2022年度共同利用研究報告書

2023年01月27日

所属・職名 九州大学大学院数理学研究院・教授

増田 弘毅

		整理番号	2022a010	
1.研究計画題目	データサイエンスにおける統計科学			
2.新規・継続	新規			
3.種別	一般研究			
4.種目	研究集会（Ⅱ）			
5.開催方法	ハイブリッド開催			
6.研究代表者	氏名	増田 弘毅		
	所属 部局名	九州大学大学院数理学研究院	職名	教授
7.研究実施期間	2022年11月26日(土曜日)～2022年11月26日(土曜日)			
8.キーワード	統計科学、諸科学との協働			
9.参加者人数	55人			

10.本研究で得られた成果の概要

データサイエンス分野において有機的な異分野協働の芽を育てるためには、統計学を共通言語とした議論が建設的な役割を果たす。多様なバックグラウンドを持つ講演者が集まった本研究集会を通じてそのことを再認識できた。解釈可能性、計算可能性、科学的発見の引き金となる異分野交流、そして統計数理に基づく普遍性、これら全ての相乗的な発展が喫緊の課題である。今後も継続してこのような場を持ち、統計学的視点を基盤とした異分野横断型データサイエンスを発展させてゆく必要があるだろう。

本研究集会をつうじて、異分野協働の可能性が芽生えつつある。研究代表者は森氏の講演内容について、確率過程の統計学の視点から、同期現象の統計解析に関する問題定式化の糸口を探索している。これはまだ数理統計基盤が十分に練り上げられていない内容であり、現在、統計的モデリングとしての問題設定へ立ち戻りつつ共同研究打ち合わせを続けているところである。また、個人を対象にしたデータの収集・分析にあたっては、プライバシーに配慮することが求められる場面が多く、統計モデリングに基づいた方法によって問題解決につなげられる可能性を感じている。他にも、工学・科学・医学分野で、小標本・高次元データや頑健性のないデータといった複雑な構造をもつデータを扱う状況が多くなってきており、講演で紹介されたように統計的な方法やものの見方により有用な分析を行うことができる可能性がある。数理統計の理論構築のみならず、アルゴリズムやソフトウェアの開発、さらには社会実装が求められる現代において、こうした知見が、今後の異分野協働および産業への応用に役立つことが期待される。

報告書：一般研究-研究集会(II)「データサイエンスにおける統計科学」

本研究集会は、データサイエンスという巨大な科学分野において、中長期的な視野をもって統計学を共通言語とした有機的な異分野協働の芽を育てる目的で行われた。新たな融合研究領域の創成、産業界との連携、ならびに基礎数理研究へのフィードバックおよびその深化を目的とし、川野秀一氏（九州大学数理学研究院）、森史氏（九州大学芸術工学研究院）、船渡川伊久子氏（統計数理研究所）、堀磨伊也氏（鳥取環境大学）、星野伸明氏（金沢大学）、矢田和善氏（筑波大学）の六方に最新の研究内容を講演していただいた。数理・応用・融合分野と幅広いジャンルを対象にしたことにより、我々も統計学的視点の普遍性をあらためて実感・再認識できただけでなく、異分野からの視点をつうじて新たな問題意識を持てた非常に有意義な集会であった。

川野秀一氏の講演内容は、データサイエンスによって工場の水処理をできるだけ高精度に行うというものであった。工場の水処理においては得られるサンプルが限られており、小標本の問題に取り組む必要がある。そこで川野氏は、スパース推定や多変量重回帰分析によって精度よく予測し、さらにブートストラップによって安定して変数選択する方法を考案した。小標本の場合のパラメータ推定は、統計学では古典的な話題であるが、実務で使うことの難しさを改めて感じた。

森史氏の講演内容は、複数のメトロノームの振動（振動子）がシンクロするという不思議な現象を解明するための数理モデリングであった。講演者は特に「ノイズ」に着目し、２つの振動子がしっかりとシンクロしている場合、確率的なノイズのある結合位相振動子モデルを用いて、振動子間の結合とノイズの大きさを同時推定する式を導出した。統計学においてノイズ構造の推定は重要であるが、物理現象に対してノイズをも推定するという内容は興味深いものであった。このような物理モデルに基づく統計解析は、再生可能エネルギーの発電量予測や材料の物性予測など、今後ますます必要とされると考えられる。森氏の扱っている問題は、確率過程の統計学という視点で言うと、特定の停止時刻でのみ時系列データを観測する状況におけるモデルの推測に相当する。これはまだ数理統計の基盤が十分に確立されていない内容であり、現在、統計的モデリングとしても問題設定へ立ち戻って共同研究打ち合わせを続けているところである。

船渡川伊久子氏は、COVID-19 の詳細な統計解析に関する内容を講演された。実際のデータは頑健性が不十分であり、とくに、Lancet のような専門誌における解析においても慎重な結果の解釈が望まれることが説明された。さらに、肥満等の健康に関する統計、関連した妊娠中の体重増加や帝王切開などが話題にのぼった。数理統計では実はあまり出てこない「データの見方」というデータサイエンスの根幹に関する講演が主であり、今後のデータ解析における注意点等がわかった。

堀磨伊也氏は、ディープラーニング等の機械学習の結果を解釈する手法である SHAP の数学的内容を講演された。統計学においては、スパースモデリングなどによって結果を解釈

することが多いが、より複雑な問題に対する結果が解釈されることはあまりない。今後、統計学において結果を解釈するためのブレークスルーが必要となると考えられる。また、配車リクエストやMoodleを使ったラーニングアナリティクスへの応用にも触れられた。

星野伸明氏は、プライバシー保護に関する話題について講演された。プライバシーを保護することと詳細なデータを使うことがトレードオフであり、プライバシーを保護しつつデータを活用するという内容が入門的なところからわかりやすく説明された。近年、マーケティングをはじめとした経営活動では個人単位の様々なデータが集められる一方、プライバシー保護に対する要求は高まってきており、プライバシー保護を考慮した統計モデリングが今後ますます必要になると考えられる。また、理論面では一般化多項分布が説明され、非常に興味深い内容であった。

矢田和善氏の講演内容は、高次元小標本の状況における平均ベクトルの推測に関する話題であった。球面集中現象とよばれる高次元特有の特徴を活用し、ノイズを除去したり精度を向上したりする手法に関する最新の内容が紹介された。講演者の10年間の研究の集大成であり、高次元小標本データ解析の様々な内容が網羅されていた。産業への応用を考えると、このような高次元小標本データの極めて高度な解析に加え、外れ値や欠測などの問題に対応した手法の開発も望まれる。理論と実用の双方における問題提起の要素も感じられる講演であった。

コロナ禍の影響によりハイブリッドでの開催となったが、オンライン参加と対面参加を合計して５７名の多くの研究者に参加していただいた。上記のとおり極めて多岐にわたる講演内容であり、それぞれにおいて活発な議論が行われた。データサイエンスにおいては対象を様々な視点から複合的に考えることが重要である。研究者が個人で行う研究ではそのような機会を持つことは難しいため、今回の研究集会は大変意義のあるものとなった。やはり分野が異なると新鮮な視点に触れることができ、結果自身の視野を広げることにつながった。統計的機械学習という用語も定着した現在、解釈可能性、計算可能性、科学的発見の引き金となる異分野交流、そして統計数理に基づく普遍性、これら全ての相乗的な発展が喫緊の課題であることを再認識できた次第である。継続してこのような場を持ち、統計学を基盤とした異分野横断型データサイエンスを発展させてゆく必要があるだろう。

九州大学 IMI 共同利用(研究集会（Ⅱ）)

データサイエンスにおける統計科学

Statistical Science in Data Science

１１月２６日（土）　10:25～17:05

10:25-10:30　挨拶（増田 弘毅）

10:30-11:10　講演①　座長：後藤 佑一
　　講演者：川野 秀一（九州大学 大学院数理学研究院）
　　講演タイトル：データサイエンスによる水処理微生物群集の解析
11:10-11:20　フリーディスカッション

11:20-12:00　講演②　座長：後藤 佑一
　　講演者：森 史（九州大学 芸術工学研究院）
　　講演タイトル：振動の時刻データを用いた振動子間の結合強度の推定
12:00-12:10　フリーディスカッション

12:10-13:30　昼休憩

13:30-14:10　講演③　座長：倉田 澄人
　　講演者：船渡川 伊久子（統計数理研究所）
　　講演タイトル：COVID-19 と健康関連指標
14:10-14:20　フリーディスカッション

14:20-15:00　講演④　座長：倉田 澄人
　　講演者：堀 磨伊也（公立鳥取環境大学 人間形成教育センター）
　　講演タイトル：予測モデルの局所的解釈に基づくフィードバックの実現
15:00-15:20　フリーディスカッション・休憩

15:20-16:00　講演⑤　座長：廣瀬 雅代
　　講演者：星野 伸明（金沢大学 経済学経営学系）
　　講演タイトル：情報保護のための一般化多項分布族

16:00-16:10　フリーディスカッション
16:10-16:50　講演⑥　座長：廣瀬 雅代
　　講演者：矢田 和善（筑波大学 数理物質系）
　　講演タイトル：データ変換法を用いた高次元平均ベクトルの推測について
16:50-17:00　フリーディスカッション

17:00-17:05　挨拶（廣瀬 慧）

Speaker: Shuichi Kawano (Kyushu University)

Title: Analyzing water treatment microbial community via data science

Speaker: Fumito Mori (Kyushu University)

Title: Inference of interaction intensities of coupled oscillators using only spike time data

Speaker: Ikuko Funatogawa (The Institute of Statistical Mathematics)

Title: COVID-19 and Health Related Indicators

Speaker: Maiya Hori (Tottori University of Environmental Studies)

Title: Realization of feedback based on local interpretation of predictive models

Speaker: Nobuaki Hoshino (Kanazawa University)

Title: Generalized Multinomial Distributions for Information Protection

Speaker: Kazuyoshi Yata (University of Tsukuba)

Title: Inference on high-dimensional mean vectors by data transformation techniques