正準相関分析を包絡分析法に適用する研究

M2018SS009 尾﨑 友彦
指導教員：松田 眞一

1　はじめに

2 群の多次元でのデータ構造に対する分析として，包絡分析法 (DEA = Data Envelopment Analysis) と正準相関分析 (CCA = Canonical Correlation Analysis) がある．これらは同じようなデータに適用する方法であるが，分野が異なるため，両者を併用する研究は少ない．両者を併用して変数選択を行った先行研究として，上田 [7] と，上田の修正案を提案した信田 [4] がある．信田の論文では包絡分析法を R でプログラムを作成し，提案した変数選択手法を 2 つの事例に適用しているが，一般的に組み合わせて使用する場合が吟味されておらず，また，シミュレーションが行われていない．本論文では，正準相関分析を適用して変数選択した後に包絡分析法を行う方法論の確立と，シミュレーションを行うことができる汎用的な分析モデルを作成することを目的とする．

2　包絡分析法とは

包絡分析法とは，公共事業から民間企業におよぶ，さまざまな事業体の効率性を評価する OR 手法の 1 つであり，投入した入力 (資源) に対しての産出した利益 (出力) を表す比率尺度を参照して，事業体の効率性を相対評価する手法である．原則として，基本的には事業体の比率尺度が最大となるように考える．このため，より少ない入力で，できるだけ大きな出力を得ることが効率的になる．この方法によって，対象とする集団の中から優れた事業体の集団 (効率的フロンティア) の存在を明示することができ，それらの事業体を基準として，非効率な事業体の改善点を具体的に示すことができる．具体的な計算手法としては．多入力，多出力系のシステムの効率性を公平に相対評価するために線形計画法を用いる．また，想定される投入対産出の関係を示す生産関数の形に応じて，幾つかのモデルが存在する．(刀根 [6] 参照)

3　正準相関分析とは

正準相関分析とは，2 組の多次元変量の相関関係を調べる統計解析手法の 1 つである．変数の集合 $\boldsymbol{x}$ と変数の集合 $\boldsymbol{y}$ が与えられた場合に，$\boldsymbol{x}$ 変数と $\boldsymbol{y}$ 変数それぞれについての 2 つの線形結合の中で，最大の相関を持つものを見つけ出す手法である．(ニャナデシカン [5] 参照)

包絡分析法と正準相関分析の違いとして重みの付け方が挙げられる．包絡分析法は入出力の項目に重みを付け，仮想的な入出力に対しては比率尺度が最大となるような係数を考える．一方，正準相関分析は変量群の重み付き合計の相関が最大となるように，重み係数を選択する．包絡分析法は，事業体別に項目ごとに好ましい重み係数を付けることができるが，正準相関分析は全体の相関が最大となるように重み係数を付ける．

また，包絡分析法は基本的に入力が原因，出力が結果となるように変数を設定するが，正準相関分析は入力が原因，出力が結果ではなく，並列でもよい．

4　先行研究と変数選択

信田 [4] では，R 上で包絡分析法を CCR モデルを用いて解くプログラムを自作している．本研究ではこのプログラムを使用して事例研究を行った．

信田の先行研究者である，上田 [7] の変数選択の仕方では，第 1 正準変量の推定係数に非負制約を課して変数選択を行った．ただし，すべての第 1 正準変量の推定係数が正の場合，この方法では変数選択に使用することができない．

信田の変数選択の仕方では，第 1 正準変量で負の推定係数が付いた変数について逆数を取ることで対応する．その後，第 2 正準変量の推定係数について，正の方向と負の方向おいてそれぞれ推定係数の絶対値が最も大きい入力変数を選択する．この結果，変数選択をしない場合の包絡分析法と近い結果になった．

5　事例研究

5.1　事例研究に使用したデータ

今回の研究の事例研究として，食品とガンの関連の調べた吉田 [8] があり，それを参考に Web[1, 2, 3] より，ガンと特に関連の深いと思われる要素のデータを取得した．

事例研究に使用したデータは沖縄県を除く 46 都道府県について，2017 年の都道府県別 75 歳未満年齢調整死亡率 [%] の肺がん死亡率，大腸がん死亡率の 2 つを出力変数，2016 年から 2018 年の平均の二人以上の世帯の 1 世帯当たり品目別年間支出金額及び購入数量 [g] のみそ消費量，食塩消費量，マーガリン消費量，2017 年度都道府県別成人 1 人当たりの酒類消費数量 [ℓ] のブランデー消費量，ビール消費量，単式焼酎消費量，清酒消費量，2016 年度都道府県別 20 歳以上の「毎日吸っている」または「時々吸う日がある」の男性の割合 (以下では男性喫煙率とする)[%] の 8 つを入力変数, とするデータを作成し，信田 [4] のプログラムで分析した．このとき，ブランデーのみ元データに 0.0 が混じっていたが，これでは変数選択の過程で逆数が取れないため全てのブランデーの数値に 0.01 を加算することで逆数を取れるようにした．また，2016 年の男性喫煙率の熊本県のデータのみ，熊本地震の影響で調査データが存在しなかったため，代わりに 2013 年の熊本県のデータを使用した．

5.2　元データに包絡分析法を適用した結果

後述する変数選択に関連する処理を行わずに上記のデータで信田のプログラムを用いて包絡分析法を行ったところ，46 都道府県中 29 の都道府県の D 効率値が 1 となり効率的フロンティアに選ばれてしまった．これは入力変

数が多すぎたため，複数の都道府県で何かしらの要素にウェイトをかけることで効率的になってしまったのだと思われる．これではどの都道府県が優れているのか全くわからないため，元データに一定の処理を行った後に，変数選択をすることで，より優れた都道府県を効率的フロンティアとして絞り込むことを行う．

5.3 信田の手順の再現

事例研究に使用した信田の手順を再現を説明する．まず，変数選択の過程として，変数選択を行うために，データに対して，標準化を行った後，正準相関分析を行う．

次に，正準相関分析で得られた第 1 正準変量の推定係数を見ながら，元データに逆数を取る処理を行う．今回の出力変数である，肺がん死亡率と大腸がん死亡率は，いずれも値が小さい方が望ましいものであるため，値が小さい方が効率的フロンティアになる．そのため今回は出力変数の元データの逆数を取ることを行うことによって，元のがんの死亡率の値が小さくなるほど大きい値を取るようにした．

続いて，出力変数の元データの逆数を取ってから再度正準相関分析を行って得られた第 1 正準変量の推定係数を見て入力変数の逆数を取る．このとき，出力変数である肺がん死亡率と大腸がん死亡率の第 1 正準変量の推定係数は正となっているが，みそ消費量，マーガリン消費量，ブランデー消費量，ビール消費量, 男性喫煙率の第 1 正準変量の推定係数が負となっている．今回の事例では，入力変数はいずれも小さい方が効率的フロンティアになるのかが自明ではないが，入力変数の第 1 正準変量の推定係数の符号を逆転させるために今回は探索的に元データの逆数を取る．上記の 5 つの入力変数の元データの逆数を取ったところ，ブランデーの推定係数が -0.002 となり，負の推定係数となったため，ブランデーに対してもう一度逆数を取る処理を行って元に戻した．このもう一度逆数を取って元に戻す処理は信田の手順にはないものであるが，全ての入力変数の第 1 正準元に戻す変量の推定係数を出力変数と同じ符号にするために行った．この手順により全ての変数の第 1 正準変量の推定係数が正になった．

第 1 正準変量の推定係数を全て同じ符号に揃えたので，第 2 正準変量の推定係数に基づいて変数選択する．今回は出力変数の第 2 正準変量の推定係数が，正の方向と負の方向に分かれているため，入力変数の第 2 正準変量の推定係数から，正の方向と負の方向について，それぞれ絶対値が最も大きい入力変数を 1 つずつ選択する．今回の事例では表 1 の結果から，出力変数の肺がん死亡率に対する入力変数としてみそ，出力変数の大腸がん死亡率に対する入力変数としてマーガリンを選択する．表 1 は，それぞれ選択した変数を太字にしたものである．

最後に，選択された変数を用いて包絡分析法を行うこれらを変数選択後に包絡分析法を行った結果，効率的フロンティアとして，富山県，長野県，滋賀県，京都府，の 4 県が選ばれた．これにより，変数選択なしの包絡分析法では，効率的フロンティアを絞ることができなかったが，変数選択をすることにより，効率的フロンティアを絞ることができた．

表 1 がん死亡率と入出力の推定係数 : 第 2 正準変量の推定係数に基づいて変数選択する

		第 1	第 2
正準相関係数		0.471	0.337
入力に対する推定係数	**みそ**	**0.015**	**0.111**
	食塩	0.051	−0.036
	マーガリン	**0.053**	**−0.070**
	ブランデー	0.002	0.065
	ビール	0.079	−0.031
	単式焼酎	0.020	0.0004
	清酒	0.033	−0.051
	男性喫煙率	0.145	−0.048
出力に対する推定係数	**肺がん死亡率**	**0.093**	**−0.137**
	大腸がん死亡率	**0.084**	**0.142**

6 事例研究の考察

6.1 効率的フロンティアに選ばれた県の特徴

効率的フロンティアに選ばれた県の変数選択で使用した要素の数値と順位を記述すると，表 2 のようになる．

表 2 効率的フロンティアに選ばれた県の特徴 : 各要素の順位

	肺がん死亡率	大腸がん死亡率	みそ消費量	マーガリン消費量
富山県	28	44	4	7
長野県	46	30	1	31
滋賀県	40	41	37	5
京都府	42	21	40	4

6.2 事例研究から新しく発見できた課題

事例研究では変数選択をすることにより，効率的フロンティアを導くことができた．ただし，第 1 正準変量の推定係数が最も大きかった男性喫煙率が，第 2 正準変量の推定係数では，絶対値はみそやマーガリンに劣っている．これは両方の出力変数に大きく影響する変数は，中和されてしまい目立たなくなることが起きている可能性があると考えられる．実際に，みそ，マーガリン，男性喫煙率の 3 つを変数選択した場合，山形県，富山県，山梨県，長野県，滋賀県，京都府，熊本県の 7 県が効率的フロンティアに選ばれた．このような場合において知見を優先するか，機械的に処理すべきなのかは今後考えていく必要があると思われる．

また，みそは本研究では信田の手法に従って行ったところ，発がんに悪影響を及ぼすという結果になったが，吉田 [8] では，みそは発がんを抑えるという研究結果が出ている．両者の結果は一見すると矛盾しているように思えるが，今回の事例での効率的フロンティアの 1 つに，みその摂取量が 1 位であるにも拘わらず肺がん死亡率が 46

位という特異な県である長野県を導出することができており，これはみそが発がんを抑えるという吉田の研究結果に当てはまる．このような結果が得られた原因としては，みその中に地域ごとの種類の違いがあり，そこでみその摂取による大腸がん死亡率に差が出ている可能性が考えられる．

7　シミュレーション用のプログラム

シミュレーションを行う為に，5.3 節で行った手順を自動で行うプログラム DEACCAsim.R を自作した．今回のシミュレーションでは，5 節の事例研究で使用したデータ (以降 cancer.txt と記述する) から乱数データの生成を行った．

7.1　作成したプログラム

作成した DEACCAsim の流れは下記のようになる．

1. 元データの対数を取って，平均と分散共分散行列を計算する
2. mvrnorm 関数を用いて乱数データの生成を行ってから，exp 関数を用いて乱数データを指数変換する
3. 指数変換した乱数データを標準化して正準相関分析を行う
4. 2 つの出力変数の第 1 正準変量の推定係数が同じ向きか異なる向きか判断する
5. 出力変数が小さい方が望ましい場合は，指数変換した乱数データに出力変数の逆数または反転した値を取る
6. 手順 5 の処理を行った乱数データの値を標準化して，正準相関分析を行う
7. 手順 6 の正準相関分析の結果の出力変数の第 1 正準変量の推定係数の符号が同じか異なるかの確認を行う
8. 手順 6 の正準相関分析の結果の全ての入力変数の第 1 正準変量の推定係数の符号が出力変数と同じかどうかの確認を行う
9. 異なる符号だった入力変数については，第 1 正準変量の推定係数の符号を逆転させるために，元データの逆数または反転した値を取る
10. 手順 9 の処理を行った乱数データの値を標準化して，正準相関分析を行う
11. 手順 9 の処理で 1 つでも入力変数の第 1 正準変量の推定係数の符号を逆転させる処理を行った場合は手順 7 の処理に戻る処理を 5 回まで繰り返す
12. 手順 11 の処理を通過した乱数データに変数選択を行わずに包絡分析法を行う
13. 手順 11 の処理を通過した乱数データの入力変数の第 2 正準変量の推定係数に基づいて，正負の方向について絶対値が最大の入力変数をそれぞれ 1 つずつ変数選択する
14. 手順 13 の処理で選択された 2 つの入力変数と出力変数で包絡分析法を行う
15. 実行結果として，処理が手順 14 まで完了できたかどうか，各出力変数について変数選択された入力変数の番号，変数選択された入力変数が逆数や反転の処理を行ったものだったかどうか，効率的フロンティアの数を返す

7.2　プログラムの補足説明

7.2.1　元データの対数を取って乱数データを生成してから指数変換する理由

事前に，cancer.txt の各変数について，元データと対数を取ったデータの歪度と尖度を計算したところ，ほとんどの変数において，対数の方が歪度と尖度の値が 0 に近い，すなわち，正規分布に近いデータだったため対数を取った状態で乱数データを生成した．また，正規分布で乱数データを生成すると負の値を取る可能性があるが，対数正規分布ならば負の値を取ることもないため都合がよいこともある．

7.2.2　逆数，反転の処理の詳細

本研究における，変数の符号を逆転させる際に用いる逆数，反転の処理をそれぞれ記述する．逆数は，対象とする変数の元データの値を a から，$1/a$ に変換する処理を指す．反転は，対象とする変数の元データの値を a から，その変数の元データの 最大値 + 最小値 $- a$ にして大小を反転させる処理を行うことを指す．
これらの処理は出力変数が死亡率や悪性物質の生産量のように少ない方が好ましい場合に行う．入力変数に対しては出力変数との相関関係に応じて行う．
逆数の処理に関しては，刀根 [6] でも触れられており，小さい値が異常に大きく変換される欠点があることが述べられている．また，$a = 0$ の場合，逆数を取ることができない．反転に関しては欠点のようなものは述べられていないが状況に応じて考えることを推奨されている．

7.2.3　手順 4 や 手順 7 で出力変数の第一正準変量の推定係数の符号が同じなのかを確認する理由

今回のプログラムが対象とする 5.3 節で行う処理は出力変数の第 1 正準変量の推定係数の符号が同じであることを前提としており，符号が異なる場合は使用できないためである．

7.3　シミュレーション結果

DEACCAsim.R を繰り返し実行するプログラム Execution.R を作成し，符号の逆転させる処理を，逆数にした場合と反転にした場合について，set.seed で 1 から 100 まで設定して合計 100 回をそれぞれ実行させて，処理が完了した回数，肺がんに対応する効率的フロンティアとして選択され同符号だった回数，肺がんに対応する効率的フロンティアとして選択され異符号だった回数，大腸がんに対応する効率的フロンティアとして選択され同符号だった回数，大腸がんに対応する効率的フロンティアとして選択され異符号だった回数，効率的フロンティアの個数の各パターンの回数を出力するようにした．逆数の場合の結果を表 3，反転の結果を表 4 に示す．
表には尺の都合で記載していないが，反転の場合は効

表 3 逆数の場合

処理が完了した回数							
60							
肺がんに対応する効率的フロンティアとして選択され同符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
2	0	7	1	5	1	1	8
肺がんに対応する効率的フロンティアとして選択され異符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
8	6	0	3	0	8	10	0
大腸がんに対応する効率的フロンティアとして選択され同符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
18	3	0	7	6	4	2	7
大腸がんに対応する効率的フロンティアとして選択され異符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
1	2	5	2	1	2	0	0
効率的フロンティアの個数の各パターンの回数							
1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個
1	4	10	13	14	5	9	4

表 4 反転の場合

処理が完了した回数							
71							
肺がんに対応する効率的フロンティアとして選択され同符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
2	0	12	1	6	0	0	7
肺がんに対応する効率的フロンティアとして選択され異符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
9	5	0	3	0	8	18	0
大腸がんに対応する効率的フロンティアとして選択され同符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
18	5	1	10	8	4	1	7
大腸がんに対応する効率的フロンティアとして選択され異符号だった回数							
みそ	食塩	マーガリン	ブランデー	ビール	単式焼酎	清酒	男性喫煙率
1	3	8	1	1	2	1	0
効率的フロンティアの個数の各パターンの回数							
1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個
2	10	19	11	11	9	5	1

率的フロンティアに 9 個以上選択される場合があり，9 個の場合が 2 回，10 個の場合が 1 回存在した．

8 Leave-One-Out 法でのシミュレーション

46 都道府県から 1 つ取り除いたデータを用いて解析を繰り返す Leave-One-Out 法によるシミュレーションを実行した．この場合は逆数だと全ての場合で処理が完了したが，反転では 1 つ処理が完了しなかった．それでも乱数と比べると安定した結果を導出でき，反転の方が元データの解析に近い結果が多く得られた．

9 シミュレーションの考察

符号の逆転する処理を逆数で行った場合と反転で行った場合でも，みそが大腸がんに効率的フロンティアに選択されて同符号だったことが最も多かった．この手法だとみそを摂取しているにも関わらず大腸がんの死亡率が高くはないものが，効率的フロンティアに選ばれている可能性が高いと考えられる．また，処理が完了した回数が逆数に比べて反転の方が 11 回ほど多い．このことから逆数で符号を変更を行う場合は小さい値が極端に大きくなってしまい，処理が上手く完了できなくなるケースがしばしば起きてしまう可能性が考えられる．

肺がんの効率的フロンティアに選択されたマーガリン，及び，大腸がんの効率的フロンティアに選択されたみそはいずれも同符号で効率的フロンティアに選択されている．このことから，効率的フロンティアに選ばれた都道府県はみそ，または，マーガリンを多量に摂取しているにも関わらず，対応するがんの死亡率が高くない県ということになる．みその場合は死亡率が異なる原因の一つとして地域によってみその種類が異なる場合が考えられる．一方で，マーガリンは地域によってが種類が異なるとは考えにくく，効率的フロンティアに選択された県は，健康に良くないと思われるマーガリンを多量に摂取しているにも関わらず，肺がん死亡率が高くない県ということになるが，なぜ効率的にフロンティアに選択されたのかの原因の予測は簡単なことではないと思われる．

今回の方法で選択した入力変数が良い面もあれば悪い面もあるのかそうでないのかは，Leave-One-Out 法での同符号で選択された回数と異符号で選択された回数にどれだけ差があるのかで判断の基準にできるかもしれない．

10 おわりに

包絡分析法は，個々の事業体の特色を評価しつつも，それぞれの特色毎に優れた事業体を目標として設定して，非効率な事業体の改善点を具体的に示すことができる方法であるが，闇雲に入力変数や出力変数を追加しても効率的フロンティアが乱立してしまい，基準が成り立たなることも少なくないため，導入まで検討できる企業は少ないと思われる．正準相関分析を包絡分析法に適用することによって，処理が完了することができれば，効率的フロンティアを絞り込むことができ，処理が失敗する際のそれぞれの例外処理を徹底できれば，より良い経営効率を向上に役に立つ手法になると思われる．

参考文献

[1] 家計調査（二人以上の世帯）:『油脂・調味料』 https://www.stat.go.jp/data/kakei/5.html

[2] 国立がん研究センター：がん登録・統計 https://ganjoho.jp/reg_stat/statistics/

[3] 国税庁 酒のしおり :『平成 29 年度成人 1 人当たりの酒類販売 (消費) 数量表 (都道府県別)』 https://www.nta.go.jp/taxes/sake/shiori-gaikyo/shiori/2019/pdf/041.pdf

[4] 信田真佑 :『 正準相関分析と包絡分析法に関する研究』．南山大学大学院理工学研究科修士論文，2015.

[5] R ニャナデシカン『統計的多変量データ解析』日科技連出版社，1979.

[6] 刀根薫 :『経営効率性の測定と改善 -包絡分析法 DEA による-』．日科技連出版社，1993.

[7] 上田徹 :『DEA における変数選択について』．日本オペレーションズ・リサーチ学会秋季研究発表会アブストラクト集，50-51，2003.

[8] 吉田里穂 :『悪性新生物の地域性に関する統計的分析』．南山大学情報理工学部情報システム数理学科卒業論文，2014.