探索的データ解析の現代化

（株）データサイエンスコンソーシアム，慶應義塾大学　　柴田 里程

1　Tukey の時代と現代

J.Tukey による探索的データ解析 (EDA) の提唱 (1977) は，今日のデータサイエンス興隆の契機となったが，それでもすでに 40 年以上の年月が経過している．その間に起きたコンピュータの能力の画期的な向上とネットワーク化を考えれば，彼の提案した探索的データ解析も現代化が必要であることに疑いはない．40 年前にはコンピュータの能力はかなり制約的であり，さまざまなことをごく少数のスカラー量に縮約するか，ごく単純な図表に縮約し，それらを介して人間が判断するという形しかとれなかった．しかし，そのような制約から自由になった現在，探索的データ解析にかぎらず伝統的な統計学も大きな見直しが迫られている．本講演では，線形モデルを例に，スカラー量へ縮約せずに高次元線形空間を直接視覚的に眺めることで，どれだけ直感的な理解を助け，発想を豊かにするか，その現代化の道筋を明らかにする．

2　データの価値評価

データサイエンスにおいて，データ解析の果たす役割としては，データからなんらかの発見をする以外に，データの価値を見極める役割も見逃せない．至るところに寝っているデータの活用が一向に進まない理由の一つが，価値があるに違いないと思い込んでいるだけで，具体的な評価がなされていないところにある．データの価値評価は単になにに分析を行いました，ではすまない．総体としてデータを理解し，さまざまな側面からの評価を積み重ねる必要がある．

3　○○分析よさようなら

線形モデルを前提とするデータ解析手法としては，回帰分析，分散分析，ロジット分析, 正準相関分析, コレスポンデンス分析などさまざまな分析法が存在し悩ましい．しかし，これらはごく少数のスカラー量に縮約するしかないという歴史的な制約から生じたもので，今やその違いに煩わされる必要はまったくない．以下のような点に留意すれば，これらの壁は容易に乗り越えられ早期のデータの総体的な理解に結びつく．

3.1　対比

被説明変量，説明変量を問わず，カテゴリカルデータベクトルを対比 (contrast) でコーディングすることで，数値データベクトルと同等に扱うことができ，すべてを線形代数の枠組みで扱えるようになる．

3.2　ノルム，直交，射影

線形モデルのあてはめを射影として理解し，分散分析は直交化した当てはめ結果のピタゴラスでしかないと理解することで，回帰分析と分散分析の違いは霧消する．

3.3　視覚表現

線形モデルを当てはめる前に説明変量ベクトル群の様子を探るには TextilePlot, 当てはめ結果の診断には Parallel Coordinate Plot を用いれば，全体的な様子から細部の様子にいたるまで特定の縮約量に頼らない解析が自由におこなえる．