# 連続・離散変換－情報の保持と秘匿－

統計科学技術センター
特命教授　　馬場　康維

## 1　はじめに

多次元のデータの分析の過程で、

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。
３）順位データとして得られている観測値を連続量とみなして分析する。
４）順序付きカテゴリーとして得られているデータを連続量とみなして分析する。

といった方法がとられることがしばしばある。例えば、順位データに主成分分析を適用する、評点法で得られた5段階評価のデータを用いて回帰分析を行うなどがこれにあたる。

この報告では、例として、連続量を便宜的に離散化して得られたデータとオリジナルデータに主成分分析法を適用したときの結果を比較し、連続量の離散化の影響について考える。連続量を離散化した分析結果がもとのデータの結果と大きな相違がなければ、生データから個体情報が特定されるという可能性の低いデータの提供が可能になるであろう。

さらに離散化したデータに雑音を入れて連続量に変換することを考える。この変換によっても分析に必要な情報が保持されるならば、離散化－連続化という過程を通して、さらに個体情報の秘匿の可能性が高まるであろう。

## 2　離散化による影響

40人の生徒の9教科（国語、社会、数学、理科、音楽、美術、体育、技術、英語）の点数のデータ（杉山（1995）によるデータからサンプリングにより抽出した40人分）を用いて離散化の結果を比較する。主成分分析を適用すると、図１、図２の結果が得られる。図は第1主成分と第2主成分による散布図である。なお、 順位データに変換したデータを用いても結果はほぼ同じになる。

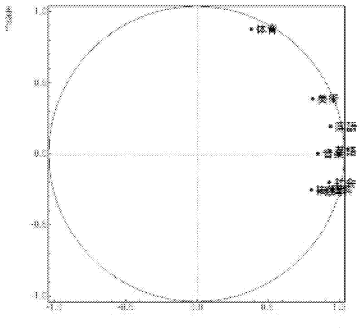

図１　オリジナルデータによる因子負荷量

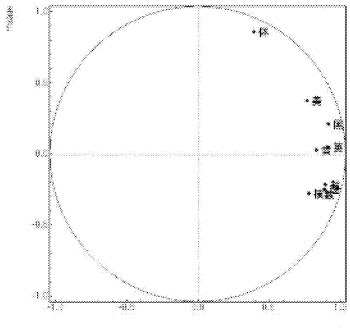

図２　５カテゴリーデータによる因子負荷量

## 3　離散化データの連続化

離散化したデータに雑音を入れて連続量に変換することを考える。そうすることによって、個体情報の特定化はさらに難しくなるであろう。ここでは上記の離散化したデータを用い、逆に連続化することを考える。

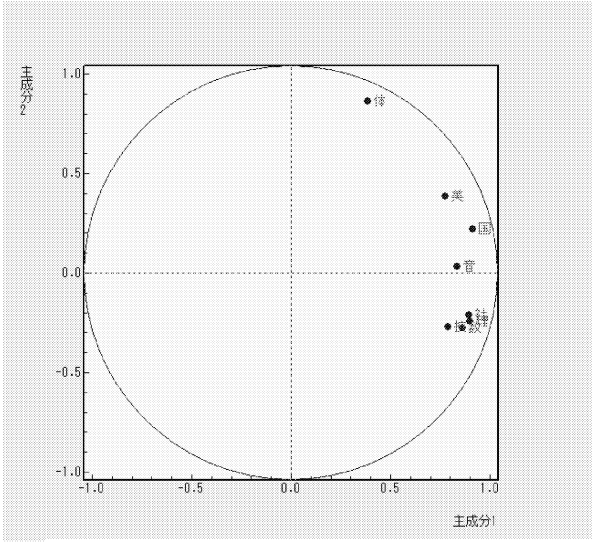


図３　最連続化したデータによる因子負荷量

以下では５カテゴリーのデータに一様乱数を割当て，カテゴリー１，２，３，４，５をそれぞれ区間[0,20]，[21,40]，[41,60]，[61,80]，[81,100]のデータに変換した結果を示す。

乱数を用いてカテゴリーを連続化したデータによる結果は図３のようになる。図 1，図２と比べても結果はほとんど変わらない。したがって，主成分分析などでは，連続データ→離散データ→再連続化といった変換によって、個体情報を秘匿しながら、必要な情報が再現できる変換ができる可能性がある。

## 4　離散化－再連続化の影響

連続型データの離散化、再連続化という変換を通して、主成分分析の結果がほぼ同じになるのは、この手法では変数の相関関係が主たる役割を果たしていることに関係している。すなわち、変換が相関構造をほぼ保持していれば、オリジナルデータを用いるかわりに変換したデータを用いてもほぼ同じ結果を得られるということである。表１、表２にはそれぞれオリジナルデータの相関行列、再連続化したデータの相関行列を示した。個々の相関係数の違いはあるが、相関関係はほぼ保持されていることが分かる。

表１　オリジナルデータによる相関行列

| | 国 | 社 | 数 | 理 | 音 | 美 | 体 | 技 | 英 |
|---|---|---|---|---|---|---|---|---|---|
| 国 | 1.00 | 0.85 | 0.77 | 0.79 | 0.74 | 0.83 | 0.52 | 0.70 | 0.89 |
| 社 | 0.85 | 1.00 | 0.82 | 0.88 | 0.76 | 0.64 | 0.19 | 0.79 | 0.89 |
| 数 | 0.77 | 0.82 | 1.00 | 0.84 | 0.70 | 0.60 | 0.12 | 0.66 | 0.79 |
| 理 | 0.79 | 0.88 | 0.84 | 1.00 | 0.77 | 0.64 | 0.15 | 0.78 | 0.82 |
| 音 | 0.74 | 0.76 | 0.70 | 0.77 | 1.00 | 0.71 | 0.29 | 0.58 | 0.78 |
| 美 | 0.83 | 0.64 | 0.60 | 0.64 | 0.71 | 1.00 | 0.55 | 0.52 | 0.72 |
| 体 | 0.52 | 0.19 | 0.12 | 0.15 | 0.29 | 0.55 | 1.00 | 0.15 | 0.38 |
| 技 | 0.70 | 0.79 | 0.66 | 0.78 | 0.58 | 0.52 | 0.15 | 1.00 | 0.71 |
| 英 | 0.89 | 0.89 | 0.79 | 0.82 | 0.78 | 0.72 | 0.38 | 0.71 | 1.00 |

表２　再連続化データによる相関行列

| | 国 | 社 | 数 | 理 | 音 | 美 | 体 | 技 | 英 |
|---|---|---|---|---|---|---|---|---|---|
| 国 | 1.00 | 0.71 | 0.70 | 0.69 | 0.70 | 0.74 | 0.50 | 0.60 | 0.82 |
| 社 | 0.71 | 1.00 | 0.76 | 0.78 | 0.68 | 0.50 | 0.17 | 0.74 | 0.80 |
| 数 | 0.70 | 0.76 | 1.00 | 0.81 | 0.62 | 0.54 | 0.15 | 0.69 | 0.71 |
| 理 | 0.69 | 0.78 | 0.81 | 1.00 | 0.69 | 0.56 | 0.21 | 0.74 | 0.76 |
| 音 | 0.70 | 0.68 | 0.62 | 0.69 | 1.00 | 0.57 | 0.32 | 0.50 | 0.74 |
| 美 | 0.74 | 0.50 | 0.54 | 0.56 | 0.57 | 1.00 | 0.49 | 0.54 | 0.64 |
| 体 | 0.50 | 0.17 | 0.15 | 0.21 | 0.32 | 0.49 | 1.00 | 0.09 | 0.36 |
| 技 | 0.60 | 0.74 | 0.69 | 0.74 | 0.50 | 0.54 | 0.09 | 1.00 | 0.65 |
| 英 | 0.82 | 0.80 | 0.71 | 0.76 | 0.74 | 0.64 | 0.36 | 0.65 | 1.00 |

## 5　課題

相関構造に基づく分析手法、たとえば重回帰分析、判別分析などでも事情は同様である。したがって、個体情報を秘匿しながら、個票を用いた結果と類似の結果が得られるデータの提供に離散化－連続化という一連の変換が有用であるといえる。個体の秘匿度の評価尺度、情報の保持の評価尺度の構成が今後の課題である。

参考文献

杉山高一（1995）『多変量データ解析入門』朝倉書店