連続変数とカテゴリー変数を含む
集約的シンボリックデータの特徴抽出

統計数理研究所　中野 純司
徳島文理大学　　山本 由和
統計数理研究所　清水 信夫

1　はじめに

シンボリックデータ解析のひとつの考え方は，個々のデータを扱うのではなくそれらを意味のあるグループに分割したとき，それぞれのグループをひとつのデータと考えそれに対して推論を行うものである．われわれはグループを表現するためのいくつかの記述統計量のことを集約的シンボリックデータと呼ぶ．この考え方は超多量のデータが与えられたときにその大まかな性質を理解するのに有用である．現実の個体データでは実数変数とカテゴリー変数の両方を含む場合が多いが，それらを同様に扱うことは容易ではない．ここではカテゴリー値を数値化することにより実数変数とみなし，それらの２次までのモーメントを可視化することによってグループの特徴を抽出することを提案する．

2　グループを記述する統計量

同じ変数が記録された多量の個体データが意味のあるグループに分類されているとき，それらのグループの特徴をとらえることを目的とする．そのためには少数の記述統計量でグループを表現する必要がある．実数変数だけが記録されている場合は，最も簡単なグループの代表値は各変数の標本平均であり，その次に重要なのは標本分散共分散であろう．これらは実数変数の２次までのモーメントである．カテゴリー変数に関しては２次までのモーメントに対応するのは２変数ごとの分割表である．従って，これらの記述統計量をグループを表すための集約的シンボリックデータと考え，以後の解析では個々の個体データは利用しない．

3　カテゴリー変数の連続値化

２つのカテゴリー変数の分割表は２つの連続変数の２次までのモーメントに比較して多くの情報を含むので，そのままではまだ理解しにくい．分割表をより理解しやすくするためには，対応分析によって各カテゴリー値に数値を対応させることが行われている．ここでも同様のことを行いたいが，われわれの問題では各グループごとに変数のペアごとの分割表があることに注意しなくてはならない．これはグループごとのバート行列の比較を行うことになる．そこで各カテゴリー値に数値を与えるときに，グループ間の差がもっとも際立つようにする．そのように定式化すると多重対応分析の場合と同様に，行列の特異値値問題を解くことで解が得られる．

4　グループの可視化

カテゴリー変数を連続値化することによって，グループごとにすべての変数の標本平均，標本分散，２変数ごとの標本相関が得られる．それらは拡張した平行座標プロットによって可視化することができる．これにより集約的シンボリックデータの直観的な特徴抽出が可能となる．