2002年日本オペレーションズ・リサーチ学会
春季研究発表会

L_p計量ノルムを用いた正準相関分析

現在申請中　日本大学　＊　畑澤　文祐　HATAZAWA Fumihiro
01205220　日本大学　　篠原　正明　SHINOHARA Masaaki

1　はじめに

私たちが意思決定をする時、多変量データ解析 (multivariate data analysis) が多く使われる。その代表的な手法に正準相関分析 (canonical correlation analysis:CCA) がある。基準変数群と説明変数群の相関を最大にするように、相互の変数群の全ての変数に重み付けをするのが正準相関分析の目的である。本研究では、L_p 計量ノルムを導入した正準相関分析を提案する。なお、本実験計算には「EXCEL 多変量解析」(エスミ)、「NUOPT」(数理システム) を用いた。

2　距離関数と L_2-CCA

2.1　距離関数

$a, b \in \mathbb{R}^n$ とする。このとき 2 点 a, b 間の距離を

$$d^{(p)}(a,b) = \left(\sum_{i=1}^{n}|a_i-b_i|^p\right)^{\frac{1}{p}} \quad (1)$$
$$(1 \leqq p < \infty)$$

で定める。これを p 乗距離、または L_p 計量ノルムと言う。

2.2　L_2-CCA

x_i:　標本 i に対する説明変数ベクトル値 $(m \times 1)$

y_i:　標本 i に対する被説明変数ベクトル値 $(s \times 1)$

X:　入力データ行列 $(m \times n)$
$X = (x_1, x_2, \cdots, x_n)$

Y:　出力データ行列 $(s \times n)$
$Y = (y_1, y_2, \cdots, y_n)$

u:　入力変数に対する評価ベクトル値 $(m \times 1)$

v:　出力変数に対する評価ベクトル値 $(s \times 1)$

従来型 L_2-CCA は以下のように定義できる。

L_2-CCA

目的関数　：　$\|v^TY - u^TX\|^2 \to \min$

制約式　：　$v^TYY^Tv = u^TXX^Tu = 1$

3　L_p-CCA

L_2-CCA で使われている $\|v^TY-u^TX\|^2$ は $p=2$ の残差二乗和最小化であるが、以下に本研究にて提案する $p=1$"絶対誤差和最小"CCA と $p=\infty$"上限誤差最小"CCA を定義する。

3.1　L_1-CCA

"絶対誤差和最小"CCA は、
$\sum_{i=1}^{n}|v^Ty_i - u^Tx_i| \to \min$ に、z_{i1} と z_{i2} に導入し、$v^TY^TYv = u^TX^TXu = 1$ は擬似的に $\sum_{i=1}^{n}u_i = \sum_{j=1}^{m}v_j = 1$ と置き換えることにより、以下のように定式化して実現する。

L_1-CCA

目的関数　：　$\sum_{i=1}^{n}\sum_{j=1}^{2}z_{ij} \to \min$

制約式　：
$$\begin{cases} v^Ty_i - u^Tx_i = z_{i1} - z_{i2} \\ z_{ij} \geqq 0 \\ \sum_{i=1}^{n}u_i = \sum_{j=1}^{m}v_j = 1 \end{cases}$$

3.2　$L_\infty-CCA$

"上限誤差最小"CCA は、すべての i において、ある 1 つの z という変数で、

$$|v^Ty_i - u^Tx_i| \leqq z$$

とし、その z を最小化する。以下に定式化すると、

$$L_\infty - CCA$$

目的関数 : $z \to \min$

制約式 :
$$\begin{cases} -z \leqq v^T y_i - u^T x_i \leqq z \\ \text{(for all i)} \\ z \geqq 0 \\ \sum_{i=1}^{n} u_i = \sum_{j=1}^{m} v_j = 1 \end{cases}$$

4 例題

表 1: 入学前の成績と入学後の成績

	入学前の成績		入学後の成績	
学生 No.	数学	英語	数学	英語
1	100	90	80	90
2	20	30	20	30
3	60	90	50	80
4	70	70	30	50
5	70	50	30	20
6	60	60	60	70
7	40	40	30	80
8	70	60	60	50
9	90	100	80	90
10	60	50	30	60

以上の問題を、データを正規化して正準相関分析をした結果を示す。

表 2: 正規化データ解析結果

	$p=2$	$p=1$	$p=\infty$
u_1	0.373885	0.494667	-0.138459
u_2	0.672741	0.505333	1.138459
v_1	0.929762	0.59265	0.572344
v_2	0.097999	0.40735	0.427656
$\sum_{i=1}^{n} \varepsilon_i^2$	3.305531	3.616365	3.814910
$\sum_{i=1}^{n} \lvert\varepsilon_i\rvert$	4.955531	4.413975	5.338241
$\sup\{\lvert\varepsilon_i\rvert\}$	1.091077	1.0349581	0.956911

5 制約付き CCA

制約条件が入る場合、従来の方法では、制約付き非線形計画法となり、計算が非常に大変であった。しかし、L_p 計量ノルム $(p=1, p=\infty)$ による手法の最大のメリットは、制約が線形であれば、制約条件を書き加えるだけで、重み係数をニーズに合わせて自由自在に変化させることができる点である。例題に関して、制約式に $u_2 \geqq 0.7$ の式を加えて解くと、

表 3: $u_1 \geqq 0.7$

	$p=1$	$p=\infty$
u_1	0.7	0.7
u_2	0.3	0.3
v_1	0.870304	0.793556
v_2	0.129696	0.206444

- 複数の最適解パターンが得られる場合
  原データをそのまま $L_1 - CCA$ にかけたところ、最適化手法を変えることにより、目的関数値が同じで最適解が複数個が存在した例があった。なお、比較資料として、$L_2 - CCA$ に $v^T Y^T Y v = u^T X^T X u = 1$ ではなく、$\sum_{i=1}^{n} u_i = \sum_{j=1}^{m} v_j = 1$ を制約式として与えて解いた解も併記しておく。

表 4: $L_1 - CCA$ での、複数の最適解パターンを持つケース

	単体法	内点法	$L_2 - CCA$
u_1	−0.238095	−0.142602	0.110070
u_2	1.238095	1.142602	0.889930
v_1	0.285714	0.285714	0.398126
v_2	0.714286	0.714286	0.601874
目的関数値	25.7143	25.7143	————

6 成果と考察

$L_p - CCA$ は、$p=2$ の場合と、解の大小関係には類似性が見られ、それぞれの条件の下で、最適解を得ることができた。また、重み係数に制約を与えても、線形計画法により、高速で解くことができた。$L_p - CCA$ の提案により、解析者のニーズにあった解析方法を選ぶことが可能になる。今後の課題として、もっと大規模なデータ数での解析、複雑な制約条件の場合について、もっと深く研究を進め、より実用的な手法に発展させて行きたい。

参考文献

[1] 篠原 正明, "CCR モデル— LP 定式化のゲーム論ならびに多変量解析的解釈" RAMP シンポジウム論文集,1997

[2] 木下 栄蔵, "わかりやすい数学モデルによる多変量解析": 近代科学社

[3] 藤沢 偉作, "楽しく学べる多変量解析法": 現代数学社

[4] 柳井 春夫、高木 廣文, "多変量解析ハンドブック":現代数学社