識別情報を利用した主成分解釈のための手法と脳波データ識別への応用

201213073 山本　智基

経営工学主専攻　　指導教員：イリチュ（佐藤）美佳　教授

1. 序論

多変量データ解析の手法の一つである主成分分析の目的は、データの主な変動を要約する主成分を得ることである。そのため得られた主成分を解釈することは重要である。従来、主成分の解釈には、得られた主成分とデータの変数との相関である負荷量を用いているが、負荷量により主成分を解釈することは必ずしも容易ではない。

そこで、本研究では、対象とするデータが何らかの外的な識別情報を持っているとき、主成分を識別情報により解釈する手法を提案する。この提案手法により、識別情報と得られた主成分との関連を示す量が得られるため、主成分を介在させることによって、データ識別のために、説明力の高い変数を選択することが可能となる。この手法を異なる思考という識別情報を持つ脳波データに対して適用した数値例を示す。

2. 主成分分析

n 個の個体について p 変数の観測値からなるデータ行列を $\boldsymbol{X} = (x_{ia})\ (i = 1,\cdots,n,\ a = 1,\cdots,p)$ とする。$\boldsymbol{X}$ の変数に関する不偏分散共分散行列を $\boldsymbol{\Sigma}$ とし、その要素を $\sigma_{ab}(a,b = 1,\cdots,p)$ とする。$\boldsymbol{\Sigma}$ の固有値を $\lambda_1,\cdots,\lambda_\alpha(\lambda_1 \geq \cdots \geq \lambda_\alpha,\ \alpha \leq p)$ とし、対応する固有ベクトルを $\boldsymbol{l}_1,\cdots,\boldsymbol{l}_\alpha$ とすると、第 f 主成分 $(1 \leq f \leq \alpha)$ は、$\boldsymbol{z}_f = \boldsymbol{X}\boldsymbol{l}_f\ (\boldsymbol{l}_f^{\mathrm{T}}\boldsymbol{l}_f = 1)$ と表される。また、$\boldsymbol{l}_1,\cdots,\boldsymbol{l}_\alpha$ ベクトルからなる $n \times \alpha$ 行列 $\boldsymbol{L}$ を次のように表す。

$$\boldsymbol{L} = (\boldsymbol{l}_1,\cdots,\boldsymbol{l}_\alpha) \tag{1}$$

主成分 z_f と変数 x_g の相関係数を主成分 z_f の因子負荷量という。因子負荷量を用いることにより、ある主成分と強く関係する変数を見つけ出すことができ、その変数により主成分の考察を行うことができる。

3. 制約つき主成分分析

制約つき主成分分析 [1] は、データが与えられたとき、これを外部情報によって説明できる部分とできない部分に分解し（外部分析）、分解された個々の部分に必要に応じて主成分分析を適用する（内部分析）方法である。

行に関する外的情報を示す情報行列を $\boldsymbol{G}(n \times \alpha)$、列に関する外的情報を示す情報行列を $\boldsymbol{H}(p \times q)$ とすると、外部分析における基本モデルは、$\boldsymbol{X} = \boldsymbol{G}\boldsymbol{M}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{B}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{G}\boldsymbol{C} + \boldsymbol{E}$ と表される。ここで、$\boldsymbol{M}(\alpha \times q)$、$\boldsymbol{B}(n \times q)$、$\boldsymbol{C}(\alpha \times p)$ はモデルのパラメータ行列、$\boldsymbol{E}(n \times p)$ は誤差行列を表す。いま、

$$\boldsymbol{X} = \boldsymbol{G}\boldsymbol{M}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{E}_1 \tag{2}$$

とし、$\boldsymbol{E}_1$ の平方和 $SS(\boldsymbol{E}_1) = tr(\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1)$ を最小化する $\boldsymbol{M}$ を推定する問題を考えると、

$$\hat{\boldsymbol{M}} = (\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-} \tag{3}$$

を得る。なお $(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-}$、$(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-}$ はそれぞれ $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$、$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$ の一般逆行列を示す。このときの残差は、$\hat{\boldsymbol{E}}_1 = \boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{H}}$ であり、$\boldsymbol{P}_{\boldsymbol{G}} = \boldsymbol{G}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-}\boldsymbol{G}^{\mathrm{T}}$ と $\boldsymbol{P}_{\boldsymbol{H}} = \boldsymbol{H}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-}\boldsymbol{H}^{\mathrm{T}}$ はそれぞれ直交射影行列である。

4. 識別情報を利用した主成分解釈のための手法

外部情報として K 個のグループへの識別情報が与えられているとき、データを主成分分析によって要約し、得られた主成分の情報と予め与えられている識別情報との関連が分かれば、情報を縮約した上で新たなデータの識別に役立てられ得る。そこで、得られた主成分と識別情報との関連を抽出する手法を提案する。

K 個のグループの識別情報を示す行列を $\boldsymbol{Y}$、各グループの重みを示す行列を $\boldsymbol{W}$ とし、(2) 式において

$$\boldsymbol{G} \equiv \boldsymbol{Y}\boldsymbol{W} \equiv \left(\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{array}\right) \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_K \end{pmatrix} \tag{4}$$

とする。ここで、$\boldsymbol{Y}$ は、$\frac{n}{K}(k-1)+1$ 行目から $\frac{n}{K}k$ 行目まで、k 列目の要素に 1、k 列目以外の要素に 0 を取る行列であり、$\boldsymbol{W}$ は、$w_1,\cdots,w_K(w_k \neq 0)$ を対角要素とする対角行列である。また、(2) 式の $\boldsymbol{H}$ に (1) 式に示す $\boldsymbol{L}$ を代入する。このとき (2) 式は、$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{W}\boldsymbol{M}\boldsymbol{L}^{\mathrm{T}} + \tilde{\boldsymbol{E}}$ となる。ここで、$\tilde{\boldsymbol{E}} = \boldsymbol{O}$ とするとき、

$$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{W}\boldsymbol{M}\boldsymbol{L}^{\mathrm{T}} \tag{5}$$

である。ここで、第 k グループのα 個の主成分に対する関連性を示すベクトルを $\boldsymbol{m}_k = (m_{k1}, \cdots, m_{k\alpha})$ $(k = 1, \cdots, K)$ とし、変数 a のα個の主成分に関する $\boldsymbol{L}$ の要素からなるベクトルを$\tilde{\boldsymbol{l}}_a = (l_{a1} \cdots l_{a\alpha})$ $(a = 1, \cdots, p)$ とする。このとき、(5) 式より、

$$\boldsymbol{X} = \left(\begin{array}{ccc} w_1\boldsymbol{m}_1\tilde{\boldsymbol{l}}_1^{\mathrm{T}} & \cdots & w_1\boldsymbol{m}_1\tilde{\boldsymbol{l}}_p^{\mathrm{T}} \\ \vdots & & \vdots \\ w_1\boldsymbol{m}_1\tilde{\boldsymbol{l}}_1^{\mathrm{T}} & \cdots & w_1\boldsymbol{m}_1\tilde{\boldsymbol{l}}_p^{\mathrm{T}} \\ \hline \vdots & & \vdots \\ \hline w_K\boldsymbol{m}_K\tilde{\boldsymbol{l}}_1^{\mathrm{T}} & \cdots & w_K\boldsymbol{m}_K\tilde{\boldsymbol{l}}_p^{\mathrm{T}} \\ \vdots & & \vdots \\ w_K\boldsymbol{m}_K\tilde{\boldsymbol{l}}_1^{\mathrm{T}} & \cdots & w_K\boldsymbol{m}_K\tilde{\boldsymbol{l}}_p^{\mathrm{T}} \end{array} \right) \tag{6}$$

となる。(6) 式より、(5) 式中の $\frac{n}{K}(k-1)+1$ 行目から$\frac{n}{K}k$ 行目までは、第 k グループを示すベクトル $\boldsymbol{m}_k$ によって説明しようとするものであることが分かる。さらに、各変数の特徴は、α個の主成分を介在させた p 個の変数に対する特徴ベクトル$\tilde{\boldsymbol{l}}_1, \cdots, \tilde{\boldsymbol{l}}_p$ により説明するものである。そのため $\tilde{\boldsymbol{E}}$を最小にするように ($\tilde{\boldsymbol{E}}$を $\boldsymbol{O}$ にするように) 推定される (5) 式における$\hat{\boldsymbol{M}}$ は、外的情報として与えられた識別情報と主成分分析により得られた主成分との関連性を示すものと考えられる。(3) 式と同様に (5) 式における $\tilde{\boldsymbol{E}}$ を最小にする推定量 $\hat{\boldsymbol{M}}$ は次のように得られる。

$$\hat{\boldsymbol{M}} = \quad \{(\boldsymbol{YW})^{\mathrm{T}}\boldsymbol{YW}\}^{-}(\boldsymbol{YW})^{\mathrm{T}}\boldsymbol{XL}(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L})^{-} \tag{7}$$

また、第 k グループに含まれるデータの重心を $\boldsymbol{v}_k = (v_{k1}, \cdots, v_{kp})$ とすると

$$\hat{\boldsymbol{M}} = \left(\begin{array}{ccc} \frac{1}{w_1} < \boldsymbol{v}_1 \cdot \boldsymbol{l}_1 > & \cdots & \frac{1}{w_1} < \boldsymbol{v}_1 \cdot \boldsymbol{l}_p > \\ \vdots & \vdots & \vdots \\ \frac{1}{w_K} < \boldsymbol{v}_K \cdot \boldsymbol{l}_1 > & \cdots & \frac{1}{w_K} < \boldsymbol{v}_K \cdot \boldsymbol{l}_p > \end{array} \right)$$

となる。ここで、$< \boldsymbol{a} \cdot \boldsymbol{b} >$ は $\boldsymbol{a}$ と $\boldsymbol{b}$ の内積を示す。

5. 数値例

12 個の電極を用いて計測された 4 種類 (A〜D) の思考に対する脳波データを用いた。[2] 一つの思考に対し、サンプリング周波数 1000Hz で 4 秒間測定されたデータ (4000 × 12) を 1 セットとし、各思考 75 回ずつ測定した。そして、思考ごとに 26 回目から 55 回目の測定結果を抽出し、順に行結合したもの (120000 × 12) をそれぞれ行結合し、解析対象とするデータ $\boldsymbol{X}$(480000 × 12) とした。

まず、作成した $\boldsymbol{X}$ に対して主成分分析を行った。求めた因子負荷量うち第 3 主成分までを図 1 に示す。次に、主成分分析により得られた方向ベクトルを $\boldsymbol{l}_1$から $\boldsymbol{l}_{12}$ とし、(1) 式の $\boldsymbol{L} = (\boldsymbol{l}_1, \cdots, \boldsymbol{l}_{12})$ とする。また、4 つの思考の識別情報を示す行列を (4) 式に示される $\boldsymbol{Y}$ とする。$\boldsymbol{Y}$ は 1〜120000 行目までは (1,0,0,0)、120001〜240000 行目までは (0,1,0,0)、240001〜360000 行目までは (0,0,1,0)、360001〜480000 行目までは (0,0,0,1) の値を取る。また、思考 k に属する個体の重心ベクトルのノルムの値を (4) 式の w_k とする。そして、(7) 式に $\boldsymbol{X}$、$\boldsymbol{L}$、$\boldsymbol{Y}$、$\boldsymbol{W}$ を代入して、推定値 $\hat{\boldsymbol{M}}$ を得た。この結果を表 1 に示す。

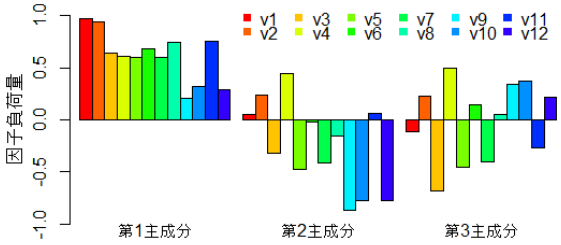

図 1　因子負荷量

表 1　思考と主成分の関連性

	第 1 主成分	第 2 主成分	第 3 主成分
思考 A	-0.987	0.027	0.046
思考 B	0.976	-0.062	-0.064
思考 C	0.828	-0.234	-0.444
思考 D	0.552	0.418	0.418

図 1 より、第 1 主成分は変数 v_1、v_2、v_8、v_{11} と相関が高く、第 2 主成分 は v_9、v_{10}、v_{12} と相関が高いことが分かる。

また表 1 より、第 1 主成分は各思考と大きく関連していることが分かる。また、主成分と各思考との関連性を主成分ごとに相対的に見ると、第 2 主成分と第 3 主成分において思考 C、D が、思考 A、B と比較し、主成分に大きく関連していることが分かる。

これらの結果から、主成分を介在させ、データ識別のために説明力の高い変数の選択を行うと、第 2 主成分を介在させることにより、v_9、v_{10}、v_{12} は思考 C、思考 D の特徴をよく説明している変数と考えられる。

6. 結論

本研究では、識別情報が与えられているデータに対し、主成分分析を行い、得られた主成分とデータの識別情報との関連性を抽出する手法を提案した。

提案手法を脳波データへ適用したところ、得られた主成分解釈に関して有効な結果が得られた。また、主成分を介在させ、識別に有効な変数の選択を行うことが可能となった。

参考文献

[1] Y. Takane, T. Shibayama, Principal Component Analysis with External Information on Both Subjects and Variables, *Psychometrika*, Vol. 56, No. 1, pp. 97-120, 1991

[2] K. Tanaka, K. Matsunaga, H.O. Wang, Electroencephalogram-Based Control of an Electric Wheelchair, *IEEE Transactions on Robotics*, Vol. 21, No. 4, pp. 762-766, 2005