双対平坦性に基づく正準相関分析と混合分布を用いた項目反応理論

数理情報学専攻 48-196235　渡邉 元和
指導教員　駒木 文保 教授

1　はじめに

本論文は 2 部構成となっている．

第 I 部では，指数型分布族の持つ双対平坦性を利用した正準相関分析の手法を提案する．本研究では，指数型分布族の持つ双対平坦性を利用して与えられる角度を考え，この角度を最小化する合成ベクトルを求めるものとして，指数型分布族における正準相関分析を定めている．第 II 部では，項目反応理論と呼ばれるテスト理論に用いられる統計モデルを扱う．本研究では，被験者能力の分布に混合正規分布を用いたモデル [3] を，問題の選択肢が順序を考えられる多肢となっている場合に用いられる段階反応モデル [6] へと拡張し，パラメータ推定手法を与えている．

本発表では主に第 I 部を扱い，第 II 部は概要を紹介する．

2　正準相関分析

正準相関分析 [4,5] は，2 つの多変量変数群が与えられたときに，変数群の間の相関構造を分析するために用いられる多変量解析手法である．

正準相関分析は，行列 $Y \in \mathbb{R}^{n\times k}$, $X \in \mathbb{R}^{n\times \ell}$ が与えられたとき，それぞれの行列を $\boldsymbol{\alpha} \in \mathbb{R}^k, \boldsymbol{\beta} \in \mathbb{R}^\ell$ を用いた線形変換をすることで得られる 2 本のベクトル $\boldsymbol{r} := Y\boldsymbol{\alpha}$, $\boldsymbol{u} := X\boldsymbol{\beta}$ の相関係数を最大化するような $\boldsymbol{\alpha}, \boldsymbol{\beta}$ を求める手法であり，以下の制約付き最適化問題

$$
\begin{aligned}
(\boldsymbol{\alpha}, \boldsymbol{\beta}) &:= \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmax}} \boldsymbol{\beta}^\top X^\top Y \boldsymbol{\alpha} \\
\text{s.t. } &\|Y\boldsymbol{\alpha}\| = \|X\boldsymbol{\beta}\| = 1
\end{aligned}
$$

を解くことで，その解が与えられる．これを幾何的に捉えると，正準相関分析は，各列ベクトルを基底ベクトルとする超平面

$$
S_1 := \{\boldsymbol{r} : \boldsymbol{r} = Y\boldsymbol{\alpha}\},\ S_2 := \{\boldsymbol{u} : \boldsymbol{u} = X\boldsymbol{\beta}\}
$$

を考え，2 本のベクトルのなす角度が最小となるような，$\boldsymbol{r} \in S_1, \boldsymbol{u} \in S_2$ を求める手法であると解釈することができる．

3　指数型分布族の双対平坦性

指数型分布族は，確率変数 $\boldsymbol{y} = (y_1, \ldots, y_n)$ の確率密度 (質量) 関数が，パラメータ $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^n)$ を用いて，$p(\boldsymbol{y}|\boldsymbol{\theta})d\mu(\boldsymbol{y}) = \exp(\theta^i y_i - \psi(\boldsymbol{\theta}))d\mu(\boldsymbol{y})$ で与えられるような確率分布族のことである．

情報幾何 [1,2] では，分布全体の集合 $\mathcal{S}$ を座標系 $\boldsymbol{\theta}$ を持つ n 次元多様体と考え，フィッシャー情報行列を計量として用いる．ここで，e-接続，m-接続と呼ばれる双対な接続に対して，座標系 $\boldsymbol{\theta}$ と，$\eta_i = E_{\boldsymbol{\theta}}[y_i]$ によって定められる座標系 $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ がそれぞれアファイン座標系となり，このように e-接続，m-接続に対するアファイン座標系が存在する空間はそれぞれ e-平坦，m-平坦と呼ばれる．また，平坦な座標系における測地線は，その座標系で直線として表され，それぞれ e-測地線，m-測地線と呼ばれる．

指数型分布族は，e-平坦かつ m-平坦な空間であることより双対平坦空間と呼ばれる．双対平坦空間では，双対な座標系を $\boldsymbol{\theta}, \boldsymbol{\eta}$ として，$\boldsymbol{\theta}$ 座標系に対する計量を g_{ij}，$\boldsymbol{\eta}$ 座標系に対する計量を g^{ij} とすれば，$\dfrac{\partial \eta_i}{\partial \theta^j} = g_{ij}$，$\dfrac{\partial \theta^i}{\partial \eta_j} = g^{ij}$ が成立し，g_{ij} と g^{ij} は逆行列の関係にある．また，同じ接空間における $\boldsymbol{\theta}$ 座標系の基底 $\boldsymbol{e}_i$ と，$\boldsymbol{\eta}$ 座標系の基底 $\boldsymbol{e}^j$ に関して，双対直交性と呼ばれる関係 $\left\langle \boldsymbol{e}_i, \boldsymbol{e}^j \right\rangle = \delta_i^j$ が成り立っている．

4　指数型分布族に対する正準相関分析

指数型分布族のなす空間における，双対平坦性を利用した角度を定め，その角度を最小化する変数を求めるものとして指数型分布族における正準相関分析を与える．

4.1　指数型分布族における角度

指数型分布族上の 3 点 $p, q, r \in \mathcal{S}$ が与えられたとき，q を頂点として p と q を結ぶ e-測地線と，p と q を結ぶ m-測地線を利用して与えられる角度 ϕ $(0 \le \phi \le \pi)$ を以下のように定める．

$$
\phi := \arccos\left[\frac{\left\langle \dfrac{d}{dt}\tilde{p}(0), \dfrac{d}{dt}\tilde{r}(0) \right\rangle}{\left\| \dfrac{d}{dt}\tilde{p}(0) \right\| \left\| \dfrac{d}{dt}\tilde{r}(0) \right\|} \right]
$$

ただし，$\frac{d}{dt}\tilde{p}(0)$ は，p と q を結ぶ e-測地線 $\tilde{p}(t) = t\boldsymbol{\theta}_p + (1-t)\boldsymbol{\theta}_q$ の $t=0$ での接ベクトル，$\frac{d}{dt}\tilde{r}(0)$ は，r と q を結ぶ m-測地線 $\tilde{r}(t) = t\boldsymbol{\eta}_r + (1-t)\boldsymbol{\eta}_q$ の $t=0$ での接ベクトルである．また，$\boldsymbol{\theta}_p, \boldsymbol{\theta}_q$ はそれぞれ p, q の $\boldsymbol{\theta}$ 座標系における座標，$\boldsymbol{\eta}_q, \boldsymbol{\eta}_r$ はそれぞれ q, r の $\boldsymbol{\eta}$ 座標系における座標を表している．

4.2 指数型分布族に対する正準相関分析の定式化

指数型分布族を考えたときの正準相関分析を，以下の手順で与える．

まず，頂点 $\boldsymbol{v} \in S$ を 1 点与え，e-平坦な部分空間 e-span(X) と，m-平坦な部分空間 m-span(Y) を

$$\text{e-span}(X) := \{\boldsymbol{\theta} : \boldsymbol{\theta} = \sum_{i=1}^{\ell} (\boldsymbol{x}_i - \boldsymbol{\theta}_{\boldsymbol{v}})\beta^i + \boldsymbol{\theta}_{\boldsymbol{v}}\}$$

$$\text{m-span}(Y) := \{\boldsymbol{\eta} : \boldsymbol{\eta} = \sum_{i=1}^{k} (\boldsymbol{y}^i - \boldsymbol{\eta}_{\boldsymbol{v}})\alpha_i + \boldsymbol{\eta}_{\boldsymbol{v}}\}$$

と定める，このとき，$\boldsymbol{u} \in \text{e-span}(X), \boldsymbol{r} \in \text{m-span}(Y)$ として，$\boldsymbol{u}$ と $\boldsymbol{v}$ を結ぶ e-測地線と，$\boldsymbol{r}$ と $\boldsymbol{v}$ を結ぶ m-測地線を利用して与えられる角度を最小とするような，$\boldsymbol{u}$, $\boldsymbol{r}$ を求めるものとして正準相関分析を与える．

4.3 指数型分布族に対する正準相関分析の解

指数型分布族に対する正準相関分析の解となる $\boldsymbol{\alpha}, \boldsymbol{\beta}$ は，一般化固有値問題

$$\begin{pmatrix} O & \widetilde{Y}^\top \widetilde{X} \\ \widetilde{X}^\top \widetilde{Y} & O \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \lambda \begin{pmatrix} \widetilde{Y}^\top g^{ij}(\boldsymbol{\eta}_{\boldsymbol{v}}) \widetilde{Y} & O \\ O & \widetilde{X}^\top g_{ij}(\boldsymbol{\theta}_{\boldsymbol{v}}) \widetilde{X} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}$$

の固有値が最大となるような固有ベクトルを選ぶことで得られる．ただし，$\widetilde{X}$，$\widetilde{Y}$ は，それぞれ，

$$\widetilde{X} := (\boldsymbol{x}_1 - \boldsymbol{\theta}_{\boldsymbol{v}}, \ldots, \boldsymbol{x}_\ell - \boldsymbol{\theta}_{\boldsymbol{v}})$$
$$\widetilde{Y} := \left(\boldsymbol{y}^1 - \boldsymbol{\eta}_{\boldsymbol{v}}, \ldots, \boldsymbol{y}^k - \boldsymbol{\eta}_{\boldsymbol{v}}\right)$$

を表し，$\boldsymbol{\theta}_{\boldsymbol{v}}$，$\boldsymbol{\eta}_{\boldsymbol{v}}$ は，それぞれ $\boldsymbol{v}$ の $\boldsymbol{\theta}$ 座標系での座標と $\boldsymbol{\eta}$ 座標系での座標を表すものとする．

5 数値実験

指数型分布族に対する正準相関分析と，従来の正準相関分析の比較をするために数値実験を行った．

指数型分布族として，平均パラメータと分散パラメータを持つ正規分布を用いて，$n = 2000$，$k = \{60, 100\}$，$\ell = \{20, 40, 60, 80, 100\}$ で実験を行った．正準変数のスケーリングは接ベクトルの長さを利用して与え，評価指標は，正準変数 $\boldsymbol{u}, \boldsymbol{r}$ に対する KL ダイバージェンス $\text{KL}(\boldsymbol{r} : \boldsymbol{u})$ の n 変数での平均値を用いた．

表 1. KL ダイバージェンスの平均値 ($\times 10^{-4}$)

	$k=60$		$k=100$	
ℓ	提案手法	従来手法	提案手法	従来手法
20	4.0370	4.0402	3.9986	4.0032
40	3.7753	3.7778	3.7343	3.7381
60	3.7020	3.7041	3.6616	3.6649
80	3.6211	3.6235	3.5848	3.5883
100	3.4893	3.5504	3.4479	3.5122

結果は表 1 のようになった．ただし，提案手法は，指数型分布族に対する正準相関分析，従来手法は従来の正準相関分析をそれぞれ表す．実験の結果，提案手法のほうが，より変数間の KL ダイバージェンスが小さくなるような正準変数の組が得られることがわかった．

6 第 II 部：混合分布を用いた項目反応理論

項目反応理論とは，テストやアンケートの回答結果から，被験者の潜在的な能力と各項目の持つ特性とを同時に推定する統計モデルである．項目反応理論では，被験者能力の分布を正規分布に仮定してパラメータ推定を行うことが多いが，その仮定が妥当とはいえない場合があることが指摘されているため，より柔軟な被験者の能力の分布を考えた研究が行われている．

本研究では，被験者の潜在能力に混合正規分布を用いたモデル [3] を，項目が順序のある多肢選択の場合に用いられる段階反応モデル [6] の枠組みに拡張し，マルコフ連鎖モンテカルロ法によるパラメータ推定手法を与えた．また，数値実験によりその性能を確かめた．

参考文献

[1] S. Amari. *Differential-geometrical Methods in Statistics*. Lecture notes in statistics. Springer-Verlag, 1985.

[2] S. Amari and H. Nagaoka. *Methods of information geometry*, Vol. 191. American Mathematical Soc., 2007.

[3] F. B. Gonçalves, B. C. C. Dias, and T. M. Soares. Bayesian item response model: a generalized approach for the abilities' distribution using mixtures. *Journal of Statistical Computation and Simulation*, Vol. 88, No. 5, pp. 967–981, 2018.

[4] H. Hotelling. Relations between two sets of variates. *Biometrika*, Vol. 28, No. 3–4, pp. 321–377, 1936.

[5] C. Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société Mathématique de France*, Vol. 3, pp. 103–174, 1875.

[6] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 1969.