高次元の統計学
青嶋 誠・矢田 和善著
https://www.kyoritsu-pub.co.jp/bookdetail/9784320112636

まえがき

かつての統計解析に用いるデータセットは，データの次元数 d と標本数 n に，$d < n$ なる大小関係が前提でした．ところが，1990 年代後半，情報化の進展に伴い，$d \gg n$ という高次元データが出現しました．当時の統計学（多変量解析）は，$d < n$ なる条件が理論の拠り所となっていましたので，高次元データの統計的推測に精度を保証することはできませんでした．2000 年代になり，$d > n$ の枠組みで高次元データの理論研究が徐々に始まりました．筆者たちも，2005 年頃からこの未開の地に足を踏み入れ，当時はまだ文献がほとんどありませんでしたので，新たな統計学の開拓を始めました．2010 年代に入り，理論と応用の両面から統計学が飛躍的に進歩し，多変量解析に替わる新たな統計学として高次元統計解析が誕生しました．標本数が次元数と比べ圧倒的に少ない状況でも，統計的な推測が可能になったのです．

高次元統計解析は，高次元ならではの新しいアイディアに基づいて，理論と方法論が構築されています．まだ新しい分野ですので，筆者たちの知る限り高次元統計解析を解説している書物は少なく，専門的な洋書が数冊出版されている程度です．初学者が効率的に学習するには，極めて困難な状況にあるといえます．そう思っていた矢先，「統計学 One Point」の執筆依頼を，編集委員長の鎌倉稔成先生からいただきました．One Point の性格上，学部生・大学院生や初学者に向けた高次元統計解析へといざなう入門書として，本書の執筆をお引き受けしました．入門書ですので，書籍名は重いものにならないように気をつけました．こうして決まった『高次元の統計学』は，筆者の一人である青嶋が 2016 年 3 月に日本数学会主催の市民講演会で講演した題目と同じものです．初学者にも親しみやすい内容で，高次元統計解析の基本的な考え方をお伝えできればと思います．

第 1 章は，高次元データの注意点や本書で扱う記号を簡単に説明しま

す．第 2 章は，高次元データを解析する上で鍵となる，高次元データ特有の幾何学的表現を解説します．高次元データは高次元空間で眺めることで，いくつかのパターンが見えてくることを説明します．第 3 章は，高次元データに対して，多変量解析の次元縮約法として知られる主成分分析 (PCA) にはいくつか問題があることを指摘します．固有値・固有ベクトル・主成分スコアは，次元の呪いを受けて誤った結果を出力してしまいます．第 4 章は，高次元データ特有の幾何学的表現に基づいて，ノイズ掃き出し法とクロスデータ行列法という 2 つの高次元 PCA を解説します．これらは色々なところに応用できますが，ここでは一例として，高次元混合分布の幾何学的表現のあぶり出しに応用し，高次元クラスター分析を扱います．第 5 章は，高次元平均ベクトルの統計的推測を考え，高次元漸近正規性や一致性を解説し，信頼領域や検定手法を与えます．第 6 章は，高次元データの判別分析を考え，高次元空間で浮き彫りになるデータのパターンを利用して，$d \gg n$ なる高次元小標本でも有用な判別方式を解説します．機械学習で知られるサポートベクターマシン (SVM) についても，高次元データを扱う上での問題点とその修正方法を解説します．

高次元統計解析は，新しい統計学です．標本数が次元数と比べて圧倒的に少ない状況であっても，統計的推測に高い精度を保証します．本書は，高次元にそびえ立つビッグデータに少ない標本数で立ち向かう，高次元統計解析の真髄をお伝えします．高次元の統計学には，従来の統計学の枠組みを超えた，新しい発想が必要になることをご覧に入れます．

最後に，原稿を閲読していただいた 2 名の先生方と，編集委員ならびに共立出版編集部の方々には，この場を借りて厚く御礼を申し上げます．

2019 年 2 月

青嶋　誠・矢田和善