岡本安晴「カテゴリカルデータの単純２段主成分分析：カテゴリ型と連続型の同時主成分分析」
日本行動計量学会第 41 回大会抄録集、2013、192－195.（原稿）

カテゴリカルデータの単純２段主成分分析

カテゴリ型と連続型の同時主成分分析

○岡本安晴
日本女子大学人間社会学部

1　はじめに

質問項目がカテゴリカル変量である場合は多くみられるが、その分析法として村上（1999）は非計量的主成分分析を提案している。この分析法では、k 番目（$\mathrm{k}=1,\cdots,\mathrm{m}$）のカテゴリカル変量に対するデータを指標行列$G_k$、$G_k$に対する重み（数量化）行列を$V_k$、$G_kV_k$の因子モデルにおける得点行列をF、負荷行列を$A_k$で表すとき、次式

$$\sum_{k=1}^{m}\|G_kV_k-FA'_k\|^2 \qquad (1)$$

を制約条件

$$V'_kD_kV_k=I_{qk}、\quad F'\mathbf{1}_n=\mathbf{0}_r、\ n^{-1}F'F=I_r$$

のもとで最小化する解が求められる。しかし、（１）の最小化では、G_kV_kがG_kの情報を最大限に反映しているという保証はない。また、G_kを数量化するときの次元数、すなわちV_kの列数q_kは、分析者があらかじめ指定するもので、必ずしも最大数p_k-1（p_kはG_kの列数）が設定されるとは限らない。ここでは、G_kの情報を最大限反映した数量化に基づいた主成分分析を行うため、次の単純２段主成分分析を考える。

2　単純２段主成分分析

各カテゴリカル変量G_k（$\mathrm{n}\times p_k$）の数量化を行い（第１段）、その後、数量化されたカテゴリカルデータと他の数量データがあればそれを加えて主成分分析を行う。

（１）カテゴリカルデータの数量化

数量化のための重みω_kを$G_k\omega_k$の分散が最大（極大）になるように選ぶ。

$$\mathbf{1}'_nG_k=\boldsymbol{f}'_k=\begin{pmatrix}f_1^{(k)} & \cdots & f_{pk}^{(k)}\end{pmatrix}、\quad H_k=G_k-\frac{1}{n}\mathbf{1}_n\boldsymbol{f}'_k$$

とおき、$\omega'_k\omega_k=1$の制約のもとで$\mathrm{tr}\big((H_k\omega_k)'(H_k\omega_k)\big)$を最大にする$\omega_k$を求めるため、目的関数を

$$\begin{aligned}\mathrm{Q}(v_k)&=tr\big((H_kv_k)'(H_kv_k)\big)-\lambda(v'_kv_k-1)\\&=tr(v'_kH'_kH_kv_k)-\lambda(v'_kv_k-1)\end{aligned}$$

とおく。

$$\frac{\partial}{\partial v_k}\mathrm{Q}(v_k) = 2{H'}_k H_k v_k - 2\lambda v_k$$

なので、求めるω_kは（２）式を満たす。

$${H'}_k H_k \omega_k = \lambda \omega_k \qquad (2)$$

（２）式に対して、例えば　$\lambda = 0$、　$\omega_k = \frac{1}{\sqrt{p_k}}\mathbf{1}_{pk}$　は１組の解である。

（２）式を満たす２組の解、λ_iとω_i、およびλ_jとω_j、に対して

$$(H_k\omega_i)'\left(H_k\omega_j\right) = \lambda_j {\omega'}_i \omega_j = 0$$

が成り立つ。したがって、

$$\mathrm{Cov}\left(H_k\omega_i, H_k\omega_j\right) = 0$$

である。また、

$$(H_k\omega_i)'(H_k\omega_i) = \lambda_j {\omega'}_i \omega_i = \lambda_j$$

より、

$$Var(H_k\omega_i) = \frac{1}{n}\lambda_i$$

となる。これより、標準化得点

$$z_i = H_k\omega_i/\sqrt{\lambda_i/n}$$

を得る。

$$G_k^z = (z_1 \quad \cdots \quad z_{pk-1})$$

とおく。

（２）主成分分析

G_kすなわちH_kの数量化による標準得点G_k^zに対して

$$G^z = (G_1^z \quad \cdots \quad G_m^z)$$

とおく。

（２a）データがカテゴリカル変量G_kのみからなる場合

G^zの主成分分析を行う。

（２b）データがカテゴリカル変量G_kと数量データXとからなる場合

$(G^z \quad X)$に対して主成分分析を行う。

3　分析例

10 人の作曲家についての属性、生国（a:ドイツ・オーストリア、b:フランス、c:ロシア）、作品の性格（a:絶対音楽、b:中間的、c:表題・劇音楽）、時期（a:19 世紀前半、b:19 世紀後半、c:20 世紀前半）、による評価データ（表１；村上、1999、Table 1 から転載）を分析してみる。各カテゴリカル変量（属性）に対する固有値λ_iと固有ベクトルω_iを求めると表２のようになった。いずれの変量も３カテゴリであるので、0 でない固有値の数は 2 個である。各固有値と固有ベクトルから算出される標準化得点z_iの名前を表２の対応する固有値と固有ベクトルの欄の下に示した。

岡本安晴「カテゴリカルデータの単純２段主成分分析：カテゴリ型と連続型の同時主成分分析」
日本行動計量学会第 41 回大会抄録集、2013、192－195.（原稿）

表１　10 人の作曲家の評価（村上（1999）より）

作曲家	生国	作品の性格	時期
Beethoven	a	a	a
⋮	⋮	⋮	⋮
Stravinsky	c	b	c

これらの数量化変数の標準化得点から構成されるデータ$G^z = (G_1^z \quad \cdots \quad G_m^z)$に対して主成分分析を行ったときの特異値を表３に示す。２次元の主成分分析を行ったときの各主成分と数量化変数との相関係数を座標値として数量化変数の布置を

表２　各カテゴリカル変量の固有値と固有ベクトル

	生国		作品		時期	
固有値	3.6	3	4	2.4	3.6	3
固有ベクトル	0.816	0.000	0.000	0.816	-0.408	0.707
	-0.408	0.707	0.707	-0.408	0.816	0.000
	-0.408	-0.707	-0.707	-0.408	-0.408	-0.707
数量化変数名	Country-0	Country-1	Charactr-0	Charactr-1	Period-0	Period-1

表３　主成分分析における特異値

1.393	1.252	1.057	0.895	0.614	0.444

表したものを図１に示す。第１主成分（Comp.1）は、数量化変数 Country-0、Charactr-1 と正の相関があり、period-0 と負の相関があることがわかる。第２主成分（Comp.2）は、数量化変数 Charactr-0 と正の相関があり、Country-1、period-1 と負の相関がある。すなわち、第１主成分は、「ドイツ・オーストリア -> フランス、ロシア」および「絶対音楽 -> 中間、表題・劇」と正の相関があり、「19ｃ後半 -> 19ｃ前半、20ｃ前半」と負の相関がある。第２主成分は、「中間的 -> 絶対音楽 -> 表題・劇」と正の相関があり、「フランス -> ドイツ・オーストリア -> ロシア」および「19ｃ前半 -> 19ｃ後半 -> 20ｃ前半」と負の相関がある。　各作曲家の主成分値を座標として布置を表したものが図２である。作曲家名の先頭の英字は評定値を表す。例えば、「acb：Wagner」は、「生国」、「作品」、「時期」の評価がそれぞれ a、c、b であることを表している。小円が作曲家の位置であり、長方形の枠内に主成分と対応する数量化変数が示されている。図１の表わす主成分に対応して作曲家がグループ分けされていることがわかる。

主成分分析は、データの散布を低次元空間に正射影するものと考えられる（Okamoto, 2005）。このとき、主成分は特殊な座標系に対する座標値である。図１および図２では、２つの主成分と数量化変数との対応は明快であるが、データによっては正射影された空間において解釈の容易な座標系を探すことも必要になると考えられる。すなわち、射影空間における回転である。回転の一般式は例えば Okamoto (2005)で取り上げられているが、具体的な回転方法は因子分析法で開発された種々の回転を用いることができる。

引用文献

村上隆（1999）．科学研究費補助金（09610114）研究成果報告書
Okamoto, Y. (2005). 日本女子大学人間社会学部紀要 (http://ci.nii.ac.jp/naid/110006223025)

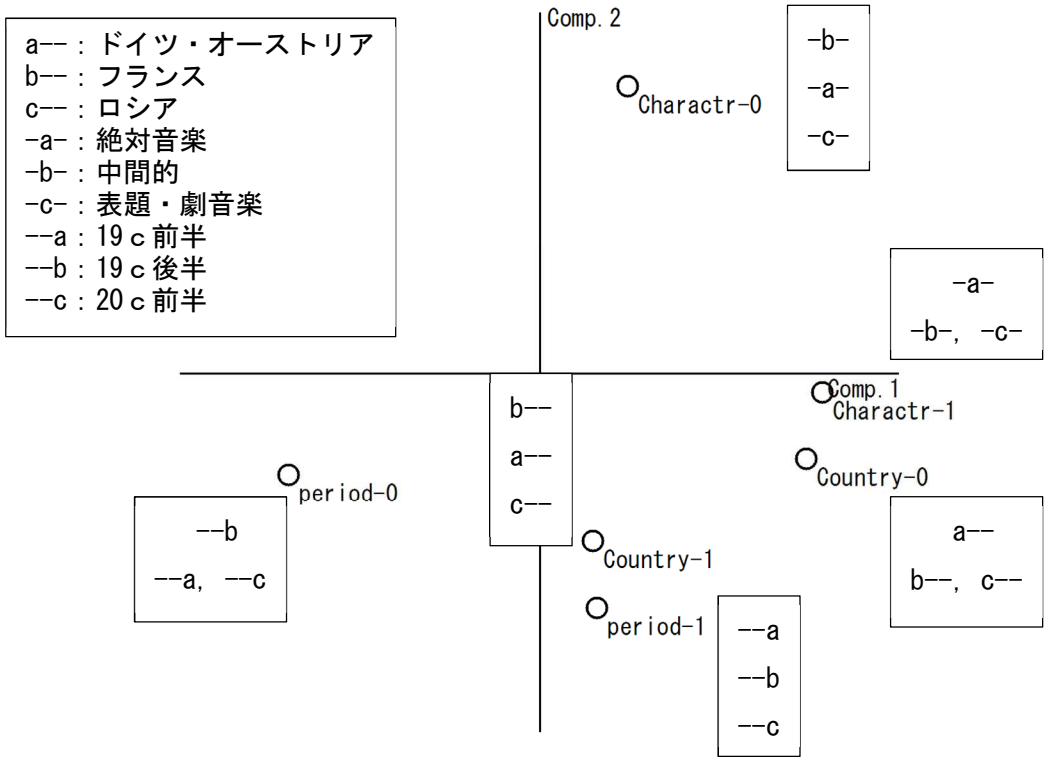

図１　数量化変数と主成分との相関係数を座標値とする布置

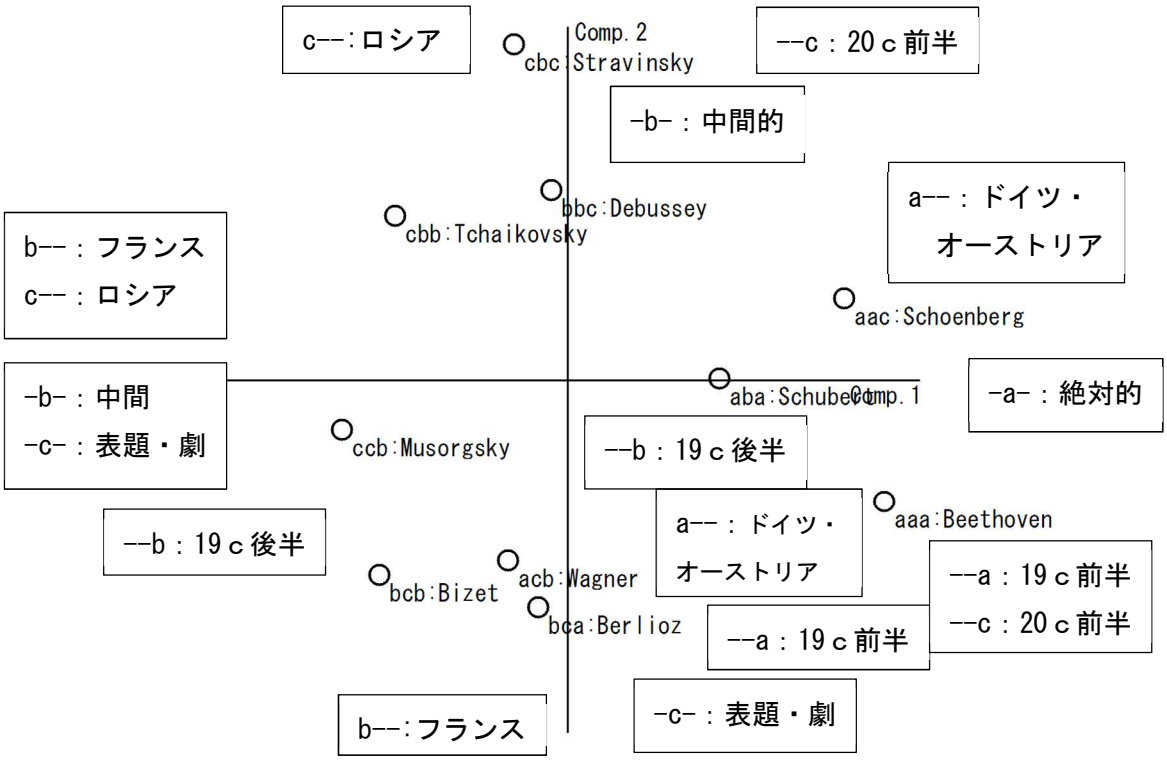

図２　主成分値を座標とする作曲家の布置