主成分分析

渡邊直樹

2021 年 10 月 6 日

手許に，m 個の質問項目について，たとえば 1 から 10 までの水準で各々 n 人の個人に回答してもらったデータがあるとしよう．質問項目が 2 つか 3 つならば，各質問項目に対する回答の平均と分散から何がしかの判断ができるかもしれない．しかし，質問項目が多い場合には（平均や分散などの）基礎統計を見たところで，データ全体の特徴を把握することは難しい．そこで，質問項目に対する回答データの特徴を 2 つか 3 つの総合的特性に集約することを考える．たとえば，

$$\text{特性 1} \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3 + \cdots + \hat{\beta}_5 X_5 \tag{1}$$

$$\text{特性 2} \quad = \quad \beta^*_0 + \beta^*_2 X_2 + \beta^*_6 X_6 + \cdots + \beta^*_m X_m \tag{2}$$

といった形で 2 つの総合的特性にデータの特徴を集約することができれば，2 次元の平面にこれらの特性をプロットすることで，データの特徴を可視化することもできる．

このように，多くの変数で表されるデータの特徴をいくつかの総合的特性にまとめる手法が主成分分析（principal component analysis）であり，データ全体のばらつきを最も高い割合で特徴づけている総合的特性を第 1 主成分，2 番目に大きな割合で特徴づけている総合特性を第 2 主成分という．主成分分は，数学的には，データの分散共分散行列または相関行列の固有値と固有ベクトルを求める作業に基づくデータ処理技法である．（どちらの行列に基づく処理を行うかによって，結果は異なる．）

各主成分には，それを表す式（たとえば，(1) 式と (2) 式）における変数の特徴から，その総合的特性にふさわしい名前を分析者自身がつける．個人 i について，(1) 式と (2) 式に変数として含まれる各質問項目に対して彼または彼女が回答した水準の値を代入したものを個人 i の主成分得点という．平面においてデータの特徴を可視化するには，n 人の個人の第 1 主成分と第 2 主成分の主成分得点の組をプロットする．

EZR (Easy R) で主成分分析を行う手続きは至って簡単である．データを読み込み，EZR のリボンに表示されているボタンを，統計量＞次元解析＞主成分分析の順に押していき，ポップアップウィンドウにてデータセットに書き込む主成分数を選択し、OK ボタンを押すだけである．次の表は EZR（Easy R）による主成分分析の出力の一部であり，8 項目の質問に対する回答データが 3 つの総合的特性，つまり，主成分（Comp. 1，Comp. 2，Comp. 3）に集約されている．

表 1: EZR の出力

Component loadings:

	Comp.1	Comp.2	Comp.3
attribute	0.4219189	0.2536643	0.056982628
compitency	-0.2793772	0.6240533	-0.203224201
discretion	-0.3421693	0.5269354	-0.003101245
firmscale	0.4125529	0.1459368	0.223776636
permanent	0.4221065	0.2278132	0.171543540
salary	0.2270193	0.2420679	-0.680518801
seniority	0.4104076	0.2493843	0.144524509
workhour	-0.2418820	0.2730662	0.626056663

Component variances:

Comp.1	Comp.2	Comp.3
4.5595113	1.0686546	0.9466214

Importance of components

	Comp.1	Comp. 2	Comp. 3
Standard deviation	2.2755309	1.0167839	0.90687098
Proportion of Variance	0.6125104	0.1222941	0.09728347
Cumulative Proportion	0.6125104	0.7348045	0.83208796

この表から，第 1 主成分は

$$\begin{aligned}\text{Comp. 1} \;=\; & 0.4219189 \times (\text{attribute}) - 0.2793772 \times (\text{compitency}) \\ & -0.3421693 \times (\text{discretion}) + 0.4125529 \times (\text{firmscale}) \\ & +0.4221065 \times (\text{permanent}) + 0.2270193 \times (\text{salary}) \\ & +0.4104076 \times (\text{seniority}) - 0.2418820 \times (\text{workhour})\end{aligned}$$

であると読み取れる．(1) 式と (2) 式では $(X_1, \ldots, X_m$ と表されている変数は次の各質問項目に対応しており，就業においてその項目をどれだけ重視するかが数値で回答されている．年功序列（seniority）: 就業年数によって確実に賃金が上がること，終身雇用（permanent）: 終身雇用制度が保証されていること，労働時間（workhour）: 労働時間が短いこと，給与（salary）: 給与が高いこと，企業規模（firmscale）: ある程度の規模の企業であること，自由裁量（discretion）: 仕事の計画や予定を自分で決められること，能力（compitency）: 自分の能力を活かせること，帰属意識（attribute）: 会社に一体感を感じること．

Standard deviation (標準偏差) は variance (分散) の正の平方根である．Proportion of Variance (寄与率) は対応する主成分がデータ全体のばらつきのうちどれくらいを説明できるかを表しており，表からは第 1 主成分の寄与率は 0.6125104，第 2 主成分の寄与率は 0.1222941 であることが読み取れる．Cumulative Proportion of Variance（累積寄与率） は，文字通り，第 k 主成分までの寄与率を足し合わせたものであり，表からは第 1 主成分と第 2 主成分でデータ全体のばらつきの約 73.5%（0.7348045）を説明していることが読み取れる．累積寄与率が 70%から x80%以上あれば，それをもたらす主成分でデータ全体の大部分を取りまと目ていると考えられるだろう．

Appendix

EZR（Easy R）を用いた主成分分析の手順をここにノートしておく．以下では，Mac 版 EZR での操作について説明するが，Windows 版 EZR のリボンに表示されている「標準メニュー」ボタンを押すと，Mac 版のリボンに表示されている選択項目が格納されているので，Windows 版でも同じ操作が可能である．ここでは，自分の計算機にデータセット（data2.csv）が保存されているとして，それを EZR で読み込むことにする．

1. データの読み込み : リボンに表示されているボタンを次の順で押す．
    - データ＞データのインポート＞テキストファイルまたはクリップボード，URL から
    - ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す．デフォルトでは「空白」が選択されている．(データファイルの場所はデフォ

ルトでローカルファイルシステムが選択されているので，そのままにしておく.）

- 新たなポップアプウィンドウが現れるので，そこで data2.csv を選択し、Open ボタンを押す.
- ホーム画面上で「データセットを表示」ボタンを押して，データが読み込まれていることを確認する.

2. 主成分分析：リボンに表示されているボタンを次の順で押す.

   - 統計量＞次元解析＞主成分分析
   - ポップアップウィンドウの「データ」の欄にて 2 つ以上の変数を選択する.
   - オプションで「データセットの主成分得点を保存」の項目にチェックを入れておく.
   - デフォルトでは「相関行列の分析」の項目にチェックが入っているので，それを外すと，分散共分散行列の分析になる.
     （相関行列での分析では，データの下処理が必要となることがある.）
   - 「適用」ボタンを押す.
   - 新しいポップアップウィンドウが現れるので，保存する主成分得点の数をスライダーで選択する.
   - OK ボタンを押す.

3. 分析結果の保存と主成分スコアを書き加えたデータセットの出力：リボンに表示されているボタンを次の順で押す.

   - ファイル＞出力ファイルに保存. ポップアップウィンドウでファイルに名前をつけ，計算機に csv 形式で分析結果を保存する.（EZR の画面にも分析結果は出力されるが，やや見にくい.）
   - データ＞アクティブデータセット＞アクティブデータセットのエクスポート.（アクティヴデータセットの保存を押すと，RData 形式で保存されてしまうので，Excel などで読み込めない.）ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す. デフォルトでは「空白」が選択されている.

以上のボタン操作などをコマンドラインからコマンドを打ち込むとなると，次のようになる．これらはEZRの画面に表示される．RStudioでRを使う場合にはこれらを自分で打ち込むことになる．かなり面倒である．

```
Dataset <- read.table("/Users/*****/Documents
/data2.csv",
header = TRUE, sep = ",", na.strings = "NA", dec = ".",
strip.white = TRUE)
\item local({
    .PC <- princomp(~attribute + compitency + discretion
    + employtype + firmscale, cor = FALSE, data = Dataset)
    cat("\nComponent loadings:\n")
    print(unclass(loadings(.PC)))
    cat("\nComponent variances:\n")
    print(.PC$sd^2)
    cat("\n")
    print(summary(.PC))
    Dataset <<- within(Dataset, {
        PC3 <- .PC$scores[, 3]
        PC2 <- .PC$scores[, 2]
        PC1 <- .PC$scores[, 1]
    })
})
```

（注）上記コマンドにおいて，*****には計算機のユーザ名が入る．筆者は自分の計算機の書類フォルダ（Documents）にデータセットを置いているので，そこからdata2.csvを読み込んだ．