1N-01

意味の加法性を有する主成分

長井　慶成†　　三浦　孝夫†

法政大学理工学部創生科学科†

東京都小金井市梶野町　3-7-2

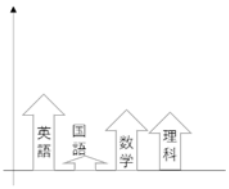
1. 前書き

テキストマイニングで用いられる代表的な統計学の分析手法の一つに主成分分析がある。主成分分析はデータ集合の特徴を捉えるのに有効な手法であり、主成分分析により得られる主成分の解釈を行うことで特徴的な部分が何かを把握することができる。主成分の解釈は変数の主成分負荷量をもとに人手で行う必要があるが、負の値をとらないデータの分析の際は、元のデータにはなかった負の概念が持ち込まれることで、直感性や一貫性のある解釈が困難になる。

この解釈しづらさを解消するために主成分負荷量の符号を正に統一することで、個々の変数の意味を単純に足し合わせることで主成分の解釈が行えるようになる。本稿ではこのような主成分を「加法的主成分」と呼び、フィシャー情報基準に基づいた加法的主成分の抽出方法を提案し、その妥当性を検証する。

表 1：加法的主成分の例

	加法的主成分
数学	0.55
理科	0.53
英語	0.65
国語	0.05



2. 加法的主成分

主成分分析の主成分は互いに直交し、分散が最大となるように取られている。そのため、新たな座標系となる加法的主成分においても無相関に近く、可能な限り分散が大きいものであることが望まれるが、分散をもとの主成分より大きくするためには無相関の制約を緩めなければならず、最大分散となりかつ最小相関となるような加法的主成分の抽出は困難である。

そこで相関と分散を同時にみることができるフィシャーの情報基準 J を次のように定義する。

$$J(q) = \frac{qRq}{||Q(q)RQ(q)||_F^2}$$

3. 加法的主成分の抽出

3.1 初期の方向ベクトルの作成

主成分分析により得られる固有ベクトルから初期値となる方向ベクトルを生成する。固有ベクトルは向きを反転させたものも固有ベクトルとなるため、表 2 の例のように 2 種類得られる。

表 2：固有ベクトル行列

	V1	V2	V3	V4
数学	-0.50	-0.24	0.83	0.08
理科	-0.60	0.08	-0.26	-0.76
英語	0.38	-0.83	0.03	-0.40
国語	-0.50	-0.50	-0.49	0.51

	V1	V2	V3	V4
数学	0.50	0.24	-0.83	-0.08
理科	0.60	-0.08	0.26	0.76
英語	-0.38	0.83	-0.03	0.40
国語	0.50	0.50	0.49	-0.51

本稿の目的は主成分負荷量の値がすべて正に統一された加法的主成分を得ることであるため、主成分負荷量が正のものを残し、負の要素をすべて 0 に置換して初期の方向ベクトルを得る。

表 3：初期の方向ベクトル

	q1	q2	q3	q4
数学	0.00	0.00	0.83	0.08
理科	0.00	0.08	0.00	0.00
英語	0.38	0.00	0.03	0.00
国語	0.00	0.00	0.00	0.51

	q1	q2	q3	q4
数学	0.50	0.24	0.00	0.00
理科	0.60	0.00	0.26	0.76
英語	0.00	0.83	0.00	0.40
国語	0.50	0.50	0.49	0.00

3.2 加法的主成分の生成

初期の方向ベクトルの長さは一部の主成分負荷量を 0 に置換したことにより長さが 1 を下回っている。この失った長さを補うためにフィシャーの情報基準 J に基づいた最急勾配法により 0 に置換した要素を新たな正の値で置き換え、加法的主成分を生成する。

例えば表 3 の q1 の更新を行う場合、まず負荷量が 0 の要素の 1 つに正の定数（例では 0.05）を加えた方向ベクトルを作成し、J の値を計算する。

表 4：方向ベクトル q1 の更新

q1-1	q1-2	q1-3
0.05	0.00	0.00
0.00	0.05	0.00
0.38	0.38	0.38
0.00	0.00	0.05
J=5	J=6	J=2

そして、J が最大となる方向ベクトルへ更新する。

表 5：q1 を更新した方向ベクトル行列 Q

	q1	q2	q3	q4
数学	0.00	0.00	0.83	0.08
理科	0.05	0.08	0.00	0.00
英語	0.38	0.00	0.03	0.00
国語	0.00	0.00	0.00	0.51

方向ベクトル行列のすべての方向ベクトルを更新するのを 1 ステップとする。また、1 ステップで各方向ベクトルの更新は一度だけ行う。これをすべての方向ベクトルの長さが 1 に達するまでステップを継続し、生成されたベクトルを加法的主成分と呼ぶこととする。

4. 実験

4.1 実験準備

本稿では、毎日新聞の記事データベース「CD-毎日新聞 2017」に収録された 2017 年 1 月 3 日から 1 月 14 日までのスポーツ記事のうち述べ語数が 6 以上の 356 文書を分析対象とする。この文書集合から形態素解析システム MeCab を用いて出現する名詞を抽出し、各文書を総出現頻度が 14 回以上の計 199 語の名詞の出現頻度を要素とする文書ベクトルで表現する。文書ベクトル集合の各名詞に対して標準化を行い、文書行列 D を生成し、その変数間の相関を表す相関行列 R を得る。こうして得られた相関行列から主成分分析により求まる固有ベクトル行列から各主成分の正、負の主成分負荷量それぞれを残した方向ベクトル行列を作成し、P,M とする。

4.2 評価指標

本稿では、主成分の解釈しやすさを次のように定義、評価を行う。

1. 「容易性」：より少ない単語で主成分の解釈が可能

$$\frac{\text{上位 k 変数の累積解釈影響率}}{\text{主成分負荷量が 0 でないの変数の割合}}$$

2. 「独自性」：他の主成分と解釈し分けることが可能

$$\frac{\text{全文書数}}{\text{変数が累積解釈影響率x\%に入る文書数}}\text{の平均}$$

4.3 実験結果

2 種類の初期の方向ベクトルからフィッシャーの情報基準を目的関数とした最急勾配法により非負の方向ベクトルが生成する。以下に主成分分析の第 2 主成分と、提案手法により得られる P の第 2 加法的主成分の主要名詞と評価を示す。

表 6：第 2 主成分の主要名詞

主成分		加法的主成分	
名詞	主成分負荷量	名詞	負荷量
トップ	0.196	主将	0.400
時間	0.170	トップ	0.196
総合	0.170	時間	0.170
記録	0.168	総合	0.170
2 位	0.157	記録	0.167
1 2 年	0.155	2 位	0.157
獲得	0.153	1 2 年	0.155
スタート	0.151	獲得	0.153
選手	0.144	スタート	0.151
目標	0.136	選手	0.144

表 7：第 2 主成分の評価

	主成分	加法的主成分
容易性	2.845	4.206
独自性	1.185	5.472

加法的主成分では、主成分にはなかった「主将」という名詞が負荷量トップになっており、負荷量も他の名詞に比べ一回り大きいことから、加法的主成分では「主将」という言葉が解釈において重要な意味を持っている。

主成分の評価においても、主成分よりも解釈のしやすいものとなっている。

続いて主成分ごとの評価を平均し、方向ベクトル行列全体の「容易性」「独自性」を評価した結果を表 8 に示す。

表 8：方向ベクトル行列の評価

	主成分	P	M
容易性	2.483	6.091	6.141
独自性	1.183	5.429	5.440

4.4 考察

生成された加法的主成分の分散と主成分間の相関を表 9 に示す。

表 9：加法的主成分の相関と分散

		P	M
総分散		732.997	743.846
相関係数	|0.75|~|1.00|	0 (0.00%)	0 (0.00%)
	|0.50|~|0.75|	2233 (11.33%)	3517 (17.85%)
	|0.25|~|0.50|	301 (1.53%)	513 (2.60%)
	0~|0.25|	17167 (87.14%)	15671 (79.54%)

総分散は直交性の制約を緩めたことで、主成分の総分散に比べ約 3.7 倍になり情報量が大幅に増加している。

また主成分間の相関を見ると、主成分の全組み合わせのうち 8 割程度は相関係数が-0.25 から 0.25 と弱い相関に抑えることができており、大半の主成分において分散と相関性は両立できている。

5. 結論

本稿では、フィッシャー情報基準を用いた再急勾配法により解釈が容易な加法的主成分の生成方法を提案した。提案手法により得られた加法的主成分は、解釈の中心的な意味をなす語の出現や解釈が少数の語から行えるようになった。

参考文献

[1] Ron Zass, Amnon SHashua : ”Nonnegative Sparse PCA”

[2] Dan Vilenchik, Rarak Yichye, Maor Abutbul: ”To Interpret or Not to Interpret PCA? This Is Our Question” ICWSM 2019