連続・離散変換－数量化と主成分分析－

馬場康維
統計数理研究所

１．はじめに

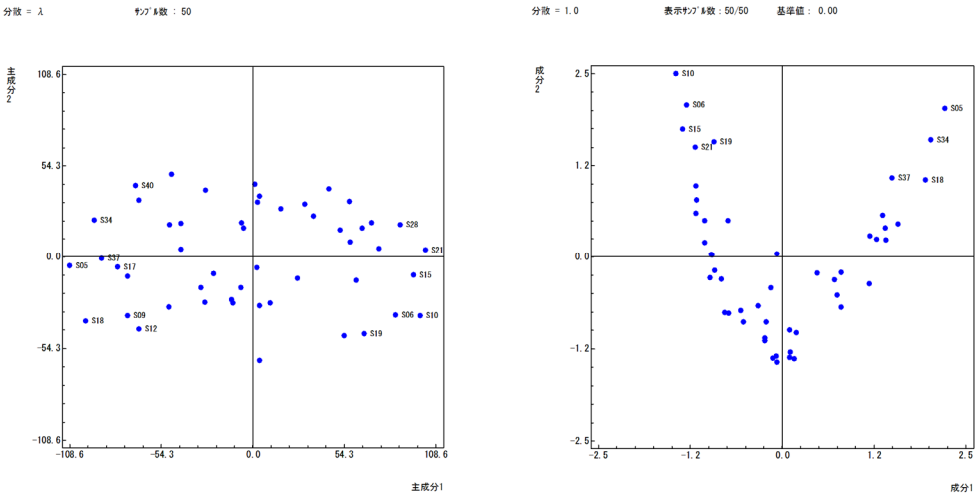
多次元のデータの分析過程で，１）観測値が連続量であるものをカテゴリーデータに変換して分析する，２）観測値が連続量であるものを順位データに変換して分析する，といった方法がとられることがある。連続量のカテゴリー化，順位化の目的は情報の縮約，データ獲得の簡易化，個体の情報の秘匿などさまざまである。

連続量で表されるデータが順序付きカテゴリーに離散化されることにより情報のロスが生じる。そのロスが，分析者にとって許容できる範囲であれば，カテゴリー化によるデメリットの主な部分は解決されたことになる。これまでの報告では，主成分分析や単回帰分析の場合を扱ってきた。

ここでは、連続量をカテゴリーに変換したときの情報の変換について考えてみる。例として主成分分析を考える。主成分分析は分散共分散行列あるいは相関行列が与えられれば結果が得られる方法である。このような方法では，たとえば５段階評価のような粗い離散化でも十分実用になる結果が得られることを以前の報告で示した。その理由は、連続量を離散化したとしても主成分は情報の主要部分を抽出する方法であるからである。

本稿では、同じデータを続量として表現し主成分分析をした結果とカテゴリーで表現したデータに数量化Ⅲ類を適用した場合の比較を行う。

２．例



データは 50 人の生徒の 9 教科の試験の成績（擬似データ）である。左の図は主成分分析の個体スコアを表している。また、右の図は数量化Ⅲ類による個体スコアのプロットであり、順序付カテゴリーの場合に 1-2 軸でプロットした場合の典型的な馬蹄形が現れている。主成分の 1 軸に対応するのが、馬蹄形にそった仮想的な曲線になる。