ビッグデータ時代の統計理論

経営学部教授
西山貴弘

にしやま たかひろ

東京理科大学理学研究科博士課程修了。博士（理学）。専門は統計科学。主要論文に Hyodo, M., Nishiyama, T. & Pavlenko, T. (2023). A Behrens–Fisher problem for general factor models in high dimensions. Journal of Multivariate Analysis, 195:e105162. など。著書に共著として『R・Python による統計データ科学』（勉誠出版、2020 年）。趣味は旅行、料理、映画鑑賞。

生田10 号館の研究室にて

はじめに

“I keep saying that the sexy job in the next 10 years will be statisticians.（これからの 10 年で最も魅力的な職業は統計家だろうって、言い続けているんだ。）”この言葉は、2009 年に、当時カリフォルニア大学バークレー校の教授で、Google のチーフエコノミストであったハル・ヴァリアン博士がある論文誌へ寄せた有名なコメントです。この当時、ビッグデータが「21 世紀の石油」として大きな注目を集めていましたが、それから既に 10 年以上が経過し、その社会での活用とサービスの実装化によって、デジタル変革の新たな潮流を生み出しています。このように、本格的なデータの時代を迎えて、コンピュータサイエンス・数理科学・統計科学を融合した「数理・データサイエンス・AI 教育」の体系化が構築されつつあります。

私の専門分野は統計科学で、その中でも多次元データを分析する方法論である多変量解析に関する研究を行っています。統計科学は、数理データサイエンス分野の中核であり、自然科学や社会科学を問わず多くの分野で用いられています。また、日常生活の中で AI や機械学習といった言葉を耳にしたことがあるかと思いますが、それらにも深く関係する分野となっています。今回は、ここ数年興味を持って研究に取り組んでいる「高次元データ分析」に関する話題を紹介します。なお、この話題に関する一連の研究を推進するために、JSPS 科研費（課題番号 26730020 若手研究（B）、課題番号 17K00056 基盤研究（C）、課題番号 20K11714 基盤研究（C））を助成いただきました。

高次元データとは？

近年の情報化の進展に伴って、諸科学のあらゆる分野や、ビジネス活動および社会活動の各所で、大規模なデータが日々取得・蓄積されています。ネットワークの普及によって、SNS、音声、画像、映像を含む様々なデータがインターネットなどを通じて大量に作成・配信され、これらが簡単に手に入るようになりました。特に現代では、日々新しく生まれるデータがますます巨大化し、それらのデータが互いに融合し複雑化するビッグデータ時代と呼ばれる時代になっています。このような情報の多様化、データ収集・蓄積技術の向上などに伴って、古典的な統計手法では取り扱うことが難しい様々な特徴をもったデータが増えてきています。

そのようなデータの中で、ゲノム科学、情報工学、金融工学といったいわゆる現代科学において、多次元データの中でも次元（変数の数）が大きいような「高次元データ」が生じる場面が増えており、高次元データ分析が脚光を浴びています。例えば、企業経営においては、これまでは蓄積された大規模な顧客データなどを分析することによって市場全体の傾向把握を行ってきましたが、現在では、それに加えて、個々の顧客属性や購買例歴のデータ分析に基づいて

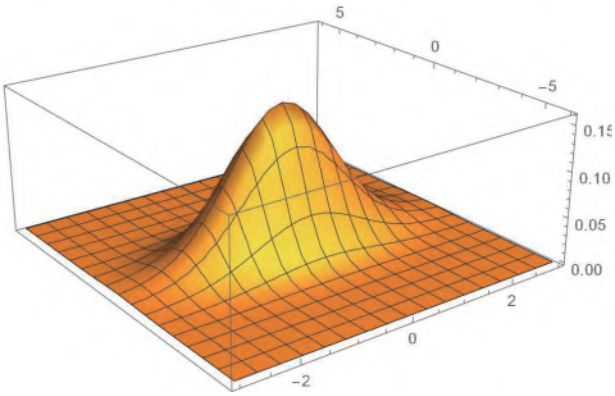

図　多変量正規分布（2 次元の場合）

↑スウェーデン王立工科大学にて

細分化されたサービスをスピーディーに提供することによって、個々の顧客ニーズを満たし、購買行動につなげる取り組みが行われています。その他、画像・音声データなどにおいても高次元データは生じますが、標本数に比べて次元が圧倒的に大きい高次元小標本のケースや、標本数と次元が同程度であるようなケースなど、様々な種類があり、これらのデータから有用な情報を引き出すことが重要な問題となっています。しかしながら、従来の多変量解析法は、一般に次元に対して標本数がかなり多い場合を想定した“大標本枠組み”の下で発展してきているため、通常の多変量解析の方法論では高次元データの推測に対する精度を保証することができません。そのため、高次元データに特化した多変量統計解析手法の開発の必要性が高まり、特に 2000 年以降、高次元データを想定した“高次元枠組み”の下での理論および方法論に対する多くの重要な研究結果がもたらされています。

これまでの研究成果のあらすじ

私はこれまでに、高次元データに対する各種の統計的仮説検定に関する研究を行ってきました。統計的仮説検定とは、例えば、「ある病気に対して新しく開発されたワクチンの薬効時間は従来のものと同じである」という仮説をたて、その仮説を否定するか否かを、確率を用いて判断する方法です。もう少し詳しく説明すると、観測されたデータを用いて検定統計量と呼ばれる統計量の値を計算し、“仮説が正しい”という仮定の下でその値がどの程度の確率で起こり得るのかを評価します。その結果、非常に稀にしか起こらないという場合には、仮説が誤っていると判断することになります。このように確率を評価するためには、検定統計量が仮説の下でどのような確率分布に従うのか（これを「帰無分布」と呼びます）を数学的に導出する必要があります。すなわち、統計的仮説検定の研究では、性能の良い検定統計量の提案やその帰無分布の導出が重要なポイントとなります。

多次元データを考える場合、データの発生メカニズム（母集団分布）として、多変量正規分布（図参照）という確率分布を仮定した下で多くの理論が構築されています。一般に、この仮定を外してしまうと理論構築がかなり難しくなることが多いのですが、残念ながら次元が大きくなるにつれ、多変量正規分布の仮定が成り立つ状況は少なくなり、この仮定の下で構築された方法論を実データに適用することは困難になってしまいます。そこで我々の研究グループは、より緩い仮定の下で、これまでに高次元枠組みにおける各種統計的仮説検定に対する新たな検定方式の提案を行い、一連の研究成果をいくつかの論文として公表しています。その中の一つとして、2021 年 11 月から 2022 年 8 月までの間、長期在外研究員としてスウェーデンの KTH Royal Institute of Technology（スウェーデン王立工科大学、写真）に滞在し、現地で Tatjana Pavlenko 准教授（当時）達と行った研究成果をまとめた論文が最近公表されました。

最後に

冒頭でも記載しましたが、現在、文系・理系を問わず、数理・データサイエンス・AI 教育が注目されており、本学でも「Si データサイエンス教育プログラム」という教育プログラムを全学的に展開しております。私も統計科学を専門とする教員として、最新の研究成果を授業にフィードバックできるよう、今後も研究と教育に取り組んでいきたいと思います。