共分散の構造化による正準相関モデルの母数推定

黒河 久*, 繁 柊 算 男*

STABLE PARAMETER ESTIMATION IN CANONICAL CORRELATION MODEL BY MEANS OF STRUCTURED COVARIANCE MATRICES

Hisashi KUROKAWA and Kazuo SHIGEMASU

Parameter estimation of the Canonical Correlation Analysis (CCA) can be often difficult in the presence of multicollinearity. The purpose of this paper is to make the parameter estimates more stable by structuring relevant covariance matrices using the prior knowledge. More concretely, we assumed factor analytic model and covariance structure model, then in terms of the estimated parameters in these models, we reproduced the covariance matrices. We substituted these covariance matrices in the CCA instead of the maximum likelihood estimates, i.e. sample covariance matrices, then obtained the estimates for the CCA vectors. In this paper, we consider only the first principal CCA vectors, because we assume here our interest is in constructing the index connecting the two sets of variables. So, the stability of the estimated parameters was judged by the cross validation techniques. The cross validation study showed that the parameter estimation using the structured covariance matrices yielded more effective prediction than the traditional method.

1. はじめに

多変量解析モデルのひとつである正準相関モデルの母数推定は, このモデルが多重共線性の影響を受けやすいため, 得られる標本数が少ないときには母数の推定値が不安定になりやすい. そこで, 事前の知識を用いて変数間の共分散行列を構造化することにより推定値の安定化を図る. 具体的には, 共分散構造に因子分析モデル, および共分散構造分析モデルを仮定し, これらのモデルに基づいて推定した共分散行列を母共分散行列の推定値として正準相関モデルの母数を推定する方法を提案する.

この推定値の良さを安定性の面から評価する. ここでは, 正準相関分析の目的を, 2群の変数間の関連を見るという目的よりも, 2つの変数群の一方から他方を合成変量によって予測するという文脈でとらえる. この観点は, 他方の変数群によってとらえられる傾向性を示すインデックスを考える上で有効であろう. 故に,

推定値の安定性を, 予測効率によって評価することにする. そこで, 実データを用いたクロスバリデーション研究によって, 提案する推定法の有効性を示す.

2. 問題点

正準相関ベクトル $\mathbf{a}$, $\mathbf{b}$ は以下の式により与えられる.

$$(\sum_{yx} \sum_{xx}^{-1} \sum_{xy} - \rho^2 \sum_{yy}) \mathbf{b} = \mathbf{o}, \quad (1)$$

$$\mathbf{a} = \sum_{xx}^{-1} \sum_{xy} \mathbf{b} / \rho. \quad (2)$$

ただし,

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \quad \sum_{zz} = \begin{bmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{bmatrix}.$$

正準相関ベクトルの通常の標本推定値は, 式(1), (2)における母共分散行列 $\sum_{zz}$ を, その最尤推定値である標本共分散行列 S_{zz} によって置き換えた, 式(3)によって求められる.

$$(S_{yx} S_{xx}^{-1} S_{xy} - \hat{\rho}^2 S_{yy}) \hat{\mathbf{b}} = \mathbf{o}, \quad (3)$$

$$\hat{\mathbf{a}} = S_{xx}^{-1} S_{xy} \hat{\mathbf{b}} / \hat{\rho}.$$

* 東京工業大学工学部 (Faculty of Engineering, Tokyo Institute of Technology)

変数群 $\mathbf{y}$, $\mathbf{y}$ に含まれる変数の数を p とすると、標本共分散行列

$$S_{zz} = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix}$$

には $p(p+1)/2$ 個の推定すべきパラメータが含まれ、得られる標本が少ない場合には、標本数に比べて多くのパラメータを含むため、推定値は標本の偏りによる影響を受けやすくなり、将来のデータに推定値を適用し、予測を行うには必ずしも適切なものとは言えない。

また、互いに相関の高い変数の組が含まれるときには、共分散行列の推定値 $\hat{\Sigma}$ の行列式が 0 に近くなるため、推定値の逆行列 $\hat{\Sigma}^{-1}$ が不安定となり、その結果、正準相関ベクトルの推定値 $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$ も不安定となるという多重共線性の問題がある。

この多重共線性の影響を抑え、正準相関ベクトルの推定値を安定化するモデルとして、リッジ回帰モデル (Hoerl-Kennard, 1970) を正準相関モデルに適用した正準リッジモデル (Vinod, 1976) があるが、このモデルは共分散行列の構造を積極的に扱うものではない。

また、寺岡(1991)は共分散行列の構造モデルにデータに関する事前の知識を積極的に取り入れることが可能なモデルである共分散構造分析モデルを重回帰モデルに適用し、予測モデルを構成することを提案している。

本研究では、正準相関モデルの母数推定において、共分散行列の構造モデルとして、因子分析モデル、および共分散構造分析モデルを仮定し、これらのモデルに基づいて推定した共分散行列を母共分散行列の推定値として正準相関モデルの母数を推定する方法を提案し、実データを用いたクロスバリデーション研究によって、提案する推定法の有効性を示す。

3. 共分散行列に対する構造モデル

ここでは、共分散行列に対して仮定した構造モデルである因子分析モデル、共分散構造分析モデルについて述べる。

3.1 因子分析モデル

観測変数が、その変数の数に比べてより少数の潜在的な共通因子の線形結合で表すことができるということのみが事前にわかっているという場合、共分散行列の構造モデルとして (探索的) 因子分析モデルを仮定

し、このモデルに基づいて推定を行うことが妥当であろう。

m 個の共通因子が仮定できるとき、 $\mathbf{z}$ の共分散行列は次のように構造化される。

$$\Sigma_{zz} = \Lambda \Lambda' + \Psi. \quad (4)$$

ただし、

$$\Psi = \text{diag}(\phi_1, \phi_2, \dots, \phi_p).$$

因子数の選択はつぎの 2 つの方法により決定することが考えられる。

- (i) データに関する事前の知識によって決定する。
- (ii) 所与のデータに対して因子分析を行い、AIC (Akaike's Information Criterion) などの基準を用いて決定する。

3.2 共分散構造分析モデル

共分散構造分析では、分析者の共分散構造に対する知識や仮説に基づいてモデルを構成し、推定を行うことが可能である。このため、あらかじめ分析者にデータの共分散構造に対する知識があるとき、共分散行列の構造モデルとして共分散構造分析モデル仮定し、このモデルに基づいて推定を行うことが妥当であろう。

観測変数間の共分散行列はモデル中の母数を用いて以下のように構造化される (記号等については、豊田(1992) 参照)。

$$\Sigma_{zz} = (KB\Gamma + \Lambda)\Phi(KB\Gamma + \Lambda)' + KB\Psi B'K' + \Delta.$$

4. 母数の推定

正準相関ベクトルの推定の安定化を、共分散行列の推定の安定化を通して実現するため、共分散構造に因子分析モデル、共分散構造分析モデルを仮定し、これらのモデルに基づいて推定した共分散行列を母共分散行列の推定値として正準相関モデルの母数を推定する。

共分散行列の推定は、構造モデルとして因子分析モデル、および共分散構造分析モデルを仮定する場合に、構造モデルを何も仮定しない場合を含めて 3 通りの方法を考える。

- (i) モデルを何も仮定しない。

この場合、共分散行列の推定値は最尤推定値である標本共分散行列となる。

$$\hat{\Sigma}_{zz} = S_{zz}.$$

- (ii) 因子分析モデルを仮定する。
因子分析によってモデルの母数の推定値 $\hat{\Lambda}$, $\hat{\Psi}$ を求め、これらを次式に代入する。

$$\hat{\Sigma}_{zz} = \hat{\Lambda} \hat{\Lambda}' + \hat{\Psi}.$$

- (iii) 共分散構造分析モデルを仮定する。
共分散構造分析によってモデルの母数の推定値を求め、これらを次式に代入する。

$$\hat{\Sigma}_{zz} = (\hat{K} \hat{B} \hat{\Gamma}' + \hat{\Lambda}) \hat{\Phi} (\hat{K} \hat{B} \hat{\Gamma}' + \hat{\Lambda})' + \hat{K} \hat{B} \hat{\Psi} \hat{B}' \hat{K}' + \hat{\Delta}.$$

共分散構造分析は、SAS の CALIS プロシージャを用いて行った。

上記の (i), (ii), (iii) によって推定された共分散行列

$$\hat{\Sigma}_{zz} = \begin{bmatrix} \hat{\Sigma}_{xx} & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_{yy} \end{bmatrix}$$

を式(1), (2)に代入して正準相関ベクトルの推定値 $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$ を求める。

5. 適用例

適用例として、表1に示したようなデータ（体力測定データ）の分析例を示す。このデータは2群に分けられ、第1群は体格測定、体力診断テストのデータ、第2群は運動能力テストのデータである。被験者は東工大生6,262人である。

体格、体力を代表するインデックスと運動能力を代表するインデックスとの関わりを調べ、前者によって後者を予測することを考える。このため、第一正準相

表1. 体力測定データ

第1群 (体格測定, 体力診断テスト)		第2群 (運動能力テスト)
身長	体重	50 m 走
胸囲	座高	走り幅跳び
握力	背筋力	ハンドボール投げ
立位体前屈		けんすい
上体反らし		1,500 m 走
垂直跳び		
サイドステップ		
踏台昇降運動

図1. Cross Validation Study

関のみを扱うこととする。

このデータを用いたクロスバリデーション研究によって提案する手法の有効性を示す。

5.1 クロスバリデーション研究

正準相関モデルの母数推定値の安定性を評価するためにクロスバリデーションを行う。クロスバリデーションを次の手順によって行う（図1参照）。

Step 1. 全データ（ $n=6,262$ ）の中から validation data（ $n_{val}=1,000$ ）をランダムに抽出する。

Step 2. 残りのデータの中から calibration data

表2. AIC の値を最少による因子数
(10 回中, 選択された回数)

標本数	因子数					
	3	4	5	6	7	8
50	2	4	2	2	0	0
100	0	2	5	3	0	0
150	0	2	1	7	0	0
200	0	0	1	7	2	0
300	0	0	1	4	3	2
400	0	0	0	2	7	1

表 3. 体力測定データの相関行列

	身長	体重	胸囲	座高	握力	背筋力	サイド	立位	上体	垂直	踏台	50 m	幅跳び	ハンド	懸垂	1,500 m
身長	1.0000															
体重	0.4958	1.0000														
胸囲	0.2619	0.7620	1.0000													
座高	0.7370	0.4289	0.1813	1.0000												
握力	0.3365	0.4479	0.4189	0.2109	1.0000											
背筋力	0.2381	0.3871	0.3787	0.1377	0.5595	1.0000										
サイド	0.1134	0.0391	0.0884	0.0143	0.2264	0.2507	1.0000									
立位	0.0097	0.0113	0.0686	0.0236	0.1182	0.1217	0.1641	1.0000								
上体	0.0950	0.0650	0.0764	0.0741	0.1102	0.1378	0.1542	0.2647	1.0000							
垂直	0.1926	0.0727	0.1156	0.0173	0.3486	0.3088	0.3619	0.2040	0.1415	1.0000						
踏台	0.0705	0.0075	0.0252	0.0018	0.1286	0.1571	0.2134	0.0920	0.0508	0.1405	1.0000					
50 m	-0.1019	-0.0359	-0.0863	-0.0098	-0.3244	-0.2891	-0.3391	-0.0904	-0.0724	-0.4930	-0.2360	1.0000				
幅跳び	0.2091	0.0666	0.1080	0.0580	0.3591	0.3247	0.3732	0.1452	0.1413	0.5021	0.2264	-0.6027	1.0000			
ハンド	0.2195	0.2705	0.2876	0.2876	0.0954	0.3532	0.3167	0.1475	0.1231	0.3376	0.2104	-0.3990	0.3844	1.0000		
懸垂	-0.1441	-0.2083	-0.0378	-0.1850	0.2735	0.2621	0.3313	0.2202	0.0999	0.3680	0.2638	-0.4361	0.4271	0.2648	1.0000	
1,500 m	-0.0531	0.0674	0.0134	0.0368	-0.1312	-0.1640	-0.2902	-0.1406	-0.0871	-0.2390	-0.4808	0.3755	-0.3894	-0.2272	-0.3740	1.0000

($n_{cat}=50, 100, 150, 200, 300, 400$) をそれぞれ 10 組ずつランダムに抽出する。

Step 3. calibration data を用いて、仮定したそれぞれのモデルに基づいて共分散行列を推定し、正準相関ベクトルの推定値を求める。

Step 4. 得られた正準相関ベクトル推定値 $\hat{a}$, $\hat{b}$ を用いて、validation data, x_i, y_i ($i=1, \dots, n_{val}$) に対する 2 つの n_{val} 次元ベクトル $\hat{a}'x_i$ と $\hat{b}'y_i$ の間の標本相関係数を計算する。

Step 5. それぞれの標本数 n_{cat} に対して、計算した 10 個の正準合成変量間の相関係数の平均をとり、この値の母数の推定値の安定性を示す指標とする。

5.2 体力測定データに対する因子分析モデル

体力測定データの共分散構造に対して仮定した因子分析モデルの因子数の選択は次の 2 つの方法により決定した。

- (i) 16 個の変数に対する理論的検討の結果、因子数を 3 とあらかじめ特定する。
- (ii) 所与のデータ (calibration data) に対して因子分析を行い、AIC 基準を用いて因子数を決定した。

その結果、各 10 組のデータセットにおいて因子数は表 2 に示すようになった。

上記 (ii) の場合に選択された因子数は、上記 (i) の場合に比べていずれも多めに選択されており、また、標本数が多くなるにつれて因子数も多くなることがわかる。

これらの因子分析モデルに含まれる自由度を考慮し

た推定すべきパラメータの数は、因子数を 3 に固定したモデルでは、61、また、AIC の値を最小にする因子数をもつ因子分析モデルでは 61 から 116 (因子数 8 のとき) であり、いずれの場合も標本共分散行列に含まれる推定すべきパラメータの数

$$\frac{16 \times (16+1)}{2} = 136$$

よりも少ないものになっている。

5.3 体力測定データに対する共分散構造分析モデル

体力測定データに対して仮定したモデルは、図 2、図 3 に示したような確認的因子分析モデル、パス解析モデルである。共分散構造分析は、これらの 2 つのモデルを包含する、より一般的なモデルを扱うことを可能にしたモデルであるが、表 3 に示したように、ここで扱う体力測定データは、各変数間の相関係数が全体的に小さく、潜在変数間に因果関係を仮定するような複雑なモデルを仮定しても効果がないと考え、共分散構造分析モデルとしては、最も簡単なモデルと考えられるこれらのモデルを仮定した。

5.3.1 確認的因子分析モデル

確認的因子分析モデルは、3 因子探索的因子分析モデルをもとに 16 個の変数に対する理論的検討により構成した。このモデルを図 2 に示す。この確認的因子分析モデルに含まれるパラメータの数は 45 である。

参考のために、すべての calibration data を用いて、このモデルに基づく分析を SAS の CALIS プロシー

図 2. 確認的因子分析モデル

図3. パス解析モデル

ジャを用いて行い、次式で示されるモデルの適合度を表す指標の一つである GFI (Goodness of Fit Index) の値を計算した。その結果、この値は 0.921 であった。

$$GFI = 1 - \frac{\text{tr}((W(S - \hat{\Sigma}))^2)}{\text{tr}((WS)^2)}$$

(W は重み行列)

5.3.2 パス解析モデル

パス解析モデルは、潜在変数を仮定しないモデルである。16 個の観測変数に対する理論的検討の結果、これらの変数間に因果関係、相関関係を仮定した。このモデルを図 3 に示す。このパス解析モデルに含まれるパラメータの数は 45 である。

このモデルについても、参考のために、すべての calibration data を用いた分析を行い、GFI の値を計算した。その結果、この値は 0.918 であった。

5.4 クロスバリデーションの結果

以上に示したような共分散行列に対する構造モデル 4 通りに、モデルを仮定しない場合を加えて次の 5 通りを考える。

- (i) モデルなし (従来の正準相関分析)
- (ii) (探索的) 因子分析モデル I (因子数 3)
- (iii) (探索的) 因子分析モデル II

図4. クロスバリデーションの結果 I

(因子数は AIC の値を最小にするもの)

- (iv) 共分散構造分析モデル I (確認的因子分析モデル)
 - (v) 共分散構造分析モデル II (パス解析モデル)
- まず、因子分析モデル I, II と、因子分析モデル I をもとに構成した共分散構造分析モデル I (確認的因子分析のモデル) の場合の結果を図 4 に示す。calibration data の数が少ないときには、従来の正準相関分析に比べて因子分析モデル I, II, および共分散構造分析モデル I を仮定したときのほうが良い結果が得られているが、標本数が多いときには従来の正準相関分析が

図5. クロスバリデーションの結果 II

最も良い結果が得られている。モデル II を仮定したときは標本数が多くなっても良い結果が得られている。

また、因子分析モデル I をもとに、データに関する知識をモデルに積極的に組み込んだ共分散構造分析モデル I は、すべての場合で因子分析モデル I よりも良い結果が得られており、また、標本数が少ないときにはこれら 3 つの推定法の中で最も良い結果が得られている。

次に、共分散構造分析モデル I (確認的因子分析モデル) と共分散構造分析モデル II (パス解析モデル) の結果を図 5 に示す。共分散構造分析モデル II を仮定したときも calibration data の数が少ないときには、従来の正準相関分析に比べて良い結果が得られている。共分散構造分析モデル I と共分散構造分析モデル II の結果を比べると、calibration data の数が 50 のときには共分散構造分析モデル I の結果のほうが良い結果を出しているが、そのほかの場合にはどちらも同じような結果となっている。

6. おわりに

正準相関モデルの母数推定において、得られる標本が少ない場合や、多重共線性の影響を受けている場合には推定値が不安定になる。この問題に対処するための方法を提案し、この方法の有効性を実データ (東工大生の体力測定データ) を用いたクロスバリデーション研究によって示した。その結果、以下のことが確認

できた。

- 共分散構造に適切なモデルを仮定して推定を行うことによって、得られている標本数が少ないときには、正準相関モデルの母数推定が安定化されている。これは、得られている標本数が少ないにもかかわらず、多くの推定すべきパラメータを含む標本共分散行列を共分散行列の推定値として用いるかわりに、パラメータ数を減らした適切なモデルに基づく推定値を用いることが有効であることを示している。
- 因子分析モデル II の場合のように、得られている標本によってモデル選択 (ここでは因子数の選択) を行うことによって、標本数が多くなっても従来の正準相関分析に比べてほぼ同程度の結果が得られる。
- データの共分散構造に関する事前の知識がある場合、共分散構造分析を用いて、この知識をモデルに組み込むことにより、良い結果が得られる。

謝 辞

本研究で用いたデータの使用を許可していただいた東工大、石井源信教授に感謝する。

参 考 文 献

- [1] Hoerl, A.E. and R.W. Kennard (1970a) Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.
- [2] Hoerl, A.E. and R.W. Kennard (1970b) Ridge regression: Application to nonorthogonal problems, *Technometrics*, **12**, 69-82.
- [3] 寺岡正也 (1991) 共分散構造分析に基づく予測, 東京工業大学修士論文
- [4] 豊田秀樹 (1992) SAS による共分散構造分析, 東京大学出版会
- [5] Vinod, H.D. (1976) Canonical ridge and econometrics of joint production, *Journal of Econometrics*, **4**, 147-166

—1993年2月8日受付,
1993年4月5日最終修正—