加法的主成分

長井　慶成†　　三浦　孝夫†

† 法政大学理工学部創生科学科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †yoshinari.nagai.8c@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし　現実世界には負の値をとらない文書行列や画像のようなデータも存在する。主成分分析の主成分の解釈は主成分負荷量が高い変数をもとに行われることが多いが、このような正の値からなるデータに対し主成分分析を行い解釈を試みた場合、負の概念が持ち込まれることで解釈が困難になることが想定される。本稿では、主成分に代わる主成分負荷量がすべて正の値となる加法的主成分をフィッシャー情報基準に基づいた最急勾配法により生成する方法を提案する。非負である加法的主成分は主成分負荷量の大きさに応じた変数の意味を足し上げたものが主成分の解釈とみなすことのできる性質である「意味の加法性」を有する主成分となり、直感的で解釈のしやすい主成分となることが期待される。提案手法により抽出した加法的主成分は主成分分析の主成分と比較して解釈しやすい主成分となっているかを検証する。

キーワード　テキストマイニング, 主成分分析, フィッシャー情報基準

1　前　書　き

インターネットの普及により情報過多社会となった現代では、大量の情報を収集することは容易となった。これは文書データにおいても例外ではなく、統計学や自然言語処理の解析手法を用いて膨大な文書データを分析し有益な情報や知識を抽出するテキストマイニングは注目を集めている。

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を把握することができる。

世の中には文書行列や画像、音響信号データのように負の値をとらないデータも存在する。主成分の解釈は変数の主成分負荷量をもとに主観的に行う必要があるが、このようなデータを分析する際は、もとのデータにはなかった負の概念が持ち込まれることや、正負の主成分負荷量が混在することで生じる変数同士の意味の打ち消しあいなどを考慮する必要があるため、直感性や一貫性のある解釈が困難になると想定される。

例として、非負値のみをとる四科目の試験データ集合を主成分分析したところ図 1 のような主成分が得られたとする。

主成分	
数学	-0.65
理科	-0.39
英語	0.65
国語	0.05

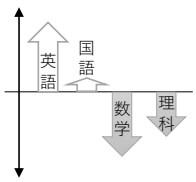

図 1　主　成　分

この主成分は、正方向は「文系科目の評価」、負方向は「理系科目の評価」というように解釈できる。英語が高得点であれば「文系科目の評価」が高いと直感的に理解できるが、数学や理科も英語と同様に高得点の場合には理系科目が「文系科目の評価」を下げる方に働き、現実との当てはまりが悪くなる。そのため、軸として一貫した一つの解釈ができることが望まれるが、正負の主成分負荷量が混在する主成分では困難である。

この解釈しづらさを解消する手段として主成分負荷量を非負値のみで表現する方法がある。主成分負荷量の符号を正に統一することで変数の意味の働く方向を揃えることができるため、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り、解釈のしやすさを向上させることができる。また、個々の変数の意味の働く大きさは主成分負荷量の大きさに比例し、相対的に大きい負荷量をとる変数ほど主成分の解釈において重要度の高い変数であると考えることができる。本稿では、個々の意味を足し合わせたものが全体の解釈と考えることができる性質を「意味の加法性」と呼び、またこの性質を持つ主成分を「加法的主成分」と呼ぶこととする。

上記の例と同様に四科目の試験データ集合の分析の結果、図 2 のような加法的主成分が得られたとする。

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

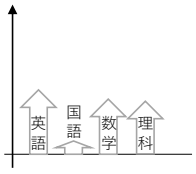

図 2　加法的主成分

先程とは異なり、主成分負荷量の符号が統一されているため意味の働く方向を揃えることができている。この加法的主成分は「理系科目と英語重視」と解釈ができ、国語以外の三科目の点数が高いほどこの主成分での評価は高くなり、現実との当てはまりの良い解釈が行える。このように、意味の加法性を有す

ることで、直感的な理解のできる解釈が行えるようになる。

既存の非負主成分の生成手法には非負値行列因子分解 [1] [2] や Nonnegative PCA [3] [4] などがある。これらと本稿の提案手法の違いとして、フィッシャー情報基準に基づき分散と相関のバランスをとりつつ、主成分分析の主成分と類似した主成分を得る点である。主成分分析の主成分と類似した主成分を生成することで、主成分分析同様に新たな空間での分散（情報量）を算出でき、新たな空間での特徴的な意味を把握できると期待される。このように、本稿ではフィッシャー情報基準に基づいた加法的主成分を生成する方法を提案し、その妥当性を検証する。

本稿では第 2 章で加法的主成分、第 3 章では加法的主成分の抽出、第 4 章では実験結果を示し、第 5 章では結論を述べる。

2　加法的主成分

主成分分析の主成分の固有ベクトルは互いに無相関であるため直交し、また分散が最大となるようにとられている。一方、加法的主成分の方向ベクトルは非負の主成分負荷量から構成されているため主成分の固有ベクトルとは異なるベクトルであり、いくら直交するようにとったとしても分散が最大となることはない。そのため、新たな座標系となる加法的主成分は無相関に近く、可能な限り分散が大きいものであることが望まれる。しかし、分散をもとの主成分より大きくするには無相関の制約を緩める必要があるため最大分散となり、且つ最小相関に抑えたような加法的主成分の抽出は困難である。

そこで本稿では、フィッシャー情報基準 J を式 1 のように定義する。q は更新する方向ベクトル、D は z 変換したデータ集合、R は D の相関行列、Q は方向ベクトル行列、$\|X\|_F^2 = \sum_{i,j} x_{ij}^2$ はフロベニウスノルムの 2 乗を表す。

$$J(q) = \frac{qRq}{\|Q(q)RQ(q)\|_F^2} \tag{1}$$

ただし、方向ベクトル行列 Q が零行列の場合、式 1 の分母が 0 となり計算することができないため、分析を行うデータ集合の各データはいずれかの変数で必ず 0 でない値をとり、またデータ集合の変数に関する相関行列は零行列ではないという条件を満たすものとする。

式 1 は、分母では新たな座標系での相関、分子では方向ベクトルの分散をとる q に関する関数である。そのため、方向ベクトル q から加法的主成分を抽出する際にこの J を用いることで、他の方向ベクトルとの相関と方向ベクトル q の分散を同時に見ることができるため、分散と相関のバランスが取れた加法的主成分を抽出することが期待される。

3　加法的主成分の抽出

本稿では、前節のフィッシャー情報基準 J を目的関数とした最急勾配法による加法的主成分の抽出方法を提案する。その過程を二節に分けて述べる。

3.1　初期の方向ベクトルの作成

主成分と類似した加法的主成分を生成するために、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同様に固有ベクトルとなるため、表 1 の左側の固有ベクトル行列が得られた時、右側のようなもう一つの固有ベクトル行列を考えることができる。

表 1　二種の固有ベクトル行列

	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符号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。しかし、この二種類の固有ベクトルは正負の負荷量が混在したものとなっている。負荷量の符号をすべて正に統一するために、固有ベクトルの主成分負荷量が正のものを残し、負の要素を 0 に置換すると表 2 のような方向ベクトル行列が得られる。

表 2　二種の方向ベクトル行列

	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　フィッシャー情報基準に基づく加法的主成分の抽出

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ッシャー情報基準 J を目的関数とし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、加法的主成分を抽出する。

この基準 J を目的関数として最急勾配法を行うが、J の勾配を計算することは困難であるため、方向ベクトルの値を変化させながら極大値を探索する。例として、表 2 の左側の行列を初期の方向ベクトル行列 Q として選択し、はじめに方向ベクトル $q1$ を更新する場合、表 3 のように方向ベクトル $q1$ の負荷量が 0 の要素の一つに正の定数（例では 0.05）を加えた方向ベクトルを作成する。今回は数学、理科、国語の負荷量が 0 のため三種類の方向ベクトルが作成される。この三種類の方向ベクトルの J の値を求め、最大となるベクトル $q1$-2 に $q1$ を更新する。方向ベクトル $q1$ を更新したため、行列 Q は表 4 のようになる。

表 3　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

表 4　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

残る $q2$、$q3$、$q4$ もランダムに決めた順番で $q1$ と同様に更新を行う。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い、他の方向ベクトルとの関連を見ながら更新する。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方向ベクトルとしてとる主成分を加法的主成分とする。

通常の最急勾配法は正負どちらの方向にも動かし極値となる座標を探索するが、本稿では符号の統一された主成分を生成することを目的としているため正の値の加算、即ち正方向の移動のみを考え、負方向への移動は行わない。

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。

サッカーの第９６回天皇杯全日本選手権は１日、大阪・吹田スタジアムで３万４１６６人を集めて決勝があり、延長戦の末、鹿島が２−１で川崎を降し、６大会ぶり５回目の優勝を果たした。優勝回数５回は、１９９３年のＪリーグ発足以降最多。鹿島はＪリーグ優勝に続く今季二つ目のタイトルで、国内３大会のタイトル総数も１９に伸ばした。…

図 3　記　事　例

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。199 名詞のうち出現頻度の高い主要な名詞を表 5 に示す。

表 5　主要な名詞

優勝	大会	メートル	相手	選手
チーム	決勝	東京	監督	東福岡

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。この相関行列に対して主成分分析を行い、求まる二種の固有ベクトル行列から正の主成分負荷量を残した初期の方向ベクトル行列 P、M を作成する。また、二種類の方向ベクトルから主成分負荷量の最も高い要素を残している方を選択した行列 Q_{one} と、0 に置換した要素の合計が少ない方を選択した行列 Q_{sum} も同時に作成し、四種の初期の方向ベクトル行列を得る。

4.2 評価方法

加法的主成分は意味の加法性を有する主成分であり、すべての変数の負荷量の大きさに応じた意味を足し合わせることで主成分の解釈が行える。そのため、各変数の負荷量の大きさから変数が主成分の解釈にどの程度の影響を与えるのかを判断でき、この影響度合いを解釈影響率、また各解釈影響率を大きいものから順に足し上げたものを累積解釈影響率と本稿では呼ぶこととする。方向ベクトル $q = (w_1, w_2, \cdots, w_n)$ としたときの i 番目の変数の解釈影響率は式 2 のように表される。

$$\frac{|w_i|}{\sum_{m=1}^{n} |w_m|} \times 100 \tag{2}$$

本稿では主成分の解釈のしやすさを次のように定義する。

まず初めに、解釈がしやすくなる要因として考えられるのが、解釈に必要となる変数の個数が少ないことである。少数の変数からの解釈では、変数の内在的な意味や類似点まで考慮する必要がなくなるため直感性のある解釈が行える。解釈影響率が少数の変数に集中しているか、解釈に影響を与えない負荷量が 0 の変数が多いときに解釈に必要となる変数は少なくなると考えられる。このような、より少ない変数で主成分の解釈が可能である性質を「容易性」と本稿では呼び、$k' = 20$ として式 3 のような評価指標で主成分を評価する。

$$\frac{\text{上位 } k' \text{変数を解釈に用いるときの累積解釈影響率}}{\text{主成分負荷量が 0 でない変数の割合}} \tag{3}$$

二つ目に解釈しやすくなる要因として考えられるのは、各主成分で独自の解釈ができることである。主成分の解釈方法のなかで最も広く使われている手法として、interpret-by-top-k [5] という方法があり、これは絶対値をとった主成分負荷量の大きい上位 k 個の変数の持つ意味から主成分の解釈を行い、負荷量の小さい変数については解釈には利用しないという方法である。通常 k は主観的に決定されるが、本稿では評価の際は累積解釈影響率を 70%までの変数を解釈に利用するとし、その個数を k とする。

主成分分析の各主成分は無相関であるため、解釈は個々の主成分で固有のものであることが望まれる。解釈に用いる k 個の変数が他の主成分では使われていない独自のものであれば、解釈し分けることが容易となる。他の主成分と解釈し分けることが可能である性質を「独自性」と名付け、評価を行う主成分で解釈に用いられる変数を x_i としたとき、式 4 のような評価指標で主成分を評価する。

$$\frac{1}{k} \sum_{i=1}^{k} \frac{\text{総主成分数}}{x_i \text{を解釈に用いる主成分数}} \tag{4}$$

4.3 実験結果

四種類の初期の方向ベクトルからフィッシャー情報基準 J に基づいた最急勾配法により加法的主成分を生成する。表 6 に主成分分析の第 2 主成分と、第 2 主成分から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を示す。

表 6　第 2 主成分、加法的主成分の主要名詞

主成分			加法的主成分		
名詞	主成分負荷量	解釈影響率	名詞	負荷量	解釈影響率
トップ	0.196	1.716	主将	0.400	4.193
時間	0.170	1.495	トップ	0.196	2.051
総合	0.170	1.491	時間	0.170	1.786
記録	0.168	1.470	総合	0.170	1.781
2 位	0.157	1.381	記録	0.167	1.757
12 年	0.155	1.364	2 位	0.157	1.650
獲得	0.153	1.340	12 年	0.155	1.629
スタート	0.151	1.324	獲得	0.153	1.601
選手	0.144	1.266	スタート	0.151	1.583
目標	0.136	1.190	選手	0.144	1.512

加法的主成分の主要名詞を見ると、主成分にはなかった「主将」という名詞が最大負荷量をとっている。また、その値も他の名詞に比べ一回り大きく、解釈影響率が高いため、加法的主成分では「主将」という名詞が解釈において重要な意味を持っていると考えられる。そのため、主成分は「上位入賞者の記録」という解釈になるが、加法的主成分は「主将」という名詞に重きを置き「主将の好記録」というように解釈ができる。

表 7　正値に更新された変数とされなかった変数

名詞	主成分負荷量	負荷量	名詞	主成分負荷量	負荷量
主将	-0.022	0.400	準々決勝	-0.112	0
展開	-0.043	0.050	全国	-0.111	0
前半	-0.013	0.050	東海大仰星	-0.109	0
大会	-0.009	0.050	東福岡	-0.101	0

第 2 主成分の初期の方向ベクトルは 199 変数のうち 55 変数が 0 に置換される。そのうち表 7 の左表の 4 変数が最急勾配法により正値に置き換えられ、右表の変数を含む 51 変数は更新されず最終的な加法的主成分においても 0 のままである。

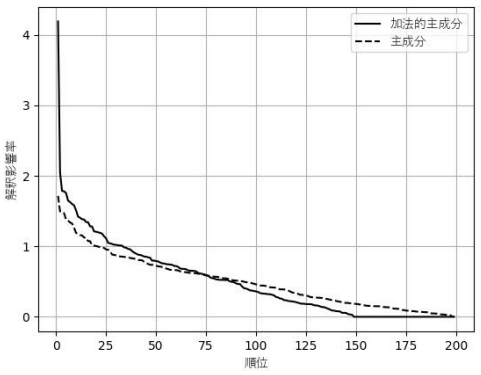

図 4　第 2 主成分、加法的主成分の解釈影響率

図 4 をみると、加法的主成分は主成分分析の主成分に比べて解釈に影響を与えない負荷量が 0 の変数が増加することで、解釈影響率が少ない語に集中している。主成分ではどれも同程度の解釈影響率であるため多くの語から総合的に解釈する必要があるが、加法的主成分では主成分よりも上位の名詞が全体の解釈のうちの大部分を説明しているため、より少ない語から全体の特徴をつかむことが可能である。

第 2 加法的主成分の容易性、独自性の評価を表 8 に、主成分ごとの評価を平均して算出した加法的主成分の方向ベクトル行列全体の評価を表 9 に示す。

表 8　第 2 加法的主成分の評価

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

表 9　方向ベクトル行列の評価

	主成分	P	M	Q_{sum}	Q_{one}
容易性	2.483	6.908	6.141	6.047	6.197
独自性	1.183	5.429	5.440	5.241	5.467

どの方向ベクトル行列でも第 2 加法的主成分と同様に「容易性」、「独自性」のどちらにおいても評価値の向上がみられ、主成分よりも加法的主成分は解釈しやすいものとなっている。

4.4　考　　察

フィッシャー情報基準 J を目的関数とすることで無相関に近く、分散が大きくなるような加法的主成分を得る工夫をした。実際に生成された主成分の分散と主成分間の相関を表 10 に示す。

表 10　加法的主成分の相関と分散

		P		M		Q_{sum}		Q_{one}	
総分散		732.997		743.846		751.317		751.662	
相関係数	$\lvert 0.75\rvert \sim \lvert 1.00\rvert$	0	(0.00%)	0	(0.00%)	0	(0.00%)	0	(0.00%)
	$\lvert 0.50\rvert \sim \lvert 0.75\rvert$	2233	(11.33%)	3517	(17.85%)	2807	(14.25%)	3198	(16.23%)
	$\lvert 0.25\rvert \sim \lvert 0.50\rvert$	301	(1.53%)	513	(2.60%)	841	(4.27%)	680	(3.45%)
	$0 \sim \lvert 0.25\rvert$	17167	(87.14%)	15671	(79.54%)	16053	(81.48%)	15823	(80.32%)

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八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最急勾配法により意味の加法性を有する加法的主成分の生成方法を提案した。得られた非負の加法的主成分は負荷量が 0 の要素の増加や解釈に大きな影響を与える変数の出現により、少数の語から解釈が行え、また主成分ごとに独自の解釈ができるという観点から主成分分析の主成分よりも解釈のしやすい主成分となった。

文　　献

[1] 澤田 宏, “非負値行列因子分解 NMF の基礎とデータ／信号解析への応用 “, 電子情報通信学会誌 Vol.95, No.9, 2012.

[2] 亀岡 弘和, “非負値行列因子分解とその音響信号処理への応用 “, 日本統計学会誌 第 44 巻, 第 2 号, pp. 383-407, 2015.

[3] Ron Zass, Amnon Shashua, “Nonnegative Sparse PCA“, Proc. Neural Information and Processing Systems, 2006.

[4] Andrea Montanari, Emile Richard, “Non-negative Principal Component Analysis: Message Passing Algorithms and Sharp Asymptotics“, IEEE Trans. Inf. Theory, vol. 62, no. 3, pp. 1458–1484, 2016.

[5] Dan Vilenchik, Rarak Yichye, Maor Abutbul, “To Interpret or Not to Interpret PCA? This Is Our Question“, Proceedings of the Thirteenth International AAAI Conference on Web and Social Media, 2019.

[6] 長井慶成, 三浦孝夫, “意味の加法性を有する主成分 “, 2022 年情報処理学会全国大会, 2022.