B-006

コレスポンデンス分析におけるデータ変動の影響
Influence of Data Variation for Correspondence Analysis

井田 正明 †
Masaaki Ida

1. はじめに

アンケート分析やテキストマイニングなどにおけるカテゴリカルデータの分析手法として，コレスポンデンス分析は重要な位置を占めている．また結果出力としての同時布置図はパターン分類やデータ可視化法として重要である．しかしながら実用に際してはデータの追加やデータの不確かさなどによるデータ変動への対応が重要な課題となっている．これまでコレスポンデンス分析におけるデータ変動についての検討が行われてきたが，本研究においてはデータの変動とカイ 2 乗値の関係を示すことによりクロス表全体での変動の影響を検討する．

2. コレスポンデンス分析におけるデータ変動

コレスポンデンス分析（対応分析）は，クロス表やバート表などのよってあらわされる複数の項目の対応関係を分析する基本的な手法である [1],[2]．この手法は，アンケート分析やテキスト分析における項目の対応関係を分析する際に頻繁に用いられる．またその結果は同時布置図によって表現されることにより全体的な対応関係を理解するのに役立つ有効な手段となっている．

しかしながら実用においては，データの追加やデータの変動についての検討が必要となってくる．すなわち，クロス表等の微小の変動によるスコアの変動と結果の解釈の再検討である．以前の考察においては，2 つの（または 3 つ以上の）項目に対するコレスポンデンス分析の結果の変動が議論されたが [3]，それらは個々の結果の変動についてであり，全体への影響については十分に検討されていない．

コレスポンデンス分析は，複数の項目とカテゴリーの関係を表現するものであり数理的には特異値分解に基づいている [1],[2]. 表 1 は，データのカテゴリーへの所属をダミー変数により表現したものである．この種の表はカテゴリカルデータの相互関係を表現するために利用されるものであり，テキスト分析やアンケート分析において頻出するものである．

この表を行列 G で表す.

$$G = (G_1, G_2).$$

行列 G_i の要素は 2 値 $\{0, 1\}$ でありその行和は 1 である．たとえば，表 1 の第 1 行目のデータはカテゴリー $G_{1,3}$ と $G_{2,1}$ に属している．クロス表 N_{12}（N_{21}） は

$$N_{12} = {G_1}^{\mathrm{T}} G_2,$$
$$N_{21} = {N_{12}}^{\mathrm{T}},$$

表 1: カテゴリカルデータのテーブル例

	$G_{1,1}$	$G_{1,2}$	$G_{1,3}$	$\cdots$	$G_{2,1}$	$G_{2,2}$	$\cdots$
1	0	0	1	$\cdots$	1	0	$\cdots$
2	1	0	0	$\cdots$	0	1	$\cdots$
3	0	1	0	$\cdots$	1	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

と表される．

コレスポンデンス分析の各項目のスコアは次式の $\boldsymbol{p}$ および $\boldsymbol{q}$ である．

$$N_{12}\boldsymbol{q} = \mu D_1 \boldsymbol{p},$$
$$N_{21}\boldsymbol{p} = \mu D_2 \boldsymbol{q},$$
$$D_1 = {G_1}^{\mathrm{T}} G_1,$$
$$D_2 = {G_2}^{\mathrm{T}} G_2,$$

$\mu\,(\mu_1, \mu_2, \cdots)$ は特異値である．

他の表現として一般化固有値問題としての定式化がある [4]. 行列 A と B を G について次のように定める．

$$A = G^{\mathrm{T}} G,$$
$$B = 2\ \mathrm{diag}(G^{\mathrm{T}} G),$$

$\mathrm{diag}(G^{\mathrm{T}}G)$ は $G^{\mathrm{T}}G$ の対角成分からなる行列である．これら行列によりコレスポンデンス分析は次の形式で表現できる．

$$A\boldsymbol{x} = \lambda B\boldsymbol{x},$$
$$\boldsymbol{x}_i B \boldsymbol{x}_i = 1,$$
$$\boldsymbol{x}_i B \boldsymbol{x}_j = 0 \quad (i \neq j),$$

$\lambda\,(\lambda_1, \lambda_2, \cdots;\ \lambda_i \geq \lambda_j \text{ for } i < j)$ は固有値（重複を含めて）である．また，$\lambda_1 = 1$ であり，$\lambda_i \geq 0$ となる．固有値 λ_i に対応する固有ベクトルを $\boldsymbol{x}_i$ と表す．$\boldsymbol{x}_i$ の要素がコレスポンデンス分析のスコアに対応することになる．

上記の 2 つの形式には次の関係がある．

$$\lambda_i = \frac{1 + \mu_i}{2}.$$

ここでは一般の摂動理論 [5] とは異なり，データの追加による χ^2 値とスコアの変化を考えることになる．$A \to A + \delta A$ かつ $B \to B + \delta B$ ならば，一次近似としての変動 $\lambda_i \to \lambda_i + \delta\lambda_i$ は次のようになる [4].

$$\delta\lambda_i = {\boldsymbol{x}_i}^{\mathrm{T}} (\delta A - \lambda_i \delta B) \boldsymbol{x}_i.$$

†大学評価・学位授与機構, National Institution for Academic Degrees and University Evaluation

この近似に関する誤差の検討が必要であるが，コレスポンデンス分析の結果の変動について，その傾向を分析するためにこれらを用いることは有用である．

また，$\lambda_i = \lambda_j (i \neq j)$, の場合への対応は一般に可能であるが複雑さを避けるためにここでは省略する．また，分析において固有値 $\lambda_1 = 1$ は検討から除かれ，$0 < \lambda_i < 1$ の場合（特に $i = 2, 3$ の場合）について考察されることが多い（一般には累積寄与率による）．

これまでは，クロス表の各要素の変化に対する固有値やスコアの変化を詳細に検討されてきたが，クロス表の全体の変動についての考察が不十分であった．そこで，表全体の傾向については χ^2 値が指標として重要であるため，次の χ^2 値の変動を考える．

$$\begin{aligned} \chi^2 &\rightarrow \chi^2 + \delta\chi^2, \\ \delta\chi^2 &= n \sum_i (4(2\lambda_i - 1)\ \boldsymbol{x}_i{}^{\mathrm{T}} (\delta A - \lambda_i \delta B) \boldsymbol{x}_i), \end{aligned}$$

n はデータ総数である．この性質を使うことによりデータ変動の表全体への影響を検討することができる．

3. 数値例

数値例として，参考文献 [6] の数値例を修正したものを取り上げる．これはカリキュラムの分析におけるコレスポンデンス分析の適用例である．

クロス表 $(G_1 \times G_2)$ の項目 G_1 のカテゴリー数は 13, また G_2 のカテゴリー数は 6 である．コレスポンデンス分析の結果のスコアとして $(\boldsymbol{x}_2, \boldsymbol{x}_3)$ の 2 種を考える．χ^2 値は 360，自由度は 60 でありクロス表に関連性がある．

このクロス表に対するデータの追加変動を考える．これによりコレスポンデンス分析のスコアおよび χ^2 値に変動が生じる．たとえば，例における $(G_{1,11}, G_{2,2})$ のカテゴリーへのデータ増加により，χ^2 値が増加し，スコアの変動が生じた結果を示したものが図 1 である（特に左側の変動）．図中での矢印はスコアの変動の方向と相対的な変動量を示している．同様に，例における $(G_{1,13}, G_{2,1})$ のカテゴリーへのデータ増加により，χ^2 値が増加する場合のスコアの変動を示したものが図 2 である．

大きな正負の χ^2 値の変動はコレスポンデンス分析の結果の解釈に対して再考の必要があることを示唆している．図 1, 2 では 2 次元での変動を図示しているが，一般にはすべての固有値およびそれに対応する固有ベクトルの変動を検討することになる．また，χ^2 値がほぼ変動しない場合に対してもその値はコレスポンデンス分析の結果の再検討の材料となる．このようにクロス表の変動の影響の指標としての活用が考えられる．

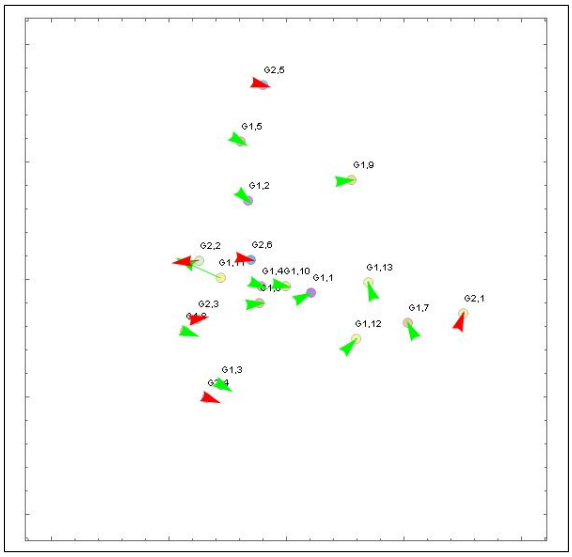

図 1: データの追加によるスコアの変動

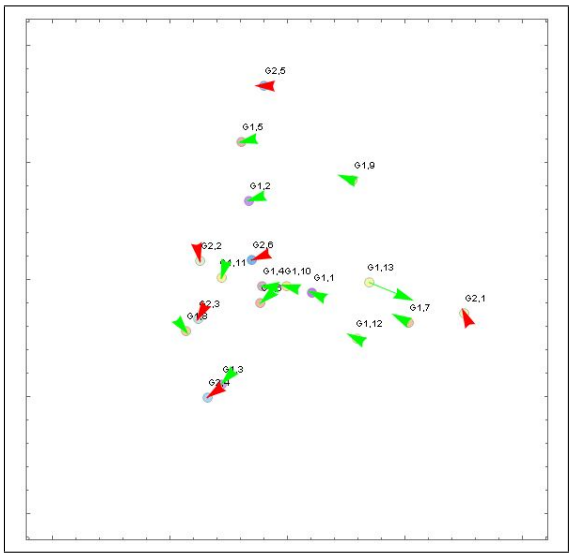

図 2: データの追加によるスコアの変動 2

参考文献

[1] J. P. Benzecri, *Correspondence Analysis Handbook*, Marcel Dekker, 1992.

[2] M. Greenacre, *Correspondence Analysis in Practice, Second Edition*, Chapman and Hall/CRC, 2007.

[3] M. Ida, Consideration on Sensitivity for Multiple Correspondence Analysis, *the International Multiconference of Engineers and Computer Scientists 2010*, pp.560–565, 2010.

[4] M. Ida, First-Order Perturbation of Correspondence Analysis with Multiple Categories, *the International Multiconference of Engineers and Computer Scientists 2015*, pp.991–994, 2015.

[5] T. Kato, *Perturbation Theory for Linear Operators, 2nd ed.*, Springer, 1980.

[6] M. Ida, T. Nozawa, F. Yoshikane, K. Miyazaki, and H. Kita, Development of Syllabus Database and its Application to Comparative Analysis of Curricula among Majors in Undergraduate Education, *Research on Academic Degrees and University Evaluation*, 2, pp.85-97, 2005.