対数線形モデルと多重対応分析による縦断的カテゴリカルデータ解析
− 女性のライフコース分析 −

M2006MM009 石河 敦子
指導教員　田中 豊

1　はじめに

本研究がとりあげる統計的手法は，おもに対応分析（Correspondence Analysis, CA）と多重対応分析（Multiple Correspondence Analysis, MCA）である．また一般化対応分析をとりいれ，これに関連付けて対数線形モデル（ Log-linear Analysis, LLA ）も扱う．これら統計的方法を社会科学の一分野，ライフコース分析のひとつの分析方法とし，女性のライフコースを分析対象として解析することで一連の分析手法としてライフコース分析に有効であるかを論じたい．CA によるデータの視覚化では分割表の行と列の同時布置が可能である．基本的な CA は 2 元分割表を行列とし，行，列ベクトルを次元を縮約したひとつの空間上に布置するという考え方だが，多変量に拡張できる．それが，CA の技法をダミー変数に適用し，個体とカテゴリー同時布置を可能にした MCA である．まずこれらの方法をブロードマトリックスやロングマトリックスに変換した職歴データの解析に用いる．次に一般化 CA をおこなう．基本的な CA は分割表独立モデルからの乖離をプロットしていると解釈できるが，CA をより一般化することで，LLA で得られるよりよいモデルからの乖離もプロット可能であるとする van der Heijden (1987) らの議論を基礎とする．

ところで，本研究では CA も LLA も記述統計学的分析の目的で使用する．しかし，記述的な分析結果が安定性をもつかにも興味がある．そこで結果の安定性を検討するべく，ポアソン分布や積多項分布などの分割表の分布で通常仮定される分布の仮定を行えないようなデータにも有効なブートストラップ法を採用する．

先行研究として，van der Heijden (1987)，Martens (1994) らが提案した分析方法をふまえる．

以上の統計的手法を『1995 年 社会階層と社会移動（SSM）調査』の職歴データ解析に適用し，解釈を行う．

2　分析方法

2.1　対応分析（CA）の基本的な理論と方法

CA はカテゴリカルデータの一分析法であり，分割表の行と列をふたつのデータのグループとみなし同時布置する手法として知られる (Benzécri, 1992), (大隈, 1994).

基本的な CA では，分割表の行と列それぞれの相対度数の組をプロフィールと呼び，プロフィールの類似性を反映させて空間に布置する．同時布置の座標を求めるには，分割表が次式のように表せると考えればよい．

$$P = \mathbf{r}\mathbf{c}^T + D_r^{\frac{1}{2}} U \Lambda V^T D_c^{\frac{1}{2}} \tag{1}$$

P が観測値の相対度数で与えられる 2 元分割表，$\mathbf{r}$ は平均行プロフィール，$\mathbf{c}$ は平均列プロフィール，$\mathbf{r}\mathbf{c}^T$ が独立モデルの期待相対度数である．$\mathbf{r}$，$\mathbf{c}$ を対角とする対角行列をそれぞれ D_r，D_c と表す．

式 1 から $\Phi = D_r^{-\frac{1}{2}} U$，$\Gamma = D_c^{-\frac{1}{2}} V$ とし，$\Phi D_r \Phi^T = I$，$\Gamma D_c \Gamma^T = I$ を満たす次のような一般化特異値分解にもとづく再生方程式が得られる．

$$D_r^{-1}(P - \mathbf{r}\mathbf{c}^T) D_c^{-1} = \Phi \Lambda \Gamma^T \tag{2}$$

式 2 の左辺は観測度数対期待度数の比率 $\frac{p_{ij}}{r_i c_j}$ を中心化した形 $D_r^{-1} P D_c^{-1} - \mathbf{1}^T\mathbf{1}$ となっており，Φ，Γ がそれぞれ行主座標，列主座標である．また行，列標準座標はそれぞれに特異値をかけた $\Phi\Lambda$, $\Gamma\Lambda$ である (Greenacre, 2007).

CA は，このプロフィール間の距離を表すという考え方である．CA において，距離はユークリッド距離ではなく，カイ 2 乗距離を使う．

2.2　多重対応分析 (MCA) とブロードマトリックス

イベントヒストリーデータは個体（個人），状態（職業），時点（年齢) からなる 3 元表で表すことができる．連続するイベントを 3 相のインディケーターマトリックス，(個体) × (状態) × (時点)，で表し，時点は離散時間とする．この行列を 0 か 1 の 2 値データが埋める．例えば，要素 z_{ijt} を持つ 3 相の行列を $\mathbf{Z}$ で表し，個体 i は時点 t において状態 j にあれば，$z_{ijt} = 1$，そうでなければ 0 とする．

MCA には上記の多元表をスーパーインディケーターマトリックスに変換した横長の行列，ブロードマトリックス (BROAD matrix) を使う．これは通常の MCA で用いられるインディケーターマトリックスの形になっており，時点と状態を組み合わせて列データとする．こうしてできた 2 相データ (個体) × (時点と状態の組み合わせ) を $\mathbf{Z}_{i[tj]}$ と表記することとする．ブロードマトリックス $\mathbf{Z}_{i[tj]}$ を MCA にかけることで行プロフィール，列プロフィールのばらつきであるイナーシャを求める．

2.3　対応分析 (CA) とロングマトリックス

イベントヒストリーデータ解析で通常の CA に適用できる分割表のうち本研究で採用したのはロングマトリックス (LONG matrix) である．イベントヒストリーデータを表す 3 元表の周辺度数から作成した分割表を個体の属性に基づきグループ分けし，各グループで周辺度数か

ら作成した (時点) × (状態) の部分分割表を作り行について連結させていく縦長の行列である．各部分分割表の行は年齢（時点），列は職業（状態）となっている．この 2 相データは (個体グループと時点の組み合わせ) × (状態) であり，個体グループ g は時点 t において状態 j にある人の数が，分割表の各要素 z_{gjt} である．行列としては個体グループと時点を行として連結させるので，$\mathbf{Z}_{[gt]j}$ と表記することとする．

2.4 対数線形モデル (LLA) と一般化対応分析 (CA)

ロングマトリックスを多元表とし，LLA による検定によりモデルをあらかじめ想定する方法を検討する．前節で説明したロングマトリックスはカテゴリー × 属性 × 時間（年齢) という 3 元表にもなる．CA を分割表の独立モデルからの乖離とみるならば，交互作用が CA に反映されると解釈できる．この性質は 2 元表であれば 2 つの元の交互作用のみが反映されるのでのぞましいが多元表では，すでにわかっている交互作用の影響は最小限におさえたい．そこで提案されているのが，van der Heijden (1987, 1989) らによる CA の一般化である．LLA であてはまりのよいモデルを選び，そのモデルとの違い (residuals) を分解 (decomposition) する．一般化 CA は次のように定式化できる (van der Heijden, 1989).

$$P = Q + S_r U \Lambda V^T S_c \tag{3}$$

式 3 の Q は通常の CA では独立モデルの期待度数による割合が行列の形で入る．また対角行列 S_r，S_c には通常の CA 同様，P の平均行列プロフィールの対角行列 $D_r^{\frac{1}{2}}$，$D_c^{\frac{1}{2}}$ を用いることとする．そうすることで，独立モデルからの乖離を表す観測値と期待値の間のカイ 2 乗距離ではなく，LLA で選択したあるモデルからの乖離を同じ尺度で表すことができる．通常 CA のイナーシャは

$$trace\left[D_r^{-1}(P - \mathbf{r}\mathbf{c}^T)D_c^{-1}(P - \mathbf{r}\mathbf{c}^T)^T\right] \tag{4}$$

と表記でき，これはカイ 2 乗統計量の 1/n 倍になっているが一般化 CA では

$$trace\left[D_r^{-1}(P - Q)D_c^{-1}(P - Q)^T\right] \tag{5}$$

である．すなわち，イナーシャによって同じスケール上で一方が独立モデルの残差平方，他方が LLA で選択したモデルの残差平方を表す．

3 データコーディング，解析方法，データの概要

SSM の調査結果から回答者 2653 人中，職歴のない人，現在学生，ある職についていた年齢が「わからない」「あてはまらない」と答えた人を除く 2383 人，1994 年 12 月 31 日時点で満 20 歳〜69 歳の有権者（男性 1115 人，女性 1268 人）について分析する．まずはブロードマトリックスによる分析を行うため，上記のデータを，職歴データとして加工しなおす．例えば**表 1** は，職歴記入用紙 3 枚に書き込みのあったある女性の職歴データである．なお，従業上の地位の短縮形は MGR=「経営者、役員」, RWR=「常時雇用されている一般従業者」, PTT=「臨時雇用・パート・アルバイト」, TMP=「派遣社員」, SLF=「自営業主、自由業者」, FML=「家族従業者」, DOM=「内職」, STU=「学生」, UNE=「無職」, MLT=「兵役」, NEX=「未就業」, FUT=「未到達」, DKNA=「わからない，あてはまらない」とした．職歴データの後ろに性別，本人の学歴などの追加要素を追加変数として加える．

表 1: 職歴データ

id	...	age21	age22	age23	...
1563F	...	NEX	RWR	RWR	...
...	age31	age32	age33	...	...
...	UNE	PTT	PTT	...	...
...	age69	age70	q1_1	q10_1	...
...	FUT	FUT	FEMALE	NWUN	...

FEMALE=女性，NWUN=新制大学
q1_1 =性別，q10_1 =本人最終学歴

4 分析結果

4.1 多重対応分析 (MCA) によるブロードマトリックスの分析結果

職歴データをブロードマトリックスに変換し，MCA を用いて解析する方法では追加要素を布置するより，むしろ追加要素の分類ごとにオブジェクトを布置したほうが特徴があらわれるようである．たとえばオブジェクトポイントをプロットする際に男女に分けたため男女の違いが読み取れた．性別カテゴリーをブロードマトリックスの追加要素として追加的に布置することも試みたがカテゴリーの平均プロフィールとしては，男性，女性どちらもその他のオブジェクトの集まる原点付近にプロットされてしまい，この結果からの示唆される情報は少ない．

4.2 職業カテゴリーの変遷

カテゴリーポイントを各次元について時系列にならべ職業カテゴリーの変遷としてグラフ化すると，特徴がとりだしやすい．前節の MCA の結果を形態を変え，年齢を横軸にとった職業カテゴリーの変遷では軸ごとに特徴がよく現れることがわかった．しかし，ここでも実際のデータが意味するところを解釈するのは困難である．

4.3 対応分析 (CA) によるロングマトリックスの分析結果

イベントヒストリーデータを表す多元表をカテゴリーごとの表にし，2.3 節で説明したように周辺度数の分割表を作る．職業を行，年齢を列にとり，さらに属性（性別，性別役割意識など）によって複数のグループに分類したものを行にそって連結したロングマトリックスとする．このロングマトリックスの行は（年齢）×（属性），列は（職業）となる．これらのロングマトリックスを CA にかけた．

結果のイナーシャは性別の分類で第 1 次元から第 4 次

元まででて 0.318, 0.152, 0.041, 0.012, それぞれ説明されるイナーシャの割合は 60.3%, 28.8%, 7.7%, 2.3%である．性別役割意識の分類では，第 1 次元から第 4 次元まででて 0.230, 0.057, 0.020, 0.006, それぞれ説明されるイナーシャの割合は 73.6%, 18.1%, 6.3%, 2.0%である．行標準座標，列主座標からなる非対称マップを**図 1** に示す．

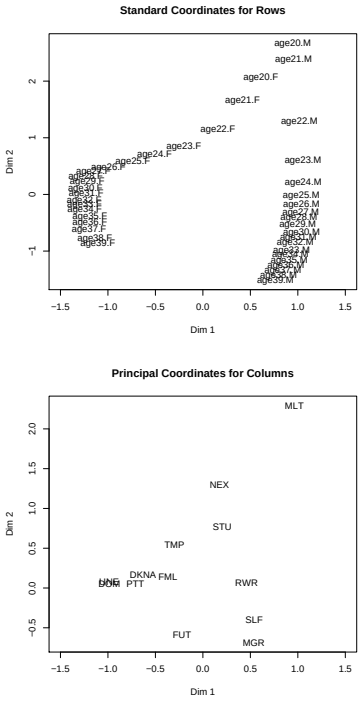

図 1: 全データ (n=2383)，20 歳から 39 歳までの職歴ロングマトリックスの非対称 CA プロット: 行は年齢と性別，年齢の後の M,F はそれぞれ男性，女性（上），列は職種（下）

4.4 対数線形モデル (LLA) による残差の分解と一般化対応分析（CA）

LLA の解釈がある程度可能であるときに，あてはまりのよい LLA からの乖離を見るために一般化 CA を用いる．

前節の個体のグループ分けから得たロングマトリックスについてイナーシャとプロットをあわせて眺めると，第 1 軸で説明されるイナーシャの割合が高い上に，年齢の変化の影響が強すぎる．このような影響があらかじめ LLA であてはまりのよいモデルとして読み取れるのであれば，CA はそのモデルからの乖離をながめるツールになりうる．モデルの適合度を調べるため，まず LLA であてはめ，あてはまりのよいモデル $[AJ][A_LC][CJ]$（A: 年齢，C: 性別役割意識カテゴリー，J: 仕事，L: 順序スコア）を用いて一般化 CA を行った．

一般化 CA にかけて求めたイナーシャの値は非常に低く，第 1 次元から第 5 次元まででて 0.0086, 0.0063, 0.0044, 0.0029, 0.0016, それぞれ説明されるイナーシャの割合は 36.1%, 26.7%, 18.6%, 12.0%, 6.6%である．最初の 2 軸で約 63% 説明される．

5 ブートストラップによる安定性の評価とその結果

CA の結果はカテゴリカルデータの分割表を幾何的に表示したものであり，オブジェクトとカテゴリーの関係を示唆的に表示するものの，それ自体から統計的推測が行えるわけではない．こうしたデータからモデルの推測，検定を行うには，LLA など統計的推測理論にかなった手法でモデルを構築し検定するという方法がある．しかしながら，イベントヒストリーデータは，その性質上，となりあう事象の独立性が保障されないのでポアソン分布や積多項分布などの分割表の分析で通常仮定される分布が仮定できず，分布を仮定するような推測を行うには注意を要する．CA ではそのような分布を仮定して厳密な統計理論を導く代わりに，結果の安定性を確かめる方法として，母集団の分布を仮定しなくとも行えるブートストラップ推定，検定を Greenacre (1984, 2007) が採用している．

5.1 ブロードマトリックスの MCA のイナーシャと列標準座標の推測

MCA の分析結果で得られるイナーシャと列標準座標の安定性について調べる．具体的な計算の流れは以下のとおりである．

1. 標本の多重対応分析からイナーシャと主成分分析ではスコアに相当する列座標を得る．
2. 主要な軸の数を決める．
3. ブートストラップ推定の回数を設定．（B=1000)
4. 標本の大きさ n に等しい n 個の行番号を復元抽出．
5. 復元抽出で得た行番号からなるクロス表，ここでは再標本で得られた行からなる職歴データを作成．ブートストラップクロス表とする．（行プロフィールの再標本）
6. 4 から 5 を B 回繰り返し B 個のブートストラップクロス表を得る．
7. ブートストラップクロス表をブロードマトリックスに変換し，多重対応分析によりイナーシャと説明されるイナーシャの割合を得る．列座標については軸の向きが不定であるため，もとの標本の列座標ベクトルとの内積を求め，正であればそのまま，負であれば-1 をかけて列座標とする．これらの値をブートストラップ推定値として格納．
8. ブートストラップ推定値からヒストグラムを作成．

5.2 ロングマトリックスの CA の列主座標の安定性

さらに前節でブロードマトリックスの MCA からのイナーシャの推測を行った際に得た同じブートストラップクロス表の周辺からロングマトリックスを作成し，ロン

グマトリックスの CA の列主座標の安定性を検討する．ここでは部分ブートストラップ法 (partial bootstrap) を用いる．計算の流れは以下のとおりである．

1. イナーシャと列標準座標のブートストラップ推定の方法と同様にリサンプリングにより B 個のブートストラップクロス表を得る．これは前節の方法で得たものを使う．それぞれ，周辺度数からロングマトリックスを作成．
2. 列プロフィールを追加要素とする．追加要素の射影のプロセスは
   (a) もとの分割表の行の標準座標を得る．
   (b) 推移方程式に行の標準座標と追加する列の標準座標を代入して列主座標を得る．
3. 射影された列プロフィールである追加要素を，もとの対応分析のプロット上に布置．
4. カテゴリーごとに布置された B 個の点の凸閉包を描く．

結果を**図 2** に示す．概ね安定した分布である．「派遣」(TMP) など，周辺度数の低い列主座標に対応する凸閉包は大きく，凸閉包の中心とも離れる傾向にある．

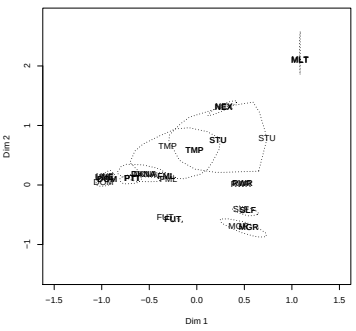

図 2: ロングマトリックスの CA の列主座標のブートストラップ分布: 太字は凸閉包の中心，細字はもとのロングマトリックスの列主座標

6 考察および結論

CA と MCA を中心とした統計的手法を使い女性のライフコースを分析し，おもにその分析手法の有効性を検証してきた．それぞれの方法に一長一短がある．MCA では個人のライフコースをブロードマトリックスというデータの形に収めることができ，直接そのデータを扱えるという利点があるにもかかわらず，結果の解釈が難しいという欠点がある．ロングマトリックスの CA から個体グループごとにキャリアの変遷に違いが明らかであったように，ブロードマトリックスを用いた MCA と比べると CA の方が結果の解釈が行い易い．また LLA を補完的に用いて一般化 CA をとりいれることで，コホートのライフコースとして解釈のしやすい結果が得られた．

職歴データをロングマトリックスに変換し CA にかける方法では結果を個体グループごとに別々に布置することでグループ間の違いが見られた．この違いがでるかぎりにおいては結果の解釈がしやすい．しかし，ロングマトリックスはブロードマトリックスの周辺度数をとったものであり，個体ごとの変遷ではない．ライフコース分析という観点からみると個人のライフコースではなく，コホートごとのおおざっぱなライフコースを見ていることになる．よって，まとまったライフコースの特徴を出すためにはグループの分け方に工夫が必要である．

個体グループ間に違いを見出すといってもロングマトリックスの CA だけでは微妙な現れ方しかしないような場合がある．とりあげた例でもすべて第 1 軸で年齢に従った変化が目立って現れてしまった．年齢による変遷は，この例の場合はとくに自明なので，あらかじめこの影響を取り除くための LLA が有効となった．LLA を補完的に用いて一般化 CA を行うことで結果の解釈がしやすくなった．

ブートストラップ法で行った結果の推測では．MCA と CA，どちらのプロットの結果も主要な軸においては安定していることが確かめられた．

謝辞

研究をすすめるにあたり，熱心にご指導いただきました指導教授，田中豊先生に感謝いたします．

〔二次分析〕に当たり、東京大学社会科学研究所附属日本社会研究情報センター SSJ データアーカイブから「1995 年 SSM 調査」(1995 年 SSM 調査研究会　代表盛山 和夫）の個票データの提供を受けました．

参考文献

[1] Benzécri, J. -P. (1992) *Correspondence analysis handbook*, Marcel Dekker.

[2] Greenacre, Michael J. (1984) *Theory and Applications of Correspondence Analysis*, Academic Press, London.

[3] Greenacre, Michael J. (2007) *Correspondence Analysis in Practice*, Chapman & Hall/CRC, Boca Raton, second edition.

[4] M. J. Greenacre and J. Blasius (eds.) (1994) *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, Academic Press, London.

[5] 大隈昇, ルバール, L. ほか (1994) 記述的多変量解析法, 日科技連.

[6] van der Heijden, Peter G. M. (1987) *Correspondence Analysis of Longitudinal Categorical Data*, DSWO Press, Leiden.

[7] van der Heijden, Peter G. M. and de Leeuw, J. (1989) Correspondence Analysis, with Special Attention to the Analysis of Panel Data and Event History Data, *Sociological Methodology*, 19, 43-87.