y)$ 次元のベクトル，$\boldsymbol{U}$ は $P_y \times f_y$ の行列，$\boldsymbol{b}_t$ は $f_x(< P_x)$ 次元のベクトル，$\boldsymbol{V}$ は $P_x \times f_x$ の行列，$\boldsymbol{\eta}_y \sim (0, \Sigma_y)$，$\boldsymbol{\eta}_x \sim (0, \Sigma_x)$ と仮定する．

さらに Brynjarsdottir and Berliner (2014)に従い，各因子モデルによる低次元の因子空間において関係があり，階層回帰モデル

$$\boldsymbol{a} = \boldsymbol{H}\boldsymbol{b} + \boldsymbol{\varepsilon} \tag{3.4}$$

で規定されると仮定する．ここで $\boldsymbol{a}$ および $\boldsymbol{b}$ は(3.3)の因子スコアベクトルをデータ $t = 1, \ldots, N$ について纏めた $f_y \times N$ および $f_x \times N$ の因子スコア行列，$\boldsymbol{H}$ は $f_y \times f_x$ の回帰係数行列，$\boldsymbol{\varepsilon}$ は $f_y \times N$ の誤差行列で各列は独立に $N(0, \sigma^2 I)$ に従うと仮定する．(3.2)–(3.4)を纏めて**階層因子回帰モデル(Hierarchical Factor Regression)**と呼ぶ．

いま(3.2)でのデータを纏めた $\boldsymbol{Y}$ と $\boldsymbol{X}$ の同時分布は，共通パラメータ $\boldsymbol{H}$ を条件付として独立であると仮定する．このとき，(3.2)，(3.3)のモデルの尤度関数は

$$p(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{U}, \boldsymbol{a}, \boldsymbol{V}, \boldsymbol{b}) = p(\boldsymbol{Y} | \boldsymbol{U}, \boldsymbol{a}, \boldsymbol{H})\, p(\boldsymbol{X} | \boldsymbol{V}, \boldsymbol{b}, \boldsymbol{H})\, p(\boldsymbol{H} | \boldsymbol{a}, \boldsymbol{b}). \tag{3.5}$$

と書かれる．

3.3 高次元空間への構造の回復

Brynjarsdottir and Berliner (2014)は，この圧縮次元空間を「結晶化空間」と名付けたが，高次元空間の構造 $\boldsymbol{F}$ を推定する提案は行われていない．$\boldsymbol{X}$ が与えられたとき，構造方程式(3.1)は誤差の条件付期待値 $E_{|x}[\boldsymbol{e}] = 0$ の仮定の下での $\boldsymbol{Y}$ の条件付期待値は

$$E_{|x}[\boldsymbol{Y}] = \boldsymbol{F}\boldsymbol{X} + E_{|x}[\boldsymbol{e}] = \boldsymbol{F}\boldsymbol{X} \tag{3.6}$$

であり，さらに $\boldsymbol{X}$ の確率測度に関して期待値をとって無条件期待値の関係として下記が得られる．

$$E_x\left\{E_{|x}[\boldsymbol{Y}]\right\} = \boldsymbol{F}E_x[\boldsymbol{X}] \quad \text{i.e. } \boldsymbol{\mu}_y = \boldsymbol{F}\boldsymbol{\mu}_x \tag{3.7}$$

他方，因子モデルによる圧縮次元空間上の変量の期待値をとれば，$E_x\{E_{|x}[\boldsymbol{Y}]\} = E[\boldsymbol{Y}] = \boldsymbol{U}\boldsymbol{a}_t$，$E_x[\boldsymbol{X}] = \boldsymbol{V}\boldsymbol{b}_t$ であり，(3.7)式から

$$\boldsymbol{U}\boldsymbol{a}_t = \boldsymbol{F}\boldsymbol{V}\boldsymbol{b}_t \tag{3.8}$$

が得られ，データを纏めた $\boldsymbol{a}$ と $\boldsymbol{b}$ の表記により $\boldsymbol{F}$ は次式により求められる．

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{a}\boldsymbol{b}'\boldsymbol{V}'\left(\boldsymbol{V}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{V}'\right)^{-1} \tag{3.9}$$

次にモデルの全パラメータの同時事後確率が下記のように表現される．

$$\begin{aligned} & p\left(\boldsymbol{U},\boldsymbol{a},\boldsymbol{V},\boldsymbol{b},\boldsymbol{H},\sigma^2,\boldsymbol{F},\Sigma_y,\Sigma_x,\Lambda_h|\boldsymbol{Y},\boldsymbol{X}\right) \\ & \quad \propto p\left(\boldsymbol{U},\boldsymbol{a}|\boldsymbol{H},\boldsymbol{Y}\right)p\left(\boldsymbol{V},\boldsymbol{b}|\boldsymbol{H},\boldsymbol{X}\right)p\left(\boldsymbol{H}|\boldsymbol{a},\boldsymbol{b},\sigma^2\right)p\left(\boldsymbol{F}|\boldsymbol{U},\boldsymbol{a},\boldsymbol{V},\boldsymbol{b}\right) \\ & \qquad \times p\left(\sigma^2|\boldsymbol{a},\boldsymbol{b},\boldsymbol{H}\right)p\left(\boldsymbol{U}|\boldsymbol{a},\Sigma_y\right)p\left(\Sigma_y\right)p\left(\boldsymbol{V}|\boldsymbol{b},\Sigma_x\right)p\left(\Sigma_x\right) \\ & \qquad \times p\left(\boldsymbol{a}|\Lambda_a\right)p\left(\Lambda_a\right)p\left(\boldsymbol{b}|\Lambda_b\right)p\left(\Lambda_b\right)p\left(\boldsymbol{H}|\Lambda_h\right)p\left(\Lambda_h\right)p\left(\sigma^2\right) \end{aligned} \tag{3.10}$$

これは因子モデルと多変量回帰モデルの事後分布の組み合わせであり，それぞれ共役な事前分布の利用のもとで，ギッブスサンプリングにより効率的に分布評価が可能である．さらに高次元空間での市場反応係数行列 $\boldsymbol{F}$ の復元は，(3.9)式の関係を利用して MCMC 過程での副産物として，周辺事後分布 $p(\boldsymbol{F}|\boldsymbol{Y},\boldsymbol{X})$ が評価できる．MCMC アルゴリズムについては Terui and Li (2017)に詳細が記載されている．

3.4 実証分析(2)

2 節と同じデータを用いた実証分析の結果を紹介する．

表 1 には，式(3.9)の関係を利用して回復した構造 $\boldsymbol{F}^{(k)}$ の事後分布を評価し，95%HPD 領域の意味で有意な回帰係数推定値(事後平均)と説明変数名の一部が記載されている(紙面の都合上トピック 1–5 までを記載した．全トピックについては Terui and Li (2017)を参照)．回帰係数は弾力性を意味し，商品 ID，カテゴリーの名前が続いて表記されている．これにより，下記のようなマネジメントに有用な知見が得られる．

(i) 各トピックにおける同時購買の説明変数は有意となるケースが多い結果になった．トピックモデルにより，同時購買された商品で回帰モデルを構築したためと考えられる．これらの情報を基に同時購買しやすい商品の組み合わせの発見が可能になり，適切なプロモーションにより特定商品の売上の向上が期待できる．

(ii)「価格」変数はトピック 6 および 10 で多く有意になっている．目的変数は他の商品の価格の影響を多く受け，商品自身の価格の影響は著しくない．

(iii) トピック 5，6，7，9，と 10 のプロモーションの影響は強い．

(iv) トピック 9 では，多くの食料品が有意な説明変数として抽出されている．このトピックにおける目的変数の牛乳は飲み物ではなく，調理品として買われていると推測できる．

(v) 売上数が一番多いトピック 4 では，水やジュースカテゴリーの商品は目的変数の牛乳と同期している反面，クッキーカテゴリーの商品とは逆の関係性を持っている．これは 6

表 1. トピックごとの説明変数の抽出.

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
1	[4891]miso	2.318	[3064]rice cracker	0.079	[3064]rice cracker	-0.019	[2092]tomato juice	-0.089	-	-
	[4929]coffee drink	2.184	[3523]cup noodle	0.030	[1639]cup noodle	0.010	[235]yogurt	-0.043	-	-
	[5025]rice cracker	2.141	[7211]barbecue sauce	-0.001	-	-	-	-	-	-
	[5577]suger	2.093	-	-	-	-	-	-	-	-
	[3249]energy drink	1.946	-	-	-	-	-	-	-	-
	[7034]coffee drink	1.942	-	-	-	-	-	-	-	-
	[1927]family ice	-1.885	-	-	-	-	-	-	-	-
	[3287]fruit juice	1.764	-	-	-	-	-	-	-	-
	[494]sauce	1.709	-	-	-	-	-	-	-	-
	[7384]premium ice	1.649	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
2	[5790]biscuit and cookie	3.000	[6012]cheese	-0.014	[2245]snack	0.154	[5550]cup noodle	-0.260	-	-
	[4919]family ice (cream)	1.948	-	-	[2553]chocolat	-0.090	[3128]cup noodle	-0.123	-	-
	[3846]cup noodle	1.764	-	-	-	-	[2112]instant soup	0.108	-	-
	[1345]chocolat	1.764	-	-	-	-	-	-	-	-
	[2727]soy	1.680	-	-	-	-	-	-	-	-
	[1826]dressing	1.674	-	-	-	-	-	-	-	-
	[89]coffee drink	1.532	-	-	-	-	-	-	-	-
	[3587]personal ice	1.462	-	-	-	-	-	-	-	-
	[6860]instant coffee	-1.460	-	-	-	-	-	-	-	-
	[1413]biscuit and cookie	-1.389	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
3	[1287]cooked curry	-2.088	-	-	[3225]salad oil	0.486	[2031]dressing	-0.178	-	-
	[4825]lactic acid bacteria b	1.913	-	-	-	-	[2030]dressing	0.145	-	-
	[1138]mayonnaise	-1.607	-	-	-	-	-	-	-	-
	[494]sauce	1.516	-	-	-	-	-	-	-	-
	[6402]chocolat	1.379	-	-	-	-	-	-	-	-
	[3116]candy	1.352	-	-	-	-	-	-	-	-
	[469]dessert	-1.316	-	-	-	-	-	-	-	-
	[166]personal ice (cream)	1.294	-	-	-	-	-	-	-	-
	[3559]cheese	1.272	-	-	-	-	-	-	-	-
	[883]carbonated flavor	1.128	-	-	-	-	-	-	-	-

表 1.（つづき）

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
4	[258]water	3.963	[3182]cup noodle	0.165	-	-	[869]carbonated flavor	-0.049	-	-
	[5304]biscuit and cookie	-3.253	-	-	-	-	-	-	-	-
	[4309]instant curry	3.042	-	-	-	-	-	-	-	-
	[3731]carbonated flavor	3.008	-	-	-	-	-	-	-	-
	[4792]family ice (cream)	2.941	-	-	-	-	-	-	-	-
	[5184]chocolat	-2.626	-	-	-	-	-	-	-	-
	[2334]snack	-2.54	-	-	-	-	-	-	-	-
	[4094]fruit juice	2.431	-	-	-	-	-	-	-	-
	[6122]chocolat	2.328	-	-	-	-	-	-	-	-
	[6832]cola	2.314	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
5	[7897]instant noodle	2.174	[5046]dessert	0.036	[4710]biscuit and cookie	0.451	[5255]chocolat	-0.328	[2227]rice cracker	-0.108
	[6815]black tea drink	-1.998	[4983]health food	-0.015	[3419]snack	0.276	[5916]family ice	-0.047	-	-
	[3020]dressing	1.781	[2297]snack	0.000	166]Principle of Chinese foo	-0.187	[3015]mayonnaise	-0.020	-	-
	[4297]cooked curry	1.584	-	-	-	-	[6451]100% fruit juice	-0.006	-	-
	[6735]miso	1.538	-	-	-	-	-	-	-	-
	[6862]	1.528	-	-	-	-	-	-	-	-
	[7717]energy drink	1.528	-	-	-	-	-	-	-	-
	[2486]candy	-1.443	-	-	-	-	-	-	-	-
	[505]candy	-1.413	-	-	-	-	-	-	-	-
	[6211]candy	1.410	-	-	-	-	-	-	-	-

月から 8 月までの夏の季節性の影響であると推測できる．

紹介したモデル分析では，実務や Hanssens et al.（2001）などで行われてきたカテゴリーに限定した小変数回帰では得られず，全商品のデータを使うことによってはじめて得られる意外な商品の組み合わせの発見がある．1970 年代の人工知能ブームにおいて注目された「紙おむつとビール」の同時購買のデータマイニングによる発見に類似した知見である．他方，これに加えて変数間の関係性の構造が弾力性の形で評価され，店舗マネジメントに役立つマーケティング戦略への情報を提供できる利点を持っている．

4. おわりに

マーケティングにおいては，主体（消費者）の異質性がいち早く重視され，個別対応のための「個」のモデリングが展開され実用化してきた．本研究詳解では，「買い物状況（トピック）」の異質性をトピックモデルで表現し，買い物状況ごとに異なる購買要因の存在とその関係の構造を仮定し，集計 POS データを用いてきめ細かいマーケティングを実現するモデルを提案する研究を紹介した．全商品データを分析に取り入れることで，目的変数を説明する意外な商品の発見のみならずその量的関係が弾力性の形で測定可能となり，実店舗のきめ細かいマーケティング戦略に有用な情報が提供可能である．

本稿ではトピック数の選択問題は取り上げていないが，無限ディリクレ過程を利用したノンパラメトリックベイズの適用によるトピック数の推定が可能であり，階層因子回帰モデルにおける因子数の推定も同様である．また本研究では対象企業を広範囲とするため，小企業でも保有している集計 POS データの活用を念頭に置いた．対象企業の範囲は狭まるが，レシートデータが扱える状況を想定すれば，1 回の買い物トリップでの同時購買情報を利用して購入者の異質性を反映させたモデル分析が可能である．さらに個人が特定されるメンバーシップ顧客データベースを対象とすれば，さらに消費者異質性と購買状況の異質性の相互作用の分析も可能であろう．これらは今後の研究課題として有望であろう．

謝　辞

JSPS 科研費 Grant Number（A）25245054 の助成を受けた．

参 考 文 献

Blei, D. M. (2012). Introduction to probabilistic topic models, *Communications of the ACM*, **55**, 77–84.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models, *Proceedings of the ICML*, **6**, 113–120.

Blei, D. M. and McAuliffe, J. (2007). Supervised topic models, *Neural Information Processing Systems*, **3**, 993–1022.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**(2), 123–140.

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

Brynjarsdottir, J. and Berliner, L. K. (2014). Dimension-reduced modeling of spatio-temporal process, *Journal of American Statistical Association*, **109**, 1647–1659.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis, *Genomics*, **99**(6), 323–329.

Christidis, K., Apostolou, D. and Mentzas, G. (2010). Exploring customer preferences with probabilistic topics models, *In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2010*.

Çiğdem Şahin, B. (2009). An evaluation and an application of using canonical correlation analysis in marketing research, *International Journal of Economic and Administrative Studies*, **1**(3), 41–68.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences*, **101**, 5228–5235.

Hanssens, D. M., Parsons, L. J. and Shultz, R. L. (2001). Market response models, *Econometric and Time Series Analysis*, 2nd ed., Kluwer Academic Press Inc., Boston, MA.

Ishigaki, T., Takenaka, T. and Motomura, Y. (2011). Customer behavior prediction system by large scale data fusion in a retail service, 人工知能学会論文誌, **26**(6), 670–681.

Ishigaki, T., Terui, N., Sato, T. and Allenby, G. (2017). Personalized market response analysis for a wide variety of products from sparse transaction data, *International Journal of Data Science and Analytics*, **5**(4), 233–248.

Iwata, T. and Sawada, H. (2013). Topic model for analyzing purchase data with price information, *Data Mining and Knowledge Discovery*, **26**, 559–573.

Iwata, T., Watanabe, S., Yamada, T. and Ueda, N. (2009). Topic tracking model for analyzing consumer purchase behavior, *International Joint Conference on Artificial Intelligence '09*, 1427–1432.

Klami, A., Virtanen, S. and Kaski, S. (2013). Bayesian canonical correlation analysis, *Journal of Machine Learning Research*, **14**, 965–1003.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis, *Statistica Sinica*, **14**(1), 41–67.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data, *The Annalsof Statistics*, **37**(1), 246–270.

Morimoto, T. and Kawasaki, Y. (2016). Forecasting financial market volatility using a dynamic topic model, *Asia-Pacific Financial Markets*, **24**(3), 149–167.

Rossi, P. E., Allenby, G. and McCulloch, R. (1996). The value of purchase history data in target marketing, *Marketing Science*, **15**(4), 321–340.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction, *Publisher Association for Computational Linguistics*, **2**, 24–29.

照井伸彦 (2008).『ベイズモデリングによるマーケティング分析』, 東京電機大学出版局, 東京.

Terui, N. and Dahana, W. D. (2006). Price customization using price thresholds estimated from scanner panel data, *Journal of Interactive Marketing*, **20**(3), 58–70.

Terui, N. and Li, Y. (2017). Measuring large scale market responses from aggregated sales regression model for high dimensional sparse data, Discussion Paper of DSSR No.66, Graduate School of Economics and Management, Tohoku University, Sendai.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, **58**(1), 267–288.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis, *Journal of the Royal Statistical Society, Series B*, **21**(3), 611–622.

Wang, H., Chen, Z., Fei, G., Liu, B. and Emery, S. (2016). Targeted topic modeling for focused analysis, *Association for Computing Machinery*, **13**(17), 1235–1244.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm, *Bayesian Statistics*, **7**, 723–732.

Zeng, J., Liu, K., Huang, W. and Liang, J. (2017). Sparse probabilistic principal component analysis model for plant-wide process monitoring, *Korean Journal of Chemical Engineering*, **34**(8),

2135–2146.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.

High-dimensional Sparse Modeling of Large-scale Aggregated POS Data

Yinxing Li and Nobuhiko Terui

Graduate School of Economics and Management, Tohoku University

Micro-marketing based on consumer heterogeneity using disaggregated ID-POS data has been well studied in the literature, but the implementation and operation of this approach remain limited, particularly in real store management. On the other hand, while aggregated POS data are collected by most retailers, it is widely recognized that these data have never been well utilized.

We discuss the possibility of using aggregated POS data by applying new techniques in high-dimensional sparse data modeling and machine learning. In particular, we propose a procedure comprising two sub-models: the topic model first decomposes the aggregate number of sales to several different market baskets, and then hierarchical factor regression is used to reduce dimensionality and ultimately recover from the reduced dimension to the original space in order to detect the marginal effect among all products in each market basket.

The proposed model, which uses a large amount of product data, not only makes it possible to discover unexpected predictors, but also measures the quantitative relation in the form of elasticity for managerial implications.

Key words: Aggregated POS data, heterogeneity in shopping context, topic model, high-dimensional sparse-data, hierarchical factor regression.

統計数理(2018)
第 66 巻 第 2 号 249–265

特集「サービス科学の今」
［研究ノート］

統計モデルによる消費者理解の可能性

佐藤 忠彦†

（受付 2017 年 10 月 30 日；改訂 2018 年 3 月 11 日；採択 4 月 6 日）

要　　旨

統計モデルの高度活用は，サービス科学を発展させるための必須のツールであり，今後さらに重要度の増すアプローチである．本稿は，サービスを高度化する際に必須の「消費者理解を深めるには」という観点で統計モデル(特にベイズモデル)をどのように活用すべきか？に対する提言を目的とし，周辺研究を整理し，既存研究の紹介を基本として整理した．事例で紹介した 2 つの研究は，マーケティングを題材としたものであるが，サービス研究に対しても十分な示唆を供するものである．

キーワード：ベイジアンモデリング，消費者異質性，時間的異質性，潜在変数．

1. はじめに

本論は，統計モデルを消費者理解にどのように使いうるかを議論することが主たる目的である．2 つの既発表の事例のエッセンスを紹介し，統計モデルが消費者理解に活用できる可能性を理解してもらいたいと考えている．

現代の需要と供給の関係において，「供給は需要を規定できるのか？」という問いを考えなければならない．この問いは，供給者サイドと需要サイドが出会う場である市場において，供給者サイドと需要サイドのどちらに主導権があるのかという問いに読み替えてもらってもよい．市場がどんどん拡大していく右肩上がりの経済の時代，消費者は人と同じものを持つ，人と同じサービスを享受することでその心は満たされていた．しかしながら，現代の消費者は人と少しでも違うものをもち，他人とは別のサービスを受けることに喜びを感じる．サービスの市場における主たるプレイヤーである消費者のニーズは，細分化，多様化という形式で変化してきている．これらの事実を踏まえると，サービスの市場は「需要が供給を規定する市場」だという認識を持たなければならない．「需要サイドが考える “良いもの” を作れば売れるはず」や「需要サイドが考える “良いサービス” を提供すれば使ってもらえる」といった川下発想に基づく戦略立案が求められているのである．

前段の川下発想に基づく戦略立案のキーは，「消費者を理解すること」である．問題は，「消費者を理解する」の含意は何か？である．「消費者を理解する」ことは，消費者の行動の真のメカニズムを解明することと即断してよいものであろうか？ 結論を述べれば，消費者の真の行動メカニズムを解明することは容易ではなく，基本的に不可能である．

企業は，最終的に消費者の真の行動メカニズムの解明を求めているわけではない．もちろん，真の行動メカニズムを解明できるのであればそうしたいが，前段に示したようにそれは基本的に無理であり，その代わりにマーケティング意思決定をする際に最も必要な，時を得た情

† 筑波大学 ビジネスサイエンス系：〒 112–0012 東京都文京区大塚 3–29–1

(1)「消費者を理解する」=「マーケティングを高度化しうる仮説的役割を担う情報を発見する」
(2)「消費者を理解する」≠「消費者の真の行動メカニズムを明らかにする」

図 1.「消費者を理解する」の構図.

報を，データ(行動の結果データやアンケートによる態度データ)から抽出したいと考えているのである．換言すれば，企業がデータに基づき行っている「消費者理解のための活動」は，真の構造を解明することを狙っているわけではなく，消費者行動の仮説的役割を担う高次情報の抽出を狙いとしている．以上の検討に基づけば，「消費者を理解する」の含意は何か？に対する答えは，図 1 のように整理できる．

前段までの議論に基づくと，消費者の真の行動メカニズムを明らかにすることはできないし，そもそもそれを考えることは無意味だといっているように思われてしまうかもしれない．しかし，それは誤解である．消費者の真の行動メカニズムが解明できれば，それにこしたことはないし，当然，可能であるならばそれを実現したい．しかし現実的には，消費者の真の行動メカニズムを解明するのは容易ではないし，実際には何が真の構造なのかさえ分からないため評価もできない．そういった状況でも，「消費者の理解」を深める活動を行わない限り，市場に存在する様々なサービスを高度化できない．

前段までに示した消費者理解をさせるためのキーとなる技術が，ベイジアンモデリングである．以降には，ベイジアンモデリングの概要を示す．なお，$\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\psi}$ は，観測データ，パラメータ，超パラメータをそれぞれ示すものとし，議論を進める．従来の統計モデリングでは少ないパラメータで表現されるモデルが良いモデルとされてきた．所謂，「けちの原理」である．どのようなデータであったとしても平均と分散のみで規定される正規分布から得られたと仮定するようなものである．一方，パラメータ数を増やせば統計モデルの記述能力は向上するが，汎化能力と呼ばれる将来のデータの予測能力が低下する．この問題への対策として，パラメータ $\boldsymbol{x}_t$ についても統計モデル $p(\boldsymbol{x}_t|\boldsymbol{\psi})$ を想定するのがベイズモデルである．$p(\boldsymbol{x}_t|\boldsymbol{\psi})$ をベイズ統計では事前分布と呼ぶ．この事前分布を導入し，ベイズの定理

$$p(\boldsymbol{x}_t, \boldsymbol{\psi}|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t, \boldsymbol{\psi})}{p(\boldsymbol{y}_t)} \propto \underbrace{p(\boldsymbol{y}_t|\boldsymbol{x}_t)}_{\text{尤度関数}} \quad \underbrace{p(\boldsymbol{x}_t|\boldsymbol{\psi})}_{\boldsymbol{x}_t\text{の事前分布}} \quad \underbrace{p(\boldsymbol{\psi})}_{\boldsymbol{\psi}\text{の事前分布}} \tag{1.1}$$

を用いることで，想定した事前分布 $p(\boldsymbol{x}_t, \boldsymbol{\psi})$ がどのように修正されるのか，つまりパラメータに関する不確実性がデータによりどの程度修正されたのかを観察するのである．ここで $p(\boldsymbol{x}_t, \boldsymbol{\psi}|\boldsymbol{y}_t)$ を事後分布と呼び，我々の興味はこの事後分布に集約される．なお，データ $\boldsymbol{y}_t$ の発生確率 $p(\boldsymbol{y}_t)$ は $\boldsymbol{x}_t, \boldsymbol{\psi}$ によらない数値になるので，事後分布 $p(\boldsymbol{x}_t, \boldsymbol{\psi}|\boldsymbol{y}_t)$ は(1.1)式右辺 2 行目に比例する．この仕組みにより，多数のパラメータも安定して推定できるようになり，結果として高い予測能力とデータ記述能力を同時に持つ総合的な統計モデルを構成できる．この一連のモデル化の行為を，通常，ベイジアンモデリングと呼ぶ．(1.1)式を用いる際に t が時点を示すならば $\boldsymbol{x}_t$ は時変パラメータを，個人を示すとすれば個人毎のパラメータとなる．

消費者を理解するためには，「消費者異質性」，「時間的異質性」および「潜在変数」といった観点を意識しなければならない．以降には，それらが何を意味するのかを概説する．

初めに，「消費者異質性」と呼ぶ概念を説明する．現代のマーケティング活動では，マイクロ・マーケティングと呼ぶ「個」に焦点を当てた活動が脚光を浴びている．この活動の前提は，消費者一人一人は異質だという認識であり，一人一人の違いにこそマーケティングを高度化するための情報が含まれており，無視すべきではないと考える点である．消費者は，同一日に同一商品を同一の場所で購買したとしても，人が違えば購買に至るメカニズムに差が生じ，その反応も異なる．t が消費者を表すとすれば，(1.1)式の $\boldsymbol{x}_t$ が消費者異質性の直接的なモデル表現となる．消費者の行動を規定する説明変数として「商品の価格」を考え，$\boldsymbol{x}$ がその反応を示すと考えると，消費者の異質性の仮定の下では，$t \neq t'$ ならば $\boldsymbol{x}_t \neq \boldsymbol{x}_{t'}$ となる．実際には，推定された個人ごとのパラメータ値に大きな差があれば異質性が強いと，逆の場合は異質性は小さいと判断する．

次に，「時間的異質性」と呼ぶ概念を説明する．時間的異質性は，現代のマーケティングにおいて上述の消費者異質性と並び重要な概念である．時間的異質性の仮定の下では，同一の個人であったとしても時点が違えば，その反応には差が生じると考える．t が時点を表すとすれば，(1.1)式の $\boldsymbol{x}_t$ が時間的異質性の直接的なモデル表現となる．消費者の行動を規定する説明変数として「商品の価格」を考え，$\boldsymbol{x}$ がその反応を示すと考えると，時間的異質性の仮定の下では，$t \neq t'$ ならば $\boldsymbol{x}_t \neq \boldsymbol{x}_{t'}$ となる．実際には，推定された時点ごとのパラメータ値に大きな差があれば時間的異質性が強いと，逆の場合は異質性は小さいと判断する．

3つ目として，今日的な視点でマーケティングデータの解析をする際に重要な役割を担う「潜在変数」を説明する．現在，企業には消費者の行動の結果を示すビッグデータが多量に蓄積されている．そのためややもすると，「ビッグデータに含まれる観測変数間の関係性を明らかにすれば消費者の行動は解明できる」といった考えが出現している．しかし，消費者の行動は顕在変数間の関係性の抽出のみで解明できるものではなく，「消費者の行動に至る要因は全てデータとして観測できるわけではない」という認識が不可欠である．拠り所をビッグデータにおくだけではなく，消費者行動理論から経験や勘に至るまで，様々な情報を上手に使って，前述の意味での「消費者の理解」を実現しなければならない．問題は，消費者の行動に影響する潜在的な変数は何か？ということである．特定の商品の購買を想起した場合は，「家庭内在庫量」，「ブランドロイヤルティ」，「参照価格」などが行動に影響する潜在変数になる．このように具体的行動を想定すればそれに影響しそうな潜在変数の候補は様々想定できる．もちろん，ここで例示した変数以外にも商品購買に影響する可能性を有する潜在変数は数多くある．しかし，行動に影響する潜在変数は，対象とする消費者の行動が何かによって変化するのは明らかである．実際にモデル化する場合は具体的に示すことになるが，一般的に「消費者の嗜好」，「消費者の好き・嫌いや興味がある・ないといった態度」，「消費者が直面する状況」，「消費者のこれまでの経験」と「消費者の将来を予測する能力」などが消費者の行動に影響する潜在変数の抽象的な説明になる．消費者異質性や時間的異質性に加えて，潜在変数(潜在構造)も考慮し消費者の行動を評価できて初めて，実フィールドで高度に活用可能な情報を手にしうる．これまでマーケティング分野では，本項で説明したような個人ごと，あるいは時刻ごとのような細かい潜在変数を意識することは少なかった．これは，モデルの複雑化に対応可能なデータがなかったのが主たる理由である．しかし，ビッグデータの蓄積が進んだ現在，行動の背後に存在する潜在変数は積極的に考慮していかなければならない．もう少し言えば，潜在変数を考慮しビッグデータの活用を進めない限り，「消費者の理解」は進まないのである．

本節の最後に本特集号のテーマであるサービス研究と以降の事例で取り上げるマーケティング研究の共通点と相違点を提示しておくことにする．サービス研究は，個々のサービスを総合的にデザインするうえで必要な知識と思考の提供を目指している．日本学術会議経営学委員会・総合工学委員会合同サービス学分科会 (2017)によると，サービス研究には次の5つの領域

が存在する．一つ目は，「サービスマーケティング」であり，共創される価値の解明とその価値のコミュニケーションとデリバリーをデザインする．二つ目は，「サービスオペレーション」であり，価値を具現化し，サービスの仕様・意匠・設計から提供方法を企画・運営する．三つ目は，「サービスマネジメント」であり，サービスを事業体として持続的かつ効果的・効率的に運営するための資源配分と組織内外の体制を設計する．これら 3 つの領域は，経営学や工学等の領域で個別に教育．研究されてきた．近年，統計を含む情報学の技術を用いて効率的に観察・分析・適用を行う「サービスサイエンス」も重要な領域となっており，これが四つ目の領域となる．最後の五つ目は，ICT や IoT の進展に伴い，個々のサービスを有機的に連携させることで，社会全体により高水準で複合的な価値の提供を目指す「サービスエコシステム」である．それらのサービスの領域では，サービスの提供サイドと需要サイド間とで価値を共創する関係をサービスの一単位(医療サービス，小売りサービス，WEB サービス等)と考え，それぞれサービス全体を対象として研究されることが多い．いずれのサービスでもその高度化は，提供者と需要者間で価値共創を促進することで実現できると考えられている．しかし，サービス提供者が需要サイドの態度，行動を的確に把握できなければ，そもそも共創を促進することはできず，結果需要サイドに支持されるサービスは構築できない．マーケティング研究は，サービス研究のようにある特定サービス全体を対象とすることは少ないが，サービスにおける上述の「提供者による需要サイドの態度，行動の把握」は，やはり最も重要な課題であり，ほとんどのマーケティング研究のテーマとなっている．その意味で，本稿のマーケティング研究の事例は，部分的であるにしろ，サービスの根幹部分に対する重要な知見を供するアプローチとして活用できる．重要な点なので言及しておく．

本稿の構成は以下のとおりである．2 節には，当該分野における(特に)ベイジアンモデリングによる研究事例を，3 節には，本稿の目的に合致する既発表の研究事例のエッセンスをそれぞれ示す．4 節は，まとめと課題である．

2. 先行研究

本節では，基本的にマーケティング研究を対象とし 1 節に示した「消費者異質性」，「時間的異質性」および「潜在変数」の観点を含む研究を概観する．

2.1 消費者異質性を含む研究

消費者異質性は，マーケティングにおける重要概念であり，当該概念を取り込んだ研究の数は膨大である．Allenby and Rossi (1998)が指摘したように，マーケティングにおける意思決定において，消費者選好の分布はその中心的な役割を果たす．現代のマーケティング研究では，消費者異質性のモデリングが最重要の課題なのである．隣接分野である計量経済学分野では，異質性を局外母数として捉え，積分消去することにより興味の対象外として取り扱うこととは対照的である．Allenby and Rossi (1998)と同一著者により整理された Rossi and Allenby (2003)は，マーケティング・サイエンス領域における階層ベイズモデルの有用性を示した先駆け的な論文である．その後マーケティング分野で階層ベイズモデルを用いた研究が増大することになる．マーケティング分野における最重要モデルの 1 つである離散選択モデルを用いて，消費者異質性を取り扱った研究として，Rossi and Allenby (2003), Terui and Dahana (2006), Terui and Ban (2008), 井上 (2010), Terui et al. (2011), 山田・佐藤 (2012), 日高・佐藤 (2016)等がある．また，ポアソン回帰や負の二項回帰などを含む回帰モデルによって消費者の異質性を取り扱ったものとしては，Manchanda, and Chintagunta (2004), Manchanda et al. (2004), 宮津・佐藤 (2015), 山田・佐藤 (2016)等がある．これらの研究は，全てマルコフ連鎖モンテカ

ルロ法(MCMC 法)により推定されている．その技術的詳細に関しては，Chib and Greenberg (1996), Geman, S. and Geman, D. (1984), Hastings (1970)等を参照してほしい．その他，階層ベイズを用いた消費者異質性モデリングの方法論や応用事例に関しては，Rossi et al. (2005), 照井 (2008), 佐藤・樋口 (2013)に整理されている．

2.2 時間的異質性を含む研究

時間的異質性を取り扱ったマーケティング研究は，2.1 項に示した消費者異質性を取り扱った研究に比べて極端に少ない．その理由は二つある．一つ目は，マーケティング研究者の動的消費者行動に関する興味の低さであり，二つ目はそもそも動的消費者行動に活用可能な時系列データが少なかったことである．しかしながら，近年の IT 技術の革新に伴い粒度の細かい時系列データの蓄積が進み，消費者の動的行動特性の解明にも脚光が当たるようになった．マーケティング分野でなされる時間的異質性を取り扱う研究は，基本的に当該分野で市場反応分析と呼ばれる範疇にカテゴライズされるものが多い．Naik et al. (1998), Kondo and Kitagawa (2000), 山口 他 (2004), 佐藤・樋口 (2008a), Bass et al. (2007) といった研究は，目的変数に売上等の集計量を採用する集計型の動的市場反応を取り扱ったものである．これらのモデルは，基本的に線形・ガウス型の状態空間モデルの枠組みでモデル表現されており，その推定にはカルマンフィルタと最尤法が採用されている．一方で Bruce (2008), 佐藤・樋口 (2008b, 2009), 本橋・樋口 (2013), 本橋 他 (2012)は，目的変数に来店やブランド選択等を採用する非集計型の動的市場反応を取り扱った研究である．観測モデルが離散選択型の非ガウス型モデルになるため，線形・ガウス型の状態空間モデルでは表現できず，一般状態空間モデルの枠組みでモデル化され，モデルの推定は，Bruce (2008)では MCMC +粒子フィルタ，佐藤・樋口 (2008b, 2009), 本橋・樋口 (2013), 本橋 他 (2012)では粒子フィルタと最尤法が用いられている．集計型の動的市場反応モデルは，その推定が容易にできるという利点もあり，当該コミュニティでも研究が増えつつあるが，非集計型の動的市場反応モデルは，モデル推定に様々な困難が生じるため，まだ発展途上である．しかしながら，いずれの枠組みであったとしても時間的異質性を捉えようとするマーケティング研究は，その仮定の自然さから今後研究が進められなければならない領域となっている．佐藤・樋口 (2013)には，マーケティングにおける状態空間モデルの理論と応用事例が解説されている．

線形・ガウス型・状態空間モデルの具体的なモデリング方法やその周辺技術については，北川 (2005)に詳しい．非線形性や非ガウス分布を内包可能な一般状態空間モデルについても近年研究が進み，逐次モンテカルロ法の一種である粒子フィルタのアルゴリズムが与えられている．一般状態空間モデルと粒子フィルタに関しては，Doucet et al. (2001), 北川 (2005), 樋口 (2011)などに詳しく解説されている．

2.3 潜在変数を含む研究

マーケティングでは，POS データや ID 付 POS データ，WEB の行動履歴等大規模かつ多次元のデータが蓄積されるようになった．これらビッグデータは大規模であったとしても消費者の行動の一端を測定しているに過ぎないといった認識をもたなければならない．消費者の行動は観測されている変数のみでは説明できず，その背後に存在する潜在的な変数や潜在的な構造までをも含めてモデル化，評価しなければならないのである．

Terui and Dahana (2006)(参照価格，価格閾値), Terui and Ban (2008)(広告ストック), Terui et al. (2011)(広告ストック，考慮集合), 井上 (2010)(ディテールストック), 佐藤・樋口 (2008a)(参照価格), 佐藤・樋口 (2009)(家庭内在庫量，消費量), 山田・佐藤 (2012)(パフォーマンス評価／期待), 宮津・佐藤 (2015)(心理的財布), 青柳・佐藤 (2015)(参照価格，広告ストック)等

表 1. 周辺研究の概観.

文献	枠組み	消費者異質性	時間的異質性	潜在変数
Allenby and Rossi (1998)	階層ベイズモデル	○		
青柳・佐藤 (2015)	状態空間モデル		○	○
Bass et al. (2010)	状態空間モデル		○	
Bruce (2008)	状態空間モデル	○	○	
井上 (2010)	階層ベイズモデル	○		○
日高・佐藤 (2016)	階層ベイズモデル	○		
宮津・佐藤 (2015)	階層ベイズモデル	○		○
本橋・樋口 (2013)	状態空間モデル		○	
本橋　他 (2012)	状態空間モデル		○	
Kondo and Kitagawa (2000)	状態空間モデル		○	
Manchanda, and Chintagunta (2004)	階層ベイズモデル	○		
Manchanda et al. (2004)	階層ベイズモデル	○		
Naik et al. (1998)	状態空間モデル		○	
Rossi and Allenby (2003)	階層ベイズモデル	○		
佐藤・樋口 (2008a)	状態空間モデル		○	○
佐藤・樋口 (2008b)	状態空間モデル	○	○	
佐藤・樋口 (2009)	状態空間モデル	○	○	○
Terui and Ban (2008)	階層ベイズモデル	○		○
Terui and Dahana (2006)	階層ベイズモデル	○		○
Terui et al. (2011)	階層ベイズモデル	○		○
山口 他 (2004)	状態空間モデル		○	
山田・佐藤 (2012)	階層ベイズモデル	○		○
山田・佐藤 (2016)	階層ベイズモデル	○		

の研究では，前段に示した問題意識のもと，消費者異質性あるいは時間的異質性の概念に加えて，( )内に示すような潜在変数や潜在構造を取り込んだモデル化がなされている．個々の潜在変数に関してここでは説明を割愛するが，それら変数，構造を捉えることが出来ればマーケティング実務の戦略構築に有用な知見を提供できる．その意味で，本項に示した事項はマーケティング分野やサービス分野で深化させなければならない領域である．

2.4 先行研究のまとめ

表 1 には，2.1 項〜2.3 項に示した研究を総括的に整理した．紹介した研究のみではあるが，消費者異質性，時間的異質性，潜在変数のいずれか，あるいは複数が研究の視点として含まれている．これらの研究群は，マーケティングを対象としたものであるが，サービス研究に対しても重要な示唆を供給する．重要な点なので注記しておく．

3. 事例紹介

本節では，ベイジアンモデリングにより消費者の深い理解を狙いとして実施した既存研究(宮津・佐藤, 2015; 青柳・佐藤, 2015)のエッセンスを紹介する．

3.1 宮津・佐藤 (2015)

3.1.1 研究の概要

当該研究は，心理的財布と関連した消費者の心的状況を考慮した購買点数(バスケットサイズ)の生起メカニズムをモデル化し，その現象を明らかにすることを目的として実施した．当該研究では，2 節で議論した「消費者異質性」を個人ごとのパラメータとして，さらに心的負荷

を「潜在変数」として表現したモデル化となっている．具体的には①消費者の購買時の心的状況を心的負荷として表現し，②その逼迫状況を心的負荷と閾値パラメータの大小関係として表現する階層ベイズ閾値ポアソン回帰モデルを提案している．

3.1.2 購買点数のモデル

消費者の小売店での購買点数(1 度の買い物でバスケットに入る商品点数)は，小売店頭での値引きなどのプロモーション活動や慣性行動(平日に来店しやすい，週末に来店しやすい等)に影響されて規定される．ID 付 POS データを用い，観測変数だけで消費者異質性に配慮しなければ容易にモデル表現，推定できる．しかし，消費者の買い物行動では，「今月はちょっと使いすぎたから出費を抑えよう」や「今日は給料日だからちょっとだけ贅沢をしよう」といった心理的な状況の変化で，購買量や購買の質に変化が生じる．しかし，そういった心理的な状況は観測データで直接測定できず，何らかの形でモデル表現しなければ，必要な情報抽出はできず，さらに消費者行動を妥当に評価できない．本研究は，そのモチベーションのもとで実施したものである．以降では，本研究のモデルを中心にキーとなる事項だけに限定し，説明する．

心的負荷のモデル　はじめに本研究の肝である消費者の心的負荷のモデル化を説明する．消費者個々は，前述したように，物理的な財布のほかに心の中にも心理的な財布を持っていると考えられている(Thaler, 1985)．しかし，それは単純に計量化できるものではない．本研究では，給料日からの累積購買金額を基に心的負荷のモデル化を行った．ただし，このアイデアには一つの課題がある．いつが給料日なのか？に関して，データが存在しないのである(そもそも顧客ごとにそれを調査することもできない)．その課題に対応するために，本研究では(3.1)式に示すように代表的な 3 つの給料日(25 日，5 日，17 日)を設定する．

$$\text{累積購買集計期間} \begin{cases} l=1, \text{ 前月 25 日から今月 24 日まで} \\ l=2, \text{ 前月 5 日から今月 4 日まで} \\ l=3, \text{ 前月 17 日から今月 16 日まで} \end{cases} \tag{3.1}$$

(3.2)式は，消費者 i の累積購買集計期間 l における購買期間 t_i までの累積購買金額を示す．

$$cumm_{i,t_i,l} = \begin{cases} \sum_{j=1}^{trans^l(t_i)} M_{i,j}, & trans^l(t_i) \neq 1 \\ 0 & , \ trans^l(t_i) = 1 \end{cases} \tag{3.2}$$

$trans^l(t_i)$ は，消費者 i の集計期間 l における集計起点日から t_i までの来店回数を示す．$M_{i,j}$ は消費者 i の購買機会 j における購買金額を示す．本研究では心的負荷を(3.3)式で表現する．

$$CummM_{i,t_i} = \alpha_i^{*(1)} cumm_{i,t_i,1} + \alpha_i^{*(2)} cumm_{i,t_i,2} + \alpha_i^{*(3)} cumm_{i,t_i,3} \tag{3.3}$$

$\alpha_i^{*(k)}, k = 1,2,3$ は $0 \leq \alpha_i^{*(k)} \leq 1$ および $\sum_{k=1}^{3} \alpha_i^{*(k)} = 1$ の制約を満たすパラメータであり，$\boldsymbol{\alpha}_i^* = (\alpha_i^{*(1)}, \alpha_i^{*(2)})$ とする(2 つのパラメータが決まれば制約から 3 つ目のパラメータは自動的に決まる)．この定式化により，例えば $\alpha_i^{*(1)} = 1$ と推定された場合，消費者 i の世帯は給料日が 25 日であると，$\alpha_i^{*(1)} = 0.5, \alpha_i^{*(2)} = 0.5$ と推定された場合，消費者 i の世帯には異なる給料日(25 日と 5 日)の人がいる(ダブルインカム)等々，類推できることになる．

固体内モデル(観測モデル)　(3.4)式が，本研究における個体内モデル(観測モデル)である．y_{i,t_i} は消費者 i の購買機会 t_i の購買点数を示す．また，$\boldsymbol{z}_{i,t_i}^{(k)}$，$\boldsymbol{\beta}_i^{(k)}$ は，消費者 i の購買機会 t_i の第 k レジームの説明変数ベクトルと消費者 i の第 k レジームの反応係数を示す(回帰構造に関しては(3.5)式)．y_{i,t_i} は非負のカウントデータであるため，(3.4)式に示すように観測モデル

の基本構造にはポアソン回帰モデルを採用する．ただし，モデルには閾値パラメータ γ_i を導入し，(3.3)式で定式化した心的負荷と比較することで，連続的にモデルが切換る構造を表現する．$CummM_{i,t_i} \geq \gamma_i$ のときは心的負荷が閾値を超えていることを示し，心的負荷が高い状態にあること(レジーム 1)を，一方で $CummM_{i,t_i} < \gamma_i$ のときは心的負荷が閾値を下回っていることを示し，心的負荷が低い状態にあること(レジーム 2)を表現する．すなわち，このモデル表現により心的負荷の状況により購買点数の生起メカニズムが切換る構造を定式化できたことになる．

$$\Pr(Y_{i,t_i} = y_{i,t_i} | \lambda^{(1)}_{i,t_i}, \lambda^{(2)}_{i,t_i}, \gamma_i, \boldsymbol{\alpha}^*_i) = \begin{cases} \dfrac{(\lambda^{(1)}_{i,t_i})^{y_{i,t_i}} \exp(-\lambda^{(1)}_{i,t_i})}{y_{i,t_i}!}, \ CummM_{i,t_i} \geq \ \gamma_i \\ \qquad \text{Regime1：心的負荷が高い状態} \\ \dfrac{(\lambda^{(2)}_{i,t_i})^{y_{i,t_i}} \exp(-\lambda^{(2)}_{i,t_i})}{y_{i,t_i}!}, \ CummM_{i,t_i} < \gamma_i \\ \qquad \text{Regime2：心的負荷が低い状態} \end{cases} \tag{3.4}$$

$$\log(\lambda^{(k)}_{i,t_i}) = \boldsymbol{z}^{(k)t}_{i,t_i} \boldsymbol{\beta}^{(k)}_i, k = 1, 2 \tag{3.5}$$

以降には得られた重要な知見を概説する．なお，モデルの全体，推定法の詳細は宮津・佐藤(2015)を確認してほしい．

3.1.3 得られた知見

本提案モデルの推定結果からは，様々な議論が出来る．しかしながら，紙幅の都合もあり，本研究のポイントでもある $\alpha^{*(k)}_i$, $k = 1, 2, 3$ と γ_i の推定結果のエッセンスのみを紹介する．

表 2 には，消費者ごとに算出した閾値パラメータ(γ_i)と心的負荷の構成パラメータ($\boldsymbol{\alpha}^*$)の事後平均(サンプルサイズ $N = 1{,}000$)の基本統計量を示した．閾値パラメータの事後平均の平均は 1.7 万円であり，消費者の平均累積購買金額が 2.5 万円/月であることを考慮すれば，平均累積購買金額の 68% を超えると心理的財布の切換が生じる換算となる．また，構成パラメータの各平均値は 0.35 程度であり，各累積購買金額からの心的負荷への寄与は全体的に見ればほぼ均等である．図 2 には，$\alpha^{*(1)}, \alpha^{*(2)}$ の事後平均の分布状況を示した。図中 $(1, 0)$，$(0, 1)$，$(0, 0)$ 近傍のプロットは購買期間がそれぞれ 1，2，3 のみ，$\alpha^{*(1)} + \alpha^{*(2)} = 1$ の直線上のプロットは購買期間が 1 と 2，$\alpha^{*(1)}$ 軸上のプロットは購買期間が 2 と 3，$\alpha^{*(2)}$ 軸上のプロットは購買期間が 1 と 3，図中の三角形内のプロットは全ての購買期間で構成されるものであり，消費者ごとにその構成に差が生じている．

表 3 には，構成パラメータの推定値から，心的負荷の構成パターンの割合を示した．ただし，各購買期間で 1% に満たないと推定されたものは切り捨てて算出している．ここで各構成要素は，心的負荷を構成している期間における累積購買金額の重みを表している．本モデルの設定

表 2. 推定値の基本統計量($N = 1{,}000$).

統計量	γ	$\alpha^{*(1)}$	$\alpha^{*(2)}$
平均値	0.1713	0.3538	0.3605
中央値	0.1073	0.0087	0.0079
最大値	1.8238	1.0000	1.0000
最小値	0.0055	0.0000	0.0000
標準偏差	0.2014	0.4347	0.4376

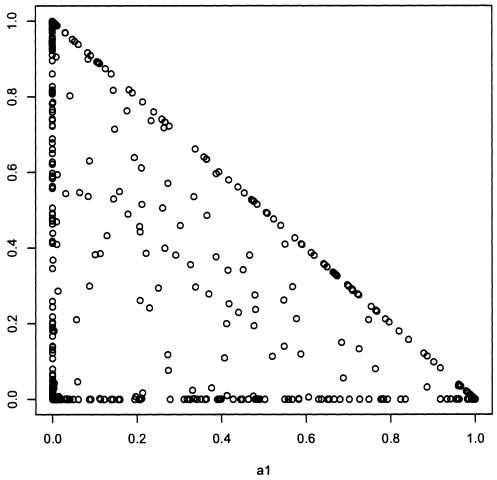


図 2. 構造パラメータ $(\alpha^{*(1)}, \alpha^{*(2)})$ の分布.

表 3. 構造パラメータの割合.

心理的財布の構成	割合 (%)	$(\alpha^{*(1)} : \alpha^{*(2)} : \alpha^{*(3)})$
購買期間 1 のみ	24.8	1.00 : 0.00 : 0.00
購買期間 2 のみ	24.6	0.00 : 1.00 : 0.00
購買期間 3 のみ	16.2	0.00 : 0.00 : 1.00
購買期間 1 と 2	8.4	0.58 : 0.42 : 0.00
購買期間 2 と 3	9.4	0.00 : 0.51 : 0.49
購買期間 3 と 1	9.3	0.52 : 0.00 : 0.48
全ての購買期間	7.3	0.32 : 0.36 : 0.32

では，期間 1，2，3 は，給料日がそれぞれ 25 日，5 日，17 日に相当する．異なる給与支給日からの累積購買金額で心的負荷が構成される場合，世帯に複数の給与所得者が存在すると推察できる．実証分析に基づけば，心的負荷が各期間単独で構成される世帯が全体の 65% を占めており，単独ないしは複数の給与所得者が存在しても同一給料日である世帯がこれに相当すると考えられる．35% 近くは心的負荷の構成に複数の累積購買金額が影響している．これらは，例えばパートタイムを含む給料日が異なる共働き世帯の構造を反映している．

本研究は，心的負荷をモデル化し，観測モデルに閾値構造を取り込むことにより，観測データを単純に用いるだけでは獲得しえない知見を抽出できた．消費者を理解するためには，こういった統計モデルの技術が有用であるとイメージしてもらえればと考える．

3.2 青柳・佐藤（2015）

3.2.1 研究の概要

当該研究は，セールスプロモーションの売上に対する動的効果をテレビ広告ストックと参照価格で階層化し，その動的進展メカニズムの解明を目的として実施した．当該研究では，2 節で議論した「時間的異質性」を時変パラメータとして，さらに集計レベルの参照価格とテレビ広告ストックを「潜在変数」として表現したモデル化となっている．具体的には，①集計レベルの

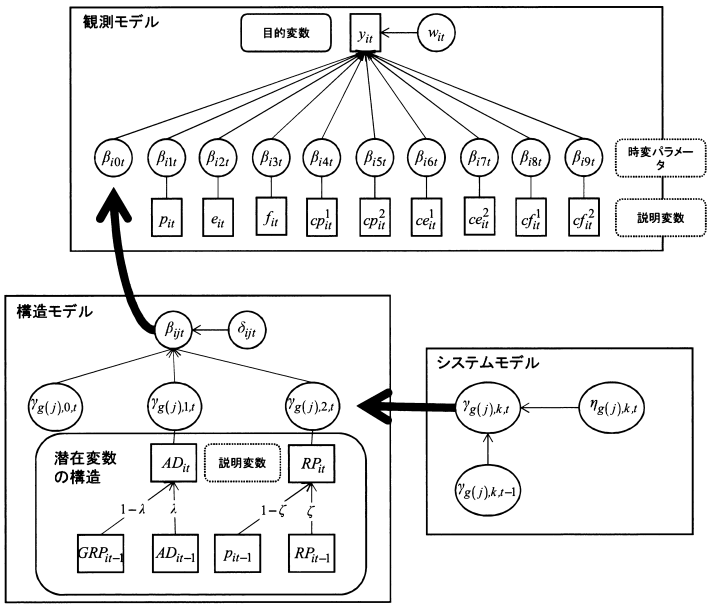


図 3. モデルの全体像.

対数点数 PI(来店客千人当たり販売点数：3 商品)をセールスプロモーション(価格，山積み，チラシ：自身，競合)で説明する動的多変量回帰(観測モデル)，②①の時変係数を目的変数とし，広告ストックおよび参照価格(両変数とも潜在変数)で説明する動的多変量回帰(構造モデル)，③②の動的多変量回帰の時変係数の時間進展メカニズム(システムモデル)，の 3 つをモデル化している．

3.2.2 動的集計型市場反応モデル

モデルの全体像 本研究におけるモデルは，外形的に非常に複雑に見えるため，はじめに図によりモデルの全体像を説明する．図 3 にはモデルの全体像を示した．本研究のモデルは，基本構造として通常の状態空間モデル(観測モデル＋システムモデル)を 3 層の状態空間モデル(観測モデル＋構造モデル＋システムモデル)に拡張したものである．3 階層状態空間モデルにおいて，観測モデルとシステムモデルの役割は，基本的に通常の状態空間モデルと同様である．構造モデルは，観測モデルに含まれる時変係数の時間進展の「理由」を記述するモデルであり，このモデル化が本研究のポイントになっている．マーケティング分野において時変係数をモデル化する場合，平滑化事前分布を用いることが多い．そのアプローチは，簡便に時変係数を表現できる一方で，時間変化のメカニズムが分からないといった批判にさらされることが多い．本提案モデルはその批判に応えうるアプローチとなっている．以降には，提案モデルの個々をごく簡単に説明する．なお，本研究は 3 商品($i = 1, 2, 3$)を対象としている．表 4 には観測モデルで用いた変数を示す．

観測モデル 観測モデルは，動的市場反応を多変量回帰モデルの枠組みで表現する．(3.6)式は商品 $i(i = 1, 2, 3)$ 個々の動的市場反応モデルを示す．実際には(3.6)式の 3 商品をベクトル表現し，解析に用いる．

表 4. 観測モデルの変数一覧.

記号	内容
y_{it}	点数 PI（対数）
p_{it}	自商品価格掛率 (対数)
e_{it}	自商品エンド陳列実施
f_{it}	自商品チラシ掲載
cp^1_{it}	競合商品 1 価格掛率 (対数)
cp^2_{it}	競合商品 2 価格掛率 (対数)
ce^1_{it}	競合商品 1 エンド陳列実施 (対数)
ce^2_{it}	競合商品 2 エンド陳列実施 (対数)
cf^1_{it}	競合商品 1 チラシ掲載
cf^2_{it}	競合商品 2 チラシ掲載
β_{ijt}	観測モデルにおける回帰係数 (市場反応係数)
w_{it}	観測ノイズ
σ^2_i	観測ノイズの分散

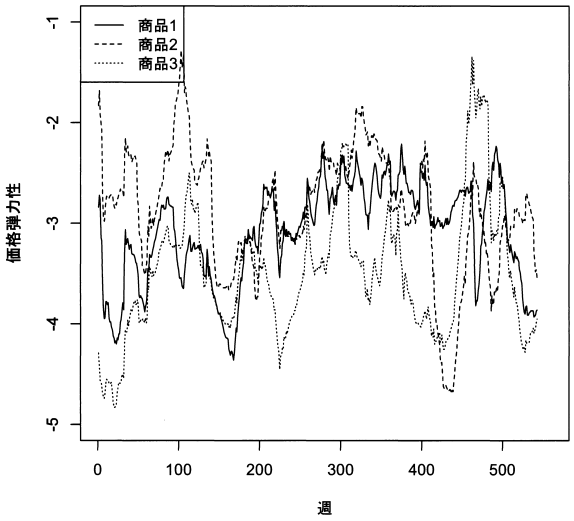


図 4. 動的価格弾力性.

$$y_{it} = \beta_{i0t} + p_{it}\beta_{i1t} + e_{it}\beta_{i2t} + f_{it}\beta_{i3t} + cp^1_{it}\beta_{i4t} + cp^2_{it}\beta_{i5t} + ce^1_{it}\beta_{i6t} + ce^2_{it}\beta_{i7t} + cf^1_{it}\beta_{i8t} + cf^2_{it}\beta_{i9t} + w_{it}, w_{it} \sim N(0, \sigma^2_i) \tag{3.6}$$

構造モデル　構造モデルは，図 4 に持式的に示したように β_{ijt} が自身の広告ストックと参照価格 AD^{own}_{it}，RP^{own}_{it} および競合の広告ストックおよび参照価格，AD^{comp}_{it}，RP^{comp}_{it} それぞれから影響を受け，時間発展する様子を表現する．(3.7)式が提案の構造モデルになる．

$$\beta_{ijt} = \log(AD^{own}_{it})\gamma_{g(j),1,t} + \log(RP^{own}_{it})\gamma_{g(j),2,t} + \log(AD^{comp}_{it})\gamma_{g(j),3,t} + \log(RP^{comp}_{it})\gamma_{g(j),4,t} + \delta_{ijt}, \delta_{ijt} \sim N(0, \tau^2_{ij}) \tag{3.7}$$

(3.7)式で用いる広告ストックと参照価格は，自商品分，競合商品分を区別して AD_{it}^{own}，AD_{it}^{comp} と添え字を用いて表現している．以降の説明では簡単のために，これらを区別せず AD_{it}，RP_{it} と記載する．(3.8)式が広告ストックの，(3.9)式が参照価格の更新式を示す．

$$AD_{it} = \lambda AD_{i,t-1} + (1-\lambda)GRP_{i,t-1} \tag{3.8}$$

$$RP_{it} = \zeta RP_{i,t-1} + (1-\zeta)Price_{i,t-1} \tag{3.9}$$

λ, ζ は，更新の程度を規定する平滑化パラメータで，0 以上 1 以下の値をとるものと仮定する．この値が 1 に近いほど前の時点のストックが残存し，逆に 0 に近いほど残存しないことを意味する．

システムモデル (3.10)式は，変数 k の時間進展を示すシステムモデルである．$g(j)$ が同じ数値になっている場合，(3.10)式は共通のものを用いる．本研究では，システムモデルを平滑化事前分布の考え方に基づき，滑らかさの仮定のもとでモデル化する．

$$\gamma_{g(j),k,t} = \gamma_{g(j),k,t-1} + \eta_{g(j),k,t}, \eta_{g(j),k,t} \sim N(0, \xi^2_{g(j),k}) \tag{3.10}$$

3 階層状態空間モデル 本研究では，上述の観測モデル，構造モデル，システムモデルは 3 階層状態空間モデルの枠組みで統合できる．3 階層状態空間モデルは(3.11)式～(3.13)式の 3 つのモデルで表現されるモデルであり，(3.6)式～(3.10)式のモデルをベクトル表現すれば得られる．その詳細については，青柳・佐藤 (2015)を参照してほしい．

$$\boldsymbol{y_t} = \boldsymbol{H}_{1,t}\boldsymbol{x}_{1,t} + \boldsymbol{w}_{1,t}, \boldsymbol{w}_{1,t} \sim \mathrm{MVN}(\boldsymbol{0}, R_1), \quad \text{(観測モデル)} \tag{3.11}$$

$$\boldsymbol{x}_{1,t} = \boldsymbol{H}_{2,t}\boldsymbol{x}_{2,t} + \boldsymbol{w}_{2,t}, \boldsymbol{w}_{2,t} \sim \mathrm{MVN}(\boldsymbol{0}, R_2), \quad \text{(構造モデル)} \tag{3.12}$$

$$\boldsymbol{x}_{2,t} = \boldsymbol{x}_{2,t-1} + \boldsymbol{w}_{3,t}, \quad \boldsymbol{w}_{3,t} \sim \mathrm{MVN}(\boldsymbol{0}, R_3), \quad \text{(システムモデル)} \tag{3.13}$$

3.2.3 得られた知見

本提案モデルの推定結果からは，様々な議論が出来る．本研究のポイントでもある商品ごとの動的価格弾力性 $\beta_{i1t}, i = 1, 2, 3$ とその動的構造パラメータ $\gamma_{2,k,t}, k = 1, \cdots, 4$ の推定結果に絞り，そのエッセンスを紹介する．

本研究では，インスタントカレーの 3 商品のデータを用いた．分析には 3 商品ので POS データとコーザルデータ(売価，山積み陳列，チラシ掲載，広告投下量(GRP))を用いた．表 5 には，変数ごとの要約統計量を示した．

表 5. 使用データの要約統計量.

項目	商品 A	商品 B	商品 C
平均販売個数	3.357	8.101	5.085
平均点数 PI	1.182	1.772	2.449
販売個数欠測日数	109	106	33
最大売価 (定価)	218	198	198
平均価格掛率	0.867	0.824	0.855
エンド陳列回数	72	161	119
チラシ掲載回数	7	19	28
GRP 総和	4,809	4,915	19,457

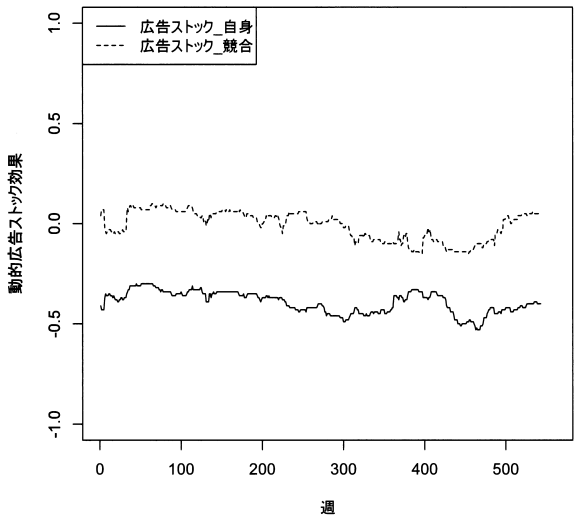


図 5．動的広告ストック効果.

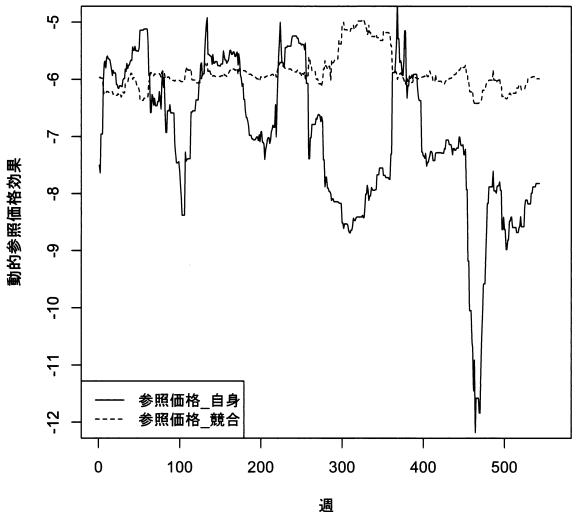


図 6．動的参照価格効果.

図 3 には，推定した商品ごとの価格弾力性消費者 $\beta_{i1t}, i = 1,2,3$ を示した．商品ごとにその推移に差が生じている．時間的異質性を捉えるという観点で，時変価格弾力性は十分に意味がある．しかし，マーケティングを高度化するにはさらなる踏み込みが必要である．上述しているが，$\beta_{i1t}, i = 1,2,3$ の変動が生じる理由をも評価したいのである．そこで重要なのが，(3.7) 式に示した構造モデルである．図 5 には広告ストック(自身，競合)の $\beta_{i1t}, i = 1,2,3$ に与える動的影響を，図 6 には参照価格(自身，競合)の $\beta_{i1t}, i = 1,2,3$ に与える動的影響をそれぞれ示した．これら図から，自商品の広告ストックが増えるほど，また自商品の店舗レベル参照価格が低下するほど，価格弾力性が負に大きくなる．ここに示したような，時間的異質性を捉え，

しかもその変動理由をも同時に捉えることは，実務マーケティングにおける「何を，何のために，コントロールするのか」といった課題に対して示唆を与えられる．今回の事例でいえば，その構造は，「価格や広告投下量は参照価格や広告ストックの変動に影響し，それら潜在変数がセールスプロモーションの動的市場反応に影響する．最終的にはその市場反応により売上が規定される」というものに対応する．これは通常考えられる「売価や広告が直接売上を規定する」といったものと大きく異なる．消費者の行動を理解し，それに応じた形式でマーケティングを高度化するには，理由までをも含めたモデリングが必要だし，有効だと考える．本研究は一つの事例でしかないが，消費者の理解を深めるには，この種のアプローチを進展させなければならない．

4. まとめと今後の課題

本研究は，消費者理解を深めるにはという観点で統計モデル(特にベイズモデル)をどのように活用すべきか？に対する提言を目的とし，既存研究の紹介を基本として整理した．現象が変われば必要なモデリングに違いが生じるが，本稿の内容は統計サイドから研究対象が主として消費者のサービスやマーケティング分野に接近する際の参考にしてもらえるのではないかと考えている．

サービス研究の源流は，ほぼマーケティング研究にあるものと考えられる．マーケティング研究から派生したサービス研究は，その重要性が叫ばれ，日本学術会議におけるサービスの参照規準小委員会でも「人間がその系に含まれる連続的システムにおいて，感情や知識を含む様々な価値を共創的かつダイナミックに生産する行為」といった形式でサービスの定義を与えている．伝統的なマーケティングが表 6 に示す Goods Dominant Logic(GDL)の考え方に基づいている一方で，サービスは表 6 に示す Service Dominant Logic(SDL)の考えに基づき進展している．その進展における重要なポイントは，GDL が一方向の価値提供なのに対し，SDL は双方向の価値共創になっている点である．SDL の考え方に基づくと価値は供給者サイドが生み出すものではなく，供給者と消費者が共創することによって生み出されるのだと，考えることになる．この発想自体が誤りであるわけではないが，共創という発想を強く意識するがゆえに，本来供給者サイドがしなければいけない「消費者の理解」がなおざりになってしまっている．価値共創の仮定のもとでも供給者は消費者理解を進めなければならないし，それをしない限り競争優位性は構築できない．その際有効に機能するのが本項で紹介したアプローチであり，統計モデルである．サービス科学分野では，本稿で指摘した事項を意識した形で消費者理解のための研究を進めなければならない．また，それが実現できれば社会に大きな貢献をもたらすものと確信している．

表 6. Goods Dominant Logic と Service Dominant Logic.

視点	GDL：Goods Dominant Logic 一方向的な価値提供 「交換価値」の最大化	SDL：Service Dominant Logic 双方向の価値共創 「使用価値」の最大化
価値の生産	企業	企業と顧客が共に
顧客の役割	価値を消費する	価値を生産し，消費する
イノベーションの対象	製品や技術	顧客とのインタラクション
価値の源泉	新しい機能や性能（交換価値）	新しい顧客行動や顧客経験（使用価値）
企業と顧客との接点	購買時	購買前，購買時，購買後

参 考 文 献

Allenby, G. M. and Rossi, P. E. (1998). Marketing models of consumer heterogeneity, *Journal of Econometrics*, **89**(1), 57–78.

青柳憲治, 佐藤忠彦 (2015). 3 階層多変量状態空間モデルによる動的市場反応形成メカニズムの解明, 日本オペレーションズ・リサーチ学会和文論文誌, **58**, 70–100.

Bass, F., Bruce, N., Majumdar, S. and Murthi, B. (2007). Wearout effects of different advertising themes: A dynamic Bayesian model of the advertising-sales relationship, *Marketing Science*, **26**(2), 179–195.

Bruce, N. (2008). Pooling and dynamic forgetting effects in multitheme advertising: Tracking the advertising sales relationship with particle filters, *Marketing Science*, **27**(4), 659–673.

Chib, S. and Greenberg, E. (1996). Markov chain Monte Carlo simulation methods in econo-metrics, *Econometric Theory*, **12**(3), 409–431.

Doucet, A., Freitas, N. de and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*, Springer, New York (etc.).

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**(1), 97–109.

日高徹司, 佐藤忠彦 (2016). 消費者とブランドとの関係を考慮した階層ベイズモデルによるクロスメディア効果推定, 日本オペレーションズ・リサーチ学会和文論文誌, **59**, 106–133.

樋口知之 (2011).『予測にいかす統計モデリングの基本―ベイズ統計入門から応用まで』, 講談社, 東京.

井上友彦 (2010). 医師の異質性を考慮した医薬品業界における営業訪問効果の分析, マーケティングサイエンス, **18**, 49–73.

北川源四郎 (2005).『時系列解析入門』, 岩波書店, 東京.

Kondo, F. N. and Kitagawa, G. (2000). Time series analysis of daily scanner sales: Extraction of trend, day-of-the-week effect and price promotion effect, *Marketing Intelligence & Planning*, **18**(2), 53–66.

Manchanda, P. and Chintagunta, P. (2004). Responsiveness of physician prescription behavior to salesforce effort: An individual level analysis, *Marketing Letters*, **15**(2), 129–145.

Manchanda, P., Rossi, P. and Chintagunta, P. (2004). Response modeling with nonrandom marketing-mix variables, *Journal of Marketing Research*, **41**(4), 467–478.

宮津和弘, 佐藤忠彦 (2015). 心理的財布と購買行動の関係性のモデル化―階層ベイズ閾値ポアソン回帰モデルの提案, 応用統計学, **44**(3), 161–182.

本橋永至, 樋口知之 (2013). 市場構造の変化を考慮したブランド選択モデルによる購買履歴データの解析, マーケティング・サイエンス, **21**(1), 37–59.

本橋永至, 磯崎直樹, 長尾大道, 樋口知之 (2012). 状態空間モデルによるインターネット広告のクリック率予測, オペレーションズ・リサーチ：経営の科学, **57**(10), 574–583.

Naik, P., Mantrala, M. and Sawyer, A. (1998). Planning media schedules in the presence of dynamic advertising quality, *Marketing Science*, **17**(3), 214–235.

日本学術会議経営学委員会・総合工学委員会合同サービス学分科会 (2017). 報告，大学教育の分野別質保証のための教育課程編成上の参照基準, サービス学分野, 1–28.

Rossi, P. and Allenby, G. (2003). Bayesian statistics and marketing, *Marketing Science*, **22**(3), 304–328.

Rossi, P., Allenby, G. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, John Wiley & Sons., England.

佐藤忠彦, 樋口知之 (2008a). 動学的売上反応モデルによる POS データの解析, マーケティングサイエンス, **15**(1), 1–26.

佐藤忠彦, 樋口知之 (2008b). 動的個人モデルによる消費者来店行動の解析(議論付き), 日本統計学会論文誌, **38**(1), 1–38.

佐藤忠彦, 樋口知之 (2009). 動的個人モデルによる購買生起行動の解析, マーケティング・サイエンス,

16(1, 2), 49–73.
佐藤忠彦, 樋口知之 (2013).『ビッグデータ時代のマーケティング—ベイジアンモデリングの活用』, 講談社, 東京.
照井伸彦 (2008).『ベイズモデリングによるマーケティング分析』, 東京電機大学出版局, 東京.
Terui, N. and Ban, M. (2008). Modeling heterogeneous effective advertising stock using single-source data, *Quantitative Marketing and Economics*, **6**(4), 415–438.
Terui, N. and Dahana, W. D. (2006). Research note — Estimating heterogeneous price thresholds, *Marketing Science*, **25**(4), 384–391.
Terui, N., Ban, M. and Allenby, G. M. (2011). The effect of media advertising on brand consideration and choice, *Marketing Science*, **30**(1), 74–91.
Thaler, R. (1985). Mental accounting and consumer choice, *Marketing Science*, **4**, 199–214.
山口類, 土屋映子, 樋口知之 (2004). 状態空間モデルを用いた飲食店売上の要因分解, オペレーションズ・リサーチ：経営の科学, **49**(5), 52–60.
山田浩喜, 佐藤忠彦 (2012). 階層ベイズモデルによる百貨店の態度ベース店舗満足化構造に関する解析, マーケティングサイエンス, **20**(1), 17–41.
山田浩喜, 佐藤忠彦 (2016). 百貨店顧客の来店回数生起メカニズムの構造異質性の解析, 行動計量学, **43**(1), 53–68.

Proceedings of the Institute of Statistical Mathematics Vol. 66, No. 2, 249–265 (2018) 265

Possibility of Achieving Consumer Understanding Using Statistical Models

Tadahiko Sato

Faculty of Business Sciences, University of Tsukuba

Advanced utilization of statistical models is an indispensable tool for the development of service sciences, and its importance will increase further in the future. The purpose of this paper is to explain how to use a statistical model (specifically, a Bayesian model) to achieve deep consumer understanding, which is essential for improving services. Specifically, we will organize related research on this issue and provide details of our two research. Although these two studies were in marketing research, they also provide suggestions relevant to service research.

Key words: Bayesian modeling, consumer heterogeneity, time heterogeneity, latent variable.

統計数理(2018)
第 66 巻 第 2 号 267–278

特集「サービス科学の今」
［研究ノート］

地域健康政策へのベイジアンネットワークの応用

鳥海 航[1]・生方 裕一[1]・久野 譜也[2,4]・岡田 幸彦[3,4]

（受付 2017 年 12 月 31 日；改訂 2018 年 3 月 15 日；採択 3 月 16 日）

要　　旨

本稿は，データ中心科学としてのサービス科学の新たな展開である地域健康政策のためのサービス科学のあり方について，国立研究開発法人日本医療研究開発機構の「AI を活用した保健指導システム研究推進事業」として採択された筑波大学の取り組み事例をもとに議論している．自治体が行う地域健康政策では説明責任が強く求められるため，担当職員にとって説明容易性の高い統計手法を用いる必要がある．また，どの自治体，どの疾病に対しても応用可能な分析方法論を確立する必要がある．本稿では，これらの必要性を満たす統計手法として，制約ベースアプローチで条件付き独立性を χ^2 検定によって行い，より効率的な構造学習が可能な Local to Global アプローチのアルゴリズムを採用したベイジアンネットワークが有用であることを主張する．そして，自治体 A の実際の健康関連ビッグデータを用いて，どの自治体，どの疾病に対しても応用可能な疾病発症ベイジアンネットワークの試行を行っている．そして，本稿で紹介した取り組み事例をふまえ，地域健康政策におけるサービス科学のあり方と今後の研究課題について議論している．

キーワード：地域健康政策，人工知能，説明責任，説明容易性，ベイジアンネットワーク．

1. はじめに

データ中心科学としてのサービス科学は，わが国が抱える社会課題への貢献も期待されている．わが国は，人口減少・少子高齢化の先進国である．すでに 2013 年度にわが国の国民医療費は 40 兆円を超え，今後さらに増大することが見込まれている．さらに，地域間の健康格差の問題が指摘され始めている．これからの超高齢時代において日本国憲法第 25 条が規定する国民の健康で文化的な最低限度の生活を営む権利をどのように保障していくかは，わが国が直面する最も重要な社会課題の 1 つである．

この社会課題に対してわが国は，総務省の支援のもと「健幸長寿社会を創造するスマートウエルネスシティ総合特区」として先導的に構築された健幸クラウドや，厚生労働省の支援のもと構築された国民健康保険中央会の KDB（国保データベース）のように，健康関連ビッグデータ（個人の国民健康保険，介護保険，特定健診等のデータ）の蓄積を政策的に進めてきた．そして，これらの健康関連ビッグデータを基礎として，厚生労働省はデータヘルス計画の策定・実

[1] 筑波大学大学院 システム情報工学研究科：〒 305–8573 茨城県つくば市天王台 1–1–1
[2] 筑波大学 体育系：〒 305–8573 茨城県つくば市天王台 1–1–1
[3] 筑波大学 システム情報系：〒 305–8573 茨城県つくば市天王台 1–1–1
[4] 筑波大学 人工知能科学センター：〒 305–8573 茨城県つくば市天王台 1–1–1

行を，総務省は証拠に基づく政策立案(EBPM; Evidence-Based Policy Making)を推進し，地域の健康課題に即した健康政策の実施が促されている．さらに，国立研究開発法人日本医療研究開発機構(AMED)は，2017 年度に「AI を活用した保健指導システム研究推進事業」を公募し，健康関連ビッグデータと人工知能を活用した健康・医療戦略の推進のための基盤研究を開始した．

健康関連ビッグデータの蓄積・活用と人工知能への注目は，データ中心科学としてのサービス科学の新たな地平を示す．それは，社会課題解決型の応用統計数理についての「学」への期待とも言えよう．そこで本稿では，地域健康政策のためのサービス科学のあり方について，上述した AMED の基盤研究として採択された筑波大学の取り組み事例[1)]をもとに議論したい．ここでの重要な論点は，地域健康政策の立案・実行を担う自治体職員(以下，担当職員)の知識・スキル水準との整合性である．なぜなら，どんなに高度な統計手法を用いたとしても，合意形成や説明責任を果たすために担当職員が説明・説得することができなければ，人工知能は無用の長物となってしまうのである．

そこでまず本稿では，地域健康政策における人工知能の使い手である担当職員の目線から，「説明・説得のしやすさ」としての説明容易性を重視した統計手法の選定を考える．その結果，地域健康政策における制約ベースアプローチのベイジアンネットワークの有用性が導かれる．次いで，健康関連ビッグデータを所与として，どの自治体，どの疾病にでも広く応用可能なベイジアンネットワークの構築方法について考察する．そして，自治体 A の実際の医療レセプトおよび特定健診データを用いた疾病発症ベイジアンネットワークの構築を行う．なお本稿では，紙幅の関係から，全ての疾病の中で代表的な高血圧と糖尿病の結果のみを取り上げる．最後に，本稿での取り組み事例をふまえ，地域健康政策におけるサービス科学のあり方と今後の研究課題について議論したい．

2. 地域健康政策で有用な統計手法の考察

2.1 行政運営に求められる説明容易性

わが国の地方行政は，日本国憲法および地方自治法・関連諸規則に従い，都道府県・市町村(以下，自治体)による地域別の行政執行が基本となっている．自治体の執行機関としての長は住民による直接選挙で選ばれ，同じく直接選挙で選ばれた議事機関としての議会との相互牽制のもと，予算を編成し，政策および事務事業を実行する．予算編成，政策および事務事業の実行等を所管するのは，自治体の職員である．一般的に，自治体の職員は部局別の組織構造に配属され，当該組織は部から課，そして係へと細分化されている．

地域健康政策に限定すると，住民の健康生活を保障することを主目的とし，自治体には健康推進課や保健推進課といった名称の担当部局が存在する．そして，それらの中に設置された係単位で予算案の作成や事務事業の実施等が行われることが一般的である．つまり，地域健康政策の立案と実行は，自治体の基本計画や首長の公約等を基礎として，主として健康推進関係の係の担当職員が事務事業を所管する．そのため，担当職員にとって，内的には上長，財政課のような予算査定担当，首長，そして議会に対する逐次的な合意形成を行うことが，外的には住民に対して説明責任を果たすことが，地域健康政策上で不可欠となる．

以上のわが国地域健康行政の特徴は，前述した国が推進する健康関連ビッグデータの蓄積・活用に際し，応用統計数理の観点で重要な注意点を喚起する．統計学の初歩的知識しか有さない担当職員が，統計学の知識がない多様な利害関係者に説明・説得しなければならない姿を想定する必要があるのである．そして前述のとおり，この文脈で人工知能の活用が期待されている．ここでの避けられない利害関係者からの問いは，「なぜその健康事務事業を行う必要があ

るのか？」である．この問いに対し，健康関連ビッグデータから当該地域の健康課題の因果メカニズムを導いたうえで，その重要な原因候補を特定し，それらを証拠として健康事務事業を立案する，という EBPM が求められる時代になった．

地域健康政策における EBPM において，人工知能の役割は大きい．なぜなら，「健康関連ビッグデータを用いた因果推論 ⇒ 重要な原因の特定 ⇒ 政策立案」という一連の流れの大部分を，人工知能が代替できる可能性が高いからである．一方で，内的な説明・説得プロセスを勝ち抜くとともに，住民への説明責任を果たすためには，「どうやってその証拠が生み出されたのか？」という追加的な利害関係者からの問いに担当職員が答える準備をしておく必要がある．つまり，地域健康政策における EBPM では，証拠の科学性や可読性だけでなく，証拠とその生成過程の説明容易性が強く求められるのである．

説明容易性は，データ中心科学としてのサービス科学がこれまで見すごしてきた重要な論点であると考えられる．内的な説明・説得プロセスも含む広い意味での説明責任が求められない場合には，予測・判別精度のみを重視し，ブラックボックス型とも呼称される計算論的機械学習を用いることは有用となろう．しかしながら，説明責任が強く求められる場合には，注意を要する．なぜなら，証拠とその生成過程の説明をも求められるからである．以上を鑑みると，地域健康政策の文脈では説明責任が強く求められることから，人工知能の利用者の知識・スキル水準と整合する説明容易性の高さを基準とした統計手法の選定が望ましいと考えられる．

ここで，地域健康課題の因果メカニズムを推論でき，説明容易性が相対的に高い手法として，地域健康政策におけるベイジアンネットワークの有用性を主張できよう．ベイジアンネットワークはグラフィカルモデルの一種であり，有向グラフと条件付き確率表によって因果メカニズムを表現することができるため，不確実性の伴う複雑な社会現象を柔軟にモデル化することが可能である（本村, 2000; Pearl, 2003; Kalisch et al., 2010）．また，変数間の確率的な関連性を生かしたソフトウェア開発や病気における要因分析（Park and Kim, 2013; Harris et al., 2017），疾病の診断・管理（Velikova et al., 2014），映画のレコメンドシステム（Ono et al., 2007）など，様々な用途での活用が提案されてきた．

そして近年，ベイジアンネットワークは政策立案現場での応用可能性が議論されるようになった．例えば，環境政策の効果分析を行った Carriger et al.（2016）は，ベイジアンネットワークを用いた確率的推論を行うことで，不確実性が高い状況下でも EBPM が容易になることを示している．また上野（2010）は，ベイジアンネットワークを用いて過疎地域での人口減少要因分析を行い，政策提案を行っている．ここで上野（2010）は，ベイジアンネットワークは独立変数を変化させたときの目的変数の変化を確率として予測できるため，政策立案の際に有用性が高いことを指摘している．そして，ベイジアンネットワークを用いることで社会現象全体の構造が可視的に明らかになり，より深い洞察を得ることが可能となる（鶴田・寒河江, 2015）．

説明容易性が必要とされる地域健康政策では，地域健康課題の因果メカニズムを可視的かつ確率的に洞察できるベイジアンネットワークの有用性が高いものと考えられる．

2.2 説明容易性と制約ベースアプローチ

ベイジアンネットワークの構造学習は，スコアベースアプローチと制約ベースアプローチという 2 つのアプローチが存在する（Koller and Friedman, 2009）．スコアベースアプローチは，データセット D から導出される構造 G のスコアを $\mathrm{Score}(G|D)$ と定義すると，式(2.1)のようにスコア関数が最大となる構造を探索するようなアプローチである．

$$G^* = \mathrm{argmax}_G \, \mathrm{Score}(G|D) \tag{2.1}$$

ここで使用されるスコア関数は，情報理論アプローチとベイジアンアプローチが存在し，

基本的にいずれのスコア関数も対数尤度を基礎としている．式(2.2)は最も素朴な対数尤度によるスコア関数であるが，この方法は過学習の問題が指摘されてきた（Liu et al., 2012）．そこで，情報理論アプローチでは BIC（Schwartz, 1978)等を，ベイジアンアプローチでは BDeu（Heckerman et al., 1995)等を，式(2.2)に対する罰則項として付け加えることで，よりよい構造推定を目指すスコア関数が式(2.3)と式(2.4)である．ここで式(2.2)の D_{ij} は，データセット D における $i(i=1,\ldots,n)$ 番目の変数の $j(j=1,\ldots,N)$ 番目のデータサンプルを表している．また，PA_{ij} は，i 番目の変数の親ノード集合を表している．式(2.3)におけるは，i 番目の変数における状態数である．式(2.4)における r_i と q_i は，i 番目の変数の状態数とその親ノード集合の状態数である．同式中の D_{ij} は，データセット D における i 番目の変数の親ノード集合の状態が j となるデータサンプルを表している．その中でも D_{ijk} は，i 番目の変数の状態が k となるデータサンプルを表している．α はモデル構築者が決めるパラメータである．

$$LL(D|G)=\sum_{j=1}^{N}\log P(D_j|G)=\sum_{i=1}^{n}\sum_{j=1}^{N}\log P(D_{ij}|PA_{ij}) \tag{2.2}$$

$$BIC(D|G)=LL(D|G)-\sum_{i=1}^{n}\frac{\log N*p_i}{2} \tag{2.3}$$

$$BDeu(D|G)=LL(D|G)-\sum_{i=1}^{n}\sum_{j}^{q_i}\sum_{k}^{r_i}\log\frac{P(D_{ijk}|D_{ij})}{P(D_{ijk}|D_{ij},\alpha_{ij})} \tag{2.4}$$

このスコアベースアプローチの構造探索は，変数が増加すると指数関数的に計算量が増加する欠点が指摘されてきた(Chickering et al., 2004)．一方で，後述する制約ベースアプローチと比較して，サンプル数が少ない場合でも構造推定の精度を高く維持できるメリットがある（Tsamardinos et al., 2006)．つまり，変数が比較的少なく，サンプル数も比較的少ない場合には，スコアベースアプローチが技術的に優位であると考えられる．しかし前述のとおり，わが国ではすでに多変量大規模サンプルの健康関連ビッグデータが蓄積されていることから，低サンプル数の問題は重要ではなく，むしろ多変量に係る計算量の問題が重大となるため，わが国地域健康政策におけるスコアベースアプローチの優位性があるとは言い難い．むしろ，地域健康政策で求められる説明容易性の観点からすると，後述の制約ベースアプローチと比較して，大きな課題があることを指摘せざるを得ない．統計学の初歩的知識しか有さない担当職員が，統計学の知識がない多様な利害関係者に対して，スコア関数にもとづく構造推定について説明するのは容易ではないのである．

相対的に説明容易性が高いと考えられる制約ベースアプローチは，条件付き独立性に注目した構造推定を行う．ここで，有限個の要素からなる確率変数集合 V に対して，$P(V)$ を V の同時確率分布とし，X，Y，Z を V の部分集合としよう．この時，$P(y,z)>0$ に対して式(2.5)が成り立つ場合，X と Y は Z を所与とした条件付き独立であると考える．

$$P(x|y,z)=P(x|z) \tag{2.5}$$

制約ベースアプローチは，この条件付き独立性を χ^2 検定，G^2 検定，相互情報量の大小などで判定する点が特徴的である．その中で最も古典的な方法は，PC アルゴリズム（Spirtes et al., 2000; Madsen et al., 2017)や TPDA アルゴリズム（Cheng et al., 2002)のように，無向完全グラフや全域木から条件付き独立性を基準として構造を絞り，方向付けを行う Global アプローチである．Global アプローチは，条件付き独立性を基準として構造を絞ることと，方向付けを

行うこととを再帰的に行うことで，より効率的に構造学習を行う RAI アルゴリズム（Yahezkel and Lerner, 2009）が提唱されるに至っている．

一方で，制約ベースアプローチには，GS アルゴリズム（Margaritis and Thrun, 2000）のような，まず部分的に局所グラフを作成し，それぞれの局所グラフをつなげることにより全体の無向グラフを作成し，方向づけを行う Local to Global アプローチ（Gao et al., 2017）も存在する．さらに，Local to Global アプローチを構成する要素技術の発展（Aliferis et al., 2010b）として，MMPC アルゴリズム（Tsamardinos et al., 2006）や HITON-PC アルゴリズム（Aliferis et al., 2003; Aliferis et al., 2010a; Aliferis et al., 2010b）のように，ターゲットノードに対する辺候補を親子関係の蓋然性から導出し，局所構造を探索する方法も提案されている．

一般的に Local to Global アプローチは，Global アプローチと比較して，条件付き独立性検定の試行回数について，効率的に構造学習をすることが可能であるとされる（Tsamardinos et al., 2006）．この検定における効率性は，次の 2 点において効果的である．まず Global アプローチで必要となる高次の条件付き独立性検定を必要としないため，高次の検定結果による信頼性や効率性の低下を防ぐことができる．次に，検定の試行回数自体を少なくすることで，構造推定の精度を維持できる（Koller and Friedman, 2009）．このように効率性が高く，信頼性も維持できる点が，Local to Global アプローチの利点としてあげられる．一方で，Local to Global アプローチは，関係の強い部分的な変数群に関する局所構造の推定を行い，探索空間を制限する（Gao et al., 2017）．つまり，Global アプローチと比較して，Local to Global アプローチには全てのグラフを構造推定の候補としない欠点が存在している．

ここで注意すべきは，地域健康政策のためにベイジアンネットワークを用いる文脈である．統計学の初歩的知識しか有さない担当職員にとって，最もなじみがあり，説明容易性が高いのは，条件付き独立性を χ^2 検定で判断することであると考えられる．さらに，自治体は国の政策によって健康関連ビッグデータをすでに有し，今後もデータがさらに大規模に蓄積されていくことを考えると，多変量大規模データからいかに効率的に構造推定を行い，地域健康政策上の情報ニーズに迅速に応えていくかが重要となろう．こうした理由から，わが国の地域健康政策においてベイジアンネットワークを適切に応用するためには，説明容易性の観点から制約ベースアプローチで条件付き独立性を χ^2 検定によって行い，より効率的な構造学習が可能な Local to Global アプローチのアルゴリズムを採用することが望ましいと考えられる．

3. 疾病発症ベイジアンネットワークの基本設計と使用データ

本稿の以降では，前章で導いたわが国地域健康政策において有用だと考えられる統計手法の条件に従い，実際の健康関連ビッグデータを用いた地域健康課題の因果メカニズムのモデリングを試行したい．この試行では，前述の条件を満たす HITON-PC アルゴリズム（Aliferis et al., 2003; Aliferis et al., 2010a; Aliferis et al. 2010b）を用い，条件付き独立性は χ^2 検定で判断し，Verma and Pearl（1990）の Inductive Causation アルゴリズムによる因果モデルの構築を行う．このとき，非巡回制約を満たす方向付けを，Meek（1995）の伝統的なオリエンテーションルールによって行う．本稿で試行する因果モデリングは，国が政策的に蓄積を進めてきた健康関連ビッグデータの活用を想定し，わが国全ての自治体で応用可能な方法を追求する．具体的には，ほぼ全ての自治体が登録し，活用していると言われる KDB を前提とし，KDB に収録されている住民の国民健康保険（以下，国保）の医療レセプトデータと特定健診の結果データを用いる．それらのデータの概要は，以下のとおりである．

わが国では，被用者保険と国保のいずれかの医療保険に加入することが義務付けられている．自治体を保険者とする国保への加入者数は 3,303 万人であり，その平均年齢は 51.5 歳であ

る(厚生労働省, 2016)．KDB では，被保険者である住民個人の疾病区分ごとの医療レセプトが含まれており，ここでの疾病区分は厚生労働省が定める「社会保険表章用疾病分類」の区分に従い，糖尿病や高血圧などの 121 疾病が存在している．この国保の医療レセプトデータを用いれば，ある住民が特定の疾病を発症しているか否かの判定が可能となる．そして，こうして判定された住民別・疾病別の発症の有無のデータは，個別自治体における健康課題についての因果メカニズムを特定する際の，欠かせない結果変数となる．

一方，特定健診は「高齢者の医療の確保に関する法律」に従い，「医療保険者(国保・被用者保険)が，40〜74 歳の加入者(被保険者・被扶養者)を対象として，毎年度，計画的に(特定健康診査等実施計画に定めた内容に基づき)実施する，メタボリックシンドロームに着目した検査項目」(厚生労働省, 2013, p.5)による健康診査である．特定健診における項目は，必須項目と詳細な健診項目から構成される．必須項目は，質問表(服薬歴，喫煙歴等)，身体測定(身長，体重，BMI，腹囲)，血圧測定，理学的検査，検尿(尿糖，尿蛋白)，血液検査(脂質検査，血糖検査，肝機能検査)からなる (厚生労働省, 2009)．

これら特定健診データの中で，疾病発症ベイジアンネットワークで用いるために離散変数として取り扱うことが可能な投入変数として，少なくとも表 1 の 33 変数を作成・使用可能である．これらの変数は，KDB に参加する全ての自治体で作成・使用可能である．本稿は全自治体への応用可能性を目指した疾病発症ベイジアンネットワークの試行であり，本稿ではこの試行のために匿名でご協力くださった自治体 A の 2015 年度のデータを用いる．自治体 A はわが国における典型的な中小自治体であり，国保の医療レセプトと特定健診のデータが個人単位で揃っている全住民は 3,391 人であった．この本稿で用いる 33 個の離散変数に関する全 3,391 人の分布は，表 1 のとおりである．

ここで本稿では，疾病発症ベイジアンネットワークを構築するにあたり，表 1 の 33 個の投入変数を，政策的に改善が可能な原因候補として位置付けられる生活習慣系・食習慣系・健康意欲系・身体指標系，政策的に変化させることは難しいがターゲットセグメント化には有用である個人属性系・病歴系，という 6 つにグループ化した．これらに加えて，自治体内の地域特性を反映すべく，どの自治体でも使用可能な地区変数として，自治体 A 内の 8 つの小学校区を地域特性系として用意した．これら 7 つのグループと各グループ内の変数リストは，図 1 のとおりである．そして本稿では，疾病発症ベイジアンネットワークの構築に際し，グループ内の変数の間の関係性を考慮しない制約を設けた．加えて，疾病発症の有無が結果変数となるように，疾病発症の有無から全ての投入変数に対して矢印が引かれることがない制約を設けた．

なお，本稿の疾病発症ベイジアンネットワークは，全ての疾病区分に対して応用可能である．紙幅の関係から，本稿では，代表的な生活習慣病である高血圧と糖尿病の 2 疾病を取り上げることとする．

4. 試行結果と議論

前述の基本設計とデータによって，高血圧と糖尿病の疾病発症ベイジアンネットワークが構築された．疾病へと矢印が向かわない投入変数と矢印を省略した簡易図は，図 2(高血圧)と図 3(糖尿病)のとおりである．なお，図中の符号は関連する 2 変数(全て二値データである)についての条件付き確率表をもとに付され，+ は正の関係を，− は負の関係を示している．

図 2 と図 3 から，自治体 A では，メタボ判定基準に該当する住民が高血圧と糖尿病の発症確率が高いことがわかる．これが因果関係を示していることが医学的に確からしいのであれば，自治体 A ではメタボ対策の事務事業を重点的に行うことが有効となりそうである．そして，糖尿病発症に関しては，このメタボ対策の事務事業を，特に男性に対して重点的に実施すること

表 1. 本稿で使用する 33 個の投入変数と分布(全 3,391 人).

変数名	内容	人
性別	女性	1,864
	男性	1,527
40~44歳	該当者	90
45~49歳	該当者	85
50~54歳	該当者	103
55~59歳	該当者	181
60~64歳	該当者	664
65~69歳	該当者	1,170
70~74歳	該当者	1,098
BMI（Underweight）	該当者	269
BMI（Normal）	該当者	2,341
BMI（Pre-obese）	該当者	696
BMI（Obese）	該当者	85
メタボ判定	基準該当	606
_基準該当	基準該当ではない	2,785
メタボ判定	予備群該当	272
_予備群	予備群該当ではない	3,119
既往歴	有り	129
（脳血管）	無し	3,262
既往歴	有り	185
（心血管）	無し	3,206
既往歴	はい	13
（腎不全）	いいえ	3,378
貧血	はい	450
	いいえ	2,941
喫煙	はい	396
	いいえ	2,995

変数名	内容	人
体重変化	はい	945
	いいえ	2,446
運動習慣	はい	2,002
不十分	いいえ	1,389
歩行習慣	はい	1,318
不十分	いいえ	2,073
歩行速度	はい	1,646
	いいえ	1,745
食習慣（早食い）	早い	704
_早い	早くない	2,687
食習慣（早食い）	遅い	245
_遅い	遅くない	3,146
食習慣	はい	464
（就寝前夕食）	いいえ	2,927
食習慣	はい	338
（夜食）	いいえ	3,053
食習慣	はい	169
（朝食抜き）	いいえ	3,222
飲酒	毎日	919
_毎日	毎日ではない	2,472
飲酒	時々	787
	時々ではない	2,604
睡眠不十分	はい	590
	いいえ	2,801
改善意識	健康改善するつもりはない	1,174
	健康改善するつもりである	2,217
改善行動	健康改善行動をしていない	1,953
	健康改善行動を始めている	1,438

が，自治体 A では有効かもしれない．

一方，図 2 の高血圧発症に関しては，45 歳から 74 歳までの喫煙する男性が，そして貧血で食事が遅い 70 歳から 74 歳までの女性が，さらに男女を問わず太り気味の住民が，自治体 A では高血圧の発症確率が高そうである．もしこれが自治体 A の実態に即していると担当職員が感じるのであれば，自治体 A ではメタボ対策だけでなく，これら 3 つの領域に特化した高血圧対策の事務事業に注力することが有効となりそうである．

これらの結果と解釈はあくまでも 2015 年度の自治体 A のデータをもとにした試行であり，実用化を目指すうえでは，医学的見地からの補強や解釈が困難な矢印への対応方法の検討など，さらなる研究開発が必要である．しかし，本稿で主張したいことは，わが国地域健康政策の文脈に即し，説明容易性を重視した統計手法の選定を行い，かつどの自治体，どの疾病にも広く応用可能な方法論を設計することで，データ中心科学としてのサービス科学の新たな地平

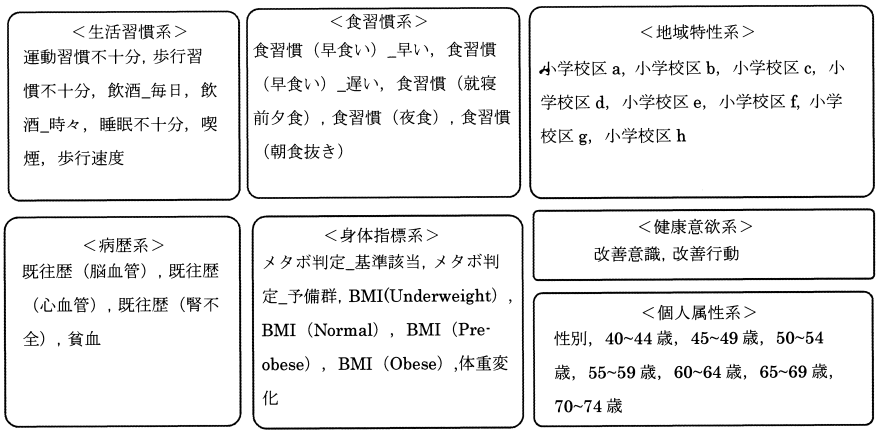


図 1．7 つのグループと変数.

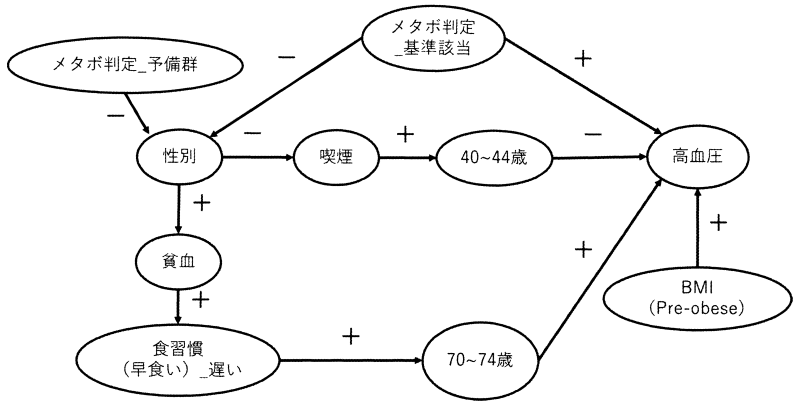


図 2．高血圧の疾病発症ベイジアンネットワーク.

が切り拓かれる可能性があることである．その中で，本稿が考察し試行した疾病発症ベイジアンネットワークの構築方法論は，今後のわが国の地域健康政策において 1 つの有用なやり方であると考えられる．

5. おわりに

本稿では，地域健康政策のためのサービス科学のあり方について，AMED の基盤研究として採択された筑波大学の取り組み事例をもとに議論した．自治体が行う地域健康政策では説明責任が強く求められるため，統計学の初歩的知識しか有さない担当職員にとって説明容易性の高い統計手法を用いる必要がある．また，国が政策的に蓄積してきた健康ビッグデータを前提とし，どの自治体，どの疾病に対しても応用可能な分析方法論を確立する必要がある．本稿では，これらの必要性を満たす統計手法として，制約ベースアプローチで条件付き独立性を χ^2 検定によって行い，より効率的な構造学習が可能な Local to Global アプローチのアルゴリズムを採用したベイジアンネットワークが有用であることを主張した．

次いで本稿では，自治体 A の実際の健康関連ビッグデータを用いて，どの自治体，どの疾病

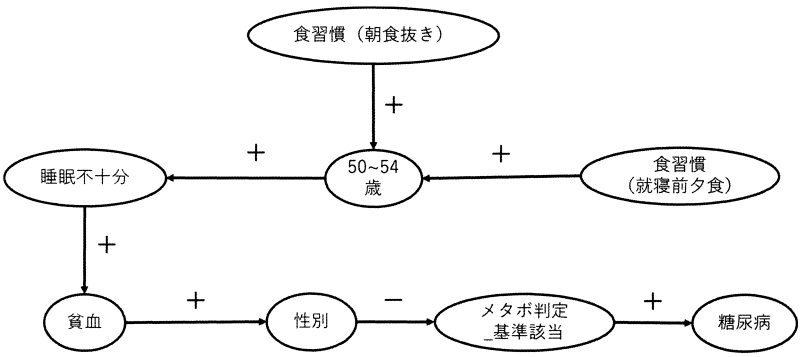


図 3. 糖尿病の疾病発症ベイジアンネットワーク.

に対しても応用可能な疾病発症ベイジアンネットワークの構築を試行した．その結果，因果とは言い難い関係性が一部含まれること，医学や公衆衛生分野の専門家による追加的な因果検証が必要であること，等の課題はあるものの，自治体 A ならではの疾病発症因果モデリングを行うことができた．そして，それを政策立案に活かす糸口を示すことができたと考える．

本稿での取り組み事例をふまえると，地域健康政策におけるデータ中心科学としてのサービス科学は，今後さらなる発展を遂げるであろう．なぜなら，本稿では個別疾病の発症の有無にしか焦点をあてていないが，合併症を含む複数の疾病の発症因果モデリング，各疾病の医療費増減の因果モデリング，社会保険・後期高齢・介護保険のデータをも包括した地域健康因果モデリング，追加的なライフスタイル・アンケートや運動データ等の取得による健康で幸せな生活のための因果モデリングなど，国民の健康長寿と国民医療費の削減との両立に貢献するデータ中心科学の発展が期待できるからである．

健康関連ビッグデータの蓄積・活用と人工知能への注目は，社会課題解決型の応用統計数理についての「学」への期待であり，本稿はその端緒を開こうとする萌芽的な研究ノートにすぎない．超高齢時代のサービス科学として，今後さらなる理論的・実証的研究が蓄積されることが求められる．

注.

[1)] AMED の「AI を活用した保健指導システム研究推進事業」は，2 つ以上の自治体との共同研究開発を条件として公募が行われた．採択結果は 2 件であり，広島大学の「自治体等保険者レセプトデータと健康情報等を基盤に AI を用いてリスク予測やターゲティングを行う保健指導システムの構築に関する研究」と，筑波大学の「自治体における保健指導の施策力に応じた最適な保健指導モデルを提示できる AI の開発研究」が採択された．いずれの取り組みも健康関連ビッグデータと人工知能の活用が想定されており，前者は住民個人への働きかけを想定したミクロ・アプローチ，後者は地域健康政策を想定したマクロ・アプローチとなっている．後者の筑波大学の取り組みは，筑波大学を代表研究機関とし，筑波大学発ベンチャーであるつくばウエルネスリサーチが自治体と連携して構築してきた 75 万人以上の健康関連ビッグデータを基盤に，筑波大学人工知能科学センターと NTT グループの人工知能技術を融合させることにより，自治体の地域健康政策を支援する AI システムを研究開発する．

参 考 文 献

Aliferis, C. F., Tsamardinos, I. and Statnikov, A. (2003). HITON: A novel Markov blanket algorithm for optimal variable selection, *AMIA Annual Symposium Proceedings*, 21–25.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. and Koutsoukos, X. D. (2010a). Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation, *Journal of Machine Learning Research*, **11**, 171–234.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. and Koutsoukos, X. D. (2010b). Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions, *Journal of Machine Learning Research*, **11**, 235–284.

Carriger, J. F., Barron, M. G. and Newman, M. C. (2016). Bayesian networks improve causal environmental assessments for evidence-based policy, *Environmental Science & Technology*, **50**(24), 13195–13205.

Cheng, J., Greiger, R., Kelly, J., Bell, D. and Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach, *Artificial Intelligence*, **137**, 43–90.

Chickering, D. M., Heckerman, D. and Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard, *Journal of Machine Learning Research*, **5**, 1287–1330.

Gao, T., Fadnis, K. and Campbell, M. (2017). Local-to-global Bayesian network structure learning, *Proceedings of the 34th International Conference on Machine Learning*, 1193–1202.

Harris, M. J., Stinson, J. and Landis, W. G. (2017). A Bayesian approach to integrated ecological and human health risk assessment for the south river, Virginia mercury contaminated site, *Risk Analysis*, **37**(7), 1341–1357.

Heckerman, D., Geiger, D. and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, **20**(3), 197–243.

Kalisch, M., Fellinghauer, B. A., Grill, E., Maathuis, M. H., Mansmann, U., Bühlmann, P. and Stucki, G. (2010). Understanding human functioning using graphical models, *BMC Medical Research Methodology*, **10**(1), 14–24.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques*, MIT Press, Cambridge.

厚生労働省 (2009). 政策レポート(特定健康診査(いわゆるメタボ健診)・特定保健指導).

厚生労働省 (2013). 特定健康診査・特定保健指導の円滑な実施に向けた手引 Ver2.0.

厚生労働省 (2016). 我が国の医療保険について.

Liu, Z., Malone, B. and Yuan, C. (2012). Empirical evaluation of scoring functions for Bayesian network model selection, *BMC Bioinformatics*, **13**(15), 1–16.

Madsen, A. L., Jensen, F., Salmerón, A., Langseth, H. and Nielsen, T. D. (2017). A parallel algorithm for Bayesian network structure learning from large data sets, *Knowledge-Based Systems*, **117**, 46–55.

Margaritis, D. and Thrun, S. (2000). Bayesian network induction via local neighborhoods, *Advances in Neural Information Processing Systems*, **12**, 505–511.

Meek, C. (1995). Causal inference and causal explanation with background knowledge, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 403–410.

本村陽一 (2000). ベイジアンネットワーク, 電子情報通信学会誌, **83**(8), 645–646.

Ono, C., Kurokawa, M., Motomura, Y. and Asoh, H. (2007). A context-aware movie preference model using a Bayesian network for recommendation and promotion, *Proceedings of the 11th International Conference on User Modeling*, 247–257.

Park, H. J. and Kim, S. H. (2013). A Bayesian network approach to examining key success factors of mobile games, *Journal of Business Research*, **66**(9), 1353–1359.

Pearl, J. (2003). Causality: Models, reasoning and inference, *Econometric Theory*, **19**, 675–685.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464.

Spirtes, P., Glymour, C. N. and Scheines, R. (2000). *Causation, Prediction and Search*, MIT Press, Cambridge.

Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm, *Machine Learning*, **65**(1), 31–78.

鶴田康人, 寒河江雅彦 (2015). ベイジアンネットワークを用いた階層型少子化因果モデルの構築, 金沢大学ディスカッションペーパー, No.24.

上野眞也 (2010). 地域政策の効果を予測する―ベイジアンネットワーク分析の応用, 熊本大学政策研究, **1**, 29–40.

Velikova, M., Van Scheltinga, J. T., Lucas, P. J. and Spaanderman, M. (2014). Exploiting causal functional relationships in Bayesian network modelling for personalized healthcare, *International Journal of Approximate Reasoning*, **55**(1), 59–73.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models, *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, 255–270.

Yehezkel, R. and Lerner, B. (2009). Bayesian network structure learning by recursive autonomy identification, *Journal of Machine Learning Research*, **10**, 1527–1570.

Applicability of Bayesian Network to Regional Health Policy in Japan

Wataru Toriumi[1], Yuichi Ubukata[1], Shinya Kuno[2,4] and Yukihiko Okada[3,4]

[1]Graduate School of Systems and Information Engineering, University of Tsukuba
[2]Faculty of Health and Sport Sciences, University of Tsukuba
[3]Faculty of Engineering, Information, and Systems, University of Tsukuba
[4]Center for Artificial Intelligence Research, University of Tsukuba

This paper discusses the service science for regional health policy and helps to contribute to the development of service science as data centric science. Because of strong accountability towards residents, it is necessary for municipal officials to use the statistical methods when planning the regional health policy as this would make it easy to explain how and why they choose it. Also, it is necessary to establish a statistical method applicable to any municipality and any disease. In this paper, we propose that the Bayesian network adopting the algorithm of Local to Global approach which enables more efficient structural learning is useful in fulfilling these needs. This algorithm is one of the constraint-based approaches and tests conditional independence with test.

In addition, we develop a disease-causing Bayesian network applicable to any municipality and any disease.

Key words: Regional health policy, artificial intelligence, accountability, ease of explanation, Bayesian network.

統計数理(2018)
第 66 巻 第 2 号 279–294

特集「サービス科学の今」
[原著論文]

集約的シンボリックデータのカイ2乗統計量を用いた非類似度とその不動産情報データへの適用

清水 信夫[1]・中野 純司[1]・山本 由和[2]

(受付 2017 年 11 月 9 日；改訂 2018 年 9 月 18 日；採択 9 月 19 日)

要　　旨

近年，サービス科学においては連続変数とカテゴリー変数が混在している大量のデータが得られることが多い．そしてそれらの個体データはいくつかの自然なグループに分かれる場合がある．そのとき，個々の個体データそのものではなく，その集合であるグループに対する推論および解析に興味があることがある．われわれは，そのようなグループを表すためにいくつかの記述統計量の集合をデータと考え，それを集約的シンボリックデータ(Aggregated Symbolic Data，ASD)と呼ぶ．ここでは，連続変数とカテゴリー変数がともに含まれる場合に，2 次以下のモーメントに関する統計量を ASD と考える．また，連続変数をカテゴリー化することによりすべての変数について同様の基準によるカイ 2 乗統計量を考えた上で，それらの和として ASD 間の非類似度を構成する手法を提案する．そして，この方法を東京都区部の不動産情報データに適用し，各区ごとのデータの集合を考え，それらの ASD を計算する．さらに各 ASD の値から区の間の非類似度を求め，各区の階層的クラスタリングおよび多次元尺度構成法による分析を行う．

キーワード：Burt 行列，カイ 2 乗統計量，階層的クラスタリング，多次元尺度構成法，ビッグデータ．

1. はじめに

近年，サービス産業においては Web システムを用いたデータの収集が多用されており，その活動の詳細なデータが計算機上に連続的に蓄積されるようになっている．それらのデータは連続変数とカテゴリー変数が混在した多次元データであることが多い．また，それらのデータ数は非常に多く，いわゆる“ビッグデータ”の代表例となっている．

このようなデータの全体像を見るためには，個体データに着目するこれまでの方法ではその個数や変数の多さのために計算が困難な場合がある．ただし，そのようなときには個体データは意味のある自然な比較的少数のグループに分かれることが多い．したがって，個体データそのものではなく，個体がまとめられたグループに関心を向けた手法が必要である．その方法の一つとして Diday (1988)はシンボリックデータ解析を提唱した．

シンボリックデータ解析においては，データとして各連続変数ごとに 1 つの値ではなく，ある値を中心としてばらつきをもつデータ(区間データや分布値データ)などで表されるものが考

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 徳島文理大学 理工学部：〒 769–2193 香川県さぬき市志度 1314–1

表 1. 東京都区部の不動産情報データ(一部).

No.	区	賃料 (万円)	面積 (m^2)	物件種別	構造種別	···	管理形態
1	荒川区	8.25	26.83	マンション	鉄筋コンクリート	···	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
557	北区	19.80	89.84	一戸建て	木造	···	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
4588	港区	22.30	71.28	マンション	鉄筋コンクリート	···	巡回管理
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
17512	中央区	21.20	75.46	マンション	鉄骨鉄筋	···	巡回管理
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
498088	足立区	6.40	33.34	アパート	軽量鉄骨	···	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
714202	新宿区	16.40	55.64	マンション	鉄骨鉄筋	···	常駐管理
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮

えられ，それらに従来の各種多変量解析手法を拡張する研究が Bock and Diday (2000)，Billard and Diday (2006)，Diday and Noirhomme-Fraiture (2008)などにまとめられている．それら以外にも，シンボリックデータに対するクラスタリングに関しては Verde (2004)や Irpino and Verde (2006)など，多次元尺度構成法に関しても Denœux and Masson (2000)や Groenen et al. (2006)などの研究がある．これまでのシンボリックデータ解析においては，データは最初から区間のような形で与えられている場合が多く，そこではグループ内の複数の変数間の関係は無視される．例えば，2 つの連続変数間の相関関係は考慮されない．

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

われわれはそのようなビッグデータとして，表 1 で示される大規模な不動産情報データを持っている．このデータはいくつかのグループに自然に分けることができ，ASD の適用が有効なデータと考えられる．

本論文では，グループ内の個体データのそれぞれの変数および複数の変数の組み合わせに関して 2 次までのモーメントに関する統計量を用い，それをグループを表す ASD と考える．第 2 節で連続変数とカテゴリー変数が混在する多次元データにおける ASD を定義し，その意味を考える．第 3 節では ASD 間の非類似度をカイ 2 乗統計量を用いて表す手法を提案する．第 4

節では，提案した手法を表 1 で示される東京都区部における不動産情報データに適用し，23 区ごとのデータを 23 個の ASD として考え，その相互間の非類似度を算出して階層的クラスタリングや多次元尺度構成法を行った結果について考察する．第 5 節では，本研究に関するまとめについて述べる．

2. 集約的シンボリックデータ

ここでは p 個の連続変数および q 個のカテゴリー変数からなる n 個の個体データが与えられている場合を考える．それらの個体データは G 個の自然な意味のあるグループに分かれると仮定する．グループ g に含まれる個体データの数を $n^{(g)}$ とし，グループ g の個体 i の連続変数 l の値を $x_{il}^{(g)}$ と書く．カテゴリー変数 k は m_k 個のカテゴリー値を取るとすると，それは m_k 個のダミー変数で表すことができる．すなわちグループ g の個体 i のカテゴリー変数 k が j 番目のカテゴリー値を取るとき $x_{ij}^{(g,k)}$ は 1，それ以外は 0 とする．するとグループ g のすべての個体データは

$$X^{(g)} = \begin{bmatrix} x_{11}^{(g)} & \cdots & x_{1p}^{(g)} & x_{11}^{(g,1)} & \cdots & x_{1m_1}^{(g,1)} & \cdots & x_{11}^{(g,q)} & \cdots & x_{1m_q}^{(g,q)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{n^{(g)}1}^{(g)} & \cdots & x_{n^{(g)}p}^{(g)} & x_{n^{(g)}1}^{(g,1)} & \cdots & x_{n^{(g)}m_1}^{(g,1)} & & x_{n^{(g)}1}^{(g,q)} & \cdots & x_{n^{(g)}m_q}^{(g,q)} \end{bmatrix} \tag{2.1}$$

と表すことができる．$X^{(g)}$ の最初の p 列からなる部分行列 $X_1^{(g)}$ は p 個の連続変数に対応する．また部分行列

$$X_2^{(g,k)} = \begin{bmatrix} x_{11}^{(g,k)} & \cdots & x_{1m_k}^{(g,k)} \\ \vdots & & \vdots \\ x_{n^{(g)}1}^{(g,k)} & \cdots & x_{n^{(g)}m_k}^{(g,k)} \end{bmatrix} \tag{2.2}$$

はカテゴリー変数 k のダミー変数からなる行列であり，q 個のカテゴリー変数に対応するのは $X_2^{(g)} = [X_2^{(g,1)} \cdots X_2^{(g,q)}]$ である．これらを用いると $X^{(g)} = [X_1^{(g)}\ X_2^{(g)}]$ である．

個体数 $n^{(g)}$ が非常に大きいとき，$X^{(g)}$ を保持し続けるのは記憶領域の制約などのために困難を伴う．またそのような状況で $X^{(g)}$ をそのまま用いた詳細な計算は長い時間がかかり，データの全体像を捉えるという面での意義も乏しい．そこでこのグループを表すいくつかの記述統計量を考え，それを用いてグループに対する統計的推論を行うことにする．

ひとつの連続変数データを集約する簡単な方法は標本平均と標本分散を用いることである．さらに詳しい情報として尖度，歪度を用いることもある．複数の連続変数に関しては標本相関係数も用いられる．これらはモーメントを表す記述統計量である．各グループにおけるデータの情報をモーメントに関する記述統計量で表す場合，低次のモーメントによる統計量だけでは多くの情報が抜け落ちてしまうし，高次のモーメントによる統計量を多く用いると，情報の脱落は少なくなるものの保持すべき値が多くなり扱いが難しくなる．そこで，われわれは，2 次以下のモーメントにより表される記述統計量の集合を考え，これを集約的シンボリックデータ (Aggregated symbolic data，ASD) と呼ぶことにする．

まず，重要な情報としてグループ g の個体数 $n^{(g)}$ が考えられる．これはデータの値の 0 乗 (=1) の合計と考えると 0 次のモーメントと言える．

次に 1 次のモーメントは各変数ごとの和に対応する．これは $X^{(g)}$ に関しては

$$\mathbf{1}'_{n^{(g)}} X^{(g)}/n^{(g)} = [\bar{x}_1^{(g)}, \ldots, \bar{x}_p^{(g)}, \hat{p}_1^{(g,1)}, \ldots, \hat{p}_{m_1}^{(g,1)}, \ldots, \hat{p}_1^{(g,q)}, \ldots, \hat{p}_{m_q}^{(g,q)}] \tag{2.3}$$

と同じ情報である．ここで $\mathbf{1}'_{n^{(g)}}$ はすべての成分が 1 である $n^{(g)}$ 次元横ベクトルを表す．明らかにこれらは各連続変数の平均および各カテゴリー変数の周辺分布のパラメータである．

さらに 2 次のモーメントは(2.1)式より

$$X^{(g)'}X^{(g)} = \begin{bmatrix} X_1^{(g)'}X_1^{(g)} & X_1^{(g)'}X_2^{(g)} \\ X_2^{(g)'}X_1^{(g)} & X_2^{(g)'}X_2^{(g)} \end{bmatrix} \equiv \begin{bmatrix} S_{11}^{(g)} & S_{12}^{(g)} \\ S_{21}^{(g)} & S_{22}^{(g)} \end{bmatrix} \tag{2.4}$$

を考えることになる．$S_{11}^{(g)}$ は連続変数データの積和行列である．$X_2^{(g,k_1)'}X_2^{(g,k_2)} = S^{(g,k_1k_2)}$ がカテゴリー変数 k_1 とカテゴリー変数 k_2 に対する分割表となることを考えると，$S_{22}^{(g)}$ はそれらを部分行列とする Burt 行列である．$S_{12}^{(g)}$ は連続変数とカテゴリー変数に関する 2 次のモーメントを表す $p \times (m_1 + \cdots + m_q)$ 行列であるが，その第 k 部分行列 $X_1^{(g)'}X_2^{(g,k)}$ の (l, j) 成分はカテゴリー変数 k の値がカテゴリー値 j を取る個体における連続変数 l の合計である．

われわれはこれらの記述統計量で各グループの特徴が表されていると考え，これを用いてグループの関係を調べることにする．

3. ASD 間の非類似度

ここで考えているデータは連続変数とカテゴリー変数の両方を含む．これらを統一的に考える方法の一つとして，連続変数をカテゴリー変数に変換する．そして 2 つのグループのカテゴリー変数の分布の差を考えるためにカイ 2 乗統計量を用いることにする．

3.1 2 つのカテゴリー変数の組み合わせに関する非類似度

まず，2 つの ASD g_1, g_2 における異なる 2 つのカテゴリー変数 (k_1, k_2) の組により形成される分割表の間の非類似度を考える．2 つのグループの分割表は $S^{(g_1,k_1k_2)} = [s_{j_1j_2}^{(g_1,k_1k_2)}]$，$S^{(g_2,k_1k_2)} = [s_{j_1j_2}^{(g_2,k_1k_2)}]$ であり，$j_a = 1, \ldots, m_{k_a}$ $(a = 1, 2)$ である．2 つの ASD が同じ性質を持つ場合，それぞれの分割表の各セルの出現確率は等しい．この仮定が正しい場合，分割表のセル (j_1, j_2) の出現個数の期待値の推定量は

$$\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})} = \frac{s_{j_1j_2}^{(g_1,k_1k_2)} + s_{j_1j_2}^{(g_2,k_1k_2)}}{n^{(g_1)} + n^{(g_2)}} n^{(g_a)} \quad (a = 1, 2)$$

と考えられる．一方，2 つの ASD が異なる場合にはそれぞれの分割表の各セルにおける出現個数 $s_{j_1j_2}^{(g_a,k_1,k_2)}$ と $\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})}$ よりカイ 2 乗統計量を求めることができ，その総和を非類似度と考える．この場合，$\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})}$ が分母になるので，これが 0 になるときは無視しなくてはならない．すなわち $s_{j_1j_2}^{(g_1,k_1k_2)} = s_{j_1j_2}^{(g_2,k_1k_2)} = 0$ となるセルは無視してカイ 2 乗統計量を考える．これより

$$\chi^{2(g_1g_2,k_1k_2)} = \sum_{\substack{a=1}}^{2} \sum_{j_1=1}^{m_{k_1}} \sum_{j_2=1}^{m_{k_2}} \frac{\left\{ s_{j_1j_2}^{(g_a,k_1,k_2)} - \widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})} \right\}^2}{\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})}} \tag{3.1}$$
$$s_{j_1j_2}^{(g_1,k_1k_2)} + s_{j_1j_2}^{(g_2,k_1k_2)} \geq 1$$

を (k_1, k_2) の組による分割表における ASD 間の非類似度と考えることができる．これを $k_1 < k_2$ なる全ての (k_1, k_2) に関して考え，その総和をとったものが Burt 行列における ASD 間の非類似度

$$d_{(cc)}^{(g_1g_2)} = \sum_{k_1=1}^{q-1} \sum_{k_2=k_1+1}^{q} \chi^{2(g_1g_2,k_1k_2)} \tag{3.2}$$

と考えられる．なお，Burt 行列の対角成分の違いは考えていないことに注意する．それは Burt 行列の対角成分は各カテゴリー変数の周辺分布を表し，その情報は Burt 行列の非対角成分にある分割表の列和，行和として含まれているからである．

3.2　2 つの連続変数の組み合わせに関する非類似度

2 つの連続変数 l_1, l_2 のデータは 2 次元平面上にプロットできる．その平面を格子状に分割し，それぞれの格子をカテゴリー値と考える．2 次元平面 $(-\infty,\infty)\times(-\infty,\infty)$ を各次元ごとに N 個ずつの区間に分割するとし，その各区間の境界値 $h_j^{(l)}$ $(j=0,1,\ldots,N)$ について

$$-\infty = h_0^{(l)} < h_1^{(l)} < \cdots < h_{N-1}^{(l)} < h_N^{(l)} = \infty$$

とする．われわれは ASD のみを保持していると考えるので，各格子内に何個のデータがあるかの情報は持っていない．持っているのはグループ g に含まれる個体数 $n^{(g)}$ および標本平均 $\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)} = \begin{bmatrix} \hat{\mu}_{l_1}^{(g)} \\ \hat{\mu}_{l_2}^{(g)} \end{bmatrix}$ と標本分散共分散行列 $\hat{\Sigma}_{l_1l_2}^{(g)} = \begin{bmatrix} \hat{\sigma}_{l_1l_1}^{(g)} & \hat{\sigma}_{l_1l_2}^{(g)} \\ \hat{\sigma}_{l_2l_1}^{(g)} & \hat{\sigma}_{l_2l_2}^{(g)} \end{bmatrix}$ である．したがって，これらを用いて各セルの個体数を推定することを考える．

これだけの情報からだと，連続変数 l_1, l_2 の実現値 $\boldsymbol{x}_{l_1l_2} = [x_{l_1}, x_{l_2}]'$ の同時分布は標本平均が $\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)}$，標本分散共分散行列が $\hat{\Sigma}_{l_1l_2}^{(g)}$ である 2 変量正規分布に従うと仮定するのが自然である．なお，連続変数が従う確率分布について非対称性などのより複雑な状況を考えるには，3 次以上のモーメントにあたる情報が必要となるので，ここでの ASD を用いる限り考慮することはできない．たとえば，各個体の連続変数の値の分布の非対称性が強い場合は，変数変換を行うなどして非対称性を可能な限り弱め，対称性が担保されているとみなせる変数にした方がよい．そのような変換により，ある程度は外れ値の影響を軽減することができる．

この 2 変量正規分布の密度関数を $\varphi(\boldsymbol{x}_{l_1l_2}|\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)}, \hat{\Sigma}_{l_1l_2}^{(g)})$ と書くと，ASD g において領域であるカテゴリー $[h_{j_1}^{(l_1)}, h_{j_1+1}^{(l_1)}] \times [h_{j_2}^{(l_2)}, h_{j_2+1}^{(l_2)}]$ における出現確率は

$$\hat{p}_{j_1j_2}^{(g,l_1l_2)} = \iint_{[h_{j_1}^{(l_1)}, h_{j_1+1}^{(l_1)}]\times[h_{j_2}^{(l_2)}, h_{j_2+1}^{(l_2)}]} \varphi\left(\boldsymbol{x}_{l_1l_2}|\,\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)}, \hat{\Sigma}_{l_1l_2}^{(g)}\right) d\boldsymbol{x}_{l_1l_2}$$

となる．これより，異なる 2 つの連続変数 (l_1, l_2) の組による $N\times N$ 分割表をそれぞれ $S^{(g_1,l_1,l_2)} \simeq [\hat{p}_{j_1j_2}^{(g_1,l_1l_2)} n^{(g_1)}]$，$S^{(g_2,l_1,l_2)} \simeq [\hat{p}_{j_1j_2}^{(g_2,l_1l_2)} n^{(g_2)}]$ と近似できる．

2 つの ASD が同じ性質を持つ場合，それぞれの分割表の各セルの出現確率は等しいと仮定できる．これが正しい場合，セル (j_1, j_2) の出現個数の期待値の推定量は

$$\widehat{E(s_{j_1j_2}^{(g_a,l_1l_2)})} = \frac{\hat{p}_{j_1j_2}^{(g_1,l_1l_2)} n^{(g_1)} + \hat{p}_{j_1j_2}^{(g_2,l_1l_2)} n^{(g_2)}}{n^{(g_1)} + n^{(g_2)}} n^{(g_a)} \quad (a=1,2)$$

と書ける．カテゴリー変数同士の組み合わせの場合と同様の基準でカイ 2 乗統計量を考える場合，$\widehat{E(s_{j_1j_2}^{(g_a,l_1l_2)})}$ で割ることになるので，これが 0 もしくは極端に小さな値となるときは無視しなくてはならない．そこで，$\hat{p}_{j_1j_2}^{(g_1,l_1l_2)} n^{(g_1)} + \hat{p}_{j_1j_2}^{(g_2,l_1l_2)} n^{(g_2)} < 1$ となるセルは無視してカイ 2 乗統計量を考える．これより

$$\chi^{2(g_1g_2,l_1l_2)} \simeq \sum_{a=1}^{2}\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}{}_{\hat{p}^{(g_1,l_1l_2)}_{j_1j_2}n^{(g_1)}+\hat{p}^{(g_2,l_1l_2)}_{j_1j_2}n^{(g_2)}\geq 1} \frac{\left\{\hat{p}^{(g_a,l_1l_2)}_{j_1j_2}n^{(g_a)} - E(\widehat{s^{(g_a,l_1l_2)}_{j_1j_2}})\right\}^2}{E(\widehat{s^{(g_a,l_1l_2)}_{j_1j_2}})} \tag{3.3}$$

を非類似度と考えることができる．これを $l_1 < l_2$ なる全ての (l_1, l_2) に関して考え，その総和をとったものが連続変数に関する ASD 間の非類似度

$$d^{(g_1g_2)}_{(rr)} = \sum_{l_1=1}^{p-1}\sum_{l_2=l_1+1}^{p}\chi^{2(g_1g_2,l_1l_2)} \tag{3.4}$$

と考えられる．

残った問題は各区間の境界値 $h^{(l)}_j$ $(j = 0, 1, \ldots, N)$ の定め方である．非類似度はすべてのグループのペアに対して計算しなければならないので，統一性のためにもこの境界値は同一のものを利用するのがよい．そこでわれわれはすべてのデータに関する連続変数 l に対する平均と分散を考え，それを用いた正規分布の確率が同じになるように境界値を取ることにする．すなわち分割数 N に関して

$$\hat{\mu}_l = \frac{1}{n}\sum_{g=1}^{G} n^{(g)}\hat{\mu}^{(g)}_l$$

$$\hat{\sigma}_{ll} = \frac{1}{n}\sum_{g=1}^{G} n^{(g)}\hat{\sigma}^{(g)}_{ll} + \frac{1}{n}\sum_{g=1}^{G} n^{(g)}(\hat{\mu}^{(g)}_l - \hat{\mu}_l)^2$$

を用いた 1 次元正規分布 $\varphi(x_l|\hat{\mu}_l, \hat{\sigma}_{ll})$ が各区間でそれぞれ $1/N$ ずつの確率を持つように $h^{(l)}_j$ を定める．すなわち $h^{(l)}_j = \hat{\mu}_l + \hat{\sigma}_{ll}\boldsymbol{\Phi}^{-1}(j/N)$ となるように取る．ただし $\boldsymbol{\Phi}(x_l)$ は標準正規分布の分布関数である。N の値については，小さくしすぎると分布の特徴がとらえられず，一方で大きくしすぎると各セル内のデータ個数が少なくなり，セル数の増加に伴い計算時間の増大につながる．そのため，適当な範囲内でいくつかの場合に対する結果を求め，その中で適当なものを選べばよい．

3.3 連続変数とカテゴリー変数の組み合わせの場合

連続変数 l とカテゴリー変数 k のペアを考える．連続変数に関しては前節と同様にカテゴリー化する．このペアの場合の 2 次モーメントは(2.4)式の $S^{(g)}_{12}$ に対応するが，この中にはカテゴリー変数 k の各カテゴリー値が j_2 となる個体における連続変数 l の標本分散に対応する値は含まれない．そのため，この場合の標本分散に関しては全て同一の値，すなわち $\hat{\sigma}^{(g)}_{ll}$ を使用せざるを得ないことに注意する．

ここで保持する情報からだと，カテゴリー変数 k のカテゴリー値が j_2 である場合の連続変数 l の実現値 $x^{(g,k)}_{j_2l}$ の分布は，標本平均が $\hat{\mu}^{(g,k)}_{j_2l}$，標本分散が $\hat{\sigma}^{(g)}_{ll}$ である 1 変量正規分布に従うと仮定するのが自然であり，その密度関数を $\varphi(x^{(g,k)}_{j_2l}|\hat{\mu}^{(g,k)}_{j_2l}, \hat{\sigma}^{(g)}_{ll})$ と書く．すると ASD g において区間であるカテゴリー $[h^{(l)}_{j_1}, h^{(l)}_{j_1+1}]$ における出現確率は

$$\hat{p}^{(g,lk)}_{j_1j_2} = \int_{h^{(l)}_{j_1}}^{h^{(l)}_{j_1+1}} \varphi\left(x^{(g,k)}_{j_2l}\middle|\hat{\mu}^{(g,k)}_{j_2l}, \hat{\sigma}^{(g)}_{ll}\right) dx^{(g,k)}_{j_2l}$$

となる．ただし，$h^{(l)}_{j_1}$ は前節と同じものである．これより，連続変数とカテゴリー変数 (l, k) の組による分割表をそれぞれ $S^{(g_1,lk)} \simeq [\hat{p}^{(g_1,lk)}_{j_1j_2}n^{(g_1,k)}_{j_2}]$，$S^{(g_2,lk)} \simeq [\hat{p}^{(g_2,lk)}_{j_1j_2}n^{(g_2,k)}_{j_2}]$ で近似できる．

2 つの ASD が同じ性質を持つ場合，それぞれの分割表の各セルの出現確率は等しい．これが正しい場合，セル (j_1, j_2) の出現個数の期待値の推定量は

$$\widehat{E(s_{j_1j_2}^{(g_a,lk)})} = \frac{\hat{p}_{j_1j_2}^{(g_1,lk)} n_{j_2}^{(g_1,k)} + \hat{p}_{j_1j_2}^{(g_2,lk)} n_{j_2}^{(g_2,k)}}{n_{j_2}^{(g_1,k)} + n_{j_2}^{(g_2,k)}} n_{j_2}^{(g_a,k)}$$

と考えられる．前の 2 つの節と同様に，カイ 2 乗統計量を考える場合，$\widehat{E(s_{j_1j_2}^{(g_a,lk)})}$ で割ることになるので，これが 0 もしくは極端に小さな値となるときは無視しなくてはならない．カテゴリー変数同士の組み合わせの場合と同様の基準で考えるため，$n_{j_2}^{(g_1,k)}$ と $n_{j_2}^{(g_2,k)}$ のうち少なくとも 1 つが 0 である場合，およびそれらがいずれも正であっても $\hat{p}_{j_1j_2}^{(g_1,lk)} n_{j_2}^{(g_1,k)} + \hat{p}_{j_1j_2}^{(g_2,lk)} n_{j_2}^{(g_2,k)} < 1$ となる場合のセルは無視してカイ 2 乗統計量を考える．これより

$$\chi^{2(g_1g_2,lk)} \simeq \sum_{\substack{a=1}}^{2} \sum_{j_1=1}^{N} \sum_{j_2=1}^{m_k} \frac{\left\{ \hat{p}_{j_1j_2}^{(g_a,lk)} n_{j_2}^{(g_a,lk)} - \widehat{E(s_{j_1j_2}^{(g_a,lk)})} \right\}^2}{\widehat{E(s_{j_1j_2}^{(g_a,lk)})}} \tag{3.5}$$

$$\hat{p}_{j_1j_2}^{(g_1,lk)} n_{j_2}^{(g_1,k)} + \hat{p}_{j_1j_2}^{(g_2,lk)} n_{j_2}^{(g_2,k)} \geq 1,$$
$$n_{j_2}^{(g_1,k)} n_{j_2}^{(g_2,k)} > 0$$

を非類似度と考えることができる．これを全ての (l, k) に関して考え，その総和をとったものが 2 つのグループの連続変数とカテゴリー変数間の非類似度

$$d_{(rc)}^{(g_1g_2)} = \sum_{l=1}^{p} \sum_{k=1}^{q} \chi^{2(g_1g_2,lk)} \tag{3.6}$$

と考えられる．

3.4　すべての変数を用いた非類似度

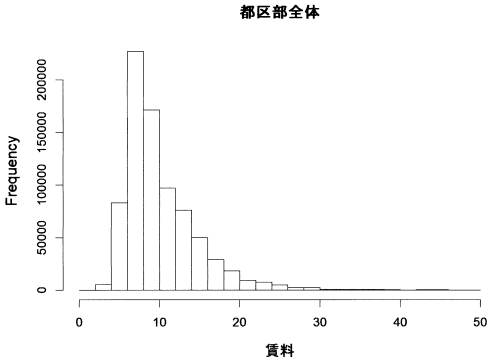
連続変数をカテゴリー変数化して考えることにより，(3.2)，(3.4)，(3.6)式は全てカテゴリー変数同士の組み合わせによる非類似度と考えられるので

$$d^{(g_1g_2)} = d_{(cc)}^{(g_1g_2)} + d_{(rr)}^{(g_1g_2)} + d_{(rc)}^{(g_1g_2)}$$

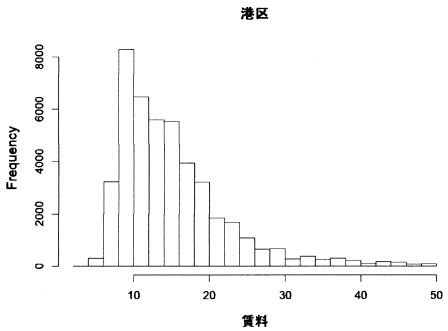
が ASD 間の全体のカイ 2 乗統計量に基づく非類似度となる．これは $X^{(g_1)'}X^{(g_1)}$ と $X^{(g_2)'}X^{(g_2)}$ の間の非類似度を計算したことになる．これらの行列はいずれも対称行列であり，対角成分よりも上側に位置する成分だけを考えていることに注意する．また，対角成分の情報は，対角成分よりも上側に位置する成分にほとんどが含まれていることより，陽には計算されていないことも注意しなければならない．

4.　不動産情報データへの適用

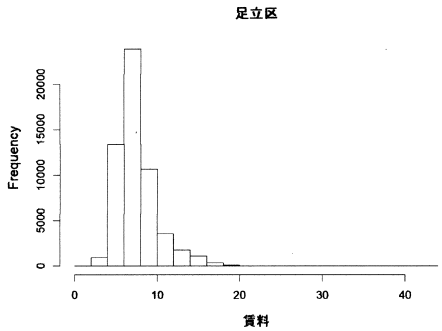
ここでは，各変数において欠測値が含まれる物件，書き間違いおよび外れ値と考えられる値が含まれる物件，他の変数への従属性が高いと考えられる変数をあらかじめ削除した後の東京都区部の不動産情報データにわれわれが提案した手法を適用し解析する．表 1 はそのデータの一部であるが，全体では 2 個の連続変数および 79 個のカテゴリー変数からなる，合わせて約 79 万件の賃貸物件情報である．なお，変数のうち「管理費」「礼金月数」「敷金月数」については元の数値がいずれも賃料を基準とする月数で表されているが，値が 0 である物件がいずれの変数でも全体の 14% 以上存在し，連続変数として考えるための適切な変数変換が難しいことか



(a)都区部全体



(b)港区



(c)足立区

図 1. 賃料のヒストグラム.

ら，ここでは各変数を少数のカテゴリー値をもつカテゴリー変数として考える．

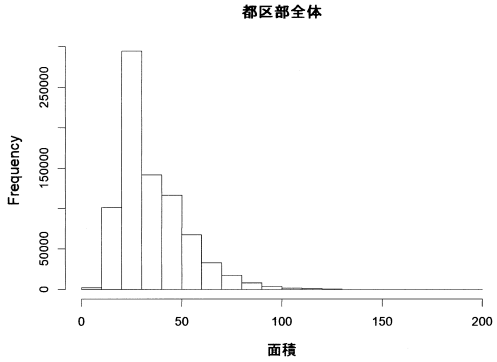
このデータにおける 2 個の連続変数「賃料」および「面積」に関し，例として都区部全体，港区，足立区それぞれの場合におけるヒストグラムを図 1(a)～(c)および図 2(a)～(c)に示す．

図 1 および図 2 より，両変数の値の分布には少なからず非対称性がみられる．そこで，このような各連続変数値に対し，各変数ごとに対数をとることを考える．対数変換後のヒストグラムの例を図 3(a)～(c)および図 4(a)～(c)に示す．

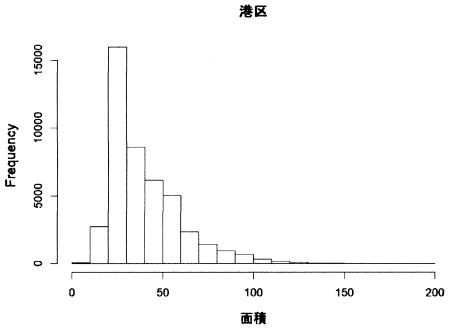
図 3 および図 4 より，対数変換後の各連続変数の値は，元の値の場合よりも非対称性が弱くなっている．そこで，以下ではこれらの値を連続変数の値として ASD を考えるものとする．

データを探索的に解析するにあたり，「区」というカテゴリー変数により各区ごとに 23 のグループに分けて考え，その ASD 間の非類似度を，連続変数をカテゴリー変数化するための分割数 N が 3 ～ 9 の場合においてそれぞれ計算した．そして，それらを用いてまず階層的クラスタリングを行う．ここでは最長距離法 (Defays, 1977)，最短距離法 (Sibson, 1973)，群平均法 (Sokal and Michener, 1958)の 3 種類の手法を用いる．なお，連続変数を含む部分の非類似度に関しては N の値により異なる値が得られることに注意する．

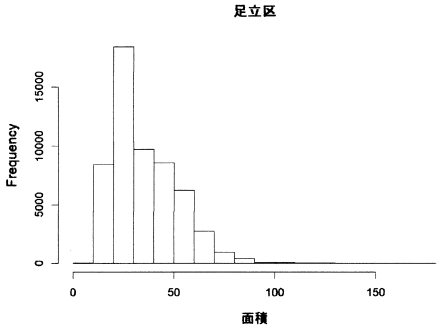
階層的クラスタリングにおいては，連続変数を含む部分のみの非類似度を用いた場合，全ての変数の組み合わせの非類似度を用いた場合についてのいずれでも，$N = 4$ 以下では N の値により大きな変化があったのに対し，$N = 5$ 以上ではどの N についても手法ごとの結果に大きな変化がみられなかった．そのため，ここでは $N = 5$ の場合を主として考えることにする．



(a)都区部全体



(b)港区



(c)足立区

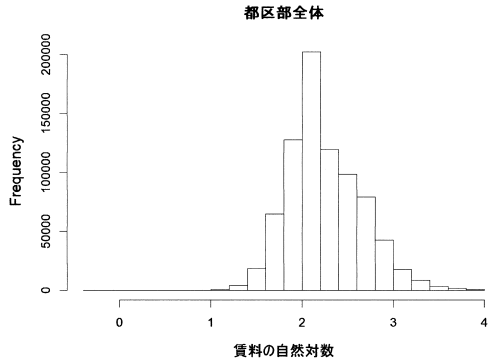
図 2．面積のヒストグラム．

この場合の，全ての変数の組み合わせの非類似度にそれぞれの手法を適用した結果が図 5(a)～(c)である．

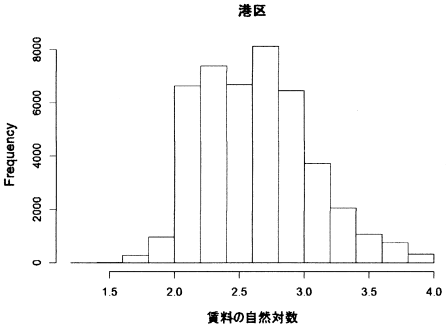
次に，$N = 5$ の場合の全ての変数の組み合わせの非類似度に多次元尺度構成法を適用する．その結果を図 6 に示す．$1 \leq g_1 < g_2 \leq 23$ に対し，図 6 における各 ASD 間のユークリッド距離 $\hat{d}^{(g_1 g_2)}$ と元々の非類似度行列における非類似度 $d^{(g_1 g_2)}$ の相関係数の 2 乗値(決定係数)は $R^2 = 0.996$ となり，非類似度行列による布置が高い精度で保持されていることがわかる．

階層的クラスタリングにおいてはクラスターをまとめるための基準により手法が特徴づけられる(Lance and Williams, 1966)ため，図 5(a)～(c)にも示されている通り，デンドログラム全体の形状については手法により多少の差異が見られる部分があるものの，どの手法でも共通する組み合わせが複数存在し，それらはそれぞれクラスターとみなせる．また図 6 における布置の全体的な形状からいくつかのクラスターが読み取れる．これらを合わせて考えると，概ね(1)中央区および港区，(2)足立区，(3)千代田区，新宿区，文京区，台東区，墨田区，江東区，品川区および渋谷区，(4)目黒区，大田区，豊島区および荒川区，(5)世田谷区，中野区，杉並区，北区，板橋区，練馬区，葛飾区および江戸川区，がそれぞれクラスターとみなせる．

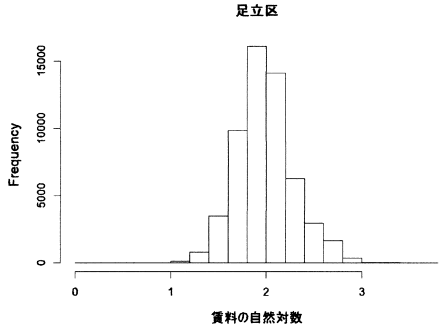
クラスター(1)およびクラスター(2)は他のクラスターと特に大きく異なる特徴があると考えられることから，後に詳細に考察する．それら以外のクラスターについて，クラスター(3)は東京都区部の中心から外側に向けて伸びる各鉄道路線の起点駅のうち特に乗降客の多い各駅(新宿駅，渋谷駅，上野駅など)が含まれる区およびその隣接区域に位置する各区が集まっている．



(a)都区部全体



(b)港区



(c)足立区

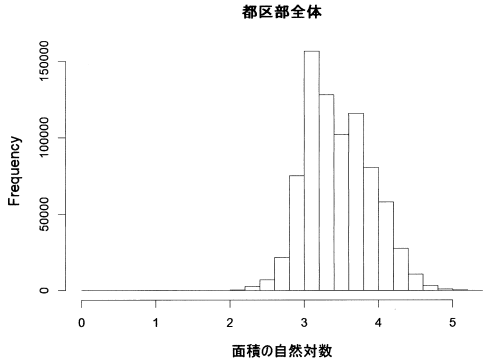
図 3. 賃料の対数のヒストグラム.

ただし，このクラスターには西武池袋線や東武東上線の起点駅として乗降客の多い池袋駅が含まれる豊島区が含まれていないのが興味深い．クラスター(5)は各区の地理的な位置より，東京都区部の中での外縁部およびその隣接区域に位置する区が集まっている．クラスター(4)はクラスター(3)とクラスター(5)の中間に位置する区の集合と考えられる．

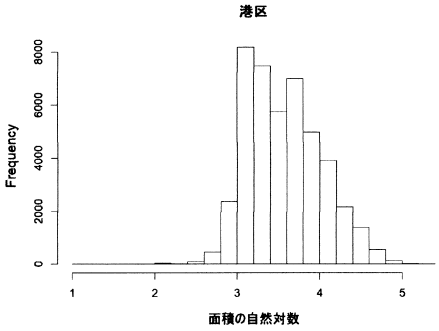
次に，中央区，港区，足立区の 3 区を選び，この不動産情報データにおけるいくつかの特徴的な 2 変数の組に関して考察する．

まず，カテゴリー変数のうち物件種別および構造種別の 2 つの組による分割表の各区ごとの違いを図示したものが図 7(a)～(c)である．これは物件数が多い組み合わせほど濃い色となるように表示した，いわゆるヒートマップである．また(b) (c)それぞれの図の下に，この 2 つの変数の組に関して中央区を基準とした各区への非類似度を記す．

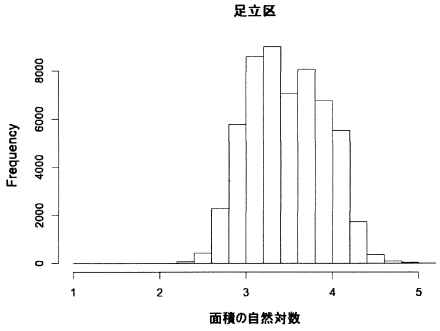
物件種別(当初のカテゴリー数は 5)においては，カテゴリー値 1 がマンション，カテゴリー値 2 がアパートであり，それ以外の種別を全てその他としてカテゴリー値 3 にまとめた．構造種別(当初のカテゴリー数は 10)においては，カテゴリー値 1 が鉄筋コンクリート造り，カテゴリー値 2 が鉄骨鉄筋造りなど，カテゴリー値 3 が鉄骨造りなど，カテゴリー値 4 が軽量鉄骨造りなど，カテゴリー値 5 が木造，カテゴリー値 6 がその他の 6 つのカテゴリーに集約した．この図より，中央区と港区における物件は鉄筋コンクリート造りもしくは鉄骨鉄筋造りなどのマンションが大半を占め，類似性が高いことがわかる．また，区の物件全体に占める鉄筋コンクリート造りの物件の割合は港区の方が中央区よりも高いことが読み取れる．一方で，足立区は



(a)都区部全体



(b)港区



(c)足立区

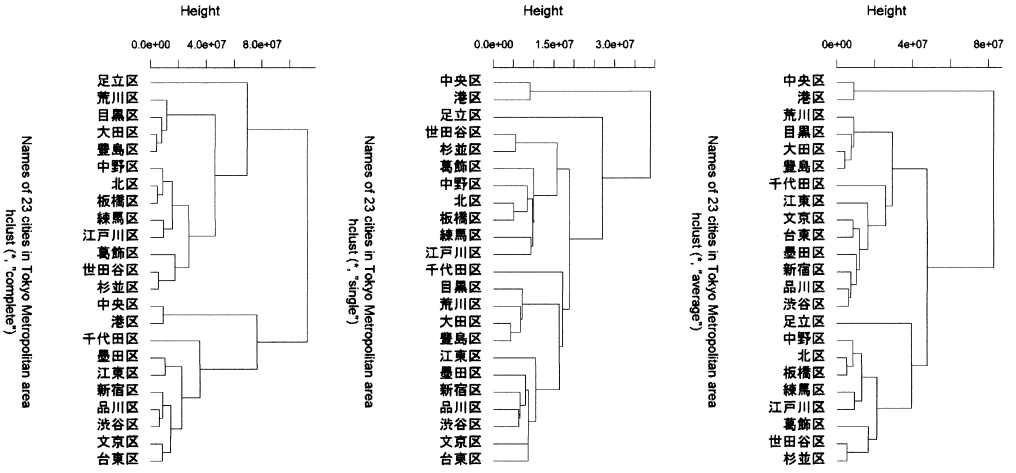
図 4. 面積の対数のヒストグラム.

鉄筋コンクリート造りもしくは鉄骨鉄筋造りなどのマンションの割合が中央区や港区と比べて低く，鉄骨造りのマンション，軽量鉄骨造りのアパート，木造アパートがそれぞれ一定の割合を占めている．すなわち，足立区においては物件の状況が中央区や港区とは大きく異なることがこの組み合わせからわかる．

さらに，中央区を基準とした各区への非類似度は，中央区と港区との間の値よりも中央区と足立区との間の値の方が極めて大きくなっており，足立区が中央区や港区と異なる状況に対応していると考えられる．

次に，連続変数に関して考える．表 2 は面積および賃料それぞれの対数に関し，各区ごとの各変数の平均および標準偏差，2 つの変数の相関係数，およびこの 2 つの変数の組に関して中央区を基準とした各区への非類似度をまとめたものである．これより，各変数の平均値は中央区および港区が足立区より高く，2 つの変数の相関係数も中央区と港区において足立区よりも大きいことがわかる．そして，この 2 つの変数に関する非類似度は，中央区と港区との間の値よりも中央区と足立区との間の値の方が極めて大きくなっており，ここでも足立区が中央区や港区と異なる状況に対応していると考えられる．

さらに，連続変数とカテゴリー変数の組に関しても考察する．表 3 はカテゴリー変数「物件種別」における各カテゴリー値ごとの件数と賃料の対数の平均，およびこの 2 つの変数の組に関して中央区を基準とした各区への非類似度を示したものである．この表より，3 区いずれにおいてもマンションの物件数の比率が高く，マンションの賃料の対数の平均が各区ごとの全体



(a)最長距離法 (b)最短距離法 (c)群平均法

図 5. 不動産情報データの集約的シンボリックデータの階層的クラスタリング($N = 5$).

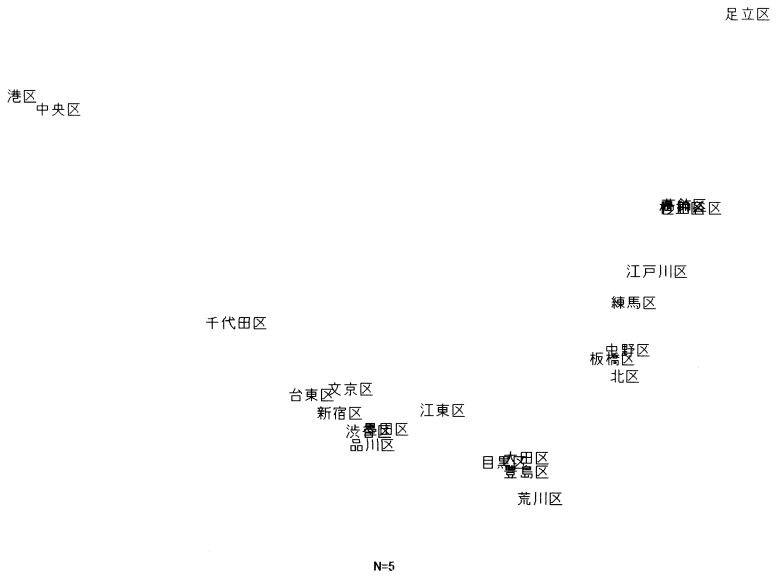
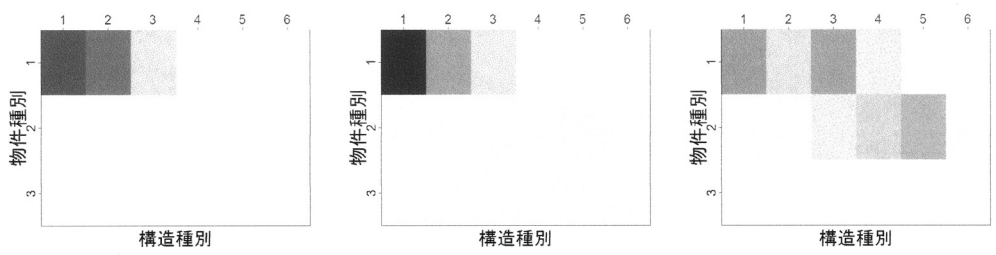


図 6. 不動産情報データの集約的シンボリックデータの多次元尺度構成法の布置($N = 5$).

の物件の賃料の対数の平均(表 2 参照)に近くなっている．各カテゴリー値ごとの賃料の対数の平均に注目すると，どのカテゴリーでも港区および中央区の値が足立区よりも高くなっている．そして，この 2 つの変数に関する非類似度も，中央区と港区との間の値よりも中央区と足立区との間の値の方が極めて大きくなっており，ここでも足立区が中央区や港区と異なる状況



-　　$\chi^{2(g_1g_2,k_1k_2)} = 2.32\times10^3$　　$\chi^{2(g_1g_3,k_1k_2)} = 3.47\times10^4$

(a) 中央区　　(b) 港区　　(c) 足立区

図 7. 物件種別と構造種別による分割表のヒートマップおよび中央区を基準とした非類似度（k_1：物件種別，k_2：構造種別，g_1：中央区，g_2：港区，g_3：足立区）．

表 2. 賃料と面積それぞれの対数に関する各区ごとの平均および分散，各区ごとの相関係数，および中央区を基準とした非類似度（$N = 5$）（l_1：賃料の対数，l_2：面積の対数，g_1：中央区，g_2：港区，g_3：足立区）．

区	賃料の対数 (万円)		面積の対数 (m^2)		相関係数	非類似度
	平均	標準偏差	平均	標準偏差		
中央区	2.492	0.3412	3.549	0.4045	0.9328	-
港区	2.629	0.4203	3.560	0.4557	0.9205	$\chi^{2(g_1g_2,l_1l_2)} = 5.128\times10^3$
足立区	1.990	0.2087	3.471	0.4365	0.8449	$\chi^{2(g_1g_3,l_1l_2)} = 5.796\times10^4$

表 3. 物件種別の各カテゴリー値ごとの件数，賃料の対数の平均（単位：万円）および中央区を基準とした非類似度（$N = 5$）（l：賃料の対数の平均，k：物件種別，g_1：中央区，g_2：港区，g_3：足立区）．

区	マンション		アパート		その他		非類似度
	件数	賃料の対数の平均	件数	賃料の対数の平均	件数	賃料の対数の平均	
中央区	29886	2.492	25	2.322	27	2.655	-
港区	43824	2.634	587	2.234	86	3.045	$\chi^{2(g_1g_2,lk)} = 1.275\times10^3$
足立区	37738	2.064	17408	1.817	707	2.303	$\chi^{2(g_1g_3,lk)} = 1.085\times10^4$

に対応していると考えられる．

なお，本論文で解析した不動産情報データについては，本節の最初に述べた通り，連続変数の数に対してカテゴリー変数の数が圧倒的に多いため，どの g_1 および g_2 についても，$d^{(g_1g_2)}_{(rr)}$ および $d^{(g_1g_2)}_{(rc)}$ よりも $d^{(g_1g_2)}_{(cc)}$ の値の方がずっと大きくなる．すなわち，カテゴリー変数同士の

みの非類似度から導出される構造が大きく影響している．

5. おわりに

本論文では，連続変数とカテゴリー変数が混在する多次元データ集合がいくつかのグループに分かれていると考え得る場合に，各グループを単位とする解析手法として集約的シンボリックデータ(ASD)解析法を提案し，その適用例として東京都区部の不動産情報データを解析した．そして，各 ASD 間の非類似度をカイ 2 乗統計量を用いて表す手法を提案し，これに対する階層的クラスタリングおよび多次元尺度構成法を，不動産情報データにおける各区ごとのグループを ASD と考えて適用し，いくつかの特徴的なクラスター構造を発見した．また特徴的ないくつの区に関して各変数の統計量および区の間の非類似度を計算し，区の間の差異について考察した．

連続変数とカテゴリー変数が混在する多次元データ場合については，連続変数を各グループ共通の少数の領域に離散化して考えることで，連続変数を含む組み合わせについても近似した分割表の組み合わせからなる Burt 行列を各 ASD ごとに考えることができ，全ての変数の組み合わせをカテゴリー変数同士の組み合わせとして統一的に考えることが可能になる．

異なる ASD 間の全体の非類似度については，データ集合の変数の数が多くなればなるほど大きな値となるため，冗長な変数が存在していると望ましい非類似度の値よりも大きな値が算出される．そのため，非類似度を計算するための適切な変数選択の手法の開発が今後の課題と考えられる．

謝　辞

本研究は，統計数理研究所共同研究(課題番号 27-共研-4204 および 28-共研-4105，研究代表者 清水信夫)，および科学研究費補助金基盤研究(C)（課題番号 26330054，研究代表者 中野純司)より支援を受けました．本論文で使用した不動産情報データは，(株)リクルート住まいカンパニーより提供を受けました．本論文の 2 名の査読者からは，内容に関して非常に有益なコメントを多く頂きました．深く感謝致します．

参 考 文 献

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis*: *Conceptual Statistics and Data Mining*, John Wiley & Sons Ltd, Chichester, UK.

Bock, H.-H. and Diday, E. (2000). *Analysis of Symbolic Data*: *Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin.

Defays, D. (1977). An efficient algorithm for a complete link method, *The Computer Journal*, **20**(4), 364–366, doi:10.1093/comjnl/20.4.364.

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

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis: The basic choices, In: Bock, H.-H. (ed.), *Classification and Related Methods of Data Analysis*, Proceedings of IFCS-87, Aachen, July 1987, North-Holland, Amsterdam, 673–684.

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

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

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

Lance, G. N. and Williams, W. T. (1966). Computer programs for hierarchical polythetic classification ("similarity analyses"), *The Computer Journal*, **9**(1), 60–64.

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single-link cluster method, *The Computer Journal*, **16**(1), 30–34, doi:10.1093/comjnl/16.1.30.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin*, **38**, 1409–1438.

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

Dissimilarity between Aggregated Symbolic Data Using Chi-squared Statistics and Its Application to Real Estate Data

Nobuo Shimizu[1], Junji Nakano[1] and Yoshikazu Yamamoto[2]

[1]Institute of the Statistical Mathemetics
[2]Faculty of Science and Engineering, Tokushima Bunri University

In recent service science research, we often have huge amount of individual data with both continuous and categorical variables. These data sets can sometimes be divided into rather small number of naturally defined groups. In such situations, we are interested in inference and analysis for these groups, not for individual data. For describing these groups, we consider a set of descriptive statistics, and call it "aggregated symbolic data" (ASD). We propose to use up to second moments descriptive statistics for both continuous and categorical variables as ASD, and define a dissimilarity as the sum of chi-squared statistics among all variables including continuous variables. We apply our method to real estate data in Tokyo metropolitan area. We consider 23 cities in Tokyo as ASD and calculate dissimilarity among 23 ASDs, and investigate some characteristics relationships among ASDs by using hierarchical clustering and multidimensional scaling.

Key words: Big data, Burt matrix, chi-square statistics, hierarchical clustering, multidimensional scaling.

統計数理(2018)
第 66 巻 第 2 号 295–317

［原著論文］

B-スプライン及びAdaptive Group LASSOに基づく正則化非線形ロジットモデルによるデフォルト確率の推定

高部 勲[1,2]・山下 智志[3]

（受付 2017 年 12 月 31 日；改訂 2018 年 6 月 9 日；採択 6 月 20 日）

要　　旨

企業の過去のデフォルトデータを基にデフォルト確率予測モデルを構築する際には線形な 2 項ロジットモデルが用いられることが多いが，これについては従前から，(1)企業の信用スコアと財務指標との間の非線形性に対する考慮が不十分であり，また(2)多くの説明変数の候補からの変数選択に莫大な計算時間がかかるというという 2 つの課題についての指摘がある．本稿では，このような非線形性と変数選択という 2 つの課題を同時に解決することを目的として，(1)B-スプラインに基づく非線形・ノンパラメトリック回帰モデル及び(2)Adaptive Group LASSO に基づく効率的な変数選択という 2 つの手法を組み合わせることにより，従前の手法よりも効果的かつ効率的なデフォルト確率予測モデルの構築を試みた．複数の銀行のデータを統合した独自のデータベースを用いてデフォルト確率予測モデルの構築を行った結果，本稿で提案したモデルは，t 値・p 値に基づく変数選択や単純な LASSO と比較して，いずれの期間においても最も説明変数の数が少なくなっており，より効率的な変数選択を行うことができた．また AR 値などの指標の観点から，推定精度が向上していることが確認された．

キーワード：信用リスク， B-スプライン， Adaptive Group LASSO.

1. 導入

金融機関の信用リスク管理を考える際に，個別企業のデフォルト確率や倒産確率の予測精度の向上は重要な課題となっている．企業のデフォルト確率の予測モデルには，企業価値や債券価格を確率過程で記述するモデル(Merton, 1974; Duffie and Singleton, 1999)や，多変量判別分析に基づくモデル(Altman, 1968; 白田, 2008)などがあるが，金融機関における実務では企業の過去のデフォルトに関するデータを基にデフォルト確率を予測するモデルを構築することが多く，その際には線形な 2 項ロジットモデルがよく用いられている(尾木, 2017; 山下・三浦, 2011; 森平, 2009; Martine, 1977; Engelmann and Raumeier, 2006)．しかし線形な 2 項ロジットモデルについては従前から，以下の 2 つの課題があることが指摘されている．

(1)企業の信用スコアと各種財務指標との間の非線形性に対する考慮が不十分

[1] 総合研究大学院大学 複合科学研究科統計科学専攻：〒 190–8562 東京都立川市緑町 10–3
[2] 総務省統計局：〒 162–8668 東京都新宿区若松町 19–1
[3] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

(2)多くの説明変数(各種財務指標)の候補からの変数選択に莫大な計算時間がかかる

(1)の非線形性に関する課題について，従来の研究ではロジットモデルの説明変数として 2 次以上の多項式などを用いることにより対処している場合が多い．しかしそのようなモデルでは，非線形かつ多様な変動を把握するには限界があると考えられる．また，(2)の変数選択の課題については，t 値・p 値や AIC を基にしたステップワイズな変数選択により対処している事例が多いが，説明変数として用いる財務指標の数が多くなると，比較対象となるモデルの数も指数的に増大し，計算時間の面で限界があることから，より効率的な変数選択の手法が必要とされている．

上記の 2 つの課題に対して個別に対処している先行研究は存在するものの(これらの先行研究の具体的な内容については 2 節で示す)，これらの課題を同時に考慮したデフォルト確率予測モデルの事例については，調べた限りでは存在しない．そこで本稿では，これらの課題を同時に解決することを目的として，以下の 2 つの手法を組み合わせることにより，従前の結果と比較して，より精度の高いデフォルト確率予測モデルの効率的な構築を試みた．

(1)B-スプラインに基づく非線形・ノンパラメトリック回帰モデル
(2)Adaptive Group LASSO に基づく効率的な変数選択

本研究では複数の銀行のデータを統合した独自のデータベースを用いて，中小企業を対象としたデフォルト確率予測モデルの構築を行う．本研究において提案した手法はデフォルト確率との非線形な関係の合理的・効率的な構築に寄与するものであり，各種財務指標に基づく与信判断などにも資すると考えられる．

2. 先行研究と課題

2.1 ロジットモデルに基づくデフォルト確率予測モデル

企業の過去のデフォルトに関するデータを基にデフォルト確率予測モデルを構築する際に，個別企業に関する大規模なデータが活用できる場合には，線形な 2 項ロジットモデルが利用されることが多い．このような研究としては，中小企業信用リスク情報データベース協会(CRD 協会)のデータを用いた高橋・山下 (2002)や，日本政策金融公庫のデータを用いた尾木他 (2015)などがある．

線形な 2 項ロジットモデルは，企業 i $(1 \leq i \leq n)$ のデフォルト確率を P_i，対応する財務指標を x_{ij} $(1 \leq j \leq p)$ とした場合，以下のように表現される．

$$\log \frac{P_i}{1-P_i} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{2.1}$$

式(2.1)の左辺は P_i のロジット変換である．式(2.1)は以下のように表現することもできる．

$$\begin{aligned} P_i &= \frac{1}{1+\exp(-Z_i)} \\ Z_i &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \end{aligned} \tag{2.2}$$

ここで Z_i は企業 i の信用スコアを表しており，一般的にこの数値が大きくなるほど企業の信用力が低くデフォルト確率 P_i が高くなる．信用スコア Z_i を基に企業のデフォルトの可能性を予測することができる．

上記の 2 項ロジットモデルにおける回帰係数 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ は最尤法により推定する．

具体的には以下の対数尤度 $L_1(\beta)$ を最大化するような $\hat{\beta}$ を回帰係数の推定値とする．

$$
\begin{aligned}
L_1(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \log[{P_i}^{\delta_i}(1-P_i)^{1-\delta_i}] \\
&= \sum_{i=1}^{n} [\delta_i \log(P_i) + (1-\delta_i)\log(1-P_i)] \qquad (2.3) \\
\delta_i &= \begin{cases} 1 & (\text{企業 } i \text{ がデフォルトしている場合}) \\ 0 & (\text{企業 } i \text{ が非デフォルトである場合}) \end{cases}
\end{aligned}
$$

P_i には，式(2.2)で表されるデフォルト確率を代入する．このような単純な 2 項ロジットモデルに基づくデフォルト確率予測については，回帰係数の推定が容易である一方，次節以降に示すような課題があることが指摘されている．

2.2 財務指標と信用スコアとの非線形な関係

式(2.1)による 2 項ロジットモデルでは線形なモデルを仮定している．しかし財務指標によっては信用スコアとの間に非線形な関係があることが指摘されている（Dwyer et al., 2004; 白田, 2008）．このような場合に線形なモデルを用いると，信用スコアと財務指標との関係を適切にモデリングすることができず，デフォルト確率の予測精度が低下するおそれがある．図 1 は，今回使用するデータ(詳細は 4.1 節を参照)のうち，2005 年から 2013 年までの期間に関して，いくつかの財務指標と実績デフォルト率のロジット変換値との関係を示したものである．具体的には各財務指標の大きさの順に企業を並べ，それらを財務指標の大きさに応じて 200 のクラスに分割し，各クラスにおける実績デフォルト率のロジット変換値をプロットしたものである．その際に歪みの大きい一部の財務指標に対して対数変換又は neglog 変換を行っており，さらにそれらの値が 0 から 1 の範囲に収まるように線形変換を行っている．なお neglog 変換は以下のように定義されるもので，対数変換を負の値に拡張した変換となっている(森平, 2009; 山下・三浦, 2011)．

$$
\mathrm{neglog}(x) = \begin{cases} \log(x+1) & (\ x \geq 0) \\ -\log(-x+1) & (\ x < 0) \end{cases} \qquad (2.4)
$$

また図 1 には併せて線形ロジットモデルによる予測値(点線)と，後述の式(3.1)の B-スプラインに基づく非線形ロジットモデルについて，単変量のモデルを各財務指標に当てはめて推定した予測値(実線)を示している．その際の B-スプラインの基底の数は，AIC に基づき選択した．図 1 から，財務指標によっては実績デフォルト率のロジット変換値との間に明らかに非線形な関係があり，線形なモデルではこれらの変動に対応できていないことがわかる．

式(2.1)で示した 2 項ロジットモデルにおいて非線形な効果を扱う場合には，各財務指標の多項式や対数，平方根などの項を導入することが考えられる(Hosmer et al., 2013)．しかしそれらの関数形のどれが正しいかを事前に知ることはできないため，このような方法では各種財務指標の非線形な影響の把握には限界があると考えられる．

財務指標との関係を多項式モデルのような形であらかじめ設定するのではなく，ノンパラメトリック回帰モデルの手法を用いてデータから柔軟に曲線関係を推定している先行研究も存在する．Berg (2007)では，一般化加法モデル(Generalized Additive Model, Hastie and Tibshirani, 1990)の枠組みを企業の倒産確率モデルに導入することにより，データから柔軟な形で財務指標と倒産確率との非線形な関係を推定している．そして従来の判別分析モデルや線形な 2 項ロ

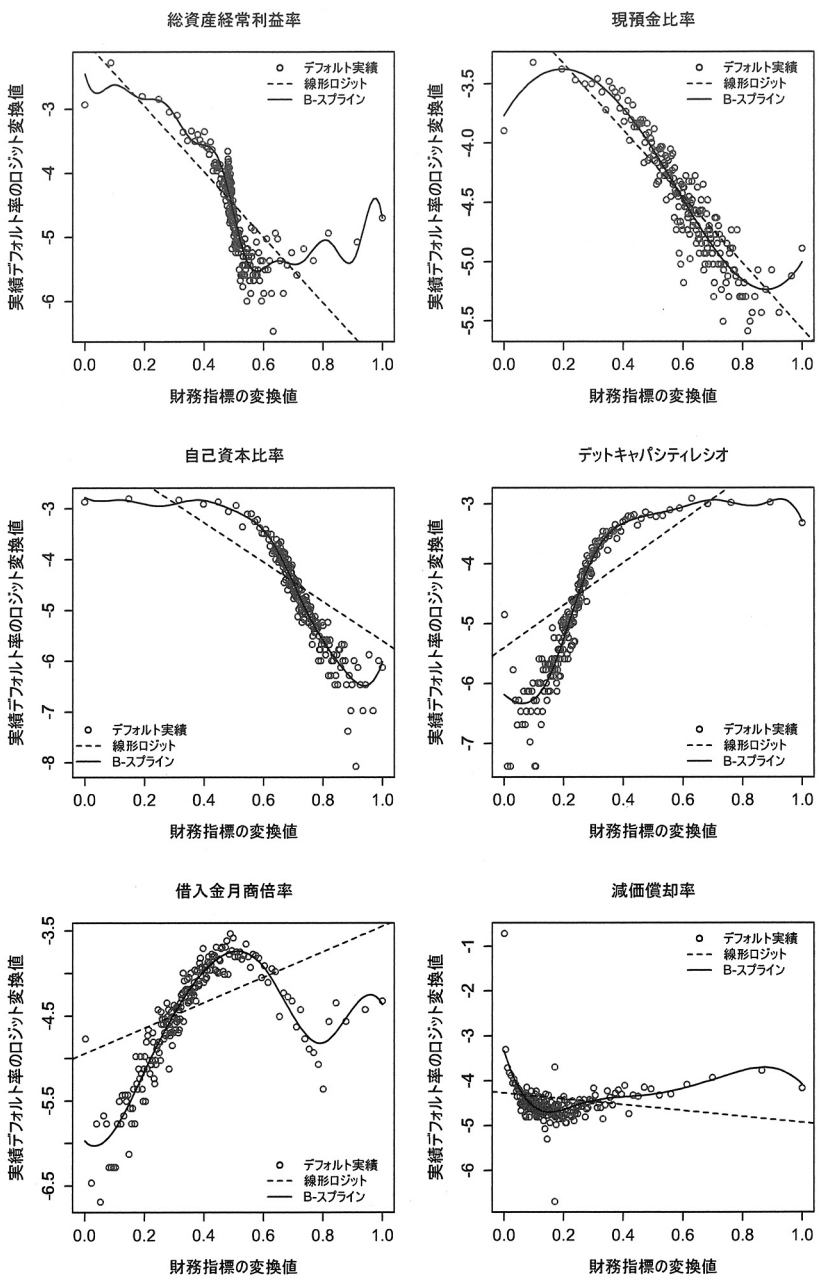


図 1. 実績デフォルト率の状況(図中の「現預金比率」,「デットキャパシティレシオ」及び「借入金月商倍率」については neglog 変換を行っている．また全ての財務指標について，0 から 1 の範囲に収まるように線形変換を行っている.)．

ジットモデルと比較して，AR 値の観点から推定精度が向上したと報告している．Giordani et al.（2014)では 2 次の自然スプラインを用いた非線形ロジットモデルを用いて個別企業の倒産

確率を分析しており，線形ロジットモデルと比較して，AR 値や疑似決定係数の観点からモデルの精度が向上し，倒産確率と各種財務指標との非線形な関係を適切に捉えることができたと報告している．山内 (2010)では財務指標を離散化したスコアリングテーブルに基づき，遺伝的アルゴリズムにより多目的最適化問題を解くことによって非線形なモデルを推定している．ただしこれらの先行研究では，いずれも非線形なモデルの構築のみに焦点を当てており，各種財務指標の中から適切なものを選択するという変数選択の観点は考慮されておらず，具体的な計算に入る前の段階でモデルに導入する財務指標の種類を，事前にある程度限定している．

2.3 複数の財務指標に関する変数選択

デフォルト確率予測モデルを構築する場合，説明変数として用いられる財務指標の候補の数は主要なものだけでも数十程度あり，場合によっては 100 を超えることもある．これらの財務指標の全ての組合せに基づくモデルを推定して比較を行う場合，対象となるモデルの数が非常に多くなるため，モデル構築にかなりの時間を要する．例えば候補となる財務指標の数が 50 である場合，$2^{50}\,(\fallingdotseq 10^{15})$ 通りのモデルの候補が考えられる．これらの候補の中から AIC 等の基準に基づくステップワイズな方式によりモデル選択を行った場合，現実的な計算時間で推定を行うことは困難であることから，より効率的な変数選択の手法が必要となる．しかし従来のデフォルト確率予測モデルの構築においては t 値・p 値を用いた単純な変数の絞込みや，何らかの先験的な知見に基づく事前の財務指標の選択が行われているのが実情である．

これに関して近年，回帰係数の推定と変数選択を同時に実行できる LASSO(Least Absolute Shrinkage and Selection Operator)に関する研究が発展しており（Tibshirani, 1996; Hastie et al., 2015; 冨岡, 2015)，この方法を適用した企業のデフォルト確率や倒産確率の推定に関する研究も行われるようになってきている（Amendola et al., 2012; Perederiy, 2009; Tian et al., 2015)．LASSO に基づくロジットモデルでは，式(2.3)の対数尤度 $L_1(\boldsymbol{\beta})$ に，$L1$ ノルムに基づく正則化項を加えた以下の罰則付きの対数尤度 $L_2(\boldsymbol{\beta})$ の最大化を行うことにより，回帰係数 $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ の推定を行う．

$$L_2(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\delta_i \log(P_i) + (1 - \delta_i) \log(1 - P_i)] - \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.5}$$

式(2.5)の最大化は，回帰係数 $\boldsymbol{\beta}$ の範囲に $\sum_{j=1}^{p} |\beta_j| \leq t$ という制約を加えた下での式(2.3)の最大化と同値である(λ と t は 1 対 1 に対応)．なお定数項 β_0 にはこのような制約を課さないのが一般的である(Hastie et al., 2015)．$L1$ ノルムに基づく正則化項の下では，値の小さい回帰係数が 0 になりやすくなる傾向があり，この性質が回帰係数の推定と説明変数の選択を同時に行うことを可能としている．ここで λ は正則化項の効果を調整するチューニングパラメータであり，交差検証法により決定することが多い．

Prederiy (2009)は企業の倒産予測における変数選択の問題について，LASSO に基づく 2 項ロジットモデルを用いて対処した先駆的な研究であり，効率的な変数選択により計算量の削減を達成するとともに，モデルの予測精度も向上したと報告している．ただし単純な線形ロジットモデルに LASSO を適用するにとどまっており，財務指標との非線形な関係を考慮しておらず，最終的に選択された財務指標の数も多くなっている．また Amendola et al. (2012)や Tian et al. (2015)では，Cox 比例ハザードモデルと LASSO を組み合わせて企業の倒産確率の長期予測を行っているが，これらの研究においても同様に単純な線形ロジットモデルが用いられており，財務指標との間の非線形な関係を考慮したモデルとはなっていない．

3. 非線形・正則化ロジットモデルに基づくデフォルト確率予測モデルの構築

3.1 本研究の目的

これまでに述べたように，信用スコアと財務指標との間の非線形性及び変数選択の問題については，双方ともモデルの構築に当たり重要な課題であるが，それぞれの課題に個別に対応する研究事例はあるものの，これらを同時に考慮したモデルに関する研究については，調べた限りでは存在しない．本研究ではこれらの課題に対し，(1)B-スプラインに基づく非線形・ノンパラメトリック回帰モデルの導入，及び(2)Adaptive Group LASSO に基づく合理的な変数選択の適用という 2 つの手法を組み合わせたデフォルト確率予測モデルを提案する．

3.2 B-スプラインに基づく非線形モデル

まず財務指標との間の非線形性を考慮したモデリングについて検討する．これについてはスプラインに基づく非線形な項を導入することにより対応する．スプラインは，説明変数に関するデータが含まれる区間をいくつかの小区間に分割し，各区間において区分的な多項式モデルを当てはめる方法である(小西, 2010; 山下・安道, 2006; 桜井, 1981)．説明変数とデフォルト確率との複雑な関係を単一の多項式モデルで把握するのではなく，隣り合う各区間における多項式モデルを滑らかに接続することにより，非線形な構造に対処する方法となっている．

本稿では B-スプラインに基づく方法を検討する．B-スプラインは局所的な台を持つスプライン関数であり，複数の多項式を滑らかに接続して基底関数を構成する．B-スプラインの導入により，特定の関数形を仮定せずに，財務指標とデフォルト確率との間の非線形な関係をデータから柔軟に推定することが可能となる．B-スプラインに基づく非線形ロジットモデルは，式(2.2)における信用スコア Z_i を以下の式(3.1)で置き換えることで得られる．

$$Z_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}), \quad f_j(x_{ij}) = \sum_{k=1}^{m_j} \beta_{jk}\phi_k(x_{ij}) \tag{3.1}$$

ここで f_j $(1 \le j \le p)$ は各財務指標に対応する非線形関数であり，ϕ_k $(1 \le k \le m_j)$ は B-スプラインの基底を表している．図 2 は B-スプラインに基づく非線形回帰のイメージを示したものである．左側の図が各説明変数に対する B-スプラインの基底を表しており(基底の数は 9 に設定)，右側の図はこれらの基底に基づく非線形回帰モデルの予測値を示している．このように非線形な基底を組み合わせることで，データから柔軟に関数を推定することが可能となる．

本稿では先行研究(Huang et al., 2010)に基づき B-スプラインの次数は 3 次とし，基底の計

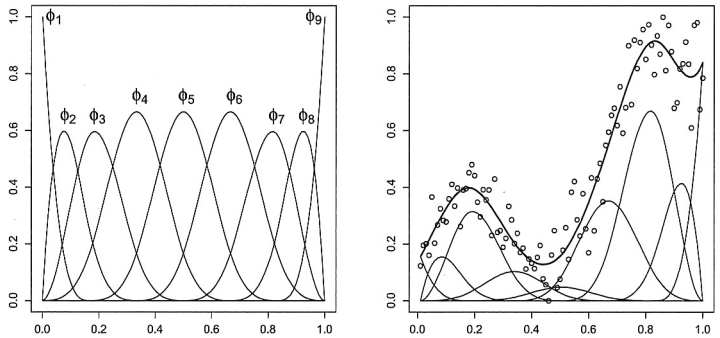


図 2. B-スプラインに基づく非線形回帰のイメージ　左図：B-スプラインの基底　右図：B-スプラインに基づく非線形モデル(点がサンプルデータ，太線が予測値)．

算には R の bs 関数を用いている．B-スプラインを構築するに当たり，区間を分割する節点を設定する必要がある．節点の位置については等間隔に設定している．また節点の数については，これを 5 から 15 の範囲で変化させて各財務指標に対して単変数の非線形ロジットモデルを当てはめ，AIC に基づき財務指標ごとにその数を事前に決定している．

3.3 Group LASSO に基づく変数選択

B-スプラインに基づく非線形モデルでは，財務指標ごとに基底を複数個用意して滑らかな非線形の曲線を表現する．このとき 1 つの財務指標に対して複数の基底が対応することになるため，変数選択の際にはこれらの複数の基底をまとめてモデルに取り込む，あるいはモデルから除去する必要がある．このように複数の変数をグループとしてまとめて扱い，変数選択を行う方法として，Group LASSO がある（Meier et al., 2008; Hastie et al., 2015）．

Group LASSO では，式(2.5)における $L1$ ノルムによる正則化項の代わりに，$L2$ ノルム $\|\boldsymbol{\beta}_j\|_2$ $(= \sqrt{{\beta_{j1}}^2 + {\beta_{j2}}^2 + \cdots + {\beta_{jm_j}}^2})$ による正則化項を用いた以下の $L_3(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p)$ を最大化することにより，回帰係数 β_0 及び $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jm_j})$ の推定値である $\hat{\beta}_0$ 及び $\hat{\boldsymbol{\beta}}_j$ $(1 \leq j \leq p)$ を得る手法である．これによりグループ単位での回帰係数の推定と変数選択を同時に行うことが可能となる．

$$L_3(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p) = \sum_{i=1}^{n} [\delta_i \log(P_i) + (1 - \delta_i) \log(1 - P_i)] - \lambda \sum_{j=1}^{p} \sqrt{m_j} \|\boldsymbol{\beta}_j\|_2 \tag{3.2}$$

P_i の式に含まれる信用スコア Z_i には，式(3.1)を代入する．なお，Yuan and Lin (2006)では式(3.2)のように，Group LASSO の重みにはグループのサイズの平方根を用いることが推奨されている．

ここで式(3.1)に関して，例えばある項 $f_j(x_{ij})$ に定数 C を加え，別の項 $f_k(x_{ik})$（あるいは定数項）から定数 C を引いても同一の信用スコア Z_i が得られることから，非線形関数の一意性が保証されないことになる．そこで非線形関数の一意性のために，Huang et al. (2010) に基づき，以下の制約を課す．

$$\sum_{i=1}^{n} \sum_{k=1}^{m_j} \beta_{jk} \phi_k(x_{ij}) = 0 \tag{3.3}$$

上記の制約については，ϕ_k を以下のように変換した新たな基底 ψ_{jk} を用いることで対応できる．

$$\bar{\phi}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \phi_k(x_{ij}), \qquad \psi_{jk}(x_{ij}) = \phi_k(x_{ij}) - \bar{\phi}_{jk} \tag{3.4}$$

スプラインと Group LASSO を組み合わせたモデルを遺伝子分野の研究に応用した事例として，Huang et al.（2010），Meier et al.（2009）がある．本稿では Huang et al.（2010）の方法をベースとしつつ，次節に示すような調整を行った上で，デフォルト確率予測モデルを構築している．

3.4 Multistep Adaptive Group LASSO に基づく変数選択

LASSO や Group LASSO では正則化項にかかるチューニングパラメータ λ を変化させることで回帰係数にかかる制約の強さをコントロールすることができるが，全ての回帰係数に同一のパラメータ λ を適用している点は改良の余地がある．そこで回帰係数の大きさの逆数を罰則とすることで絶対値の小さな係数により大きな罰則を課し，効率的に変数を選択する方法が Adaptive Group LASSO である（Bühlmann and van de Geer, 2011; Huang et al., 2010）．

Adaptive Group LASSO では，既に得られている推定値 $\hat{\boldsymbol{\beta}}_j$ を用いて計算した ω_j を基に，以下の $L_4(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p)$ を最大化することにより回帰係数の推定を行う．

$$L_4(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p) = \sum_{i=1}^{n} [\delta_i \log(P_i) + (1-\delta_i)\log(1-P_i)] - \lambda \sum_{j=1}^{p} \sqrt{m_j}\omega_j \|\boldsymbol{\beta}_j\|_2 \tag{3.5}$$

$$\omega_j = \begin{cases} \|\hat{\boldsymbol{\beta}}_j\|_2^{-1} & (\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) \\ \infty & (\|\hat{\boldsymbol{\beta}}_j\|_2 = 0) \end{cases}$$

ここで $\omega_j = \infty$ となる場合には，対応する変数をモデルから取り除くこととする．

本稿では変数の選択をより効率的に行うために，Adaptive Group LASSO を複数回適用する方法を用いる（以下ではこれを Multistep Adaptive Group LASSO と呼ぶ）．具体的には以下の手順により，係数を推定する．

(1) まず，Group LASSO を適用し，係数の推定値 $\hat{\beta}_0$ 及び $\hat{\boldsymbol{\beta}}_j$ $(1 \leq j \leq p)$ を得る．
(2) 得られた係数 $\hat{\boldsymbol{\beta}}_j$ を基に重み ω_j を計算し，Adaptive Group LASSO を適用して，係数の推定値 $\hat{\beta}_0^*$ 及び $\hat{\boldsymbol{\beta}}_j^*$ $(1 \leq j \leq p)$ を得る．
(3) 得られた係数 $\hat{\boldsymbol{\beta}}_j^*$ を基に，再度重みを計算し，Adaptive Group LASSO を適用して係数の最終的な推定値を求める．

今回の分析では計算のコストを考慮して，阪本 他 (2010) の設定を参考に，Multistep Adaptive Group LASSO における反復回数を 2 回に設定している．これらの計算の際には Adaptive Group LASSO の計算を比較的容易に行うことが可能であり，かつ高速な計算アルゴリズム (Groupwise Majorization Descent) を採用している R のパッケージ `gglasso` (Yang and Zou, 2015) を使用してモデルの構築及びパラメータの推定を行った．

4. 分析結果

4.1 データ

本稿の分析では，複数の銀行の債権に関する 2005 年から 2014 年までの統合データを用いている．またデフォルトの定義に関しては，企業の債権者区分が破たん懸念先以下に遷移する状況（破懸基準）をデフォルトとして扱っている．このデータを，モデルの構築に用いる期間と，構築したモデルの評価（バックテスト）を行う期間（アウトオブタイム）に分割して分析を行う．なお推定を行う期間の違いによって最適なモデルや結果の評価が影響を受ける可能性もあることから，分析に当たっては以下の表 1 に示すように，期間の分割の仕方を変えた 4 種類のデータセットを用意し，各データセットを対象としてモデルの構築を行い，結果を比較した．分析に用いた財務指標の一覧は表 2 に示している．

4.2 モデル構築及びパラメータ推定の際の設定

モデルの構築に当たっては，以下の設定の下でパラメータの推定等を行った．

表 1．分析に用いたデータセットの種類．

データセット	モデル構築期間	バックテスト期間
データセット 1	2005 年 〜 2013 年	2014 年
データセット 2	2005 年 〜 2012 年	2013 年 〜 2014 年
データセット 3	2005 年 〜 2011 年	2012 年 〜 2014 年
データセット 4	2005 年 〜 2010 年	2011 年 〜 2014 年

表 2. 分析に用いた財務指標の一覧.

(1) エクスポージャー	(21) 金利対現金預金比率	(41) 期末役員従業員数
(2) 融資シェア	(22) 自己資本比率	(42) 資産合計土地比率
(3) 総資産営業利益率	(23) 純資産倍率	(43) 資産合計有形固定資産比率
(4) 総資産経常利益率	(24) 固定長期適合率	(44) 資産合計前払費用比率
(5) ROA	(25) 固定比率	(45) 資産合計流動資産比率
(6) ROE	(26) 借入金依存度	(46) 資産合計固定資産比率
(7) 売上高総利益率	(27) デットキャパシティレシオ	(47) 資産合計支払手形比率
(8) 売上高営業利益率	(28) 預借率	(48) 資産合計受取手形比率
(9) 売上高経常利益率	(29) 借入金月商倍率	(49) 資産合計当座資産比率
(10) 売上高当期利益率	(30) 売上高支払利息割引料率	(50) 資産合計棚卸資産比率
(11) 売上事業キャッシュフロー率	(31) 有利子負債利子率	
(12) 総資産回転率	(32) 現金預金対利子割引料率	
(13) 売上債権回転日数	(33) 支払利息割引料対総利益率	
(14) 棚卸資産回転日数	(34) 事業キャッシュフロー有利子負債比率	
(15) 売上高有形固定資産率	(35) インタレストカバレッジ	
(16) 買入債務回転日数	(36) 事業キャッシュフロー金利負担率	
(17) 流動比率	(37) 減価償却率	
(18) 当座比率	(38) 売上高減価償却率	
(19) 支払準備率	(39) 売上高	
(20) 現預金比率	(40) 資産合計	

変数変換の適用：財務指標によっては売上高のように，少数の企業が非常に大きな値をとるような右に歪んだ分布となる場合がある．このように歪みの強い変数については，対数変換又は neglog 変換を適用し，変数の安定化を図った．その上で，さらに全ての変数に対し，0 から 1 の範囲に収まるように線形変換を行った．

はずれ値への対応：財務指標によっては，上記の変換を行ってもなお，はずれ値が存在することがある．そこで，はずれ値の影響を軽減するため，財務指標を大きさの順にソートし，分布の上下 1% で折返し処理(上下 1% を超える値に対して上下 1% における値を代入)を行った．

欠測値への対応：財務指標によっては，欠測値が存在することがある．そのような場合には中央値を代入して補完を行った．なお今回のデータセットでは欠測値がそれほど多くないため(全体の 5% 程度)，欠測値補完による分析結果への影響は，それほど大きくないと考えられる．

フラグ(ダミー)変数の導入：業種別，銀行別に関するフラグ変数を導入した．なお，これらのフラグ変数には LASSO の罰則を課していない．

チューニングパラメータ λ の決定：Adaptive Group LASSO を適用する際に，チューニングパラメータ λ を決定する必要がある．これについては AUC に基づく 5 重交差検証法により最小となる値を求め，これをベースとして最終的に 1 標準誤差ルール（Hastie et al., 2015; 川野他, 2018)により λ を決定した．AUC については 4.3 節を参照．

4.3 複数の手法に基づくモデルの比較・検証方法

本稿では，パラメータの推定と変数(財務指標)の選択に関する以下の 5 つのモデルについて，各種指標により比較を行った．

(1)**線形モデル + p 値に基づく変数選択[モデル 1]**：線形な 2 項ロジットモデルを基に，2 段階で変数の選択を行う．具体的には，まず全ての変数を用いて推定を行い，p 値が 0.1 以上の変数をモデルから除外する．そして再度パラメータの推定を行い，p 値が 0.05 以上の変数をモデルから除外して，最終的なモデルを決定した．

(2)**線形モデル + LASSO[モデル 2]**：線形な 2 項ロジットモデルを基に，式(2.5)に基づき

パラメータの推定及び変数の選択を行った．LASSO による推定には R のパッケージ `glmnet` (Friedman et al., 2010)を用いた．

(3)**線形モデル + Multistep Adaptive LASSO**[モデル 3]：線形な 2 項ロジットモデルを基に，以下の式(4.1)に基づく Adaptive LASSO を 2 回適用することにより，パラメータの推定及び変数の選択を行った．

$$L_5(\beta_0,\beta_1,\beta_2,\ldots,\beta_p)=\sum_{i=1}^{n}[\delta_i\log(P_i)+(1-\delta_i)\log(1-P_i)]-\lambda\sum_{j=1}^{p}\omega_j|\beta_j| \tag{4.1}$$
$$\omega_j=\begin{cases}|\beta_j|^{-1} & (|\beta_j|>0)\\ \infty & (|\beta_j|=0)\end{cases}$$

(4)**B-スプライン + Group LASSO**[モデル 4]：B-スプラインに基づく 2 項ロジットモデルを基に，式(3.2)に基づく Group LASSO を適用することにより，パラメータの推定及び変数の選択を行った．

(5)**B-スプライン + Multistep Adaptive Group LASSO**[モデル 5]：B-スプラインに基づく 2 項ロジットモデルを基に，式(3.5)に基づく Multistep Adaptive Group LASSO により，パラメータの推定及び変数の選択を行った．

上記の方法により推定したモデル間の比較に用いる各種指標の定義については以下のとおりである（尾木, 2017; 山下・三浦, 2011; 森平, 2009; Engelmann and Raumeier, 2006）．

AUC(**Area Under the Curve**)：AUC は，ROC 曲線(Receiver Operatorating Characteristic curve)の下側部分の面積で定義される指標である．AUC はモデルの順位性(信用スコアの低い(高い)企業ほどデフォルト率が高く(低く)なっているか)を評価するための指標であり，この値が大きいほどデフォルトの予測精度が高いといえる．AUC の計算には R の `pROC` パッケージを用いた．

AR 値(**Accuracy Ratio**)：AR 値は，CAP(Cumulative Accuracy Profiles)曲線の下側面積から計算される統計量である．AR 値と AUC との間には，AR 値 $=2\text{AUC}-1$ という関係があり，これらは同等な統計量であるが，信用リスクモデルの評価には AR 値を用いることが多い．

疑似決定係数(**Pseudo R^2**)：疑似決定係数は，$1-(L_{opt}/L_{init})$ で表される統計量であり，マクファーデンの決定係数とも呼ばれる．ここで L_{init} は定数項のみのロジットモデルの推定を行った場合の対数尤度であり，L_{opt} は財務指標を用いたロジットモデルの推定を行った場合の対数尤度である．疑似決定係数はインサンプルにおけるモデルのデータへの当てはまりを表す指標であり，この値が大きいほど当てはまりが良いといえる．

ブライアスコア：ブライアスコアは，$(1/n)\sum_{i=1}^{n}(P_i-\delta_i)^2$ で表される統計量である．ここで P_i は企業 i のデフォルト確率であり，δ_i は企業 i がデフォルトしていれば 1，非デフォルトであれば 0 となる定数である．ブライアスコアはモデルの一致性(推定されたデフォルト確率と実際のデフォルト率がどの程度近いか)を表す指標であり，この値が小さいほど一致性が高いといえる．

4.4 推定結果

期間の分割の仕方を変えた 4 つのデータセットを対象に分析を行い，説明変数として選択された財務指標について示したものが，表 3 から表 6 である．

全てのデータセットにおいて，提案手法(モデル 5)が，選択された変数の数が最も少なくなっている．また，線形モデル + LASSO(モデル 2)と提案手法(モデル 5)について，各データセットにおいて選択された変数をまとめたものが表 7 である．

表 3. 各推定方法における変数選択の結果：データセット 1.

	モデル 1	モデル 2	モデル 3	モデル 4	モデル 5
(1) エクスポージャー		✓	✓	✓	✓
(2) 融資シェア	✓	✓		✓	
(3) 総資産営業利益率	✓	✓	✓	✓	
(4) 総資産経常利益率	✓	✓	✓	✓	✓
(5) ROA	✓	✓	✓	✓	
(6) ROE	✓	✓	✓		
(7) 売上高総利益率					
(8) 売上高営業利益率	✓				
(9) 売上高経常利益率	✓	✓	✓	✓	
(10) 売上高当期利益率			✓	✓	✓
(11) 売上事業キャッシュフロー率	✓				
(12) 総資産回転率	✓	✓	✓		
(13) 売上債権回転日数	✓	✓		✓	✓
(14) 棚卸資産回転日数	✓	✓			
(15) 売上高有形固定資産率	✓				
(16) 買入債務回転日数	✓	✓		✓	✓
(17) 流動比率	✓			✓	✓
(18) 当座比率	✓			✓	
(19) 支払準備率	✓	✓	✓	✓	
(20) 現預金比率				✓	✓
(21) 金利対現金預金比率	✓		✓	✓	
(22) 自己資本比率				✓	✓
(23) 純資産倍率	✓	✓	✓	✓	✓
(24) 固定長期適合率					
(25) 固定比率					
(26) 借入金依存度		✓			
(27) デットキャパシティレシオ		✓		✓	✓
(28) 預借率	✓				
(29) 借入金月商倍率	✓	✓		✓	✓
(30) 売上高支払利息割引料率	✓			✓	
(31) 有利子負債利子率	✓	✓	✓	✓	✓
(32) 現金預金対利子割引料率	✓	✓	✓	✓	✓
(33) 支払利息割引料対総利益率	✓	✓	✓	✓	✓
(34) 事業キャッシュフロー有利子負債比率	✓	✓	✓	✓	✓
(35) インタレストカバレッジ	✓	✓	✓	✓	
(36) 事業キャッシュフロー金利負担率					
(37) 減価償却率	✓	✓	✓	✓	✓
(38) 売上高減価償却率	✓	✓	✓	✓	
(39) 売上高	✓			✓	
(40) 資産合計	✓			✓	✓
(41) 期末役員従業員数	✓				
(42) 資産合計土地比率	✓	✓		✓	
(43) 資産合計有形固定資産比率	✓	✓	✓	✓	✓
(44) 資産合計前払費用比率	✓	✓	✓	✓	✓
(45) 資産合計流動資産比率					
(46) 資産合計固定資産比率	✓	✓	✓		
(47) 資産合計支払手形比率			✓	✓	
(48) 資産合計受取手形比率	✓	✓	✓	✓	✓
(49) 資産合計当座資産比率	✓		✓		
(50) 資産合計棚卸資産比率	✓	✓	✓	✓	✓
選択された変数の数 (ダミー変数除く)	38	29	25	34	21

モデル 1：線形モデル + p 値に基づく変数選択
モデル 2：線形モデル + LASSO
モデル 3：線形モデル + Multistep Adaptive LASSO
モデル 4：B-スプライン + Group LASSO
モデル 5：B-スプライン + Multistep Adaptive Group LASSO

表 4. 各推定方法における変数選択の結果：データセット 2.

	モデル 1	モデル 2	モデル 3	モデル 4	モデル 5
(1) エクスポージャー		✓	✓	✓	✓
(2) 融資シェア	✓	✓		✓	
(3) 総資産営業利益率	✓	✓	✓	✓	
(4) 総資産経常利益率	✓	✓	✓	✓	
(5) ROA	✓	✓	✓	✓	✓
(6) ROE	✓	✓	✓		
(7) 売上高総利益率				✓	
(8) 売上高営業利益率	✓	✓			
(9) 売上高経常利益率	✓	✓	✓	✓	✓
(10) 売上高当期利益率					
(11) 売上事業キャッシュフロー率	✓				
(12) 総資産回転率	✓	✓	✓		
(13) 売上債権回転日数	✓	✓		✓	✓
(14) 棚卸資産回転日数	✓	✓	✓		
(15) 売上高有形固定資産率	✓				
(16) 買入債務回転日数	✓	✓	✓	✓	✓
(17) 流動比率	✓			✓	✓
(18) 当座比率	✓	✓	✓	✓	
(19) 支払準備率	✓	✓	✓	✓	
(20) 現預金比率				✓	✓
(21) 金利対現金預金比率	✓	✓	✓	✓	
(22) 自己資本比率				✓	✓
(23) 純資産倍率	✓	✓	✓	✓	✓
(24) 固定長期適合率		✓			
(25) 固定比率					
(26) 借入金依存度		✓			
(27) デットキャパシティレシオ		✓		✓	✓
(28) 預借率	✓				
(29) 借入金月商倍率	✓	✓		✓	✓
(30) 売上高支払利息割引料率	✓	✓		✓	
(31) 有利子負債利子率	✓	✓	✓	✓	✓
(32) 現金預金対利子割引料率	✓	✓	✓	✓	✓
(33) 支払利息割引料対総利益率	✓	✓	✓	✓	✓
(34) 事業キャッシュフロー有利子負債比率	✓	✓	✓	✓	✓
(35) インタレストカバレッジ	✓	✓	✓		
(36) 事業キャッシュフロー金利負担率					
(37) 減価償却率		✓	✓	✓	✓
(38) 売上高減価償却率		✓	✓	✓	
(39) 売上高	✓	✓		✓	
(40) 資産合計	✓		✓	✓	✓
(41) 期末役員従業員数	✓	✓			
(42) 資産合計土地比率	✓	✓	✓	✓	
(43) 資産合計有形固定資産比率	✓	✓	✓	✓	✓
(44) 資産合計前払費用比率	✓	✓	✓	✓	✓
(45) 資産合計流動資産比率					
(46) 資産合計固定資産比率	✓	✓	✓		
(47) 資産合計支払手形比率		✓	✓	✓	
(48) 資産合計受取手形比率	✓	✓	✓	✓	✓
(49) 資産合計当座資産比率	✓		✓		
(50) 資産合計棚卸資産比率	✓	✓	✓	✓	✓
選択された変数の数 (ダミー変数除く)	36	37	29	33	21

モデル 1：線形モデル + p 値に基づく変数選択
モデル 2：線形モデル + LASSO
モデル 3：線形モデル + Multistep Adaptive LASSO
モデル 4：B-スプライン + Group LASSO
モデル 5：B-スプライン + Multistep Adaptive Group LASSO

表 5. 各推定方法における変数選択の結果：データセット 3.

	モデル 1	モデル 2	モデル 3	モデル 4	モデル 5
(1) エクスポージャー		✓	✓		✓
(2) 融資シェア	✓	✓		✓	
(3) 総資産営業利益率	✓	✓	✓	✓	
(4) 総資産経常利益率		✓	✓	✓	
(5) ROA	✓	✓	✓	✓	✓
(6) ROE	✓	✓	✓		
(7) 売上高総利益率					✓
(8) 売上高営業利益率	✓				
(9) 売上高経常利益率	✓	✓	✓	✓	✓
(10) 売上高当期利益率		✓	✓		
(11) 売上事業キャッシュフロー率	✓				
(12) 総資産回転率	✓	✓	✓		
(13) 売上債権回転日数	✓	✓		✓	✓
(14) 棚卸資産回転日数	✓	✓			
(15) 売上高有形固定資産率	✓				
(16) 買入債務回転日数	✓	✓	✓	✓	✓
(17) 流動比率	✓			✓	✓
(18) 当座比率	✓				
(19) 支払準備率	✓	✓	✓	✓	
(20) 現預金比率				✓	✓
(21) 金利対現金預金比率	✓		✓	✓	
(22) 自己資本比率				✓	✓
(23) 純資産倍率	✓	✓	✓	✓	✓
(24) 固定長期適合率					
(25) 固定比率					
(26) 借入金依存度		✓	✓		
(27) デットキャパシティレシオ		✓		✓	✓
(28) 預借率	✓				
(29) 借入金月商倍率	✓	✓		✓	✓
(30) 売上高支払利息割引料率	✓		✓	✓	
(31) 有利子負債利子率	✓	✓	✓	✓	✓
(32) 現金預金対利子割引料率	✓	✓	✓	✓	✓
(33) 支払利息割引料対総利益率	✓	✓	✓	✓	✓
(34) 事業キャッシュフロー有利子負債比率	✓	✓	✓	✓	✓
(35) インタレストカバレッジ	✓	✓	✓		
(36) 事業キャッシュフロー金利負担率					
(37) 減価償却率		✓	✓	✓	✓
(38) 売上高減価償却率		✓	✓	✓	
(39) 売上高	✓				
(40) 資産合計	✓		✓	✓	✓
(41) 期末役員従業員数	✓	✓			
(42) 資産合計土地比率	✓	✓			
(43) 資産合計有形固定資産比率	✓	✓	✓	✓	✓
(44) 資産合計前払費用比率	✓	✓	✓	✓	✓
(45) 資産合計流動資産比率					
(46) 資産合計固定資産比率	✓	✓	✓		
(47) 資産合計支払手形比率			✓	✓	
(48) 資産合計受取手形比率	✓	✓	✓	✓	✓
(49) 資産合計当座資産比率	✓		✓		
(50) 資産合計棚卸資産比率	✓	✓	✓	✓	✓
選択された変数の数 (ダミー変数除く)	35	31	29	28	22

モデル 1：線形モデル + p 値に基づく変数選択
モデル 2：線形モデル + LASSO
モデル 3：線形モデル + Multistep Adaptive LASSO
モデル 4：B-スプライン + Group LASSO
モデル 5：B-スプライン + Multistep Adaptive Group LASSO

表 6. 各推定方法における変数選択の結果：データセット 4.

	モデル 1	モデル 2	モデル 3	モデル 4	モデル 5
(1) エクスポージャー		✓	✓		
(2) 融資シェア		✓		✓	
(3) 総資産営業利益率	✓		✓		
(4) 総資産経常利益率		✓	✓	✓	
(5) ROA	✓	✓	✓	✓	✓
(6) ROE		✓	✓		
(7) 売上高総利益率					✓
(8) 売上高営業利益率	✓				
(9) 売上高経常利益率	✓	✓	✓	✓	✓
(10) 売上高当期利益率		✓	✓		
(11) 売上事業キャッシュフロー率	✓				
(12) 総資産回転率	✓		✓		
(13) 売上債権回転日数	✓	✓		✓	✓
(14) 棚卸資産回転日数	✓	✓			
(15) 売上高有形固定資産率	✓				
(16) 買入債務回転日数	✓	✓		✓	✓
(17) 流動比率		✓		✓	✓
(18) 当座比率	✓				
(19) 支払準備率	✓		✓		
(20) 現預金比率				✓	
(21) 金利対現金預金比率	✓		✓		
(22) 自己資本比率				✓	✓
(23) 純資産倍率	✓	✓	✓	✓	✓
(24) 固定長期適合率					
(25) 固定比率					
(26) 借入金依存度		✓	✓		
(27) デットキャパシティレシオ		✓		✓	✓
(28) 預借率					
(29) 借入金月商倍率	✓	✓		✓	✓
(30) 売上高支払利息割引料率	✓		✓	✓	
(31) 有利子負債利子率	✓	✓	✓	✓	✓
(32) 現金預金対利子割引料率	✓	✓	✓	✓	✓
(33) 支払利息割引料対総利益率	✓	✓	✓	✓	✓
(34) 事業キャッシュフロー有利子負債比率	✓	✓	✓	✓	
(35) インタレストカバレッジ	✓	✓	✓	✓	
(36) 事業キャッシュフロー金利負担率					
(37) 減価償却率		✓		✓	✓
(38) 売上高減価償却率		✓	✓		
(39) 売上高	✓			✓	
(40) 資産合計	✓		✓	✓	✓
(41) 期末役員従業員数	✓				
(42) 資産合計土地比率	✓	✓			
(43) 資産合計有形固定資産比率	✓	✓	✓	✓	✓
(44) 資産合計前払費用比率	✓	✓	✓	✓	✓
(45) 資産合計流動資産比率					
(46) 資産合計固定資産比率	✓	✓	✓		
(47) 資産合計支払手形比率			✓	✓	
(48) 資産合計受取手形比率	✓	✓	✓	✓	✓
(49) 資産合計当座資産比率			✓		
(50) 資産合計棚卸資産比率	✓			✓	✓
選択された変数の数 (ダミー変数除く)	30	27	26	26	19

モデル 1：線形モデル + p 値に基づく変数選択
モデル 2：線形モデル + LASSO
モデル 3：線形モデル + Multistep Adaptive LASSO
モデル 4：B-スプライン + Group LASSO
モデル 5：B-スプライン + Multistep Adaptive Group LASSO

表 7. モデル 2 及びモデル 5 において選択された変数.

モデル	モデル 2				モデル 5			
データセット	1	2	3	4	1	2	3	4
(1) エクスポージャー	✓	✓	✓	✓	✓	✓	✓	
(2) 融資シェア	✓	✓	✓	✓				
(3) 総資産営業利益率	✓	✓	✓					
(4) 総資産経常利益率	✓	✓	✓	✓	✓			
(5) ROA	✓	✓	✓	✓		✓	✓	✓
(6) ROE	✓	✓	✓	✓				
(7) 売上高総利益率							✓	✓
(8) 売上高営業利益率		✓						
(9) 売上高経常利益率	✓	✓	✓	✓		✓	✓	✓
(10) 売上高当期利益率			✓	✓	✓			
(11) 売上事業キャッシュフロー率								
(12) 総資産回転率	✓	✓	✓					
(13) 売上債権回転日数	✓	✓	✓	✓	✓	✓	✓	✓
(14) 棚卸資産回転日数	✓	✓	✓	✓				
(15) 売上高有形固定資産率								
(16) 買入債務回転日数	✓	✓	✓	✓	✓	✓	✓	✓
(17) 流動比率				✓	✓	✓	✓	✓
(18) 当座比率		✓						
(19) 支払準備率	✓	✓	✓					
(20) 現預金比率					✓	✓	✓	
(21) 金利対現金預金比率		✓						
(22) 自己資本比率					✓	✓	✓	✓
(23) 純資産倍率	✓	✓	✓	✓	✓	✓	✓	✓
(24) 固定長期適合率		✓						
(25) 固定比率								
(26) 借入金依存度	✓	✓	✓	✓				
(27) デットキャパシティレシオ	✓	✓	✓	✓	✓	✓	✓	✓
(28) 預借率								
(29) 借入金月商倍率	✓	✓	✓	✓	✓	✓	✓	✓
(30) 売上高支払利息割引料率		✓						
(31) 有利子負債利子率	✓	✓	✓	✓	✓	✓	✓	✓
(32) 現金預金対利子割引料率	✓	✓	✓	✓	✓	✓	✓	✓
(33) 支払利息割引料対総利益率	✓	✓	✓	✓	✓	✓	✓	✓
(34) 事業キャッシュフロー有利子負債比率	✓	✓	✓	✓	✓	✓	✓	
(35) インタレストカバレッジ	✓	✓	✓	✓				
(36) 事業キャッシュフロー金利負担率								
(37) 減価償却率	✓	✓	✓	✓	✓	✓	✓	✓
(38) 売上高減価償却率	✓	✓	✓	✓				
(39) 売上高		✓						
(40) 資産合計					✓	✓	✓	✓
(41) 期末役員従業員数		✓	✓					
(42) 資産合計土地比率	✓	✓	✓	✓				
(43) 資産合計有形固定資産比率	✓	✓	✓	✓	✓	✓	✓	✓
(44) 資産合計前払費用比率	✓	✓	✓	✓	✓	✓	✓	✓
(45) 資産合計流動資産比率								
(46) 資産合計固定資産比率	✓	✓	✓	✓				
(47) 資産合計支払手形比率		✓						
(48) 資産合計受取手形比率	✓	✓	✓	✓	✓	✓	✓	✓
(49) 資産合計当座資産比率								
(50) 資産合計棚卸資産比率	✓	✓	✓		✓	✓	✓	✓
選択された変数の数 (ダミー変数除く)	29	37	31	27	21	21	22	19

モデル 2：線形モデル + LASSO
モデル 5：B-スプライン + Multistep Adaptive Group LASSO

線形モデル + LASSO(モデル 2)ではデータセットによって(特にデータセット 2 とそれ以外で)選択される変数が大きく異なる場合があるのに対し，提案手法(モデル 5)による推定結果で

表8．各推定方法における推定結果の比較(太字は最も良いもの)．

データセット1	交差検証法					アウトオブタイム				
	モデル1	モデル2	モデル3	モデル4	モデル5	モデル1	モデル2	モデル3	モデル4	モデル5
AUC	0.84064	0.83690	0.83469	0.84803	**0.84923**	0.88175	0.87759	0.87140	0.88360	**0.88417**
AR値	0.68127	0.67380	0.66938	0.69606	**0.69847**	0.76350	0.75517	0.74285	0.76724	**0.76835**
ブライアスコア	0.01189	0.01192	0.01193	0.01181	**0.01180**	0.00663	0.00666	0.00680	**0.00658**	**0.00658**
疑似決定係数	0.15752	0.15080	0.14901	0.16742	**0.17023**			—		
サンプルサイズ			642,025					67,250		
デフォルト件数			7,980					458		
デフォルト率			1.24%					0.68%		

データセット2	交差検証法					アウトオブタイム				
	モデル1	モデル2	モデル3	モデル4	モデル5	モデル1	モデル2	モデル3	モデル4	モデル5
AUC	0.83760	0.83574	0.83235	0.84571	**0.84626**	0.86960	0.86801	0.84470	**0.87380**	0.87319
AR値	0.67519	0.67147	0.66471	0.69141	**0.69253**	0.73919	0.73603	0.68934	**0.74763**	0.74637
ブライアスコア	0.01233	0.01236	0.01238	**0.01225**	**0.01225**	0.00742	0.00743	0.00766	**0.00739**	**0.00739**
疑似決定係数	0.15529	0.15209	0.14801	0.16586	**0.16774**			—		
サンプルサイズ			572,772					136,503		
デフォルト件数			7,392					1,046		
デフォルト率			1.29%					0.77%		

データセット3	交差検証法					アウトオブタイム				
	モデル1	モデル2	モデル3	モデル4	モデル5	モデル1	モデル2	モデル3	モデル4	モデル5
AUC	0.83516	0.83137	0.83062	0.84118	**0.84379**	0.86500	0.86266	0.85120	0.87050	**0.87116**
AR値	0.67032	0.66273	0.66123	0.68236	**0.68758**	0.73001	0.72533	0.70245	0.74097	**0.74231**
ブライアスコア	0.01253	0.01257	0.01258	0.01246	**0.01244**	0.00866	0.00869	0.01158	**0.00861**	**0.00861**
疑似決定係数	0.15340	0.14729	0.14617	0.16109	**0.16577**			—		
サンプルサイズ			500,270					209,005		
デフォルト件数			6,563					1,875		
デフォルト率			1.31%					0.90%		

データセット4	交差検証法					アウトオブタイム				
	モデル1	モデル2	モデル3	モデル4	モデル5	モデル1	モデル2	モデル3	モデル4	モデル5
AUC	0.83042	0.82684	0.82450	0.83693	**0.83994**	0.86212	0.85889	0.84480	0.86570	**0.86571**
AR値	0.66085	0.65368	0.64900	0.67386	**0.67987**	0.72424	0.71778	0.68965	0.73136	**0.73143**
ブライアスコア	0.01275	0.01279	0.01280	0.01268	**0.01266**	0.00935	0.00938	0.01647	0.00930	**0.00929**
疑似決定係数	0.14938	0.14310	0.14101	0.15686	**0.16215**			—		
サンプルサイズ			427,697					281,578		
デフォルト件数			5,708					2,730		
デフォルト率			1.33%					0.97%		

は，選択された変数にそれほど大きな違いはなく，安定した推定結果となっている．

交差検証法(モデル構築期間)及びバックテスト(アウトオブタイム)における推定結果を示したものが表8である．

データセット2(モデル構築期間：2005年～2012年)のアウトオブタイムのサンプルにおけるAUC及びAR値を除いて，いずれのデータセットにおいても，提案手法(モデル5)が最も良い性能を示しており，他のモデルと比較して，AR値や疑似決定係数などの観点から推定精度が向上していることがわかる．

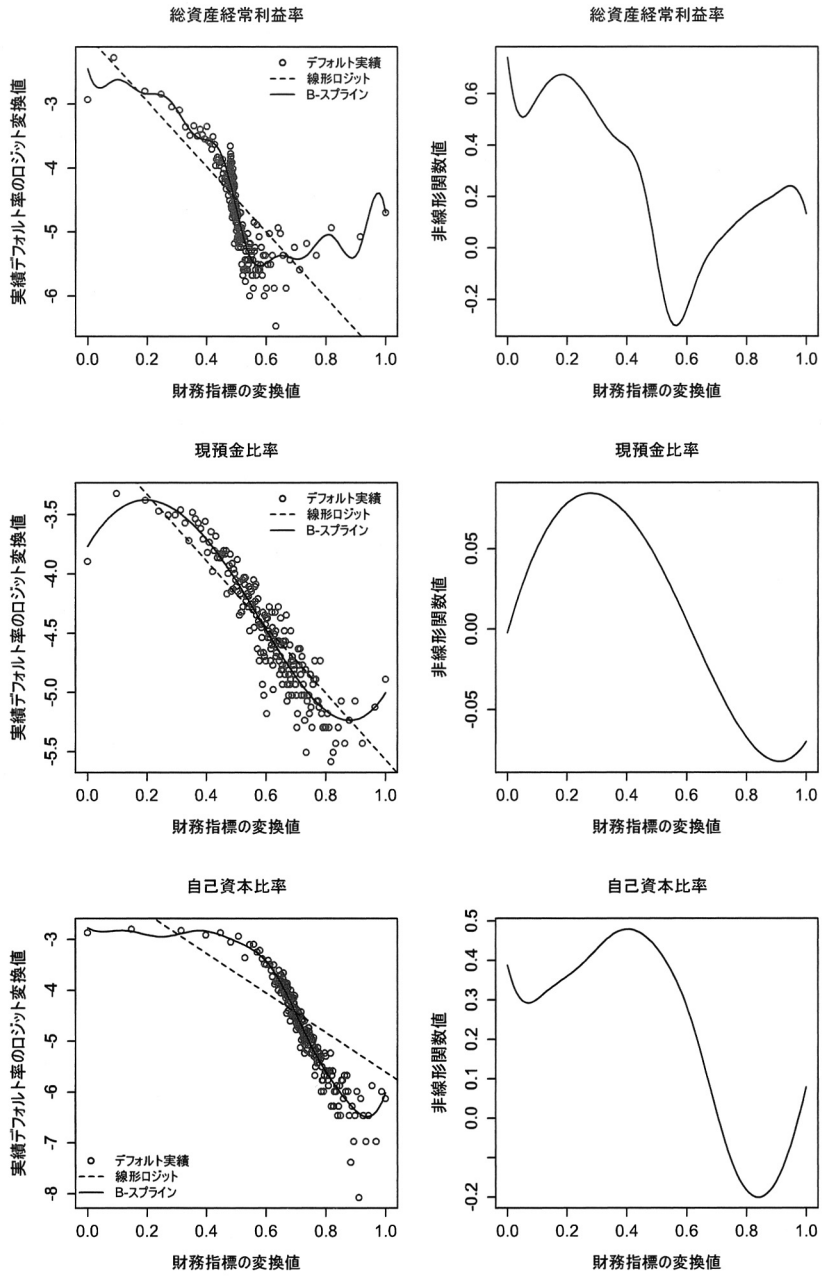


図 3. 非線形関数の推定結果(1)左列：図 1 再掲　右列：非線形関数の推定値(図中の「現預金比率」については neglog 変換を行っている．また全ての財務指標について，0 から 1 の範囲に収まるように線形変換を行っている．)．

提案手法(モデル 5)に基づき，データセット 1(モデル構築期間：2005 年〜2013 年)に対して推定された一部の財務指標に関する非線形関数(式(3.1)における f_j)を示したものが図 3 及び図 4 である．実績デフォルト率との比較のため，図 1 を再掲している．推定された非線形関数

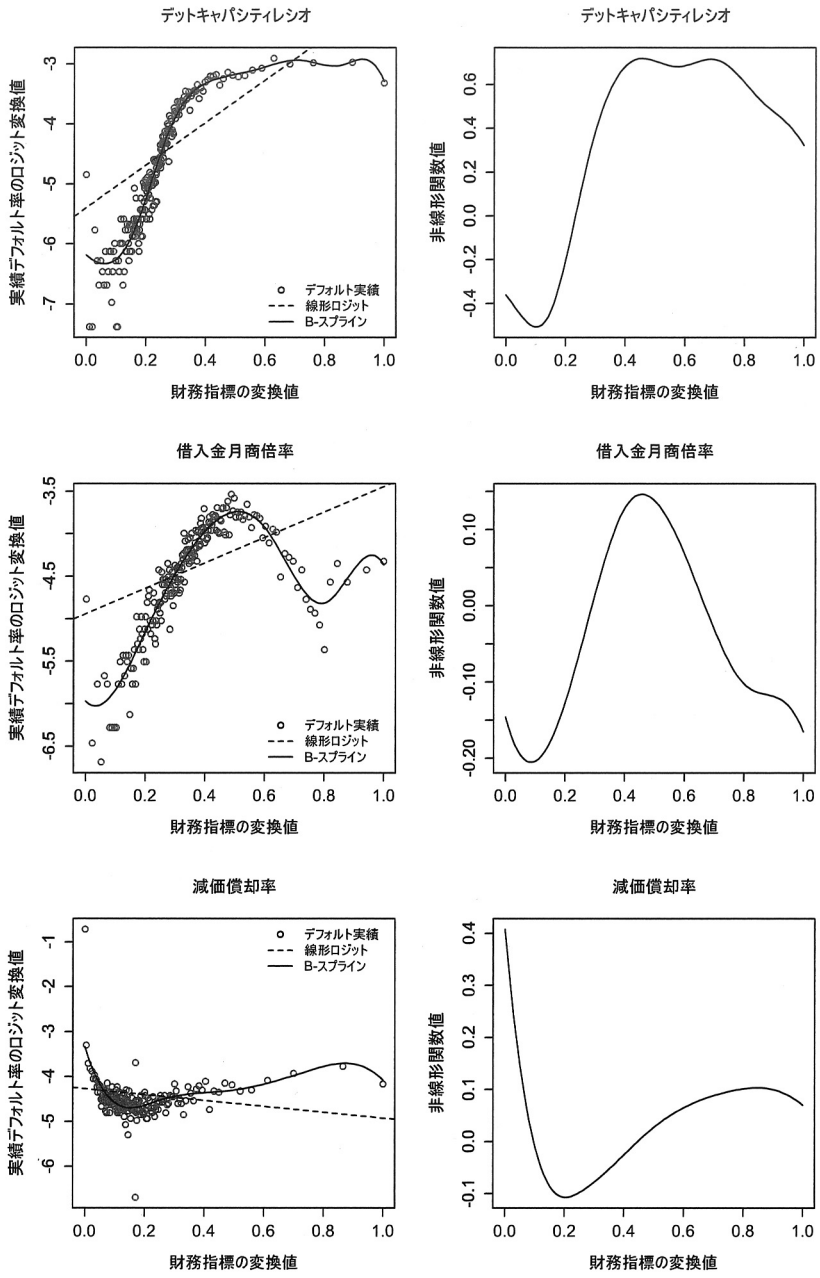


図 4. 非線形関数の推定結果(2)左列：図 1 再掲　右列：非線形関数の推定値(図中の「デットキャパシティレシオ」及び「借入金月商倍率」については neglog 変換を行っている．また全ての財務指標について，0 から 1 の範囲に収まるように線形変換を行っている.)．

(右の列)は，実績デフォルト率の変動(左の列)を，ある程度捉えていることがわかる．ただし横軸で 0 又は 1 に近い領域では，サンプルサイズが小さいため，変動に幅があることに注意する必要がある．

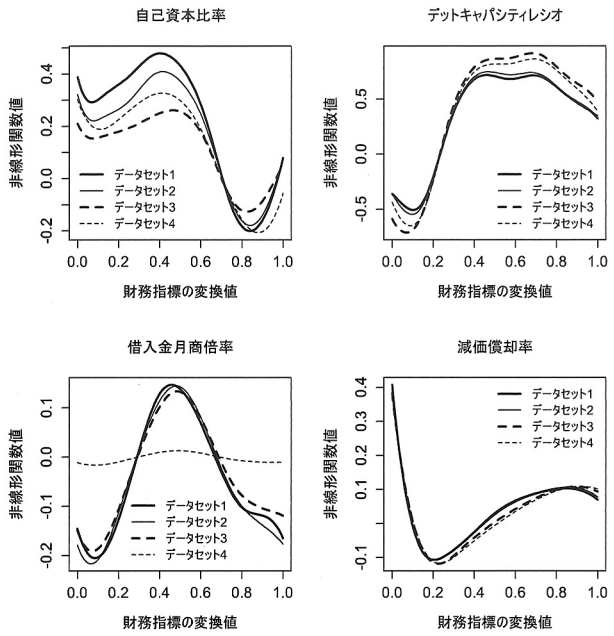


図 5. 各データセットにおける非線形関数の推定結果(図中の「デットキャパシティレシオ」及び「借入金月商倍率」については neglog 変換を行っている．また全ての財務指標について，0 から 1 の範囲に収まるように線形変換を行っている.)．

図 1 に示した財務指標の中で，提案手法(モデル 5)において，全てのデータセットで変数として選択されている「自己資本比率」,「デットキャパシティレシオ」,「借入金月商倍率」及び「減価償却率」の 4 つの財務指標について，各データセットから推定された非線形モデルの予測値を重ねて表示したものが図 5 である．図 5 をみると，モデルを構築する際に用いるデータの期間の違いによって，推定される非線形関数の水準は異なるものの，期間が異なっても非線形関数の形状には大きな違いはないことがわかる．なお，借入金月商倍率に関しては，データセット 4(モデル構築期間：2005 年〜2010 年)において，非線形関数の値が他のデータセットの場合と比較して 0 に近く，フラットに近い形状であるものの，上昇・下降のパターンは他のデータセットの場合と同様である．

5. 考察

5.1 モデルの精度

本稿では複数の銀行データを統合したデータベースを基に，B-スプラインに基づく非線形モデル及び Multistep Adaptive Group LASSO に基づく変数選択の手法を導入したデフォルト確率予測モデルの構築を行った．このようにして得られたモデルは，t 値・p 値に基づく変数選択や単純な LASSO による方法と比較して，どの期間のデータセットにおいても最も変数が少なくなっており，選択された変数の種類に大きな変動がなく，効率的かつ安定的な変数選択を行うことができた．さらに AR 値などの各種指標を用いて比較を行った結果，本稿で提案したモデルが最も推定精度が高く，当てはまりの良いモデルであることが確認された．B-スプラインに基づく非線形モデルの導入により，信用スコアと財務指標との非線形な構造を捉えること

が可能となり，モデルの推定精度が向上したと考えられる．さらに Multistep Adaptive Group LASSO に基づく変数選択の手法を導入することにより，よりコンパクトなモデルを推定することが可能となり，モデルの安定性が向上したことで，アウトオブタイムにおける推定精度の向上につながったものと考えられる．

5.2 財務指標の選択

表 3 から表 6 において，提案手法(モデル 5)の説明変数として，異なるデータセットで複数回選択された変数を見ると，利益，回転率，短期支払能力といった総合的な収益性の面から「ROA」,「売上高経常利益率」,「売上債権回転日数」,「買入債務回転日数」,「流動比率」,「現預金比率」といった，実務でもよく用いられる代表的な財務指標が選択されている．これに対してデフォルト予測や与信判断に直接的に関係すると考えられる借入・資産の面からは「デットキャパシティレシオ」,「借入金月商倍率」,「有利子負債利子率」,「現金預金対利子割引料率」,「自己資本比率」,「減価償却率」などのほか,「資産合計」やこれに占める各種資産の割合など，多くの財務指標が選択されている．収益性に関する指標を代表的なものに絞りつつ，借入・資産に重点を置くという，メリハリのある変数選択が行われている．

5.3 推定された非線形関数の形状

提案手法(モデル 5)に基づき推定された，主な財務指標の非線形関数の形状について考察する．総資産経常利益率は高い方が望ましいが，資金の必要性から総資産を処分する際に高くなる可能性もあり，極端に高すぎる又は低すぎる値は望ましくないと考えられる．自己資本比率は高い方が，デットキャパシティレシオ(有利子負債と融資の担保にできる資産との比)は低い方が望ましいが，どちらもある程度の水準を満たしていればよい指標であり，一定値以上(以下)で頭打ちになると想定される．減価償却率については，早目に償却した方が安全である一方，逆に償却が進むと経費計上分が減少してしまうという観点もある．図 3 及び図 4 における非線形関数の形状には，これらの関係が表れていると考えられる．

一部の財務指標について，モデル構築に用いるデータの期間が異なる場合における非線形関数の形状の変化を見ると，図 5 に示すように，推定される非線形関数の水準は異なるものの，期間が異なっても非線形関数の形状には大きな違いはなく，安定していることが示された．このようにして推定された各財務指標の非線形関数を用いることで，財務指標ごとに信用スコアが急激に変化する点や最も高くなる点などを判別することが可能となり，与信判断に資する情報が得られるものと期待される．

本研究において提案したデフォルト確率予測モデルは，財務指標と信用スコアとの非線形な関係が「見える」モデルの合理的・効率的な構築に寄与するものであり，各種財務指標に基づいて与信判断・審査等を行う金融実務において，有益であると考えられる．

6. 今後の課題

今後の課題として，以下の点が挙げられる．今回の分析では計算のコストを考慮して，Multistep Adaptive Group LASSO における反復回数を 2 回としたが，反復回数を多くすることがモデルの推定精度の改善に寄与するかという点に関しては検討の余地が残されている．

また，今回の手法を，より大規模なデータセットに対して分析を行うことが考えられる．具体的には，複数のデータベースを結合して得られた大規模なデータベースに対して適用することで，より多くの変数から効率的に非線形な構造を抽出できると考えられる．

謝　辞

本研究は科研費(16H02013 及び 15H03390)の助成を受けています．また改稿に当たり，有益なコメントをいただいた 2 名の査読者に感謝申し上げます．

参 考 文 献

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance*, **23**, 589–609.

Amendola, A., Restaino, M. and Sensini, L. (2012). Dynamic statistical models for corporate failure prediction in Italy, *Journal of Modern Accounting and Auditing*, **8**, 1214–1224.

Berg, D. (2007). Bankruptcy prediction by generalized additive models, *Applied Stochastic Models in Business and Industry*, **23**, 129–143.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*, Springer, Berlin.

Duffie, D. and Singleton, K. J. (1999). Modeling term structures of defaultable bonds, *Review of Financial Studies*, **12**, 687–720.

Dwyer, D. W., Kocagil, A. E. and Stein, R. M. (2004). The Moody's KMV EDF RiskCalc v3.1 Model: Next generation technology for predicting private firm risk, Moody's KMV Company, San Francisco.

Engelmann, B. and Raumeier, R. (2006). *The Basel* II *Risk Parameters*: *Estimation, Validation and Stress Testing*, Springer, Berlin.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**, 1–22.

Giordani, P., Jacobson, T., von Schedvin, E. and Villani, M. (2014). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios, *Journal of Financial and Quantitative Analysis*, **49**, 1071–1099.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall/CRC, Boca Raton, Florida.

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

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. (2013). *Applied Logistic Regression*: *Third Edition*, Wiley, New York.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models, *Annals of Statistics*, **38**, 2282–2313.

川野秀一, 松井秀俊, 廣瀬慧 (2018).『スパース推定法による統計モデリング』, 共立出版, 東京.

小西貞則 (2010).『多変量解析入門：線形から非線形へ』, 岩波書店, 東京.

Martin, D. (1977). Early warning of bank failure: A logit regression approach, *Journal of Banking and Finance*, **1**, 249–276.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression, *Journal of the Royal Statistical Society Series B*, **70**, 53–71.

Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling, *Annals of Statistics*, **37**, 3779–3821.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates, *Journal of Finance*, **29**, 449–470.

森平爽一郎 (2009).『信用リスクモデリング：測定と管理』, 朝倉書店, 東京.

尾木研三 (2017).『スコアリングモデルの基礎知識：中小企業融資における見方・使い方』, 金融財政事情研究会, 東京.

尾木研三, 戸城正浩, 枇々木規雄 (2015). 小規模企業向け保善別回収率モデルの構築と実証分析,『ファイナンスとデータ解析(ジャフィー・ジャーナル：金融工学と市場計量分析)』(日本金融・証券計量・工学学会 編), 168–201, 朝倉書店, 東京.

Perederiy, V. (2009). Bankruptcy prediction revisited: Non-traditional ratios and lasso selection, European University Viadrina, Working Paper 16, Frankfurt.

阪本亘, 高橋史朗, 竹内正弘 (2010). 正則化法を用いたロジスティック回帰モデルによる多次元データでの変数選択手法に関する研究, 数理解析研究所講究録, **1703**, 32–52.

桜井明 (1981).『スプライン関数入門：情報処理の新しい手法』, 東京電機大学出版局, 東京.

白田佳子 (2008).『倒産予知モデルによる格付けの実務』, 中央経済社, 東京.

高橋久尚, 山下智志 (2002). 大規模データによるデフォルト確率の推定：中小企業信用リスク情報データベースを用いて, 統計数理, **50**, 241–258.

Tian, S., Yu, Y. and Guo, H. (2015). Variable selection and corporate bankruptcy forecasts, *Journal of Banking and Finance*, **52**, 89–100.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B*, **58**, 267–288.

冨岡亮太 (2015).『スパース性に基づく機械学習』, 講談社, 東京.

山下智志, 安道知寛 (2006). 時間依存共変量を用いたハザードモデルによるデフォルト確率期間構造の推計手法, 統計数理, **54**, 23–38.

山下智志, 三浦翔 (2011).『信用リスクモデルの予測精度：AR 値と評価指標』, 朝倉書店, 東京.

山内浩嗣 (2010). 多目的遺伝的アルゴリズムを用いたスコアリングモデルのチューニング,『定量的信用リスク評価とその応用(ジャフィー・ジャーナル：金融工学と市場計量分析)』(日本金融・証券計量・工学学会 編), 24–54, 朝倉書店, 東京.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalized learning problems, *Statistics and Computing*, **25**, 1129–1141.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*: *Series B*, **68**, 49–67.

Estimation of Default Probability Using Regularized Nonlinear Logit Model with B-spline and Adaptive Group LASSO

Isao Takabe[1,2] and Satoshi Yamashita[3]

[1]Department of Statistical Science, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies
[2]Consumer Statistics Division, Statistics Bureau, Ministry of Internal Affairs and Communications
[3]The Institute of Statistical Mathematics

Linear binomial logit models are widely used for the assessment and evaluation of a company's default probability based on a company default database. Previous studies have been criticized on the following bases: (1) insufficient attention to nonlinear relationships between default probabilities and financial indicators; and (2) too much time required for variable selection from many candidates for regressors in the models. In this study, we aimed to solve these problems simultaneously by combining the following techniques: (1) nonlinear and nonparametric logistic regression model based on the B-spline; and (2) reasonable variable selection using adaptive group LASSO. We constructed a default probability prediction model using datasets of multiple periods, based on our own database of data from Japanese banks. The proposed model achieved more effective performance than models in other related studies. Compared with the method using t-statistic (p-value) or simple LASSO, our proposed method had the smallest number of explanatory variables in any period, and achieved more efficient variable selection. Moreover, estimation accuracy was improved from the viewpoint of AR (accuracy ratio) value.

Key words: Credit risk, B-spline, adaptive group LASSO.

統計数理(2018)
第 66 巻 第 2 号 319–337

［研究ノート］

トータルパワー寄与率を用いた海洋生態システムにおける因果性推測

ソルヴァン加藤 比呂子[1]・Subbey Sam[2,3]

(受付 2017 年 3 月 16 日；改訂 2018 年 7 月 30 日；採択 8 月 2 日)

要　　旨

多変量時系列データの変量間相互関係を推測する方法として，Ozaki (2012)は，Granger (1969)と Geweke (1982)のペアワイズ因果性推測(Partial pairwise causality)と赤池 (1968)の提案した相対パワー寄与率(Total causality)を統合した因果性推測を提案した．我々は赤池情報量規準の枠組みで，因果関係の有意性に関する規準を加え，シミュレーションデータによりその手法の検証をおこなった．また，実データ分析として，バーレンツ海域の生態系における食物連鎖(food web)で重要な 4 種類の海洋生物(シシャモ，タラ，オキアミ，ニシン)のバイオマス時系列データを用い，シシャモの年齢毎に生物種間のフィードバック関係を考察した．2–3 年齢のシシャモを含むフィードバックシステムは，1 年齢もしくは 4 年齢のシシャモを含むフィードバックシステムよりもより多くの生物種間の相互関係を示した．それらはバーレンツ海の food web に関するこれまでの先行研究を裏付け，シシャモがバーレンツ海域生態システムの food web に関連する生物種間の重要な駆動源になることを明らかにした．本稿が提案する手法は，海洋学研究で対象となる複雑な生態系において，生物種間，環境因数間の因果関係を推測する一手段として有用であると考えられる．

キーワード：多変量自己回帰モデル，多変量時系列データ，フィードバックシステム，Granger の因果性，海洋生態システム，バーレンツ海．

1. はじめに

近年の地球温暖化に伴う海洋生態系と海洋生物種の資源量の変化，生物種と環境要因との因果関係を予測することは，学術的・商業的に重要である．バーレンツ海域では，その漁業管理のために，ロシアとノルウェーの共同評価 (WGIBAR, 2016)がおこなわれている．その調査を通じて，環境的要因(Abiotic)と海洋生物種(Biotic)に関する高次元の時系列データが観測されている．しかし，それらのデータ分析手法としては，時系列モデルに基づくものは殆どみられず，その多くは，International Council for the Exploration of the Sea(ICES)でスタンダードと

[1] Marine Mammals Research Group, Institute of Marine Research, P.O.Box 1870, Nordness N-5817 Bergen, Norway

[2] Research Group on Fisheries Dynamics, Institute of Marine Research, P.O.Box 1870, Nordness N-5817, Bergen, Norway

[3] Department of Natural Resources, Cornell University, Room120 Fernow Hall, Ithaca, New York 14853, USA

推奨されている主成分分析や対応分析が用いられている（WGIBAR, 2016）．時系列データにおける主成分分析は，時間領域か周波数領域に適用するかで，観測データを独立同一分布として扱えるかどうかが変わり，それらの結果の解釈は全く異なることも気をつけなければならない（川崎, 2001）．ICES で推奨している通常の主成分分析は，サンプル間を独立と仮定しているため，理論的には時間軸上のデータにそのまま適用するにはふさわしくなく，まして変量間の因果関係を推察するには適切ではない．Planque and Arneberg（2017）は，主成分分析による Abiotic と Biotic を含む高次元時系列データの分析結果の信憑性が極めて薄いことを指摘している．

多次元時系列間の因果性推測については，1956 年の Wiener の論文にさかのぼる（Wiener, 1956）．2 次元の時系列データについて，1 つの時系列に対しもう一方の時系列からの関係を考慮した方が最小二乗値が小さい場合，2 変量間に因果関係が存在するということを提唱した．その後 Granger がその考えを経済時系列データ間の因果性分析に適用した（Granger, 1969）．Granger の因果性は基本的に時間軸領域における分析で，2 変量自己回帰モデルを適用しその予測誤差の比較により推測をおこなっている．一方で，1968 年に，赤池は秩父セメントのロータリーキルンの駆動制御のために，多変量自己回帰モデルを用いたフィードバックシステム解析を考案した（Akaike, 1968）．この手法はシステムを撹乱すると考えられるノイズに注目している．実データ分析では有色ノイズであるが，モデルの導出過程で白色化を施し，最小二乗法でモデルの係数の推定を可能にした．係数と予測誤差の共分散行列を用い，時間領域では閉ループと開ループのインパルス応答を，周波数領域では相対パワー寄与率を計算することで，双方の領域から変数間の因果関係を推察できる分析手法であった．周波数領域で提案された相対パワー寄与率は，長/中/短周期において，ある変量のパワーに対し各変量からのノイズ（パワー）の寄与がどの程度の割合か算出することができる．この手法は分析対象となるシステムに関連する多次元時系列データをまとめて取り扱えるため，医学，工学，経済学，生物学等様々な分野での適用例が多数報告されている（Akaike and Kitagawa, 1994）．

多次元時系列データ間の相互関係を分析するには，このような多変量時系列モデルを用いた因果性分析の方が，主成分分析による変動の類似性に基づく手法よりも，変量から変量への方向性を踏まえた関係を推測するのに有用である．時系列間の相互関係に対し多変量自己回帰モデルを適用して，多変量間のコヒーレンスを用いた手法（Baccalá and Sameshima, 2001）も提案されているが，それらは多変量自己回帰モデルの係数のみ，すなわち伝達特性のみに注目を対象としたものである．しかし，実際にはシステムを構成するサブシステムを撹乱するノイズ項を考慮することは大切で，システムのダイナミクスを左右するイノベーションと言われている（Ozaki, 2012）．多変量間のフィードバックを想定した場合，伝達特性とイノベーションの寄与双方から考察することが時系列間の因果性を推測する上で重要である（Bosch-Bayard et al., 2012）．

Ozaki（2012）の 14 章では，赤池の相対パワー寄与率を全周波数領域で積分し（Ozaki（2012）で「赤池の Total causality」と呼んでいる），関連した全変量を対象としたパワーの寄与からある変数の関係を外した寄与について定式化している．全変量を対象としたパワーの線形和に対し，周波数領域において総和をとると，変動の分散そのものになり，その対数表示はまさにモデルの対数尤度を示す．さらに，Granger-Geweketype causality の，2 変量間因果性の有無に関して予測誤差の対数を比較する手法（Ozaki（2012）で「Partial pairwise causality」と呼んでいる）を取り上げ，赤池の Total causality の枠組みで計算し，ある変量と全変量とのペアワイズに比較する手法を提案した．これは，Granger が 2 変量のみの時系列モデルによる予測誤差を頼りにするだけでは多変量全ての間の因果推測をするには不足であるということと，赤池が全変量を対象としているが，複雑なフィードバックシステムにおける因果性推測には，ペアワイズで

の比較が予備的に因果関係を探す手段として役立つという，双方の困難な点を補える利点がある．データの背後に潜む本当の因果関係は，限られたサンプル内のデータで予備的に探られた結果と，データに関する知見や背景の情報と合わせて推測されるだろう．この Ozaki (2012)の提案手法は，実データ分析においてシステマティックで実用的な方法と期待されるが，未だ具体的な適用例が少なく，複雑な海洋生態システムのダイナミクスの解明においては皆無である．

多次元時系列データ間に内在する因果性を分析する手法が取り上げる問題は，因果関係の有無，因果関係の強度，因果関係の可視化，因果関係による機能の分析等が考えられる．先にあげた，Granger (1969)，Geweke (1982)はペアワイズのモデル化に基づいて，また，Akaike は関連した全ての変量の関係を含むモデルで因果関係の有無を調べる手法と考えられる．Granger (1969)のコヒーレンス関数や Akaike の相対パワー寄与率やインパルス応答関数のグラフ化は因果関係を可視化と言えよう．因果関係による機能の分析では，赤池・中川 (1972)のフィードバックシステムの動的制御や，加藤・石黒 (1997)によるフィードバック経路切断による経済システムの動的解析があげられる．

本論文では，Ozaki (2012)の手法に，さらに因果関係の有無を査定する規準を情報量規準の枠組みから導入し，海洋生態システムの Abiotic や Biotic の因果関係の有無や強度，可視化の可能性を提案する．まずシミュレーションデータによる検証を行い，次に実データ分析として，バーレンツ海生態システムの food web において重要な 4 種類の海洋生物のバイオマスを用い，生物種間の因果関係を推測する．

ノイズ寄与率は，赤池の他に inverse スペクトルを適用したパワー寄与率を考慮した．実データ分析として，バーレンツ海生態システムにおける food web に関する生物資源のデータに適用し，そのシステムにおける生物種間の因果関係を推測した．

2. 統計的手法

観測された k 次元時系列データ $\boldsymbol{x}_t = (x_1(t), x_2(t), \ldots, x_k(t))', t = 1, \ldots, N$ とする(ここで，$(\cdot)'$ は転置の記号)．これらのデータは，以下に示すような多変量自己回帰(Multivariate auto-regressive，MAR)過程の実現値と仮定する：

$$\boldsymbol{x}_t = \sum_{m=1}^{M} \boldsymbol{A}_m \boldsymbol{x}_{t-m} + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

ここで M は自己回帰の次数，$\boldsymbol{A}_m$ は自己回帰モデルの係数，そして $\boldsymbol{\varepsilon}_t$ は平均ゼロベクトル，分散共分散行列 Σ に従う多変量正規分布に従うとする．自己回帰係数の推定には多数の方法があり，最小二乗法や Yule-Walker 法，また，Yule-Walker 法を効率よく計算できる Levinson アルゴリズム等が提案されている (Ozaki, 2012, 4 章)．自己回帰の最適な次数 M は Akaike Information Criteria(AIC) (Akaike, 1974)等の情報量規準による統計的モデル選択で同定でき，予測誤差系列により共分散行列が計算できる．本稿におけるモデル選択では，AIC(AIC $= -2 \times$ モデルの最大対数尤度 $+ 2 \times$ モデルの自由パラメータ数)を用いる．

2.1 赤池の相対パワー寄与率

自己回帰係数とフーリエ変換から，周波数 f に対する周波数応答関数 $\boldsymbol{F}_f$ が求まり，以下のようなパワースペクトルが求まる：

$$\boldsymbol{P}_f = \boldsymbol{F}_f \Sigma \boldsymbol{F}_f^* = \begin{pmatrix} p_{11f} & p_{12f} & \cdots & p_{1kf} \\ p_{21f} & p_{22f} & \cdots & p_{2kf} \\ & \vdots & & \\ p_{k1f} & p_{k2f} & \cdots & p_{kkf} \end{pmatrix}, \quad 0 \leq f \leq 0.5\Delta, \tag{2.2}$$

ここで $\boldsymbol{F}_f^*$ は $\boldsymbol{F}_f$ の共役複素数，Δ は観測値のサンプリング間隔とする．$\boldsymbol{P}_f$ の非対角成分はクロスパワースペクトルである．もし，Σ の非対角成分が非常に小さい，すなわち各変量の予測誤差は独立であることが仮定できると，i 番目の変量 x_i のパワースペクトルは他の変量 x_j からの周波数応答関数 F_{ijf} と予測誤差の分散 σ_{jj}^2 の影響を含む項の和

$$p_{iif} = |F_{i1f}|^2 \sigma_{11}^2 + \cdots + |F_{iif}|^2 \sigma_{ii}^2 + \cdots + |F_{ikf}|^2 \sigma_{kk}^2 \tag{2.3}$$

の形で示すことができ，j 番目の変量 x_j からのノイズの影響 r_{ijf} を以下のように示すことができる：

$$r_{ijf} = \frac{|F_{ijf}|^2 \sigma_{jj}^2}{|p_{iif}|} \in [0, 1]\,. \tag{2.4}$$

(2.4)式で示された比を赤池の相対パワー寄与率と呼ぶ．予測誤差を独立としない場合の相対パワー寄与率は Tanokura and Kitagawa (2014)により提案されている．

Ozaki (2012)では，多次元時系列データに MAR モデルを適用しその推定値と予測誤差の情報を用いて(2.4)を算出するので，全時系列データの全ての因果関係を含んだ情報(total causality)を用いて(2.4)式を算出したものを，partial innovation contribution と呼んでいる．それにより推測される結果を Partial 因果性(partial causality)と呼んでいる．

2.2 Granger と Geweke 型の因果性

観測された 2 つの時系列 x_t と y_t に対して時系列モデルを考えることにする．モデル A を適用した際の時系列の予測誤差の分散を $\mathrm{Var}(*|A)$ と表すことにすると，Granger と Geweke の因果性は以下のように定義する（Ozaki, 2012）：

定義 1.（Granger の因果性）$x_{t-} = (x_{t-1}, x_{t-2}, \ldots)'$，$y_{t-} = (y_{t-1}, y_{t-2}, \ldots)'$ とすると，$\mathrm{Var}(x_t|x_{t-}) - \mathrm{Var}(x_t|x_{t-}, y_{t-}) > 0$ ならば，時系列 x_t の変動はもう一方の時系列 y_t の変動に起因している．

定義 2.（Geweke の因果性）$\log|\mathrm{Var}(x_t|x_{t-})| - \log|\mathrm{Var}(x_t|x_{t-}, y_{t-})| > 0$ ならば x_t の変動は y_t の変動に起因している．

ここで，$\mathrm{Var}(x_t|x_{t-})$ は x_{t-} によるモデルで予測された x_t の予測誤差の分散を，$\mathrm{Var}(x_t|x_{t-}, y_{t-})$ は $x_{1,t-}$ と y_{t-} を含むモデルで予測された y_t の予測誤差の分散を示す．定義 2 で示された式，左辺の第一項で示した予測誤差を算出する予測モデルを $\mathrm{Model}^{(0)}$ とし，同じく左辺第二項に示した予測誤差を算出する予測モデルを $\mathrm{Model}^{(1)}$ とすると，Ozaki (2012)では，Granger と Geweke の定義は，本質的に以下のような予測モデルの尤度の比較に相当すると示している：

定義 3. ある観測された時系列 x_t の変動が時系列 y_t に起因されているということは，モデルの対数尤度の観点から，$-2\log l^{\mathrm{Model}^{(0)}}(x_t) - (-2)\log l^{\mathrm{Model}^{(1)}}(x_t) > 0$ となる．

ここで，$l^{\mathrm{Model}^{(0)}}(x_t)$ と $l^{\mathrm{Model}^{(1)}}(x_t)$ は $\mathrm{Model}^{(0)}$ と $\mathrm{Model}^{(1)}$ で予測した x_t の尤度を示す．

2.3 Ozaki（2012）の因果性推測

さて，Kolmogorov（1941）は，予測誤差 σ^2 とパワースペクトル $p(f)$ との関係を

$$\sigma^2 = \exp\left\{\int_{-1/2}^{1/2} \log p(f) df\right\} \tag{2.5}$$

と示した．その周波数領域において積分したものをトータルパワーの全周波数領域の総和と言うことにしよう．多次元時系列データに MAR モデルを適用する話に戻り，(2.3)式でパワースペクトルが求められたとすると，今 i 番目の変数のトータルパワーの全周波数領域の総和は

$$\begin{aligned}\log \sigma_i^2 &= \int_{-1/2}^{1/2} \log p_{ii}(f) df = \int_{-1/2}^{1/2} \log \sum_{k=1}^{K} |\alpha_{ik}(f)|^2 \sigma_k^2 df \\ &= \int_{-1/2}^{1/2} \log(|F_{i1}(f)|^2 \sigma_1^2 + |F_{i2}(f)|^2 \sigma_2^2 + \cdots + |F_{iK}(f)|^2 \sigma_K^2) df\end{aligned} \tag{2.6}$$

と示すことができる．また，j 番目の変数からの影響を除いた場合のトータルパワーの全周波数領域の総和は，(2.3)式及び(2.6)式に示されている全変数の周波数応答とノイズの分散の積の線形和から，$|F_{ij}(f)|^2 \sigma_j^2$ を除いた式を計算することになる．例えば $j = 2$ とすると，

$$\begin{aligned}\log \sigma_{i\wedge 2}^2 &= \int_{-1/2}^{1/2} \log p_{ii}^{(j)}(f) df \\ &= \int_{-1/2}^{1/2} \log \left(\sum_{k=1}^{j-1} |\alpha_{ik}(f)|^2 \sigma_k^2 + \sum_{k=j+1}^{k} |\alpha_{ik}(f)|^2 \sigma_k^2 \right) df \\ &= \int_{-1/2}^{1/2} \log(|\alpha_{i1}(f)|^2 \sigma_1^2 + |\alpha_{i3}(f)|^2 \sigma_3^2 + \cdots + |\alpha_{iK}(f)|^2 \sigma_K^2) df\end{aligned} \tag{2.7}$$

となる．これらを以下のように比較すると，

$$\begin{aligned}\log \sigma_{i\wedge j}^2 - \log \sigma_{ii}^2 &= \int_{-1/2}^{1/2} \log p_{ii}^{(j)}(f) df - \int_{-1/2}^{1/2} \log p_{ii}(f) df \\ &= \int_{-1/2}^{1/2} \log \frac{p_{ii}^{(j)}(f)}{p_{ii}(f)} df = \int_{-1/2}^{1/2} \log \frac{p_{ii}(f) - |\alpha_{ij}(f)| \sigma_{jj}^2}{p_{ii}(f)} df \\ &= \int_{-1/2}^{1/2} \log \left(1 - \frac{|\alpha_{ij}(f)| \sigma_{jj}^2}{p_{ii}(f)}\right) df\end{aligned} \tag{2.8}$$

となり，定義 3 で示したような，ある変数からの影響を含むか含まないかの 2 つのモデルの比較と等価になる．また，それは(2.7)式の最終項にあるように，全変数からある 1 つの変数の関係を外した場合の Partial パワー寄与（全周波数領域の総和をとった）をみることになる．本稿では，(2.7)式で得られる値を対数尤度の差の値ということで，本稿では ΔLL 値として示すことにする．

ΔLL 値がどの値になると変数 j から変数 i に関係していると判断するか．本稿では，モデルの比較として AIC の観点からみることにする．変数 j から変数 i の関係を外した(2.6)のように示した擬似(対数)尤度とトータル寄与の擬似(対数)尤度をもつモデル，すなわち 1 変数だけ外したモデルとの比較になるので，

$$\begin{aligned}(2.9)\qquad \mathrm{AIC}_{\sigma^2_{i\wedge j}} - \mathrm{AIC}_{\sigma^2_{ii}} &= -2\times\log\sigma^2_{i\wedge j} + 2\times(k-1) - \left\{-2\times\log\sigma^2_{ii} + 2\times k\right\}\\ &= \Delta\mathrm{LL} - 2\end{aligned}$$

となり，(2.7)式の値が 2 より小さい場合は差があるとは言い難くなる．

モデルの対数尤度の比較という点からは，ΔLL を尤度比検定の枠組みに拡張することもできるかもしれない．その場合帰無仮説は変数 j から変数 i に関係はない，すなわち $|F_{ij}(f)|^2\,\sigma^2_j = 0$ ということになる．尤度比検定量を調べるには，リサンプリングしたデータに基づいてモデルを適用した後に ΔLL を算出し，それを何回か繰り返すことによって ΔLL の帰無分布を作成しなければならない．しかし帰無仮説が $|F_{ij}(f)|^2\,\sigma^2_j = 0$ となるには，MAR モデルに含まれる全変量の自己回帰係数が関係していて，予測誤差の共分散については，実データに適用したときに独立と仮定できてもわずかの非対角要素も存在するので，帰無分布を求めるための厳密な性質を踏まえたデータをリサンプリングすることは難しい．その点に関しては，理論的な背景を踏まえ別稿で議論することにする．従って，本稿では，因果性を調べたい対象となるある変量の Total パワーに対し，関係を調べたい変量とのペアごとに ΔLL を計算し，(2.9)をもとにこの値が 2 から 3 以上の場合，有意な因果関係を推測できるものとする．

2.4 Bosch-Bayard の pNCR

赤池のパワー寄与率と似たアプローチで，Bosch-Bayard により partial noise contribution ratio(pNCR)が提案されている（Bosch-Bayard et al., 2012）．この手法は(2.2)式で示したパワースペクトルの逆行列 $\boldsymbol{P}_f^{-1}$ を用いて(2.4)式と同様な比率を算出する．ある対象とする変数のパワースペクトルの逆行列に対し，他の変数からの寄与の割合を計算する．すなわち

$$(2.10)\qquad r^b_{ijf} = \frac{(F^*_{ijf}F_{ijf})/\sigma^2_{jj}}{|p_{iif}|^{-1}}$$

と表される．赤池のパワー寄与率と同様，多次元時系列データに対して MAR モデルから推定された係数と予測誤差の共分散を用い，対象とする変数とそれへの影響を調べるための変数のペアで周波数領域において r^b_{ijf} をプロットすることで因果関係の可視化が可能となる．本稿では，赤池の手法によるトータルパワーに加え，Bosch-Bayard のトータルパワーも検討することにする．

3. シミュレーションデータ分析

ここでは，2 変量のみ対象とする Granger（1969）と Geweke（1982）の方法と，全変量の情報を用いる Ozaki（2012）の ΔLL について，3，5 変量シミュレーションデータにより比較する．また，ΔLL について，サンプル数に関する Sensitivity と，因果関係の強度との関係について検討する．

3.1 Granger と Geweke の因果性分析と ΔLL の実装の比較—3 変量データの場合

データは以下に示される多変量自己回帰モデルから生成された 200 時点の 3 変量（x_t，y_t，z_t，$t = 1,\ldots,200$）時系列，sim1 と sim2 を用いる：

$$\begin{aligned}\text{sim1:}\quad & x_t = 0.3x_{t-1} - 0.45y_{t-1} + \varepsilon_{x,t}\\ & y_t = 0.4y_{t-1} + 0.35z_{t-1} + \varepsilon_{y,t}\\ & z_t = 0.25z_{t-1} + \varepsilon_{z,t}\end{aligned}\qquad\qquad \begin{aligned}\text{sim2:}\quad & x_t = 0.3x_{t-1} - 0.45z_{t-1} + \varepsilon_{x,t}\\ & y_t = 0.4y_{t-1} + 0.35z_{t-1} + \varepsilon_{y,t}\\ & z_t = 0.25z_{t-1} + \varepsilon_{z,t}\end{aligned}$$

ここで $\varepsilon_{x,t}$，$\varepsilon_{y,t}$，$\varepsilon_{z,t}$ は平均 0 分散 1 の独立同一分布に従うものとする．生成されたデータの

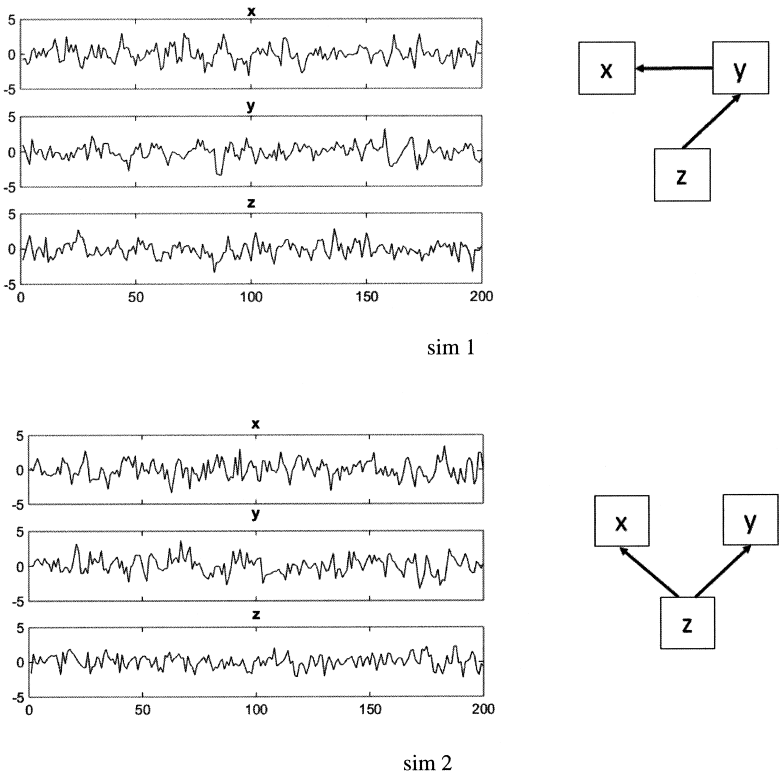


図 1. MAR (1) model によって生成された 3 変量シミュレーションデータ sim1 と sim2 とその変量間の関係.

プロットと変量間に仮定しているの関係を示したダイアグラムを図 1 に示す.

Granger と Geweke による方法では，各変量に対して以下のようなモデルを適用する：

x_t に対し，x_{t-1} からのみ説明されるモデル　$x_t = a_{xx}x_{t-1} + \varepsilon_{xx,t}$

x_{t-1} と y_{t-1} から説明されるモデル　$x_t = a_{xx}x_{t-1} + a_{xy}y_{t-1} + \varepsilon_{xy,t}$

x_{t-1} と z_{t-1} から説明されるモデル　$x_t = a_{xx}x_{t-1} + a_{xz}z_{t-1} + \varepsilon_{xz,t}$

y_t に対し，y_{t-1} からのみ説明されるモデル　$y_t = a_{yy}y_{t-1} + \varepsilon_{yy,t}$

y_{t-1} と x_{t-1} から説明されるモデル　$y_t = a_{yx}x_{t-1} + a_{yy}y_{t-1} + \varepsilon_{yx,t}$

y_{t-1} と z_{t-1} から説明されるモデル　$y_t = a_{yy}y_{t-1} + a_{yz}z_{t-1} + \varepsilon_{yz,t}$

z_t に対し，z_{t-1} からのみ説明されるモデル　$z_t = a_{zz}z_{t-1} + \varepsilon_{zz,t}$

z_{t-1} と x_{t-1} から説明されるモデル　$z_t = a_{zx}x_{t-1} + a_{zz}z_{t-1} + \varepsilon_{zx,t}$

z_{t-1} と y_{t-1} から説明されるモデル　$z_t = a_{zy}y_{t-1} + a_{zz}z_{t-1} + \varepsilon_{zy,t}$

a_{**} はモデルの係数で，最小二乗法から推定される．それぞれのモデルに対する予測誤差の分散を計算し，定義 1 に示したように分散間の差と，定義 2 に示したように分散の対数間の差を算出する．その結果を表 1 にまとめた.

まず予測誤差の分散の差はどの場合も関係が入らない変量との差が 0 より大きくなる．sim1 については，各変量ごとでそれぞれの誤差間の分散の差を比較すると，x_t に対して y_{t-1} を含む

表 1. sim1 と sim2 データにおける Granger と Geweke の方法を適用した結果.

	変量	予測誤差の分散	誤差間の差	誤差の対数の差
sim1	x_t	$\mathrm{Var}(x_t \mid x_{t-1}) = 1.1781$	−	−
		$\mathrm{Var}(x_t \mid x_{t-1}, y_{t-1}) = 0.9770$	$1.1781 - 0.9770 = 0.2011$	0.1872
		$\mathrm{Var}(x_t \mid x_{t-1}, z_{t-1}) = 1.1771$	$1.1781 - 1.771 = 0.0010$	8.6002×10^{-4}
	y_t	$\mathrm{Var}(y_t \mid y_{t-1}) = 0.8913$	−	−
		$\mathrm{Var}(y_t \mid x_{t-1}, y_{t-1}) = 0.8888$	$0.8913 - 0.8888 = 0.0024$	0.0027
		$\mathrm{Var}(y_t \mid y_{t-1}, z_{t-1}) = 0.8359$	$0.8913 - 0.8359 = 0.0554$	0.0642
	z_t	$\mathrm{Var}(z_t \mid z_{t-1}) = 0.9563$	−	−
		$\mathrm{Var}(z_t \mid z_{t-1}, x_{t-1}) = 0.9534$	$0.9563 - 0.9534 = 0.0029$	0.0030
		$\mathrm{Var}(z_t \mid z_{t-1}, y_{t-1}) = 0.9551$	$0.9563 - 0.9551 = 0.0012$	0.0012
sim2	x_t	$\mathrm{Var}(x_t \mid x_{t-1}) = 1.4536$	−	−
		$\mathrm{Var}(x_t \mid x_{t-1}, y_{t-1}) = 1.4470$	$1.4536 - 1.4470 = 0.0066$	0.0045
		$\mathrm{Var}(x_t \mid x_{t-1}, z_{t-1}) = 1.2011$	$1.4536 - 1.2011 = 0.2525$	0.1908
	y_t	$\mathrm{Var}(y_t \mid y_{t-1}) = 1.2777$	−	−
		$\mathrm{Var}(y_t \mid x_{t-1}, y_{t-1}) = 1.2769$	$1.2777 - 1.2769 = 8.5100\times10^{-4}$	6.6625×10^{-4}
		$\mathrm{Var}(y_t \mid y_{t-1}, z_{t-1}) = 1.2369$	$1.2777 - 1.2369 = 0.0408$	0.0325
	z_t	$\mathrm{Var}(z_t \mid z_{t-1}) = 0.8519$	−	−
		$\mathrm{Var}(z_t \mid z_{t-1}, y_{t-1}) = 0.8510$	$0.8519 - 0.8510 = 9.2259\times10^{-4}$	0.0011
		$\mathrm{Var}(z_t \mid x_{t-1}, z_{t-1}) = 0.8489$	$0.8519 - 0.8489 = 0.0030$	0.0035

モデルの方が z_{t-1} を含むモデルより誤差間の差が大きく，y_t に対しては z_{t-1} から説明されるモデルの方が x_{t-1} を含むモデルより誤差間の差が大きくなるため，x_t と y_t の変動には，y_{t-1} からと z_{t-1} からの関わりがあると推察される．しかし，z_t に関しては，他の変量からの関わりがあると言えるのか言えないのか不明である．誤差間の差の大小関係からだと x_{t-1} からの関わりが若干ありそうである．sim2 については，各変量毎にそれぞれの誤差間を比較するとかなり大きな差がみられるので，x_t と y_t に z_{t-1} から関わりがあると推察されるが，z_t に x_{t-1} と y_{t-1} からの関係はないはずだが，これらの比較では y_{t-1} からの関係があるかどうか判断が難しい．x_t に y_{t-1} からの関係もありそうである．定義 1 と 2 に沿った誤差分散の差の比較を明解にするために，2 つのモデルの残差和(誤差の総和)を比較する F 検定が有用である．本稿では R の library(vars)にある関数 causality を実装してみる．各検定から得られた p 値を表 2 (p_{xy} は変量 y_{t-1} から x_t への関係に関する値)にまとめる．

5% 未満の p 値を有意な差であるとすると，表 2 では，概ね正解に近い関係を示せているとみられるが，sim1 に関しては，z_{t-1} から x_t と x_{t-1} から z_t への関係が正解と異なっていた．

さて，ΔLL による因果推測については，Akaike のパワー寄与率(RPC)と Bosch-Bayard の pNCR を用いた場合の ΔLL を算出する．(2.9)式で言及したことを考慮し，$\Delta\mathrm{LL} > 2.5$ を満たす変量からの関係は有意であるとする．図 2 に ΔLL の値を変量毎にプロットした図と，その値から推察されるダイアグラムを示す．それぞれのグラフのタイトルに示した変量に対し，x 軸に示した変量を外した場合の ΔLL 値を × でプロットし，$\Delta\mathrm{LL} > 2.5$ の場合は○をつけてい

表 2. sim1 と sim2 データにおける Granger 因果性の F 検定の結果.

sim1	sim2
$\begin{pmatrix} p_{xy} & p_{xz} \\ p_{yx} & p_{yz} \\ p_{zx} & p_{zy} \end{pmatrix} = \begin{pmatrix} 0.0 & 2.2\times10^{-2} \\ 0.43 & 1.6\times10^{-9} \\ 9.9\times10^{-4} & 0.80 \end{pmatrix}$	$\begin{pmatrix} p_{xy} & p_{xz} \\ p_{yx} & p_{yz} \\ p_{zx} & p_{zy} \end{pmatrix} = \begin{pmatrix} 0.35 & 4.0\times10^{-10} \\ 0.72 & 1.2\times10^{-2} \\ 0.64 & 0.39 \end{pmatrix}$

る．その場合は，外された変量はタイトルの変量に寄与していることになる．どの図も，自分自身からの影響に関しては非常に大きいため，それが外されると y 軸で示している範囲を大幅に超えているのでプロットされていない．RPC も pNCR による場合も，ΔLL はシミュレーションデータを生成するときに用いた関係と同じ関係を導き出している．

3.2 Granger と Geweke の因果性分析と ΔLL の実装の比較—5 変量データの場合

3.1 節の例と比べ，もう少し複雑な関係を含む 5 変量のシミュレーションについても実験をおこなった．以下がデータを生成するモデルである：

$$
\begin{aligned}
x_{1,t} &= 0.7x_{1,t-1} + 0.5x_{2,t-1} + 0.6x_{4,t-1} + \varepsilon_{1,t} \\
x_{2,t} &= 0.7x_{2,t-1} + \varepsilon_{2,t} \\
x_{3,t} &= 0.7x_{3,t-1} + 0.6x_{5,t-1} + \varepsilon_{3,t} \\
x_{4,t} &= 0.7x_{4,t-1} + 0.6x_{5,t-1} + \varepsilon_{4,t} \\
x_{5,t} &= 0.7x_{5,t-1} + \varepsilon_{5,t}.
\end{aligned}
$$

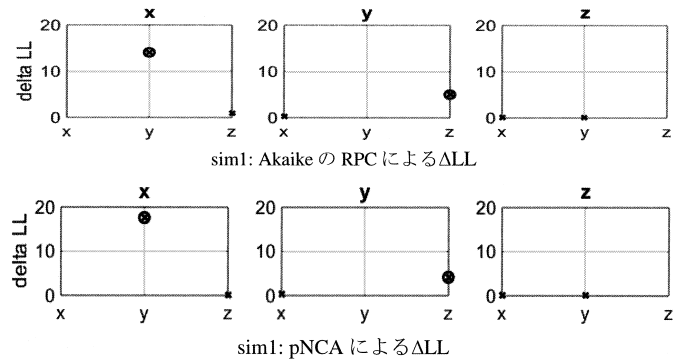
生成されたデータ 200 点について，上記と同様の Granger 因果性 F 検定をおこなうと，以下のような p 値が得られる：

$$
\begin{pmatrix} p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{34} & p_{35} \\ p_{41} & p_{42} & p_{43} & p_{45} \\ p_{51} & p_{52} & p_{53} & p_{54} \end{pmatrix} = \begin{pmatrix} 2.0\times10^{-4} & 1.93\times10^{-6} & <2.2\times10^{-16} & 1.1\times10^{-5} \\ 0.91 & 0.11 & 0.84 & 0.97 \\ 0.35 & 0.87 & 0.040 & <2.2\times10^{-6} \\ 0.0053 & 0.64 & 0.28 & <2.2\times10^{-16} \\ 0.58 & 0.96 & 0.72 & 0.69 \end{pmatrix}
$$

変量数の増加に伴い検定回数も増えることによる多重検定問題を回避するために，5% の False Discovery Rate（FDR）（Benjamini and Hockberg, 1995）を適用し有意となった関係を図 3 中央に示す．図 3 左に描かれた正解のダイアグラムと比較すると，3 箇所余計な関係が推定されていることがわかる．さて，このデータに対して ΔLL を算出してみると，図 3 右のような関係が描ける．$5 \to 4 \to 1$ を $5 \to 1$ と両方推定してしまう．このような indirect/direct な関係を最終的にどちらが正しいか見極めるには，それらの関係を含んだモデルを再度適用し，AIC で比較することによりふさわしい関係を判定することが可能である．例えばこの例の場合，$5 \to 4 \to 1$ を $5 \to 1$ を含んだモデルの AIC は 579.0（対数尤度 -264.5）$5 \to 1$ を外すと AIC は 528.8（対数尤度 -265.4）となるため，$5 \to 1$ は冗長であると判断される．3.1，3.2 の結果から，ペアワイズに基づく Granger と Geweke の方法よりも，全変量の情報に基づいた ΔLL の方が，シミュレーションデータに仮定した関係をより正しく推測していた．

3.3 サンプル数の違いによる ΔLL の Sensitivity

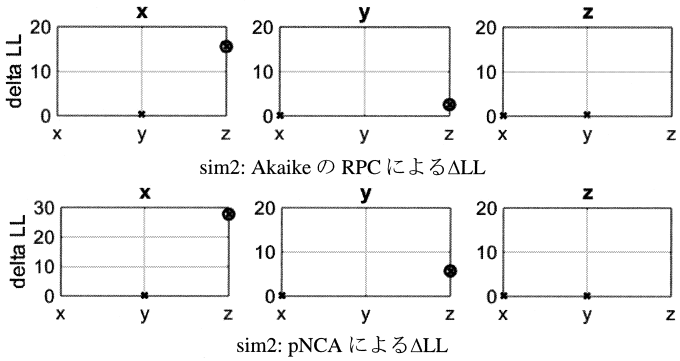
ところで，これまでのシミュレーションデータは MAR モデルを適用するには十分なサンプ



sim1: Akaike の RPC によるΔLL

sim1: pNCA によるΔLL

ΔLL 値から推察される sim1 の変量間の関係



sim2: Akaike の RPC によるΔLL

sim2: pNCA によるΔLL

ΔLL 値から推察される sim2 の変量間の関係

図 2. sim1，sim2 データにおける，Akaike の RPC と Bosch-Bayard の pNCA により算出された ΔLL 値と，それらの結果を元に作成した変数間のダイアグラム．矢印のわきには RPC と pNCA から得られた ΔLL 値を記してある．

ル数 200 点を有していた．そこで，サンプル数の変化に伴う ΔLL の Sensitivity を調べることにする．具体的には，3.2 節で生成した 5 変量データで，サンプル数を 50，100，150，200 点変え，10000 回生成することにより，ΔLL がどの程度正解を得ることができるか調べてみた．変量間の関係の正解数と誤正解数の平均 $\pm$ 標準偏差を表 3 に示す．正解数に関しては，サンプル数が 50 点の段階から正解数に近く 150 点以上ではほぼ正解数を示しているが，50 点では誤正解である関係も推定しやすくなる．150 点以上でとった誤正解として推定した関係は，3.2 節にもみられた indirect/direct の関係であった．実データの観点から考えると，例えば地球温暖化に関連する海洋調査データでは，その殆どが，温暖化を意識され始めた年前後含めて 30 年から多くて 50 年分程度の時系列データしか観測されていない場合が多い．また，生物種間によって

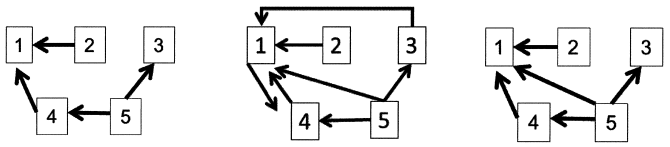


図 3. 5 変量シミュレーションデータに仮定した関係(左)，Granger 因果性 F 検定より推定された関係(中)，ΔLL により推定された関係(右)．

表 3. 5 変量データ(正解数は 4)に関して，サンプル数を変化させた場合における，関係の正解数と誤正解数．

サンプル数	正解数（平均 ± 標準偏差）	誤正解数（平均 ± 標準偏差）
50	3.9±0.26	4.5±2.35
100	4.0±0.09	1.9±1.08
150	4.0±0.00	1.2±0.56
200	4.0±0.00	1.1±0.45

は，indirect/direct の関係のどちらかを選ぶ必要がないものもある．従って限られたデータ数の関係を推定するには，ΔLL によりシステマティックに推測された変量間の関係に加え，さらに生物学的知見とその生態系の専門知識，また関連した先行研究による裏づけが必要となる．

3.4 ΔLL による因果性の強度

先に述べたように，多次元時系列データ間に内在する因果性を分析する手法が取り上げる問題として，因果関係の有無，因果関係の強度，因果関係の可視化等が考えられる．ここでは，ΔLL と変量間の寄与の強度との関係について検討する．以下の数式を用いてデータを作成し，

$$x_t = 0.3x_{t-1} - 0.15z_{t-1} + \varepsilon_{x,t}$$
$$y_t = 0.3y_{t-1} + a_{yz}z_{t-1} + \varepsilon_{y,t}$$
$$z_t = 0.3z_{t-1} + \varepsilon_{z,t}$$

a_{yz}を 0.1 から 0.9 に変化させた場合のΔLL 値を計算し，そのプロットを図 4 にまとめる．9 つ

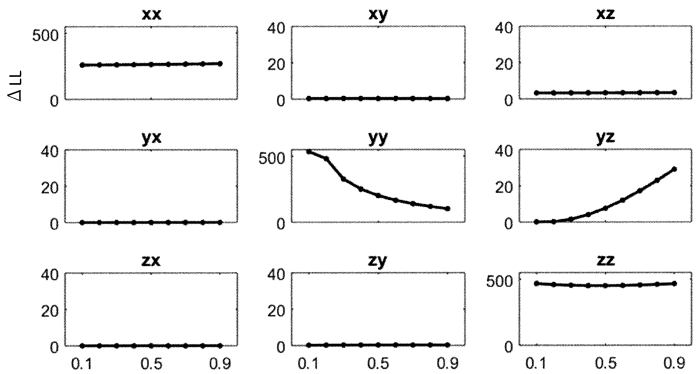


図 4. z から y への MAR モデルの係数を 0.1 から 0.9 まで変化させた場合の，ΔLL 値の変化．x 軸は変化させた係数の値，y 軸は ΔLL 値．

の図のタイトルがどの変量からどの変量への寄与かを示す．例えば xx，yy，zz は変量自身からの寄与，xy は y から x への寄与を示す．z から x への影響が xz の図に示されており，zz と比べると，xx は自身からの影響が低めに出ている．yz の図に注目すると，z から y への影響は係数の値が大きくなると ΔLL 値も大きくなるのがわかる．それに伴い，y 自身は ΔLL 値が下がってきて，y 以外の変量からの係わり合いが含まれていく変化がみられる．ΔLL は対数スケールなので，yy と yz を線形に足し合わせて一定の値にはならない．他の変量の係数については，MAR で全極モデルとして係数を推定するために，若干の変動はみられるが，y 軸のスケールを合わせると，それらの変動は yz の変化よりも微量であることがわかる．これらの結果から，多次元データが，線形で定常であることが仮定でき MAR モデルを適用した場合，ΔLL は変量からの寄与の強さを示すことができると示唆される．

4. 実データ分析

バーレンツ海における生態系では，シシャモの資源量の維持が他の生物種が関連する food web（食物連鎖）において重要な役割を示している．シシャモは 3 から 4 年程度の寿命の遠洋魚で，バーレンツ海における北東北極タラとニシンの稚魚の餌食として知られている（Gjøsæter et al., 2009; Hallfredsson and Pedersen, 2009）．シシャモの消費に伴う資源量の衰退はバーレンツ海の food web のダイナミクスに大きく影響を与えると考えられている．例えば，Hjermann et al.（2004）はバーレンツ海の food web を，シシャモを中心に栄養段階の高低レベル双方に連鎖しているという仮説を示した（図 5）．

本稿では，バーレンツ海におけるシシャモ，タラ，ニシン，オキアミ間の関係に対し，フィードバックシステムを仮定し，ΔLL によりその因果関係を分析する．時系列データは Working Group on the Integrated Assessments of the Barensts Sea（WGIBAR）（ICES, 2016）で取り上げている，1972 年から 2014 年まで収録された，シシャモ（Capelin または Cap）1 から 4 年齢の年間バイオマス，タラ（Cod）とニシン（Herring）の全年間バイオマス，オキアミ（Krill）はバーレンツ海のバイオマス密度を用いる（図 6）．モデルを適用する際，データは標準化されシシャモの各年齢におけるデータとタラ，オキアミ，ニシンを合わせて 4 次元の時系列データとして取り扱うことにする．

まず MAR モデルを適用する前に，これらのデータを標準化し相互相関を調べる．図 7 にシシャモ 1 年齢と他の 3 種類の生物種との場合の相互相関のプロットを示す．シシャモの変動に対する他の変量との相互相関は，ラグが低いところにピークがみられるが，タラやオキアミに関する相互相関ではラグの高いところにもピークがみられる．本稿の図には示していないが，他の年齢においても似たような傾向がみられた．図 6 にみられるように，タラやニシン，オキアミはシシャモの変動と比べると若干非定常性や非線形性がみられるため，長いラグに対して

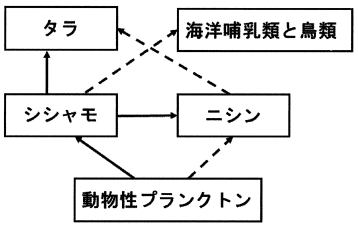


図 5. シシャモを中心としたバーレンツ海の食物連鎖に対する仮説（Hjermann et al.（2004）から図を再描画した）．

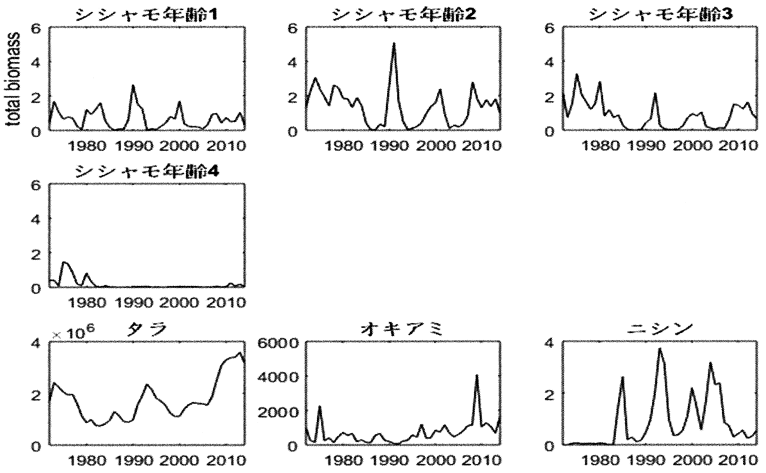


図 6. シシャモ 1–4 年齢のバイオマス(y 軸 $\times 10^6$ kg)，タラ，オキアミのバイオマス(y 軸 kg)，ニシンのバイオマス密度(y 軸 $\times 10^3$ g/m^2)，x 軸は年を示す．

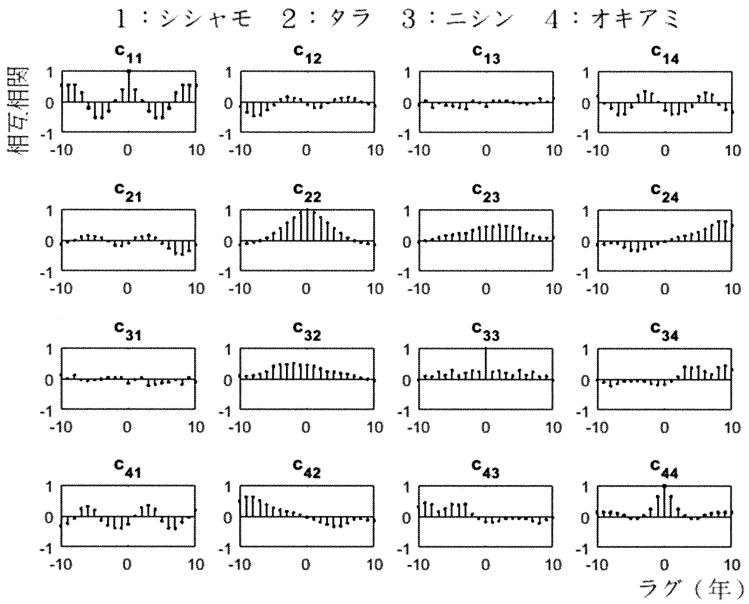


図 7. シシャモ年齢 1，タラ，ニシン，オキアミ 4 系列に関する相互相関(同じ種同士は自己相関)関数のプロット．各プロットのタイトルはどの変量間の相互相関かを示すもので，例えば c_{12} はシシャモとタラに関する相互相関を示す．下付の数字は図全体のタイトルに示した通り．x 軸に -10 から 10 までのラグを，y 軸に相互相関関数値を示す．

相互関係のピークが現れているかもしれない．

MAR モデルを適用した結果，AIC により，シシャモ 1 と 4 年齢の場合は MAR(1)モデル，シシャモ 2 と 3 年齢の場合は MAR(2)モデルが同定された（Solvang et al., 2017 の Table2 を

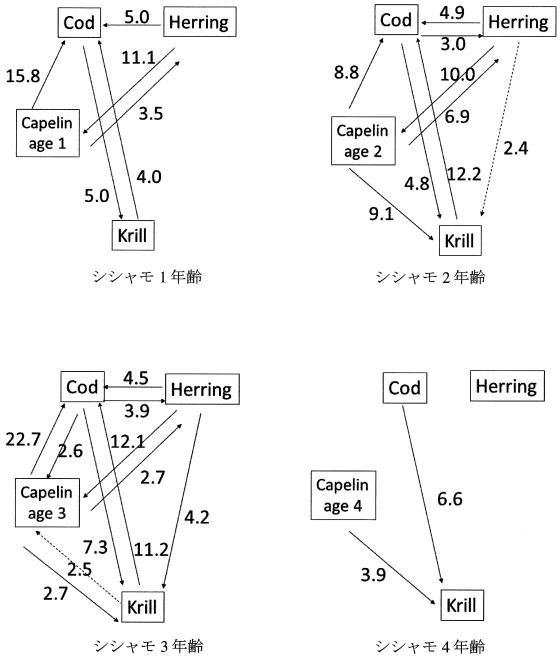


図 8. ΔLL から推察される 4 生物種(シシャモ Capelin，タラ Cod，オキアミ Krill，ニシン Herring を表す)間の関係について(右ダイアグラム，矢印横数字は ΔLL 値を示す)．

参照)．推定値より計算した ΔLL を用い，$\Delta LL > 2.5$ となる関係を実線で，ΔLL が 2.4 から 2.5 となる関係を点線で示したダイアグラムを図 8 にまとめる．シシャモ 1 年齢よりも 2，3 年齢の方が，4 生物種間の関係が多様に現れ，シシャモが 4 年齢になるとそれらの相互関係が希薄になる．これは，4 年齢のシシャモが，実データに見られるように，バーレンツ海においてドミナントではないことによると考えられる．また，シシャモ 1 から 3 年齢の変動はタラの変動に高く影響を及ぼすことを示していて，これは図 5 の仮説で示されているシシャモがタラの餌食となることに対応すると考えられる．シシャモ 2 年齢は最も高くオキアミで代表される動物性プランクトンの変動に影響を与えることを示している．これは，シシャモとオキアミが強い top-down の関係をもつこと(例えば Baum and Worm, 2009)を支持している．シシャモ 3 年齢の場合には，ニシンの変動は動物性プランクトンの変動への影響がみられるが，2 年齢の場合は小さい．ニシンの変動はタラへ影響を及ぼす方がその逆よりも若干高いが数値的にはあまり大きな差はない．図 5 に示された food web の仮説においても，動物性プランクトンとニシン，ニシンとタラは点線なので，実線で示された関係より弱い関係を示している．シシャモとニシンの関係については，先行研究で，バーレンツ海における成魚でないニシンは，シシャモの卵を餌食するので，ニシンの幼魚の資源量が 1 年齢未満のシシャモの生存を左右することが報告されている (Harme, 2000)．シシャモ 2 年齢と 3 年齢の変動がニシンの変動に関係している点については，その年齢のシシャモが動物性プランクトンやオキアミをよく餌食するので，ニシンと食糧の上で競合する可能性が影響していると考えられる．そしてニシンはシシャモ 3 年齢を餌食することはないので，さらにオキアミに関して競合関係になると考えられる (Dalpadado and Skjoldal, 1996)．

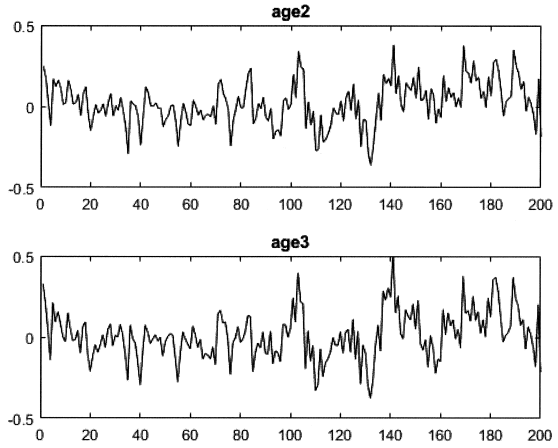


図 9. シシャモとオキアミ，ニシンとオキアミのフィードバック経路を切断した場合のオキアミのシミュレーションデータ．上図がシシャモ 2 年齢，下図がシシャモ 3 年齢の場合の推定された MAR の係数を用いてデータを生成した．

この点について，加藤・石黒 (1997) におけるシステム解析で，フィードバック経路を切断 (マスキング) した場合の各変量のシミュレーション値を比較する方法で検証してみる．

シシャモ 2 年齢と 3 年齢の場合で推定された MAR モデルの係数について，シシャモとオキアミのフィードバック経路 $(a_{13}(m), a_{31}(m), m = 1, 2)$，ニシンとオキアミのフィードバック経路 $(a_{34}(m), a_{43}(m), m = 1, 2)$ を 0 に置き換え，シミュレーションデータを生成し，図 9 にプロットした．上図はシシャモ 2 年齢，下図は 3 年齢の場合のオキアミのシミュレーション値である．若干 3 年齢の値が高く出る傾向がみられる．すなわち，シシャモ 3 年齢の場合，シシャモとニシンのフィードバック関係がないと，オキアミの資源量はシシャモ 2 年齢のときよりも上がる傾向があるということである．それは，オキアミに対する競合関係がシシャモ 3 年齢のほうがより強いという傾向を示唆している．

近年，バーレンツ海におけるシシャモの回遊や資源量が変化し，シシャモを餌食とする生物種，例えば鯨類等がオキアミを主に餌食し始めているという報告がある (Haug et al., 2002; Solvang et al., 2017)．そこで，仮にシシャモがバーレンツ海の生態システムから存在しなくなった場合を想定した分析として，シシャモを含まない他 3 生物種で構成されるフィードバックシステムを仮定し，同様の分析をおこないその結果を図 9 にまとめた．タラとオキアミのフィードバック関係が高く，タラとニシン，オキアミとニシンとの直接的な関係はそう高い値を示さなかった．Bogstad et al. (2015) によると，1–2 年齢のタラが 3–6 年齢のタラよりもより多くオキアミを食し，3–6 年齢のタラは，オキアミよりより多くのシシャモを食することを述べていた．また，北大西洋タラとニシンの稚魚がシシャモを好んで餌食する (Gjøsæter et al., 2009; Hallfredsson and Pedersen, 2009) と報告されているが，シシャモがいなくなった場合，タラやニシンの餌食としてオキアミは重要であることも指摘されている (Gjøsæter et al., 2009)．図 10 での分析では，データとしてニシンの稚魚を扱っていないが，成魚のニシンとまったく無関係という結果ではなく，また，タラとオキアミはこれらの先行研究を裏付けていたといえる．従って，タラやニシンの餌食となるシシャモの変動が，次の餌食と考えられるオキアミの変動に関連すると推察される．

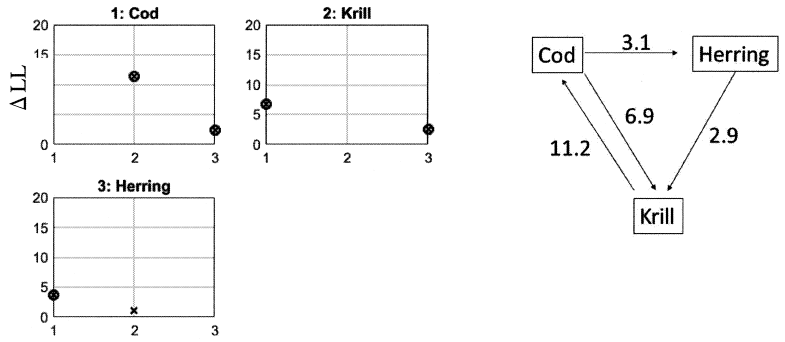


図 10. シシャモを含まない生態システムを仮定した場合における他 3 生物種のフィードバック関係.

5. おわりに

多次元時系列データ間の相互関係を調べる手法として，Ozaki (2012)は，赤池のトータルパワー寄与を基にした，Partial pairwise contribution による因果推測を提案した．我々のシミュレーション実験では，Granger と Geweke の 2 変量間だけのペアワイズ比較よりも，トータルパワーに基づく比較の方がより正確な関係をとらえやすいことを示せた．ΔLL と情報量規準に基づく有意性の規準から，因果性の有無と度合いが判断でき，可視化も可能であることが示せた．MAR の全極を対象とするモデルのため，直接や非直接的な関係や，特にサンプル数が少ない場合，本来考慮しない関係をとる場合があることも示した．サンプル数が 100 点以上の場合，得られた関係をマスキングしたモデルか，もしくはその関係に注目し係数を数値的最適化で求めたモデル間の AIC を比較することで適した関係を示すモデルを再考でき，因果推測が可能となるだろう．実データ解析では，限られたサンプル数であったが，バーレンツ海の生態系における food web で重要な 4 生物種，43 年間のバイオマス時系列データを用い，food web を駆動するシシャモの資源の重要性を裏付けた．提案手法は，複雑なフィードバックシステムの関係を予備的に予測することが可能で，あらかじめ検討をつけた関係について，さらに詳しく関係を踏まえたモデルを再考し，正確な関係をとらえるのに有効であると考えられる．また，赤池の相対パワー寄与率は，本稿で扱ったようなあまり長くない時系列データの場合，長中短期周波数領域における結果の解釈が困難な場合がある．提案手法は周波数領域に関する総和をとるため，そのような細かい解釈は必要なくなる．これまで ICES が海洋学関連の時系列データ分析に主成分分析をスタンダードとしていたが，本稿でとりあげた手法がより適切であり，より具体的に変量間の関係をとらえることができるので，海洋生態系の変化に伴う様々な要因を探ることに貢献できると考えられる．

謝　辞

本稿執筆に際し，突発的な質問にも拒まれず，常に丁寧に議論にお付き合いいただき，多くの助言と激励を頂いた尾崎統博士に，また多くの有益なコメントを頂いた査読者の方々に感謝の意を表します．

参 考 文 献

Akaike, H. (1968). On the use of a linear model for the identification of feedback systems, *Annals of the Institute of Stattistical Mathematics*, **20**, 425–439.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

Akaike, H. and Kitagawa, G. (eds.) (1994). The Practice of Time Series Analysis, Springer-Verlag, New York.

赤池弘次，中川東一郎 (1972)．『ダイナミックシステムの統計的解析と制御』，サイエンスライブラリ (情報電算機 9)，サイエンス社，東京．

Baccalá, L.A. and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination, *Biological Cybernetics*, **84**, 463–474.

Baum, J.K. and Worm, B. (2009). Cascading top-down effects of changing oceanic predator abundances, *Journal of Animal Ecology*, **78**, 699–714.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological), **57**(1), 289–300.

Bogstad, B., Gjøsæter Haug, H., T. and Lindstrøm, U. (2015). A review of the battle for food in the Barents Sea: cod vs. marine mammals, *Frontiers in Ecology and Evolution*, **25**, doi: 10.3389/fevo.2015.00029.

Bosch-Bayard, J., Wong, K.F.K., Okazaki, S., Oshio, R., Galka, A., Ozaki, T. and Sadato, N. (2012). Directed causality for non-stationary time series based on Akaike's noise contribution ratio, *FORMATH*, **11**, 121–131.

Dalpadato, P. and Skjoldal, H.R. (1996). Abundance, maturity and growth of the krill species Thysanoessa inermis and T. longicaudata in the Barents Sea, *Marine Ecology Progress Series*, **144**, 175–183.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series, *Journal of the American Statistical Association*, **77**, 304–313.

Gjøsæter, H., Bogstad, B. and Tjelmeland, S. (2009). Ecosystem effects of the three capelin stock collapses in the Barents Sea, *Marine Biology Research*, **5**, 40–53.

Granger, C.W.J. (1969). Investigating causal relations by economietric models and cross-spectral methods, *Econometrika*, **37**, 424–438.

Hallfredsson, E.H. and Pedersen, T. (2009). Effects of predation from juvenile herring (*clupea harengus*) on mortality rates of capelin (*mallotusvillosus*) larvae, *Canadian journal of Fisheries and Aquatic Sciences*, **66**, 1693–1706.

Harme, J. (2000). Capelin and herring as key species for the yield of north-east Arctic cod. Results from multispecies model runs, *Scientia Marina*, **67**(Suppl 1) 315–323.

Haug, T., Lindstrøm, U. and Nilssen K.T. (2002). Variations in minke whale (*Balaenoptera acutorostrata*) diet and body condition in reponse to ecosystem changes in the Barents Sea, *Sarsia*, **87**, 409–422.

Hjermann, D.ø., Strenseth, N.C. and Ottersen, G. (2004). Indirect climate forcing of the Barents Sea capelin: A cohort effect, *Marine Ecology Progress Series*, **273**, 229–238.

加藤比呂子，石黒真木夫 (1997). 多変量時系列モデルによる経済システムの動的解析，統計数理, **45**(2), 301–318.

川崎能典 (2001). 多変量時系列に対する主成分・因子分析, 統計数理, **49**(1), 109–131.

Kolmogorov, A.N. (1941). Stationary sequences in Hilbert space, Moscow University Mathematics Bulletin, **5**, 3–14.

Ozaki, T. (2012). Time series modeling of neuroscience data, Chapman & Hall/CRC, Boca Raton,

Florida.

Planque, B. and Arneberg, P. (2017). Principal component analyzes for integrated ecosystem assessments may primarily reflect methodological artefact, *ICES Journal of Marine Science*, doi:10.1093/icesjms/fsx223.

Solvang, H., Subbey, S. and Frank, A.S.J. (2017). Causal drivers of Barents Sea capelin (*Mallotus villosus*) population dynamics on different time scales, *ICES Journal of Marine Science*, doi:10.1093/icesjms/fsx179.

Tanokura, Y. and Kitagawa, G. (2004). Power contribution analysis for multivariate time series with correlated noise sources, *Advances and Applications in Statistics*, **4**, 65–95.

WGIBAR. (2016). Final report of the Working Group on the Integrated Assessments of the Barents Sea.

Wiener, N. (1956). The theory of prediction, *Modern Mathematics for Engineers* (ed., E.F. Beckenback), Chapter2, 165–190, McGraw-Hill, New York.

Causal Inference for Marine Ecosystems Based on Total Power Contribution

Hiroko Kato Solvang[1] and Subbey Sam[2,3]

[1]Marine Mammals Research Group, Institute of Marine Research
[2]Research Group on Fisheries Dynamics, Institute of Marine Research
[3]Department of Natural Resources, Cornell University

We introduce a statistical methodology that integrates Granger's pair-wise causal analysis and its expansion to causality basedon the log-likelihood (Partial pairwise causality), and Akaike's power contribution approach whole frequency domain (Total causality). Although the initial idea was proposed by Ozaki (2012), it has hitherto not been applied to complex marine ecosystem dynamics. In this article, we implement the approach by adding a criterion to assess significance to detect causal relationship. We perform a simulation study to verify the efficacy and sensitivity of the method, using data generated by three autoregressive models with three and five dimensions. We also applied the method to real observations to investigate causal drivers of Barents Sea capelin population dynamics. The goal of this analysis was to explore inter-species relationships, which are important food web drivers in the Barents Sea ecosystem. We present results demonstrating that the proposed methodology is a useful tool in early-stage causal analysis of complex feedback systems.

Key words: Multivariate auto-regressive model, multivariate time series data, feedback system, Granger's causality, marine ecosystem, Barents Sea.

統計数理(2018)
第 66 巻 第 2 号 339–351

［統計ソフトウェア］

P³：Pythonによる並列計算機用粒子フィルタライブラリ

中野 慎也[1,2]・有吉 雄哉[1,3]・樋口 知之[1,2]

(受付 2018 年 2 月 20 日；改訂 8 月 13 日；採択 10 月 2 日)

要　　旨

粒子フィルタ(PF)は，多数の粒子を用いたモンテカルロ計算に基づく状態推定手法であり，非線型，非ガウスの問題に適用できることから広く様々な目的で用いられるようになってきている．一方で，PF には，推定に必要となる粒子の数が状態変数の自由度に対して指数関数的に増大するため，計算量も指数関数的に増大してしまうという欠点がある．並列計算機の利用は，PF の計算量に対処する手段の一つとして有効であると考えられる．しかし，並列計算機を使うには並列プログラミングの知識が必要であり，また，PF には並列化の困難な処理が含まれているため，並列プログラミングの知識があるユーザにとっても，PF で高い並列化効率を実現するのは容易ではない．そこで，並列化効率の高い PF アルゴリズムを手軽に利用できるようにするために P³(Python Parallelized Particle Filter Library)という Python ライブラリを開発した．本稿では，P³ で利用できる PF の並列アルゴリズムについて述べ，構成の概要や特徴を紹介する．

キーワード：粒子フィルタ，並列計算，Python.

1. はじめに

粒子フィルタ(particle filter; 以下 PF)(Gordon et al., 1993; Kitagawa, 1993, 1996; Doucet et al., 2001)は，非線型・非ガウスの状態空間モデルに適用可能な状態推定手法で，非線型時系列解析や動画上のターゲット追跡，さらにはデータ同化など，様々な目的で用いられる．PF では，状態変数の確率分布を多数の粒子で表現し，モンテカルロ法の考え方に基づいて逐次ベイズ推定の計算を行う．しかし，PF では，状態変数の自由度(独立と見なせる状態変数の数)が大きくなると，いわゆる次元の呪いの問題が顕在化するという問題がある(Daum and Huang, 2003; Bengtsson et al., 2008; Snyder et al., 2008)．また，詳細な事後分布の情報を得たい場合には，さらに多数の粒子を用いる必要があり，12 変数の推定に 10^8 個の粒子を使用した事例もある(Nakamura et al., 2009)．PF の計算量は，少なくとも粒子数 N のオーダーになるため，PF では計算時間が大きな問題となってくる．

多数の粒子を用いて高速に推定を行う手段としては，並列計算機の利用が考えられる．しかし，マルチコア，マルチノードの並列計算機が安価に購入できる状況にある一方，それを使い

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 総合研究大学院大学 複合科学研究科：〒 240–0193 神奈川県三浦郡葉山町湘南国際村
[3] 現 日本文理大学 工学部：〒 870–0397 大分県大分市一木 1727

こなすには並列プログラミングの知識が必須であり，誰もが直ちに並列計算機上で PF を実装できるものではない．また，PF で通常用いられるフィルタリング手続きには，並列化の困難な処理が含まれているため，並列プログラミングの知識があるユーザにとっても，高い並列化効率を実現するのは容易ではない．フィルタリング手続きを改良し，高い並列化効率を実現する方法はいくつか提案されている（Bolić et al., 2005; Nakano, 2010; Nakano and Higuchi, 2010）ものの，その手続きは煩雑であり，プログラムの実装に掛かる手間の問題は依然として残る．

P^3（Python Parallelized Particle Filter Library）は，並列化効率の高い PF 手法を利用しやすくするために，Python ライブラリとして整備したものである．Python は，インタプリタ型のプログラミング言語ではあるが，numpy, scipy をはじめとする高速な数値計算用ライブラリや，統計計算，機械学習のライブラリが豊富に用意されている他，mpi4py のような並列計算の手段も用意されており，科学技術計算の分野にもかなり普及した感がある．一般的な状態空間モデルに適用可能な逐次ベイズ推定手法を実装した Python ライブラリとしては，すでに `FilterPy`（Labbe, 2015）などがあるが，P^3 では並列計算機を活用して比較的大きな規模の非線型・非ガウス状態空間モデルを扱うことを想定している．以下では，まず PF の基本的な形のアルゴリズムについて述べ，P^3 で実装されている並列計算用の PF アルゴリズムについて説明した後，P^3 の構成や特徴を述べる．

2. 粒子フィルタの基本的なアルゴリズム

2.1 非線型状態空間モデル

時刻 t_k の状態を $\boldsymbol{x}_k$，時刻 t_k に得られる観測を $\boldsymbol{y}_k$ と表す．時刻 t_{k-1} から時刻 t_k の間の状態遷移は，関数 $\boldsymbol{f}_k$ を用いて

$$\boldsymbol{x}_k = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}) + \boldsymbol{v}_k \tag{2.1}$$

という形のシステムモデルで記述されるものとする．但し，確率変数 $\boldsymbol{v}_k$ は確率的な変動を表し，システムノイズと呼ばれる．一方，観測 $\boldsymbol{y}_k$ は，状態 $\boldsymbol{x}_k$ との間に以下の関係があるものとする:

$$\boldsymbol{y}_k = \boldsymbol{h}_k(\boldsymbol{x}_k) + \boldsymbol{w}_k. \tag{2.2}$$

関数 $\boldsymbol{h}$ は $\boldsymbol{x}_k$ を観測に対応づける関数，$\boldsymbol{w}_k$ は観測ノイズである．式(2.1), (2.2)をまとめて，非線型状態空間モデルと呼ぶ．

この非線型状態空間モデルは，以下に示す確率分布の形に書き直すこともできる:

$$\boldsymbol{x}_k \sim p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}), \tag{2.3a}$$

$$\boldsymbol{y}_k \sim p(\boldsymbol{y}_k|\boldsymbol{x}_k). \tag{2.3b}$$

例えば，式(2.1)は，$\boldsymbol{v}_k$，$\boldsymbol{w}_k$ がガウス分布

$$\boldsymbol{v}_k \sim \mathcal{N}(\boldsymbol{v}_k; \boldsymbol{0}, \mathsf{Q}), \tag{2.4a}$$

$$\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{w}_k; \boldsymbol{0}, \mathsf{R}) \tag{2.4b}$$

に従うとき，

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{f}_k(\boldsymbol{x}_{k-1}), \mathsf{Q}), \tag{2.5a}$$

$$p(\boldsymbol{y}_k|\boldsymbol{x}_k) = \mathcal{N}(\boldsymbol{y}_k; \boldsymbol{h}_k(\boldsymbol{x}_k), \mathsf{R}) \tag{2.5b}$$

と書くことができる．

式(2.3)の形を導入すると，時刻 t_k までの観測データ $\boldsymbol{y}_{1:k} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ が与えられた時の $\boldsymbol{x}_k$ は，$t = 0$ における状態 $\boldsymbol{x}_0$ の分布 $p(\boldsymbol{x}_0)$ を与えておけば，次の 2 つの式を用いて推定することができる:

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}) = \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:k-1})\, d\boldsymbol{x}_{k-1}, \tag{2.6a}$$

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}) = \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k) p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})}{\int p(\boldsymbol{y}_k|\boldsymbol{x}_k) p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})\, d\boldsymbol{x}_k}. \tag{2.6b}$$

ここで，$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})$ は予測分布，$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ はフィルタ分布と呼ばれる．式(2.6a)では予測分布を得るのにフィルタ分布が用いられ，式(2.6b)では逆にフィルタ分布を得るのに予測分布が用いられていることに注意すると，式(2.6a), (2.6b)を時刻 t_0 から交互に適用することで，各時刻のフィルタ分布 $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ が得られる．このように式(2.6a), (2.6b)を逐次的に適用して，$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ を推定する方法を逐次ベイズ推定と呼ぶ．PF は逐次ベイズ推定の枠組みに従って $\boldsymbol{x}_k$ を推定するアルゴリズムの一つである．

2.2 粒子フィルタ(PF)の概要

PF では，確率分布 $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})$ や $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ を N 個の粒子で表し，式(2.6)を近似計算する．PF には様々な変形版が存在する（e.g., Doucet et al., 2001; van Leeuwen, 2009)が，最も基本的なアルゴリズムは次に述べるとおりである．

まず，式(2.6a)の計算をモンテカルロ法で行う．時刻 t_{k-1} において，$p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:k-1})$ に従う N 個のサンプル(粒子)$\{\boldsymbol{x}^{(i)}_{k-1|k-1}\}_{i=1}^N$ が得られていたとする．$p(\boldsymbol{x}_k|\boldsymbol{x}^{(i)}_{k-1|k-1})$ に従う乱数は，$p(\boldsymbol{v}_k)$ から生成した乱数 $\boldsymbol{v}^{(i)}_k$ を用いて

$$\boldsymbol{x}^{(i)}_{k|k-1} = \boldsymbol{f}_k(\boldsymbol{x}^{(i)}_{k-1|k-1}) + \boldsymbol{v}^{(i)}_k \tag{2.7}$$

から得られる．式(2.7)にしたがい，各 i に対して $\boldsymbol{x}^{(i)}_{k|k-1}$ を生成すれば，予測分布 $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})$ に従う N 個の粒子 $\{\boldsymbol{x}^{(i)}_{k|k-1}\}_{i=1}^N$ が得られる．

次に式(2.6b)にしたがって，フィルタ分布を求める．式(2.6b)は，$\{\boldsymbol{x}^{(i)}_{k|k-1}\}_{i=1}^N$ の各粒子に尤度 $p(\boldsymbol{y}_k|\boldsymbol{x}^{(i)}_{k|k-1})$ で重みづけすれば，フィルタ分布 $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ の形状を表現できることを示している．ここで，各粒子 $\boldsymbol{x}^{(i)}_{k|k-1}$ が $p(\boldsymbol{y}_k|\boldsymbol{x}^{(i)}_{k|k-1})$ に比例する確率

$$\beta^{(i)}_k = \frac{p(\boldsymbol{y}_k|\boldsymbol{x}^{(i)}_{k|k-1})}{\sum_{i=1}^N p(\boldsymbol{y}_k|\boldsymbol{x}^{(i)}_{k|k-1})} \tag{2.8}$$

で抽出されるように N 回復元抽出を行い，新たに N 個の粒子の集合 $\{\boldsymbol{x}^{(i)}_{k|k}\}_{i=1}^N$ を作る．そうすると，$\{\boldsymbol{x}^{(i)}_{k|k}\}_{i=1}^N$ は，元の粒子 $\boldsymbol{x}^{(i)}_{k|k-1}$ の複製を $\beta^{(i)}_k$ にほぼ比例する個数だけ含んでおり，重みづけをしなくてもフィルタ分布 $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ を表現していることになる．このように N 回復元抽出によって新たに N 個の粒子の集合を得る操作をリサンプリングと呼ぶ．

アルゴリズムをまとめると，以下のようになる：

(1) $t = 0$ における状態 $\boldsymbol{x}_0$ の分布 $p(\boldsymbol{x}_0)$ にしたがう乱数 $\boldsymbol{x}^{(i)}_{0|0} \sim p(\boldsymbol{x}_0)$ を N 個生成する．
(2) 毎時間ステップ k $(k = 1, \ldots, K)$ において以下を実行する．
  (a) 予測
    • 各 i $(i = 1, \ldots, N)$ について $p(\boldsymbol{v}_k)$ にしたがう乱数 $\boldsymbol{v}^{(i)}_k \sim p(\boldsymbol{v}_k)$ を N 個生成する．
    • 各粒子 i $(i = 1, \ldots, N)$ について

$$\boldsymbol{x}_{k|k-1}^{(i)} = \boldsymbol{f}_k(\boldsymbol{x}_{k-1|k-1}^{(i)}) + \boldsymbol{v}_k^{(i)}$$

により，$\boldsymbol{x}_{k-1|k-1}^{(i)}$ から $\boldsymbol{x}_{k|k-1}^{(i)}$ を生成する．

(b) フィルタリング

〈尤度・重みの計算〉 各粒子 i について，尤度 $p(\boldsymbol{y}_k|\boldsymbol{x}_{k|k-1}^{(i)})$ を計算し，重み

$$\beta_k^{(i)} = p(\boldsymbol{y}_k|\boldsymbol{x}_{k|k-1}^{(i)}) \Big/ \sum_{i=1}^{N} p(\boldsymbol{y}_k|\boldsymbol{x}_{k|k-1}^{(i)})$$

を求める．

〈リサンプリング〉 粒子の集合 $\{\boldsymbol{x}_{k|k-1}^{(1)}, \ldots, \boldsymbol{x}_{k|k-1}^{(N)}\}$ から，各粒子 $\boldsymbol{x}_{k|k-1}^{(i)}$ が $\beta_k^{(i)}$ の確率で抽出されるように復元抽出を N 回繰り返す「リサンプリング」を行い，$\{\boldsymbol{x}_{k|k}^{(1)}, \ldots, \boldsymbol{x}_{k|k}^{(N)}\}$ を生成する．

このように，尤度に比例する重みでリサンプリングを行い，フィルタ分布を計算するのが PF の基本的な形式（Gordon et al., 1993; Kitagawa, 1996）である．このような形式のアルゴリズムを特にブートストラップフィルタ，あるいはモンテカルロフィルタと呼ぶ場合もある．

3. 並列アルゴリズム

PF では，必要な粒子数 N が $\boldsymbol{x}_k$ の次元に対して指数関数的に増大する．そこで，多数の粒子を高速で処理する手段として，並列計算機の使用が考えられる．粒子フィルタのアルゴリズムにおいて，(a) の予測の手続きは，各粒子の処理が独立しているため，容易に並列化できる．しかし，(b) のフィルタリングを行う際に必要となるリサンプリングの手続きが，並列化の際に問題となる．

図 1 は，通常の粒子フィルタのリサンプリングを並列計算機上で行った場合の概念図である．ここでは，互いにメモリを共有しない複数のプロセスによって並列に処理が行われるマルチプロセス型の並列計算を考える．リサンプリングの手続きでは，重み $\beta_k^{(i)}$ の小さい（観測と合わない）粒子は破棄され，重み $\beta_k^{(i)}$ に応じて粒子の複製が生成されるため，個々のプロセスが保持する粒子数にばらつきが生じることになる．次の予測の手続きに進んだときに高い並列化効率を得るには，粒子数の増えたプロセスから粒子数の減ったプロセスに粒子を移動し，各プロセスの保持する粒子の数を均等に再配分する操作が必要になる．粒子数を決める重み $\beta_k^{(i)}$ の値にはランダム性があることから，各プロセスの粒子の増減数はランダムに決まる．したがって，粒子の再配分に伴うプロセス間通信には規則性がなく，並列処理が難しい．

そこで，高い並列化効率を実現する方法として，粒子をグループに分割してリサンプリングを行い，グループごとに異なる重みを持たせる方法が提案されている（Bolić et al., 2005）．単にグループに分けただけでは，各グループに割り当てられた個数の粒子のみで推定を行うのと同等になってしまうため，全体で高い精度の推定結果を得るには，グループ間の情報交換が必要になる．情報交換は，グループ間で粒子を部分的に交換する方法（Balasingam et al., 2011; Bai et al., 2016）や，グループ間の相互作用を考える方法（Hlinka et al., 2013; Savic et al., 2014）なども研究が進められているが，P^3 では，並列化効率を重視し，次に述べる 2 種類の手法を実装している．

3.1 階層的リサンプリング

P^3 で用意しているリサンプリング法の 1 つ目は，階層的リサンプリング（Nakano, 2010）である．この方法は，図 2 に示すように，まず各プロセス内でリサンプリングを行い，次に各プ

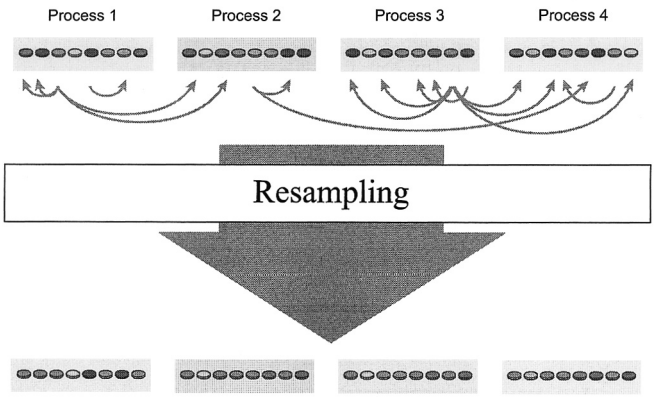


図 1. 通常のリサンプリング手法を並列計算機上で行った場合.

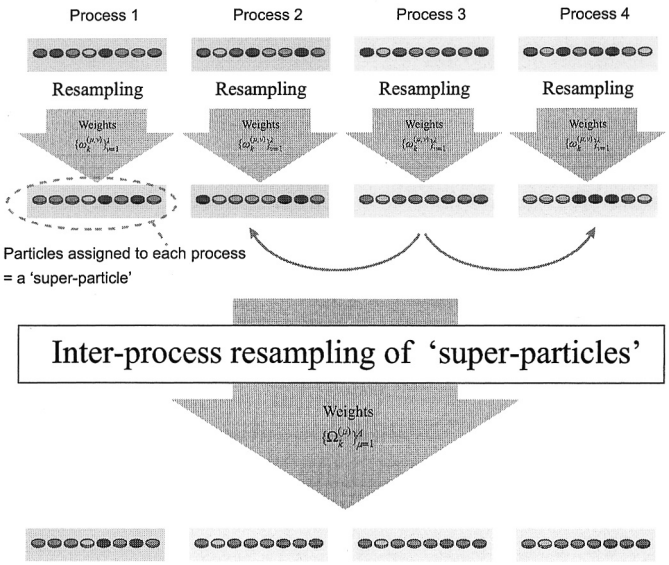


図 2. 階層的リサンプリングの概念図（Nakano, 2010）.

ロセスに属する粒子の集合を超粒子と見なして，超粒子をリサンプリングするというものである.

階層的リサンプリングを行うには，まず各プロセス μ に割り当てられた粒子の集合 G_μ の重みは

$$\Omega_k^{(\mu)} = \sum_{\nu=1}^{\lambda} \beta_k^{([\mu-1]\lambda+\nu)} \tag{3.1}$$

となる．但し，λ は各プロセスに属する粒子の数で，ここでは全プロセスに均等に λ 個の粒子が割り当てられているものとする．したがって，プロセス数を Λ として，

$$\lambda = N/\Lambda \tag{3.2}$$

である．各粒子の重み $\beta_k^{([\mu-1]\lambda+\nu)}$ を集合 G_μ の重み $\Omega_k^{(\mu)}$ で割ると，集合 G_μ 内での各粒子の重みは

$$\omega_k^{(\nu)} = \beta_k^{([\mu-1]\lambda+\nu)}/\Omega_k^{(\mu)} \tag{3.3}$$

となる．次に，各集合 G_μ の中で，重み $\{\omega_k^{(\nu)}\}$ を用いて粒子のリサンプリング (ローカル・リサンプリング) を行う．ローカル・リサンプリングの手続きは，各プロセスで閉じており，異なるプロセスのローカル・リサンプリングは並列に実行できる．最後に，各集合 G_μ を超粒子と見なし，重み $\{\Omega_k^{(\mu)}\}$ を用いてリサンプリング(メタ・リサンプリング)する．これにより，観測と合わない粒子しか保持していない集合は破棄され，観測と合う粒子の集合に置き換えられる．P^3 では，ローカル・リサンプリングの手続きを `local_resampling()` という関数で，メタ・リサンプリングの手続きを `meta_resampling()` という関数で提供している．

なお，各 μ に対する $\Omega_k^{(\mu)}$ の値のばらつきが大きくない場合は，メタ・リサンプリングを行わず，$\Omega_k^{(\mu)}$ で重み付けしたままの方がフィルタ分布 $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ をよりよく表現できると考えられる．そこで P^3 では，$\Omega_k^{(\mu)}$ のばらつきを評価するためにエントロピー

$$S_{\Omega,k} = -\sum_{\mu=1}^{\Lambda} \Omega_k^{(\mu)} \log \Omega_k^{(\mu)} \tag{3.4}$$

から

$$\Lambda_{\mathrm{eff},k} = e^{S_{\Omega,k}} \tag{3.5}$$

という量を計算し，$\Lambda_{\mathrm{eff},k}$ がある閾値より小さくなった場合のみに，メタ・リサンプリングの手続きが実行される．

3.2 Alternately lattice-pattern switching(ALPS)法

P^3 で提供するリサンプリング法の 2 つ目は，alternately lattice-pattern switching(ALPS)法 (Nakano and Higuchi, 2010, 2012)である．この方法では，複数のプロセスに割り当てられた粒子の集合をまとめてグループにし，各グループの中でローカル・リサンプリングと同じ操作を適用する．グループ分けを行う際には，まず，$2m \times 2n$ 個のプロセスを図 3 に示すような 2 次元トーラス状のグラフの各ノードに割り当てる．そして，図 3 左のように，ノード(プロセス)

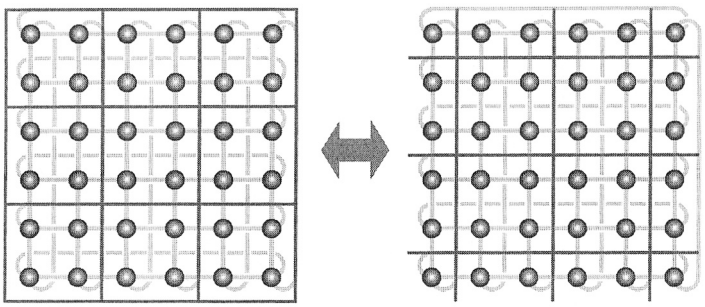


図 3. Alternate lattice-pattern switching 法の概念図 (Nakano and Higuchi, 2010)．

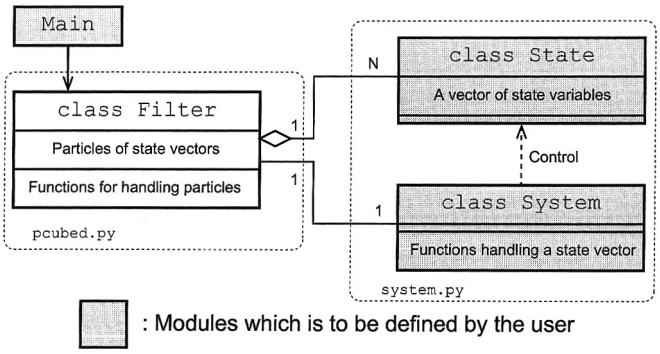


図 4. P³ の構成の概略図.

を 2 × 2 個のグループに分割する．リサンプリングは，各グループで並列に実行される．これだけでは，単なるローカル・リサンプリングと大差ないが，ALPS では，次のステップで，図 3 右に示すようにグループ分けのパターンを変え，その上でグループごとのリサンプリングを行う．ステップごとにグループ分けパターンを変えることで，観測に合う粒子の情報が全プロセスに行き渡るようにし，推定性能を向上させる．

4. P³ の構成

カルマンフィルタを適用できる線型状態空間モデルの場合，状態遷移行列，観測行列などの行列を与えれば記述できる．しかし，PF の扱う一般の非線型状態空間モデルの定義の仕方は，個々の問題によって様々であり，非線型微分方程式が使われる場合もあるし，非ガウスの確率分布が含まれることも考えられる．このような様々な場合に対応できるようにするため，P³ では，既存の Python ライブラリとの互換性には配慮せず，式(2.1), (2.2)の状態空間モデルの内容をユーザが自由に定義できるように設計されている．但し，式(2.1), (2.2)の計算を P³ のモジュールから実行できるようにするため，ユーザは，状態ベクトル，状態空間モデルの定義を `system.py` という名前のファイルに記述しておく必要がある．状態ベクトル，状態空間モデルの定義は，それぞれ `class State`，`class System` という名前で所定の形式に従って記述する．

PF を実行するために必要な関数は，ファイル `pcubed.py` 中で定義されている `class Filter` のメンバ関数の形で用意されている．並列計算やリサンプリングの処理は，`class Filter` の中に記述されており，ユーザは `class State`，`class System` を記述さえすれば，面倒なプログラミング作業を行わずとも `class Filter` 中の関数を使って状態推定の計算が実行できる．図 4 に，P³ の構成を示す．

なお，従来の逐次ベイズ推定のための計算ライブラリでは，状態ベクトルを 1 つの配列にまとめた上で計算を行う実装が多かった．しかし，高次元の状態空間モデルにおいては，状態変数のそれぞれが別々の意味を持つことが少なくない．このような場合，異なる意味を持つ状態変数は，プログラム上でも異なる変数名で扱った方が，可読性が向上すると考えられる．特に，シミュレーションモデルからシステムモデルを構成するデータ同化においては，元となるシミュレーションプログラムの変数名が流用できるため，システムモデルのプログラムを書く際にも都合がよい．そのため P³ では，状態ベクトルを配列ではなく，`class State` というクラス(構造体)の形で保持する設計になっている．

以下では，実際に P³ を使うために必要な情報として，`class State` と `class System` の定義

表 1. class System で定義すべき関数.

関数名	引数	機能
system_config	なし	事前のパラメータ設定，前処理を行う.
init_state	ydata, x	各粒子の初期値を設定する.
step_system_model	x	各粒子を 1 ステップ進めるモデル $\boldsymbol{f}_k$ に相当する.
add_noise	x	各粒子にノイズ $\boldsymbol{v}_k$ を加える.
log_likelihood	x, ydata	各粒子の対数尤度を求める.
obs_predict	x	各粒子の状態変数の値を与えたときの観測の期待値を求める.

の仕方を説明し，class Filter の内容について説明した後，メインプログラムの書き方の概要を説明する.

4.1 class State の定義

P^3 の class Filter を使う際には，予め system.py の中で class State と class System を定義する必要がある．このうち，class State は，式(2.1)に出てくる状態ベクトル $\boldsymbol{x}_k$ の全要素(つまり全状態変数)をまとめたもので，推定するべき状態変数は，すべて class State のメンバ変数として列挙する.

各メンバ変数は，numpy.ndarray とし，0 で初期化するようにする．例えば，$a_{0,k}$, $a_{1,k}$, ..., $a_{4,k}$ と $b_{0,k}$, $b_{1,k}$, $b_{2,k}$ という 8 つの変数をまとめて状態ベクトル $\boldsymbol{x}_k$ として扱いたいときは,

```
class State:
  def __init__(self):
    self.a = np.zeros((5))
    self.b = np.zeros((3))
```

のように定義する.

4.2 class System の定義

class State は，状態ベクトルの定義を与えるだけで，式(2.1)のようなシステムモデルの定義は，class System で行う．class System の中では，pcubed.py から呼ばれる関数を所定の名前で定義する必要がある．表 1 が，class System で定義すべき関数である．以下に具体的な定義の仕方を述べる.

system_config(cls)

この関数は，クラスメソッドとして定義され，事前に実行しておくべき前処理や，パラメータ値の設定をここに記述する．前処理等が必要ない場合も，system_config() をダミー関数にするなど，何らかの形で定義する必要がある.

init_state(self, ydata, x)

各粒子の初期値を設定する．引数 x は class State 型の変数であり，x のメンバ変数に状態変数の初期値が代入されるように関数 init_state() を定義する必要がある．PF において，各粒子の初期値は，初期状態の確率分布 $p(\boldsymbol{x}_0)$ にしたがう乱数で与えるので，基本的に $p(\boldsymbol{x}_0)$ にしたがう乱数が x のメンバ変数に代入されるように定義すればよい.

ydata は System.nobs の長さを持つ numpy.ndarray クラスのオブジェクトで，初期化に用いる観測データをここに与えることができる．これは，観測可能な状態変数をデータから与えるというようなことを想定している．つまり，$p(\boldsymbol{x}_0)$ ではなく，$p(\boldsymbol{x}_0|\boldsymbol{y}_0)$ から各粒子の初期値を

生成できるようになっている．

`step_system_model(self, x)`

1 ステップ分の粒子の時間発展を計算する．データ同化を行う場合，シミュレーションモデルの 1 ステップ分をここで定義することになる．

x は入力と出力を兼ねた class State の変数で，粒子 i の時刻 t_{k-1} における状態変数の値 $\boldsymbol{x}^{(i)}_{k-1|k-1}$ を x として与えると，$\boldsymbol{f}_k(\boldsymbol{x}^{(i)}_{k-1|k-1})$ の値が State クラスのオブジェクトの形で x に代入されるようにする．

`add_noise(self, x)`

粒子にシステムノイズ $\boldsymbol{v}^{(i)}_k$ を付加する．粒子の持つ状態変数の値を State クラスの形で x に入力すると，それにノイズが付加された値が x に代入されるように定義する．add_noise() を step_system() と組み合わせることで

$$\boldsymbol{x}^{(i)}_{k|k-1} = \boldsymbol{f}_k(\boldsymbol{x}^{(i)}_{k-1|k-1}) + \boldsymbol{v}^{(i)}_k$$

の計算が実行できる．

`log_likelihood(self, x, ydata)`

観測データ $\boldsymbol{y}_k$ が与えられた下での粒子 $\boldsymbol{x}^{(i)}_{k|k-1}$ の対数尤度 $p(\boldsymbol{y}_k|\boldsymbol{x}^{(i)}_{k|k-1})$ を計算する．入力として，粒子の状態変数の値を State クラスの形で x に，観測値を長さ System.nobs の numpy.ndarray クラスの形で ydata に与えるようにする．計算された対数尤度は，実数値(スカラー)の返り値として返すようにする．

4.3 class Filter の内容

class Filter は，表 2 に示す関数で構成されており，メインプログラムから，表 2 の関数を呼ぶことで粒子フィルタの計算が実現できる．以下で，各関数の役割を述べる．

`init_ensemble(nptot, yinit)`

N 個の粒子を初期化する．引数 nptot には粒子数 N(整数値)を入力し，yinit には初期化のための観測データを入力する．yinit の型は numpy.ndarray で System.nobs の長さを持つ．具体的な初期化の方法は，system.py に記述する必要がある．

`finalize()`

終了処理をする．class Filter を使い終わった後に呼ぶ．

`step_ensemble()`

各粒子を 1 ステップ進める．これは式(2.7) の $\boldsymbol{f}_k(\boldsymbol{x}^{(i)}_{k-1|k-1})$ の部分の計算に相当する．$\boldsymbol{f}_k$ の

表 2. class Filter で用意されている関数.

関数名	引数	返り値	機能
init_ensemble	iregopt, nptot, yinit	なし	N 個の粒子の初期化.
finalize	なし	なし	終了処理.
step_ensemble	なし	なし	全粒子を 1 ステップ進める.
add_noise_to_ensemble	なし	なし	全粒子にノイズを加える.
ensemble_mean	mrank	class State	全粒子の平均を返す.
dist_reweight	ydata	なし	全粒子を尤度に従って重み付けする.
local_resample	なし	なし	プロセス内でリサンプリング.
meta_resample	なし	なし	粒子の集合のメタ・リサンプリング.
regional_resample	なし	なし	ALPS によるリサンプリング.

定義は system.py に記述する．

add_noise_to_ensemble()

各粒子にシステムノイズを加える．これは式(2.7)で $\boldsymbol{v}_k^{(i)}$ を足す部分に相当する．

ensemble_mean()

粒子が保持する状態 $\boldsymbol{x}_{k|k}^{(i)}$ の平均

$$\bar{\boldsymbol{x}}_{k|k} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_{k|k}^{(i)}$$

を求める．これにより PF による $\boldsymbol{x}_k$ の推定値が得られる．

dist_reweight(ydata)

観測データ ydata を参照して，式(2.8)の重み $\beta_k^{(i)}$ を計算する．ydata は numpy.ndarray で System.nobs の長さを持つ．

local_resample()

dist_reweight で求めた重みにしたがい，各プロセス内で粒子のローカル・リサンプリングを行う．

meta_resample()

各プロセスに割り当てられた粒子の集合をまとめて超粒子と見なし，dist_reweight で求めた重みから計算される超粒子の重みにしたがって，メタ・リサンプリングを行う．但し，この関数では，超粒子の重み $\Omega_k^{(\mu)}$ のばらつきを評価するために，式(3.5)の Λ_{eff} を計算し，$\Lambda_{\text{eff}} \geq 0.5$ の時は何もしない(つまりメタ・リサンプリングを行わない)ようになっている．

regional_resample()

ALPS 法によるリサンプリングを実行する．ALPS のグループ分けパターンは，この関数を呼ぶ度に自動的に切り替えられる．

4.4 プログラムの書き方のまとめ

class Filter を用いて，PF の予測ステップを行うには，メインプログラム中で

```
Filter.step_ensemble()
Filter.add_noise_to_ensemble()
```

のように記述する．フィルタリングについては，階層的リサンプリングを用いる場合，

```
Filter.dist_reweight( ydata )
Filter.local_resample()
Filter.meta_resample()
```

のように記述すればよく，ALPS を用いる場合，

```
Filter.dist_reweight( ydata )
Filter.regional_resample()
```

のように記述すればよい．

5. おわりに

PF は，状態変数が数個程度の小規模な問題であれば容易に実装できるため，非線型・非ガウスの状態空間モデルにおいて広く活用されているが，問題の規模に対して，計算量が指数関

数的に増大するという問題がある．P^3 は，粒子フィルタの適用対象としては，比較的規模の大きい非線型の問題に対しても，並列計算機を利用して状態推定を実現できる手段を提供する．粒子フィルタを並列計算機で実行しようとすると，個々の粒子を並列処理するために並列プログラミングの知識が必要になる上，リサンプリング時に粒子を再配分する処理の実装に労力が必要となるが，P^3 では並列計算やリサンプリングの処理が抽象化されているため，ユーザは煩雑な処理を自分でプログラミングしなくても，並列化効率の高い PF アルゴリズムを利用でき，状態空間モデルの構築に専念することができると考えられる．P^3 は無償で配布している．利用を希望される方は，次のウェブサイト（http://daweb.ism.ac.jp/support/software/P-cubed/P-cubed.html）に記載の要領で申し込みいただきたい．

現在は PF のみを実装しているが，用途によっては別の手法を用いた方がよい場合もある．例えば，数百次元以上の大規模な非線型状態空間モデルを扱う場合には，PF よりもアンサンブルカルマンフィルタ（Evensen, 1994, 2003）が有効であるし，システムノイズがガウス的である場合には，混合ガウスフィルタ（Stordal et al., 2011）などが有効であると考えられる．今後は，このような他の有用な非線型逐次ベイズ推定手法を追加するなどして，機能の充実を図る予定である．

謝　辞

P^3 のプログラム開発にあたっては，科学研究費補助金基盤研究 B（課題番号: 26280010）の助成を受けた．ここに感謝の意を表する．

参 考 文 献

Bai, F., Gu, F., Hu, X. and Guo, S. (2016). Particle routing in distributed particle filters for large-scale spatial temporal systems, *IEEE Transactions on Parallel and Distributed Systems*, **27**, 481–493.

Balasingam, B., Bolić, M., Djurić, P. M. and Míguez, J. (2011). Efficient distributed resampling for particle filters, *Proceedings of IEEE International Conference of Acoustics, Speech, Signal Processing*, 3772–3775.

Bengtsson, T., Bickel, P. and Li, B. (2008). Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems, *Probability and Statistics*: *Essays in Honors of David A. Freedman*, **2**, 316–334, Institute of Mathematical Statistics, Beachwood, Ohio.

Bolić, M., Djurić, P. M. and Hong, S. (2005). Resampling algorithms and architectures for distributed particle filters, *IEEE Transactions on Signal Processing*, **53**, 2442–2450.

Daum, F. and Huang, J. (2003). Curse of dimensionality and particle filters, *Proceedings of IEEE Aerospace Conference*, **4**, 1979–1993.

Doucet, A., de Freitas, N. and Gordon, N. (eds.) (2001). *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research*, **99(C5)**, 10143–10162.

Evensen, G. (2003). The ensemble Kalman filter: theoretical formulation and practical implementation, *Ocean Dynamics*, **53**, 343–367.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings F*, **140**, 107–113.

Hlinka, J., Hlawatsch, F. and Djurić, P. M. (2013). Distributed particle filtering in agent networks, *IEEE Signal Processing Magazine*, **30**, 61–81.

Kitagawa, G. (1993). Monte Carlo filtering and smoothing method for non-Gaussian nonlinear state space model, Reserch Memo., No.462, The Institute of Statistical Mathematics, Tokyo.

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

Labbe, R. (2015). Kalman and Bayesian Filters in Python, http://filterpy.readthedocs.io (2018 年 2 月 17 日閲覧).

Nakamura, K., Yoshida, R., Nagasaki, M., Miyano, S. and Higuchi, T. (2009). Parameter estimation of in silico biological pathways with particle filtering towards peta-scale computing, *Proceedings of Pacific Symposium on Biocomputing*, **14**, 227–238.

Nakano, S. (2010). Population-based quasi-Bayesian algorithm for high-dimensional sequential problems and hierarchization of it for distributed computing environments, *Proceedings of 2010 IEEE World Congress on Computational Intelligence*.

Nakano, S. and Higuchi, T. (2010). A dynamic grouping strategy for implementation of the particle filter on a massively parallel computer, *Proceedings of 13th International Conference on Information Fusion*.

Nakano, S. and Higuchi, T. (2012). Weight adjustment of the particle filter on distributed computing system, *Proceedings of 15th International Conference on Information Fusion*, 2480–2485.

Savic, V., Wymeersch, H. and Zozo, S. (2014). Belief consensus algorithms for fast distributed target tracking in wireless sensor networks, *Signal Processing*, **95**, 149–160.

Snyder, C., Bengtsson, T., Bickel, P. and Anderson, J. (2008). Obstacles to high-dimensional particle filtering, *Monthly Weather Review*, **136**, 4629–4640.

Stordal, A. S., Karlsen, H. A., Nævdal, G., Skaug, H. J. and Vallès, B. (2011). Bridging the ensemble Kalman filter and particle filters: The adaptive Gaussian mixture filter, *Computational Geosciences*, **15**, 293–305.

van Leeuwen, P. J. (2009). Particle filtering in geophysical systems, *Monthly Weather Review*, **137**, 4089–4114.

Proceedings of the Institute of Statistical Mathematics Vol. 66, No. 2, 339–351 (2018) 351

P^3: Python Parallelized Particle Filter Library

Shin'ya Nakano[1,2], Yuya Ariyoshi[1,3] and Tomoyuki Higuchi[1,2]

[1]The Institute of Statistical Mathematics
[2]School of Multidisciplinary Science, SOKENDAI
[3]Now at Faculty of Engineering, Nippon Bunri University

Particle filter (PF) is a class of state-estimation techniques based on Monte Carlo computation that use a large number of particles. Because PF is applicable even to nonlinear and/or non-Gaussian problems, it is used for a variety of purposes. One serious problem of PF is its computational time, which is exponential in the degrees of freedom of the state vector. Parallel computing is an effective way to decrease computational time, but this approach requires skills in parallel programming. Even for experienced users, it is challenging to achieve high computational efficiency in PF computation because the PF algorithm contains a procedure difficult to parallelize. We developed a Python library named P^3(Python Parallelized Particle Filter Library), that enables us to readily use parallel-ready PF algorithms with high parallel efficiency. In this paper, we describe the parallelized PF algorithms available in P^3, as well as explaining the design and characteristics of the library.

Key words: Particle filter, parallel computing, Python.