高次元統計解析：理論・方法論とその周辺

筑波大学・数理物質系　　青嶋　誠

ゲノム科学，情報工学，金融工学などの現代科学の 1 つの特徴は，データがもつ次元数の膨大さにある．例えば，次世代シークエンサによるゲノム配列データなど，次元数が数百万を超えるデータも解析の対象になる．こういった高次元データの第 1 の特徴は，次元数が標本数を遥かに超えることである．第 2 の特徴は，高次元データは豊富な情報を有するものの，それが巨大なノイズに埋もれて見つけ難いことである．これらの理由から，通常の多変量解析法では高次元データの推測に精度を保証することができず，間違った解析結果を導くことさえある．そのため，高次元データの解析には，高次元データ特有の新しい理論と方法論が必要になり，我々はそれらを高次元統計解析と名付け，以下の 2 つの柱を打ち立てた[1]．最近では，2 つの柱を融合させ，高次元固有空間の構造に基づく新たな高次元統計解析を展開している．

1．高次元データにおける統計的推測 Aoshima and Yata (2011, SA) は，高次元データの統計的推測に幾何学的表現に基づく各種方法論を考案し，統計量の高次元漸近正規性，標本数の設計，推測の精度保証に至るまで，一連の基礎理論を拓いた先駆的論文である．高次元球面上の信頼領域 [Aoshima and Yata (2011; 2015, MCAP)]，高次元二標本問題 [Aoshima and Yata (2011, 2015)]，高次元共分散行列の推定・検定 [Aoshima and Yata (2011)，Ishii, Yata and Aoshima (2016, JSPI)]，高次元判別分析 [Aoshima and Yata (2011; 2014, AISM; 2016, MCAP), Nakayama, Yata and Aoshima (2017, JSPI)]，パスウエイ解析 [Aoshima and Yata (2011), Yata and Aoshima (2013a, JMVA; 2016, JMVA)]，高次元変数選択 [Aoshima and Yata (2011, 2016)] など，多くの統計的推測の問題に高次元データの豊富な情報を生かすための理論と方法論が与えられている．

2．高次元データにおける新しい PCA 高次元データは特有の幾何学的表現を織りなす．Aoshima and Yata (2011), Yata and Aoshima (2012, JMVA) は，高次元固有空間を双対空間で捉え，幾何学的表現を発見した．Yata and Aoshima (2012) は，ノイズの大きさを幾何学的に見積もり，固有空間を高精度に推定する「ノイズ掃き出し法」という高次元 PCA を開発した．Yata and Aoshima (2010, JMVA) は，高次元非正規データも高精度に処理する「クロスデータ行列法」という高次元 PCA を開発し，クラスタリングに応用した．これらの高次元 PCA は汎用性が非常に高く，Yata and Aoshima (2010, CSTM; 2013b, JMVA) はクロスデータ行列法を高次元潜在空間の次元推定に応用し，Yata and Aoshima (2016, EJS) はノイズ掃き出し法を高次元信号行列の推測に応用した．

高次元統計解析の最近の発展として，Aoshima and Yata (2017, SS) は強スパイクモデルというノイズモデルを提唱した．高次元データのノイズは巨大かつ非スパースであり，それゆえ潜在的な幾何学的構造が破壊され，統計的推測の精度保証はしばしば困難になる．強スパイクモデルは，そのような高次元空間のノイズを表現している．高次元統計的推測は弱スパイクモデルが前提であったが，Aoshima and Yata (2017) は強スパイクモデルのもとで新たな高次元統計解析を展開した．これは，巨大なノイズをノイズ掃き出し法で精密に解析し，強スパイクするノイズ空間を避けるようにデータを変換して潜在空間の幾何学的構造を炙り出し，高次元統計的推測の適用を可能にするというものである．

[1]解説は，青嶋・矢田 (2013a, 日本統計学会和文誌「寄稿論文」; 2013b, 数学「論説」) を参照のこと．