文学作品の統計的分析

2002MM079 酒井　美貴
指導教員　田中　豊

1　はじめに

現在、世の中には数多くの書物が存在しており、学校の教科書などでいくつかの作品を目にしたことがあるであろう。書物には小説や随筆、評論などの種類以外にも、著者の年齢や性格、好む作家などによりさまざまな系統に分類することができる。

ここでは、因子分析、主成分分析などの多変量解析を用いて、因子数や分析方法による結果の比較を行った。

2　データについて

現代数学レクチャーズ D-2・因子分析法の現代作家 100 人の作品 100 編を対象とした調査結果を用いた [?]。追加データとして、2003 年から 2005 年の芥川賞受賞作家の作品について同じ基準で調査を行った。作品は『蛇にピアス』『蹴りたい背中』『介護入門』『グランド・フィナーレ』『土の中の子供』である。

調査項目は『直喩』『声喩』『色彩語』『文の長さ』『会話文』『句点』『読点』『漢字』『名詞』『人格語』『過去止』『現在止』『不定止』『名詞の長さ』『動詞の長さ』の 15 項目である。

3　因子分析

現代作家 100 人の作品 100 編の調査データを用い、因子分析を行った。最尤推定・バリマックス回転により因子負荷量を推定した。因子の解釈については、[1] にならって解釈をすると同じ意味を持つ 3 つの因子が得られた。

- 第　因子は『名詞』『漢字』『人格語』の三つに因子負荷量の大きい。『叙述の題材に関係する因子』と名づける。
- 第　因子は色彩語』『直喩』『声喩』などに因子負荷量の大きい。『文章の修飾に関する因子』と名づける。
- 第　因子は『会話文』『句点』に因子負荷量が大きい。『文章の会話に関する因子』と名づける。

続いて、回帰推定法により因子得点を求めた。因子得点の散布図を図 1 に示す。

因子得点のプロットをみると、37、39、44 の 3 人が外れていることがわかる。37 の横光利一『日輪』と 39 の横井孝作『無限抱擁』は第　因子得点が大きいので、『叙述の題材に関する因子』の特徴を持っている作品と言える。44 の平林たい子『施療室にて』は第　、第　因子得点ともに正ではあるが第　因子得点の方が大きく、『文章の修飾に関する因子』の特徴が強いと見られる。

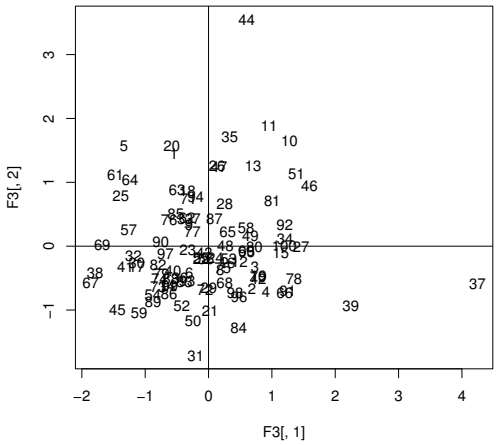

図 1　第　因子と第　因子の因子得点のプロット

4　因子数の推定を変えたときの因子分析の結果との比較

主成分分析と異なり、因子分析は因子数を多くとったときその結果は因子数を少なくとったときの結果を含まないことが知られている。因子数を 3、4、5 と推定した因子分析の結果を比較した。以下では因子数 3 と 4 の場合の比較の結果を示す。

因子数 3 の第　因子、第　因子、第　因子の因子得点を被説明変数、因子数 4 の第　因子から第　因子の因子得点を説明変数として、因子数 3 と指定したときの 3 つ因子の情報が因子数 4 と指定したときの 4 つ因子にどのくらい含まれているかを検討するために、重回帰分析を行った。それぞれの決定係数 R^2 は第　因子 0.9904、第　因子 0.9520、第　因子 0.9890 であった。

3 因子モデルの第　因子、第　因子、第　因子ともに R^2 の値が高く、いずれも値は 0.95 以上で 3 因子の情報が 4 因子で説明できていると言える。

因子数 3 と因子数 4 の因子得点の相関係数を見ると、因子数 3 の第　因子と因子数 4 の第　因子は 0.9945 でとても高い。因子数 3 の第　因子と因子数 4 の第　因子との相関は 0.8102 でこれも高いが、第　因子との相関も 0.4968 で無視できない。因子数 3 の場合の第　因子の情報の大部分は因子数 4 の第　因子に含まれるが一部は第　因子にも含まれていることがわかる。因子数 3 の第　因子と因子数 4 の第　因子は 0.9932 で非常に高くなっている。因子数 3 の因子得点は因子数 4 の第　因子から第　因子の因子得点と強い相関があることがわかる。

個々の因子の対応関係を調べるために因子数 3 および 4 と指定した場合の因子得点の相関係数を求めると次の結果が得られた。

表 1　因子数 3 と因子数 4 の因子得点の相関係数

	第　因子	第　因子	第　因子	第　因子
第　因子	0.995	0.018	-0.001	-0.043
第　因子	0.062	0.810	-0.063	0.497
第　因子	-0.004	0.005	0.993	0.037

因子数 3 の 3 因子の因子数 5 の 5 因子に対する重回帰の R^2 の値はそれぞれ第　因子 0.9886、第　因子 0.9495、第　因子 0.8530 であり、第　因子の R^2 がやや小さかった。因子の対応関係を調べると、第　、第　因子にそれぞれ 1 つの因子との相関が 0.9910、0.9160 と高かったが、第　因子は 2 つの因子との相関が 0.7033、0.6014 となり 2 つの因子に分かれていた。

5　因子分析と主成分分析の結果の比較

因子数とほぼ同じ目的で主成分分析が利用されることがある。そこで因子分析と主成分分析の結果の比較を行った。因子数 3 の第　因子、第　因子、第　因子の因子得点を被説明変数、主成分分析の第一主成分から第五主成分の主成分得点を説明変数として、因子数 3 と指定したときの 3 つの因子の情報が 5 つの主成分にどれくらい含まれるかを重回帰分析を行い検討した。

それぞれの決定係数 R^2 は第　因子 0.8784、第　因子 0.9813、第　因子 0.7493 であった。第　因子と第　因子の R^2 の値は 0.8 以上の値をとり、第　因子は他の 2 つに比べればやや低い値ではあるが 7 割以上の値であるので 3 因子の情報は 5 つの主成分でかなりの程度説明できていると言える。因子数 4、5 の場合も同様の検討を行った結果、因子数 5 は R^2 の値が 0.7344 や 0.7619 など少し低めではあるがどれも 7 割以上の値であった。しかし、因子分析に比べると決定係数は少し低く相関係数の対応もはっきりとは見られなかった。

6　新しい調査データを加えた分析

新しく調査した 5 例を加えたデータの因子分析 (3 因子) を行い、[1] の 100 例の分析結果と比較した。対応する因子得点間の相関係数は 0.9882、0.9918、0.9990 となり、どれも高い値をとっていて 100 例のデータとほぼ同じ結果が得られた。

新しく調査した 5 例のプロットを見ると、101、102、105 の第　因子の因子得点はどれも負の値をとり、『名詞』『漢字』『人格語』に因子負荷量の大きい『叙述の題材に関係する因子』の特徴は他の作品に比べると小さいことがわかる。また第　因子得点も 103、104 が負の値をとり、『文章の修飾に関する因子』を表す第　因子の特徴は小さいと言えるだろう。追加した 5 作品の中では、101 の『蛇にピアス』102 の『蹴りたい背中』105 の『土の中の子供』は『会話文』が多く『句点』も多い傾向にある。一方、103 の『介護入門』104 の『グランド・フィナーレ』は『句点』が少なく文章が長いことがわかる。

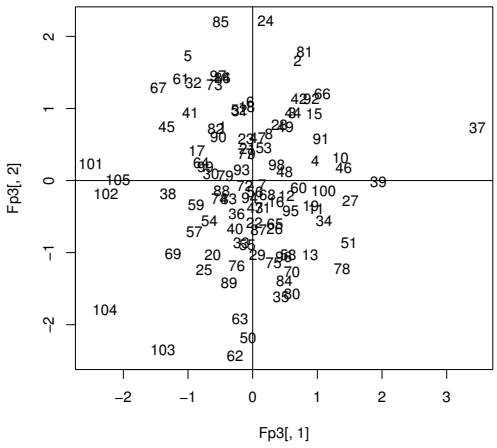

図 2　第　因子と第　因子の因子得点のプロット

7　まとめ

本研究では [1] の 100 例のデータおよび新しく調査した最近の 5 つの作品を加えた 105 例に対して因子分析を行い、1) 最近の作品がどのような特徴を持っているか、2) 因子分析において因子数の推定を変えると分析結果がどう変わるか、3) 同じ目的で主成分分析を行ったときの分析結果の比較、などについて検討した。

1）については第　、第　因子得点が負になるものも見られ、第　、第　因子の特徴が小さいことがわかった。5 作品の中では『会話文』『句点』の数により 101、102、105 と 103、104 に結果が分かれた。

2) については因子数を 4 や 5 に増やすと、意味付けが難しくなったり、重回帰分析を行ってみても 3 因子モデルの結果が 4 因子モデルの結果に含まれていた。4 因子モデルでは第　因子として『文章の結びに関する因子』が新たに加わった。

3) については得られた因子得点の情報は主成分得点の情報でほぼ説明できているという結果が得られたが、因子分析の結果や相関係数の対応と比べると、やや劣るという結果となった。

参考文献

[1] 安本美典、本田正久：現代数学レクチャーズ D-2『因子分析法』、培風館 (1981)

[2] 金明哲 (Jin Mingzhe)：フリーソフトによるデータ解析・マイニング、統計と情報の専門誌『ESTRELA』連載