

テキストマイニングを志す学生のための 「対応分析ロジック理解ドリル」の活用の試み

飯 野 秋 成¹

1. はじめに

当会の以前の年次大会において、筆者は、テキストマイニングの分析に用いられることの多い対応分析（コレスポンデンス分析）による研究事例を紹介する機会を頂戴した。筆者の担当するCG作品制作の授業において、学生に書かせた振り返り作文を分析した結果を話題提供したが¹⁾、計算のしくみや結果の解釈の方法等、手法に広く興味をもてただける機会となったのでは、と感じている。その後新たに、学内インターンシップの振り返り作文をテキストマイニングにより可視化した論文²⁾なども公表するに至っている。

いずれも、研究室の卒業生や大学院修了生たちの多くが関わったが、このような研究を足掛かりにデータサイエンスやビッグデータ解析などの分野で活躍したい、という学生も出てきている。うれしい限りではあるが、タイパ・コスパを重視する風潮ゆえか、理屈を飛ばして、早急に出力を得ることだけを意識しすぎる彼らの傾向が、ときに気になっている。統計解析パッケージソフトのロジックをブラックボックスのままで良しとするのは、この分野で身を立てたいにしては、やや心もとない。

本稿では、これまで筆者らが、テキストマイニング関連研究を志す初学者たる学生諸君に試行してきた「対応分析ロジック理解ドリル」を紹介したい。出力には計算ミスはないか。入力データにバグは含まれていないか。そもそも、自身が設定した問題に合致した計算手法を適用しているか。そのような「嗅覚」を養うべく、しばしば学生に課している1つの演習の試みである。

2. 例題の設定

「対応分析ロジック理解ドリル」の設計の基本方針は、①A3～A4判のシートに、誘導形式の複数の問を並べるとともに、解答を書き込むスペースを十分確保する、②計算量はさほど多くないものとする、③四則演算のための電卓利用は制限しない、としている。

<例題>

インターンシップから帰った3名の学生の振り返り作文を分析した結果、主に3つの単語が頻出した。表1は、そのクロス集計結果である。以下の設問に沿って計算を進めることにより、対応分析の結果を導くとともに、その意味を考察せよ。

3名の学生と3つの単語、と単純化することにより、後述するとおり、手計算でも対応可能なところまで計算量を絞り込んでいる。

表1 3名の学生がインターンシップの振り返り作文に用いた単語

	現場 b_1	デスクワーク b_2	苦労 b_3	計
学生 a_1	0	1	1	2
学生 a_2	1	0	1	2
学生 a_3	1	0	0	1
計	2	1	2	5

実務上は、インターンシップを実施する場合の1クラスの人数は数十人以上となるであろうし、少々の振り返り作文を書かせるだけでも数十～数百程度の単語が出てくる。単語の「分かち書き」を自動化し、あるいは、規模の大きい行列計算などを考えれば、専用のソフトウェア利用はもはや必須である。しかし、自ら計算手続きを行えるぐらいまで単純化することによって理論の筋が見えやすくなっている点を、以降で詳しく説明したい。

表1中の「1」は、振り返り作文内にその単語を1度以上使った、ということを示す。例えば、学生 a_3 君は「現場」という言葉を文章中に何度か使っているかもしれないが、クロス集計表上は「1」としている。3回使っていれば「3」とする考え方もあるが、各学生の言葉への思い入れや作文能力、語彙力などの影響を受けてしまう可能性もある。このため、テキストマイニングの専用アプリケーションソフトの多くは、「分かち書き後の単語は複数回カウントしない」ことがデフォルトとなっている³⁾。

3. 誘導形式の問の設定と解答例

問1 相関係数 $r=a$ と b の共分散／（ a の標準偏差）（ b の標準偏差）を数式化せよ

$a_1 \sim a_3$ の合計（正確には、 $2a_1+2a_2+a_3$ ）と、 $b_1 \sim b_3$ の合計（同じく $2b_1+b_2+2b_3$ ）が、それぞれ0となると考えれば、

$$r = \frac{(a_1b_2 + a_1b_3 + a_2b_1 + a_2b_3 + a_3b_1)/5}{\sqrt{(2a_1^2 + 2a_2^2 + a_3^2)/5} \sqrt{(2b_1^2 + b_2^2 + 2b_3^2)/5}} \quad \text{式(1)}$$

この立式には、少々統計をかじっていれば議論の余地はないが、学生のレベルによっては、統計の基礎の追加レクチャが求められる場合もある。

問2 a の分散＝1および b の分散＝1を、数式で示せ

$$(2a_1^2 + 2a_2^2 + a_3^2)/5 = 1$$

$$(2b_1^2 + b_2^2 + 2b_3^2)/5 = 1$$

以上まとめて式(2)

1 いいのあきなる 新潟工科大学工学部工学科 〒945-1195 新潟県柏崎市藤橋1719

問3 問2の条件の下で、問1の式を最大化することを考える。ラグランジュの未定乗数法を用いて、適切な数式を作成せよ

$$F = (a_1b_2 + a_1b_3 + a_2b_1 + a_2b_3 + a_3b_1)/5 \\ - \frac{1}{2}\lambda\{(2a_1^2 + 2a_2^2 + a_3^2)/5 - 1\} \\ - \frac{1}{2}\lambda\{(2b_1^2 + b_2^2 + 2b_3^2)/5 - 1\}$$

式(3)

ラグランジュの未定乗数法は大学の初等数学で学ぶが、これが役立つ貴重な機会が提供できる。最大化したい数式(=共分散の数式)から、ゼロになるべき数式の λ 倍(λ は実数、もし変数が2系統ならそれぞれの0.5 λ 倍)を差し引いた、数式Fを準備する。

ポイントは、続く問5～6の作業において、Fの前半(=共分散の数式)の偏微分が、Fの後半(=分散の数式-1)の偏微分の実数倍と一致する、という立式をさせるところにある。2つの関数のそれぞれの接線方向が一致するところに極値が存在する、という発想は、立体をイメージすればきわめて自然である。この「気持ち」の解説は、高校～大学初年次向けのYouTube動画⁴⁾にも詳しく、学生にも大いに視聴することを薦められる。

問4 問3の式を $a_1 \sim a_3$ 、 $b_1 \sim b_3$ で偏微分せよ

指示通り偏微分を行い、順次示せば、

$$\frac{\partial F}{\partial a_1} = (b_2 + b_3 - 2\lambda a_1)/5$$

$$\frac{\partial F}{\partial a_2} = (b_1 + b_3 - 2\lambda a_2)/5$$

$$\frac{\partial F}{\partial a_3} = (b_1 - \lambda a_3)/5$$

$$\frac{\partial F}{\partial b_1} = (a_2 + a_3 - 2\lambda b_1)/5$$

$$\frac{\partial F}{\partial b_2} = (a_1 - \lambda b_2)/5$$

$$\frac{\partial F}{\partial b_3} = (a_1 + a_2 - 2\lambda b_3)/5$$

以上まとめて式(4)

問5 問4の式を全て0とおき、 $a_1 \sim a_3$ を消去して、 $b_1 \sim b_3$ と λ の関係式を3つ作成せよ

式(4)の前半3つから、

$$a_1 = (b_2 + b_3) / (2\lambda)$$

$$a_2 = (b_1 + b_3) / (2\lambda)$$

$$a_3 = b_1 / \lambda$$

以上まとめて式(5)

これらを、式(4)の後半3つに代入すれば、

$$\frac{3}{2}b_1 + \frac{1}{2}b_3 - 2\lambda^2 b_1 = 0$$

$$\frac{1}{2}b_2 + \frac{1}{2}b_3 - \lambda^2 b_2 = 0$$

$$\frac{1}{2}b_1 + \frac{1}{2}b_2 + b_3 - 2\lambda^2 b_3 = 0$$

以上まとめて式(6)

問6 問5の式を、行列を用いて書き直せ

$$\begin{pmatrix} \frac{3}{2} - 2\lambda^2 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} - \lambda^2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 - 2\lambda^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

式(7)

問7 問6の左辺の行列の行列式が0となるような λ 2を2つ求めよ

$b_1 \sim b_3$ が零ベクトルではないことを前提としているので、式(7)の成立には、左辺の 3×3 行列の行列式(ヤコビアン)が0であることが求められる。 3×3 行列の行列式の求め方は、サラスの公式として知られる。詳細を省いて結果だけを示せば、以下のとおり。

$$\left(\frac{3}{2} - 2\lambda^2\right)\left(\frac{1}{2} - \lambda^2\right)(1 - 2\lambda^2) - \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2} - \lambda^2\right) \\ - \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{3}{2} - 2\lambda^2\right) = 0$$

式(8)

これを整理して、

$$16\lambda^6 - 28\lambda^4 + 13\lambda^2 - 1 = 0$$

式(9)

因数分解して整理すると、

$$(\lambda^2 - 1)(16\lambda^4 - 12\lambda^2 + 1) = 0$$

式(10)

$\lambda^2 > 0$ であることに注意して、後半の4次方程式=0から、

$$\lambda^2 = \frac{3 \pm \sqrt{5}}{8} \quad (= 0.6545 \dots, \text{および } 0.0955 \dots)$$

式(11)

なお、 $\lambda^2 = 1$ は、式(7)において $b_1 = b_2 = b_3$ の結果を招くので、採用されない。

問8 2つの λ^2 に対応する $a_1 \sim a_3$ および $b_1 \sim b_3$ を求めよ

ここからは、ロジックよりは数値自体がほしいところで、適宜電卓を用いてよい。式(7)に λ^2 を代入して整理して、 $\lambda^2 = 0.6545$ と 0.0955 の場合を順に示せば、

$$a_1 = -1.1441, 0.4370$$

$$a_2 = 0.4370, -1.1441$$

$$a_3 = 1.4142, 1.4142$$

$$b_1 = 1.1441, 0.4370$$

$$b_2 = -1.4142, 1.4142$$

$$b_3 = -0.4370, -1.1441$$

以上まとめて式(12)

問9 大きい方の λ^2 の $a_1 \sim a_3$ と $b_1 \sim b_3$ を横軸に、小さい方の λ^2 の $a_1 \sim a_3$ と $b_1 \sim b_3$ を縦軸にとり、2次元空間に配置せよ

問8の結果をグラフ化すれば、図1のようになる。学生 a_1 はデスクワークでやや苦勞をし、学生 a_2 は現場でやや苦勞したのだろうか。学生 a_3 は現場にいながら、苦勞には引っ張られていない位置に置かれる結果となった。

問10 固有値 λ ごとの寄与率を求めよ

2乗値のまま扱ってきた λ を、あらためて正の値として求めれば0.8030または0.3090。 λ は固有値とよばれ、点のバラツキの最も大きい軸方向において最大となる。また、全ての固有値の総和に対する、各固有値の割合が寄与率で、一般に、図1内に示したように大きい順に整理して示す。固有値は、少ない方のサンプルサイズから1を減じた数だけ得られるので、今回は2つとなる。一般に、寄与率の大きいもののほど軸の意味を説明しやすい。

問11 グラフ上の単語の位置の特徴から、横軸と縦軸の意味を考察せよ

この最終ステージでは、高所かつ多角的な視点の判断が求められる。横軸は「(左) デスクワーク (右) 現場実習」の軸が透けて見える。縦軸は、下に行くほど「苦勞」の度合いが大きい、もし、「楽しい」「充実」「大変」などの他のキーワードも同時布置されたなら、より興味深い考察が得られる可能性もあるだろう。

4. まとめ

ドリル形式のテキストマイニング演習の手法を提案した。分散と共分散、ラグランジュの未定乗数法、サラスの公式、因数分解、固有値と寄与率までの一連の数学の知識を総ざらいしながら、軸の解釈に至るまでを、鉛筆書きでトレーニングできる。筆者も何度となくドリルを繰り返しながら、「では、こういう設定にしたらどうなのだろう」と、思考のバリエーションを広げていくきっかけにもなった。研究室で大学院を修了した学生たちにも一度はトライさせてきた。その学習効果はたどりようもないが、結果的に、テキストマイニング系の論文をしつかり書いて卒業・修了しているところを見ると、効果はゼロではなかったのだろう。

お気づきの方もいるかもしれないが、実は、表1の数値の設定は、文献⁵⁾における数量化III類の説明事例に完全に合わせてある。当ドリルを完遂した後に、あらためてこの名著をたどれば、主成分分析や数量化III類と計算の進め方が全く同じであることも理解できるはずである。

例題の設定を発展させて、より多くの学生の振り返り作文が得られたとしたならば、学生の「カテゴリー」に基づく再集計も可能となる。表2は、実習先の業種ごとに、作文上のキーワードの出現回数をクロス集計表にまとめた例である。対応分析の計算方法を論じた文献の多くは、このようなケーススタディをスタートラインとしている。このように数字が混んでくると手計算はかなりシンドイので、エクセルシートなどをつかって計算を進めることをお勧めしたい。ヒントとして、問1相当の回答を示せば、

$$r = \frac{(24a_1b_1 + 3a_1b_2 + 15a_1b_3 + 6a_2b_1 + 15a_2b_2 + 9a_2b_3 + 10a_3b_1 + 6a_3b_2 + 3a_3b_3)/91}{\sqrt{(42a_1^2 + 30a_2^2 + 19a_3^2)/91}\sqrt{(40b_1^2 + 24b_2^2 + 27b_3^2)/91}}$$

式 (13)

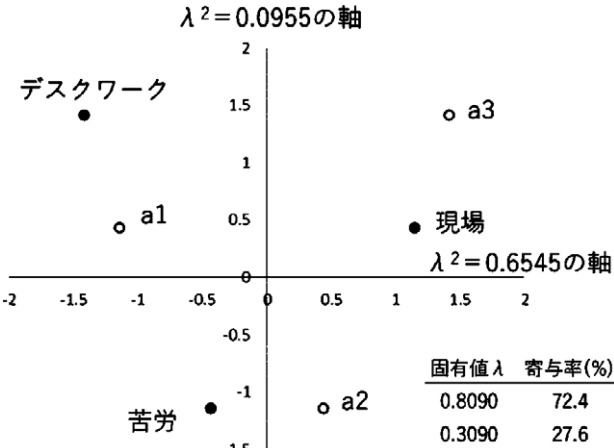


図1 学生3名とキーワード3つの布置結果

表2 インターンシップ先の業種別に集計した場合の例

	現場 b_1	デスク ワーク b_2	苦勞 b_3	計
ゼネコン a_1	24	3	15	42
設計事務所 a_2	6	15	9	30
流通 a_3	10	6	3	19
計	40	24	27	91

さらに、途中の過程を省いて、問6相当の回答を示せば、

$$\begin{pmatrix} \frac{576}{42\lambda + 30\lambda + 19\lambda} - \frac{36}{42\lambda + 30\lambda + 19\lambda} - \frac{100}{42\lambda + 30\lambda + 19\lambda} - 40\lambda & \frac{72}{42\lambda + 30\lambda + 19\lambda} - \frac{90}{42\lambda + 30\lambda + 19\lambda} - \frac{60}{42\lambda + 30\lambda + 19\lambda} & \frac{360}{42\lambda + 30\lambda + 19\lambda} - \frac{54}{42\lambda + 30\lambda + 19\lambda} - \frac{30}{42\lambda + 30\lambda + 19\lambda} \\ \frac{72}{42\lambda + 30\lambda + 19\lambda} - \frac{90}{42\lambda + 30\lambda + 19\lambda} - \frac{60}{42\lambda + 30\lambda + 19\lambda} & \frac{9}{42\lambda + 30\lambda + 19\lambda} - \frac{225}{42\lambda + 30\lambda + 19\lambda} - \frac{36}{42\lambda + 30\lambda + 19\lambda} - 24\lambda & \frac{45}{42\lambda + 30\lambda + 19\lambda} - \frac{135}{42\lambda + 30\lambda + 19\lambda} - \frac{18}{42\lambda + 30\lambda + 19\lambda} \\ \frac{360}{42\lambda + 30\lambda + 19\lambda} - \frac{54}{42\lambda + 30\lambda + 19\lambda} - \frac{30}{42\lambda + 30\lambda + 19\lambda} & \frac{45}{42\lambda + 30\lambda + 19\lambda} - \frac{135}{42\lambda + 30\lambda + 19\lambda} - \frac{18}{42\lambda + 30\lambda + 19\lambda} & \frac{225}{42\lambda + 30\lambda + 19\lambda} - \frac{81}{42\lambda + 30\lambda + 19\lambda} - \frac{9}{42\lambda + 30\lambda + 19\lambda} - 27\lambda \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

式 (14)

なお、表2の情報だけでは「分析対象の学生数」は知りえない。意外な気もするかもしれないが、注意が必要である。

参考文献

- 1) 池田将, 飯野秋成, 渡邊壮一; 創作系演習授業の振り返りにみられる工学部学生の学びの志向の分析, 工学教育, 第68巻, 第3号, pp.20-26, 2020.2
- 2) 飯野秋成, 涌井将貴, 日下部征信, 渡邊壮一, 黒木宏一, 荒川卓; 企業人との対話をもたらす工学部学生の学びの意識の変容, 工学教育, 第68巻, 第6号, pp.48-54, 2020.12
- 3) テキスト・マイニング研究会配付資料(Sep.2006).p.5, https://www.wordminer.org/wp-content/uploads/2013/04/63_14.pdf, 2024.1.12アクセス
- 4) 制約付き最適化問題 (KKT条件/ラグランジュ未定乗数法), 予備校のノリで学ぶ「大学の数学・物理」, <https://www.youtube.com/watch?v=bdWTCq98H5c>, 2024.1.12アクセス
- 5) 有馬哲, 石村貞夫; 多変量解析のはなし, 東京書籍, pp.243-258, 1989