第１２章　テキストの特徴と話題分析

茨城大学情報工学科
倉持辰洋

はじめに

- 複数のテキストについて分析を行う際にはテキストの特徴についての分析を必要とする場合がある

- この章では「用いられてる要素」、「テキストとの関連性」といった情報で分析する方法を紹介する

- 分析例として次の二つのテキストを用いる

  -「日本造船学会論文集　第193号」の13編の論文概要

  ⇒ファイル名をzousenXXとする

  -「情報処理学会平成19年度論文賞」の9編の論文概要

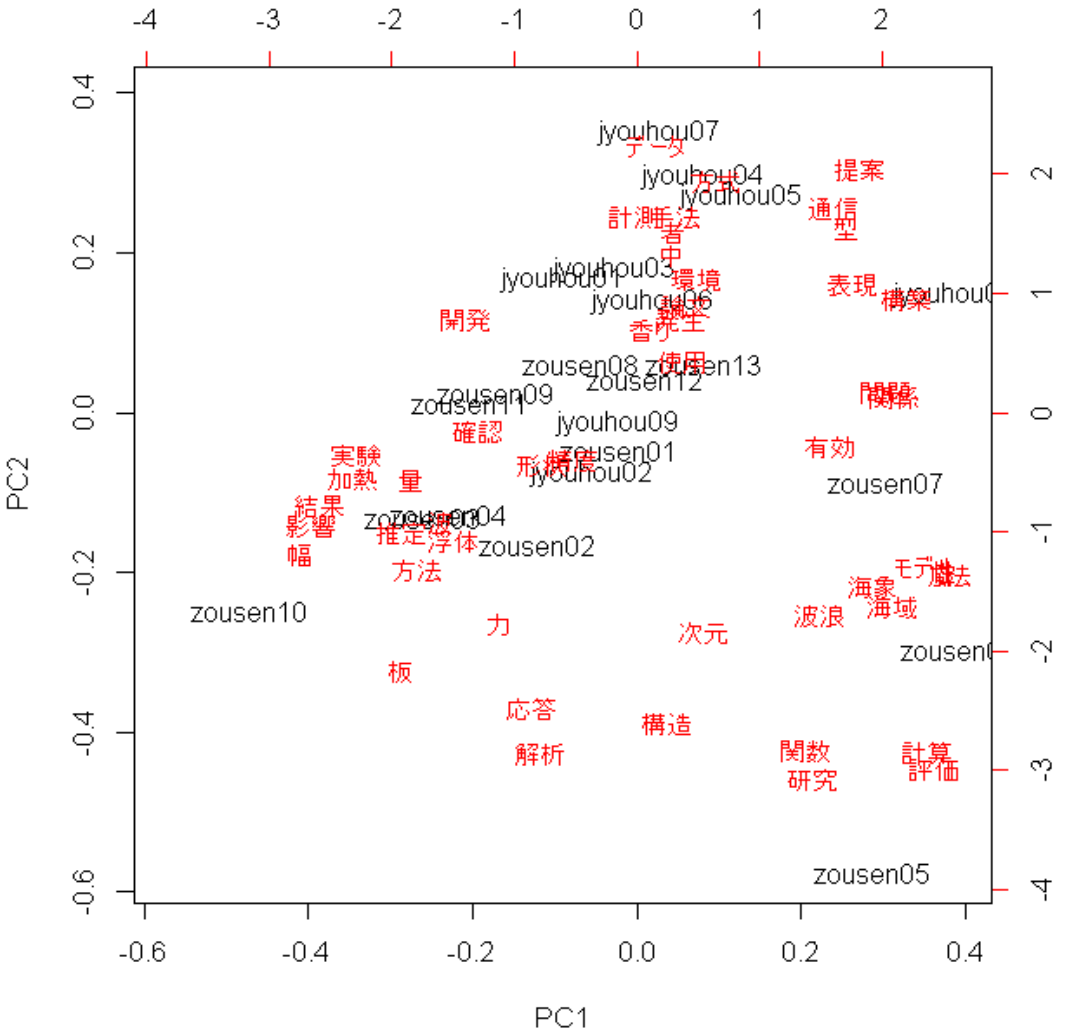
  ⇒ファイル名をjyouhou0Yとする
- 使用するデータは次のサイトより取得できる

  http://mjin.doshisha.ac.jp/iwanami/data/gaiyou.csv

特徴の分析の例(図A)

- 主成分分析によって作成された図
- 黒字がテキスト名

  赤字が名詞
- テキスト名の近くにある名詞はそのテキストで多く見られるものとなっている
- y軸でみるとjouhouは上部の方に固まり、zousenは下のほうへとなっている



データの形式と行列の表記（1）

- テキストの特徴を分析する際に用いるもの
  ⇒文字列、形態素、文節、句、共起関係など
- テキストが複数有るのでテキスト毎に要素を集計する
- 次の表Bの1行目をベクトル$x=\{x_1,x_2,\cdots,x_k,\cdots,x_p\}$とし
  2行目以降をn行p列の行列$X_{n\times p}=\{x_{ik}\}$といったデータ形式にする

ファイル名	提案	手法	加熱	応答	構造	結果	・・・	解析	情報
jyouhou01	1	3	0	0	0	1	・・・	0	0
・	・	・	・	・	・	・	・	・	・
jyouhou09	3	0	0	0	2	1	・・・	0	0
zousen01	0	0	0	0	0	0	・・・	0	0
・	・	・	・	・	・	・	・	・	・

データの形式と行列の表記（2）

- 前スライドの表をそのまま用いて分析することも可能

  ⇒ただし、分析の方法によってはデータを変換して分析に使う

  例）相対頻度（比率、百分率）

  TF-IDF（語の重み）

主な分析方法

- データの固体(=$X_{n\times p}$)と変数(x)の関連性を分析する方法でテキストの特徴を求められる
- 主に使われてる方法

  -主成分分析

  -対応分析　　⇒多変量データ解析の方法

  -因子分析
- これらの方法はデータに変換と処理を加える

  ⇒その際固有値と固有ベクトルを求める技法を利用

固有値と固有ベクトル（１）

- データ行列Xに対しXA=λAが成り立つとき

  λ：Xの固有値

  ゼロではないベクトルA：固有ベクトル

- 例) $X=\begin{bmatrix}4 & 3\\ 1 & 2\end{bmatrix}$ $\lambda=5,\quad A=\begin{bmatrix}0.9486\\ 0.3126\end{bmatrix}$ $\lambda=1,\quad A=\begin{bmatrix}-0.7071\\ 0.7071\end{bmatrix}$

- 固有値が複数ある場合もあり、大きい順に並べ第１固有値、第２固有値・・・とする

固有値と固有ベクトル（２）

- 収集したデータ表は正方行列とは限らない

  ⇒データの分散共分散行列、相関係行列、距離行列は正方行列となり固有値等を求められる
- データ$X_{n\times p}=[x_{ik}]$の分散共分散行列と相関性行列の一般式を次からのスライドで紹介する
- v_{kk}は変数kの分散、v_{hk}は変数hと変数kの共分散である

分散共分散行列の一般式

$$\Sigma=[v_{hk}]=\frac{1}{n-1}X^tX=\begin{bmatrix} v_{11} & & & & \\ v_{21} & v_{22} & & & \\ v_{31} & v_{32} & v_{33} & & \\ \vdots & \vdots & \ddots & \ddots & \\ v_{p1} & v_{p2} & v_{p3} & \cdots & v_{pp} \end{bmatrix}$$

$$v_{kk}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ik}-\bar{x_{ik}}\right)^2, \qquad v_{hk}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ik}-\bar{x_{ik}}\right)\left(x_{ik}-\bar{x_{ih}}\right),$$

$$\bar{x_{ik}}=\frac{1}{n}\sum_{i=1}^{n}X_{ik}, \qquad \bar{x_{ih}}=\frac{1}{n}\sum_{i=1}^{n}x_{ih}$$

相関係数行列の一般式

$$R=[r_{hk}]=\begin{bmatrix} 1 & & & & \\ r_{21} & 1 & & & \\ r_{31} & r_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{bmatrix}, \qquad r_{hk}=\frac{v_{hk}}{\sqrt{v_{hh}}\sqrt{v_{kk}}}$$

- データ$X_{n\times p}$を式 $\dfrac{xij-\bar{x}_j}{sj}$

  で標準化したデータの分散共分散行列は相関係数行列と等しい

主成分分析

- 図A作成するためのデータ処理方法で最もシンプルな方法が主成分分析
- どのようなもの？

  ⇒データの分散共分散か相関の情報に基づいて多くの変数のデータの情報を少ない変数へと集約して分析する方法
- 図Aでは名詞（＝変数）を2変数に集約したデータの散布図になっている

主成分分析の考え方（1）

- データ$X_{p\times h}$を$Z_{n\times m}$に射影、その際

  -Zの変数が相互に無関係

  -変数の分数（または相関）が最大

  こうなる軸を求める
- なぜ分散最大、相関最大とするのか？

  ⇒データのばらつきが大きいほどデータを考察しやすいからである

主成分分析の考え方(2)

$$X_{n\times p}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} \quad \Rightarrow \quad Z_{n\times m}=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2k} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{i1} & z_{i2} & \cdots & z_{ik} & \cdots & z_{\Im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} & \cdots & z_{nm} \end{bmatrix}$$

もし線形的な射影をおこなう前提ならデータと射影空間上の関係は

$$z_{ik}=a_{i1}x_{i1}+a_{i2}x_{i2}+\cdots+a_{ip}x_{ip}$$

となる。この式を線形結合とよぶ

- 主成分分析は係数aをどのように求めるかが大事。

  係数行列を$A_{p\times m}$とすると線形結合式は$Z_{n\times m}=X_{n\times p}A_{p\times m}$となる

Rで主成分分析を実行

```
データの読み込み
> gaiyou <- read.csv("gaiyou.csvのあるディレクトリ",head=T,row.names=1)

主成分分析および図の作成
> gaiyou.pc <- prcomp(gaiyou[,1:50], scale=TRUE)
> biplot(gaiyou.pc, var.axes = FALSE)
```

- これにより図Aが作成できる

- scale=TUREで相関係数行列、
  scale=FALSEで分散共分散行列を用いる

主成分と寄与率・累積寄与率（1）

- 前のスライドで実行した主成分分析では50変数であったデータを2変数に集約しているので、データの損失がどれくらいあるかを把握する必要がある

  ⇒寄与率と累積寄与率より読み取ることができる
- 寄与率：$\lambda_j / \sum_{i=i}^{k} \lambda_i$

  ある固有値が全ての固有値の中でどれだけ占めているか
- 累積寄与率：第1固有値の寄与率から順番に累積していったもの

主成分と寄与率・累積寄与率（2）

- 主成分分析で主成分とは固有ベクトルのこと

- 何番目の主成分まで用いて分析を行うべきなのか？

  -分散共分散を用いる場合累積寄与率が80%前後になるまで

  -相関係数行列を用いる場合は固有値が1前後

Rで寄与率と累積寄与率を示す

```
> summary(gaiyou.pc)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.5076	2.3764	2.2617	2.20348	2.12069	1.84136	1.71696
Proportion of Variance	0.1258	0.1129	0.1023	0.09711	0.08995	0.06781	0.05896
Cumulative Proportion	0.1258	0.2387	0.3410	0.43812	0.52807	0.59588	0.65484
	PC8	PC9	PC10	PC11	PC12	PC13	
Standard deviation	1.60236	1.5183	1.37427	1.33903	1.21228	1.14279	
Proportion of Variance	0.05135	0.0461	0.03777	0.03586	0.02939	0.02612	
Cumulative Proportion	0.70619	0.7523	0.79007	0.82593	0.85532	0.88144	
	PC14	PC15	PC16	PC17	PC18	PC19	
Standard deviation	1.10414	1.03871	0.96602	0.92646	0.87337	0.73043	
Proportion of Variance	0.02438	0.02158	0.01866	0.01717	0.01526	0.01067	
Cumulative Proportion	0.90582	0.92740	0.94606	0.96323	0.97849	0.98916	
	PC20	PC21	PC22				
Standard deviation	0.55083	0.48866	3.449e-16				
Proportion of Variance	0.00607	0.00478	0.000e+00				
Cumulative Proportion	0.99522	1.00000	1.000e+00				

- １行目の２乗＝固有値

  ２行目が寄与率

  ３行目が累積寄与率
- 固有値１前後を目指す場合は15までだが現実的でない⇒上位数個でやる

主成分得点

- 主成分得点とはデータ行列の行のスコア
- 行のデータと固有ベクトルとの線形結合であり、その関係は次のようになる

  -第１固有ベクトル（主成分）: $a_1, a_2, \cdots, a_p$

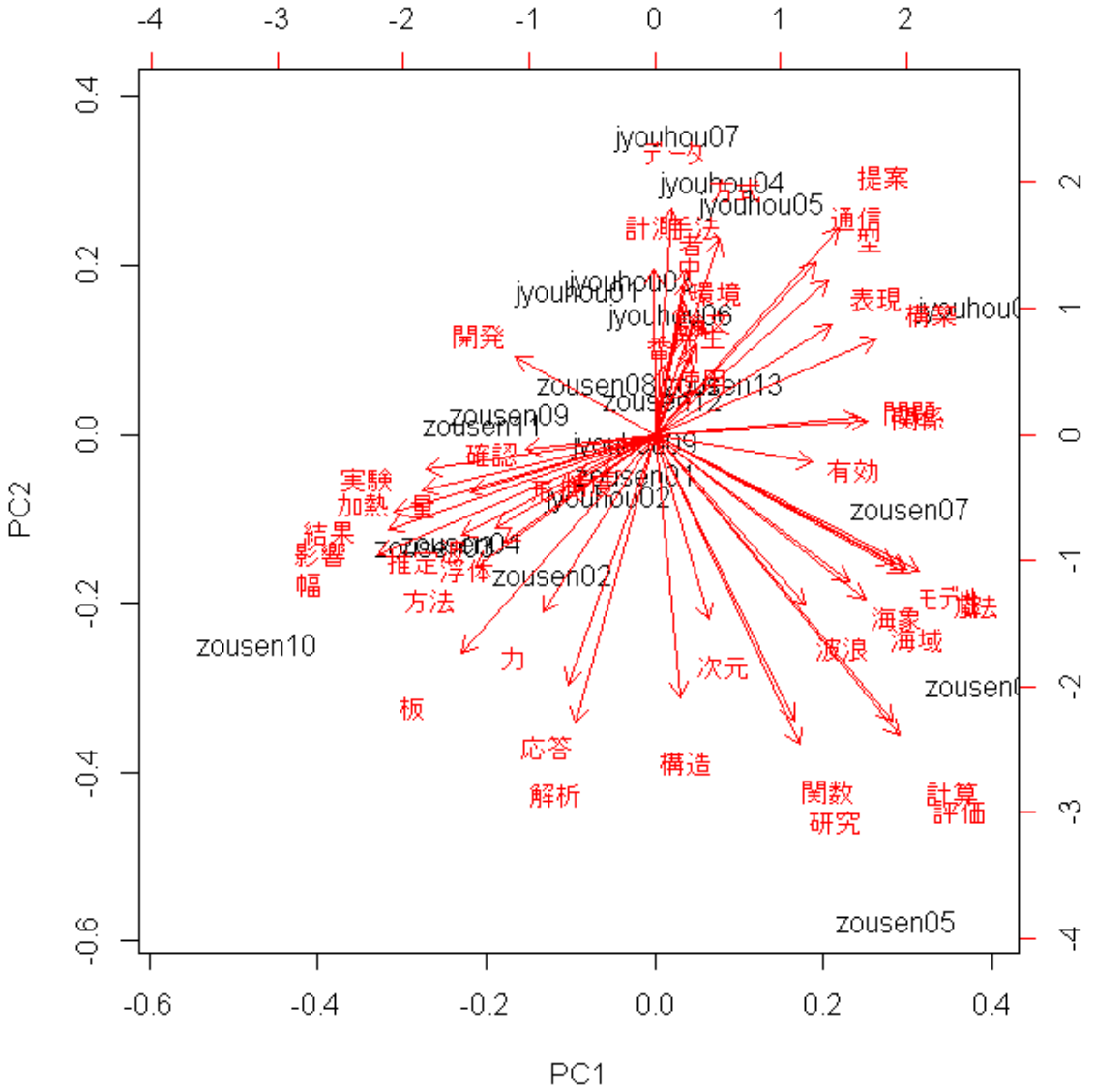
  -第i行のデータ: $x_{i1}, x_{i2}, \cdots, x_{ip}$

  -第i行の第１主成分得点:

$$z_{i1} = a_1 x_{i1} + a_1 x_{i2} + \cdots + a_1 x_{ip}$$

パイプロット図

- 第１、第２主成分の散布図と第１、第２主成分得点の散布図を重ねたもの
- Rでは$xに主成分得点が記録されコマンドは次のようになる

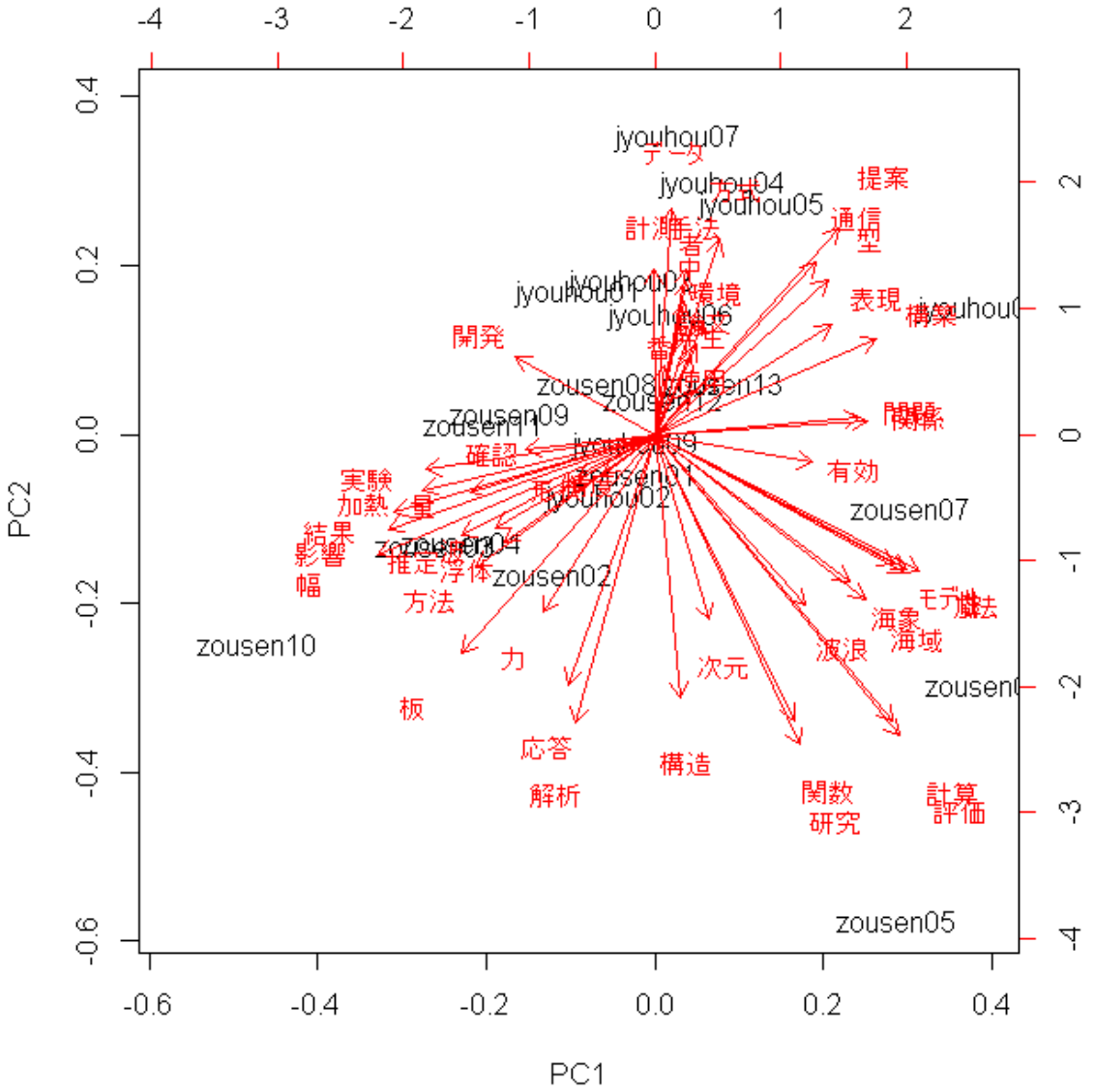


パイプロット図

- 第１、第２主成分の散布図と第１、第２主成分得点の散布図を重ねたもの
- Rではコマンドは次のようになる

```
bioplot(gaiyou.pc)
```

- テキストと変数の数が多いと考察には使いづらい

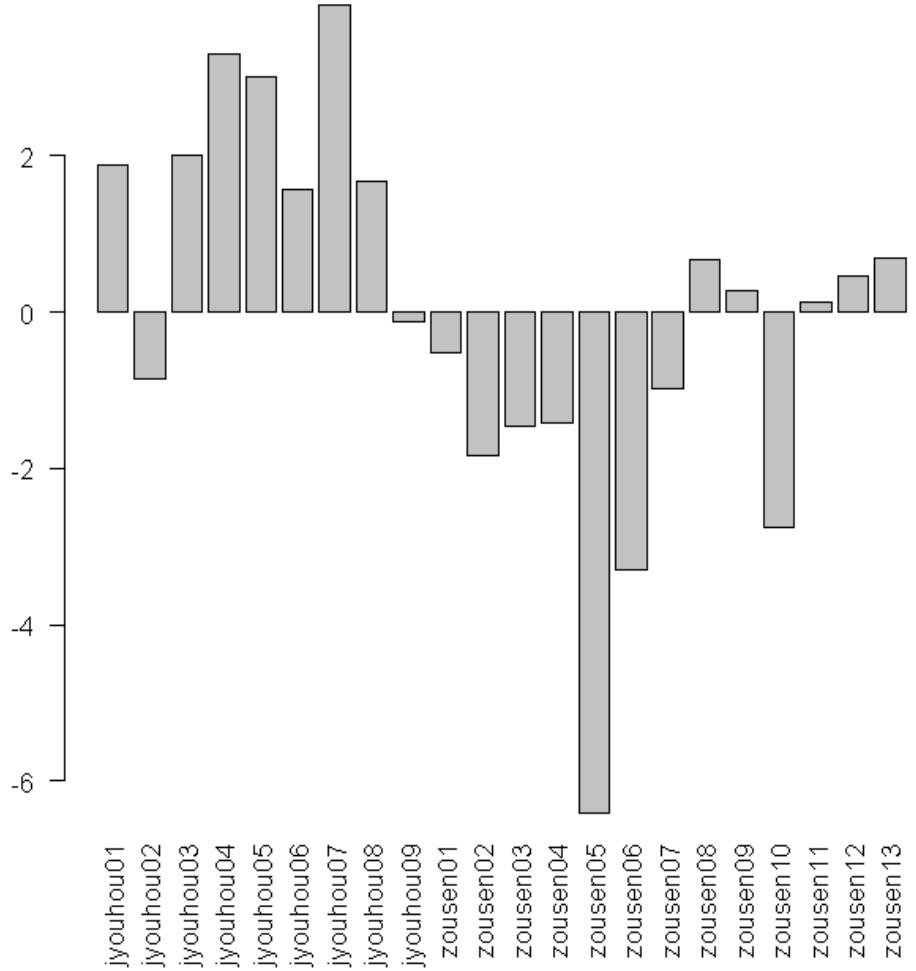


棒グラフを用いた考察（1）

- 一つの主成分と主成分得点を用いた棒グラフ
- 右図は第2主成分を用いたものでRでのコマンドは次のようになる

```
> barplot(gaiyou.pc$x[,2], las=2)
```

- 正の値の多くがjyouhou
  負の値の多くがzousen
  となっている

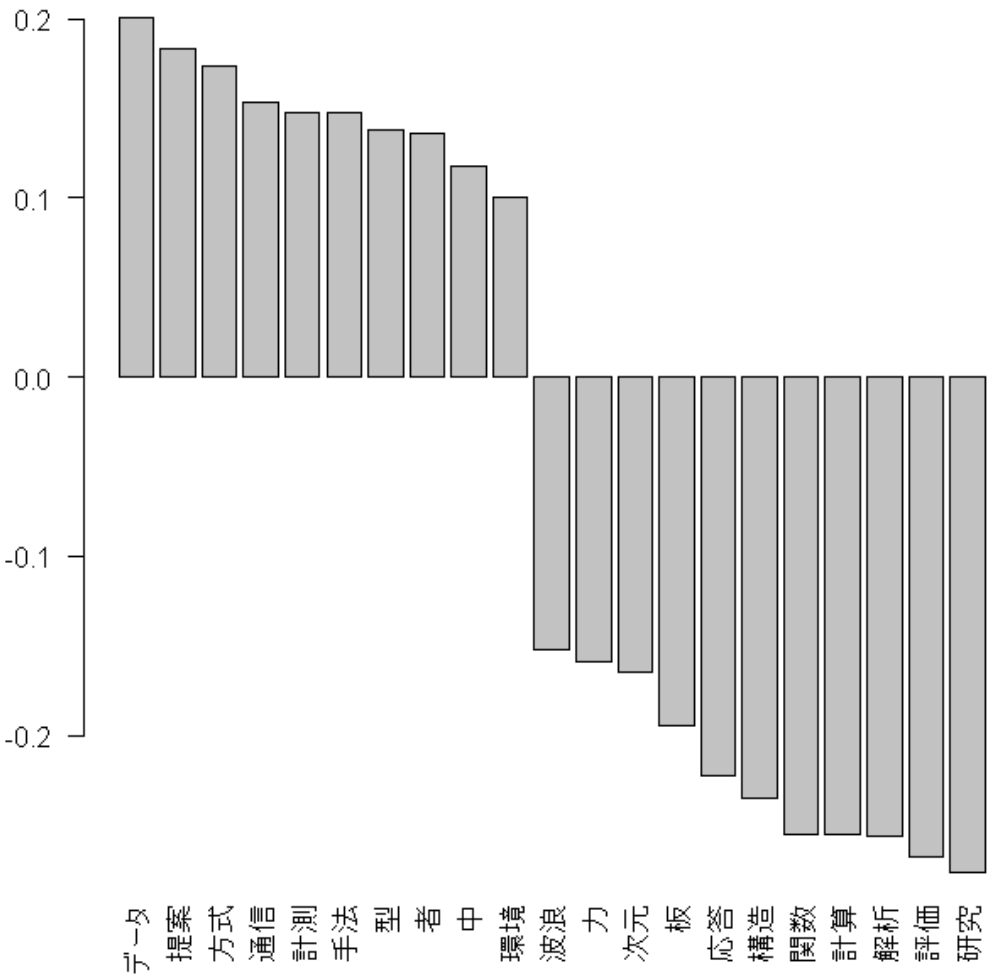


棒グラフを用いた考察（2）

- 二つの論文概要でどのような名詞が多いかを考察するには主成分の棒グラフを使う

```
> y <- gaiyou.pc$rotation[,2]
> y1 <- y[sort.list (y,dec=TRUE)]
> barplot(y1[c(1:10,40:50)],las=2)
```

- 主成分の値をソートし値の上位10個による棒グラフ
- 棒グラフの正負は一つ前のスライドの図に対応している



対応分析（1）

- 対応分析もよく使われる手法である
- 対応分析はデータの形式で名前が変わり、アルゴリズムも変わる

  -対応分析：度数データ

  ⇒この章で利用してるデータはこちら

  -多重対応分析：データが0,1、または順序尺度

対応分析（２）

- 度数データの$F_{r\times c}=[f_{ij}]$のカイ２乗統計量を用いる
カイ２乗統計量の各セルの値の平方根は

$$\chi_{ij}=\sqrt{n}\frac{f_{ij}-f_{i+}f_{+j}/n}{\sqrt{f_{i+}f_{+j}}} \quad = \quad \sqrt{n}\frac{p_{ij}-p_{i+}p_{+j}}{eqrt\, p_{i+}p_{+j}}$$

f_{i+}は$F_{r\times c}$のi行の合計、f_{+j}はj列の合計

nはデータの総度数

p_{ij}は総度数を基準とした総対数

p_{i+}は$P_{r\times c}[p_{ij}]$のi行の合計、p_{+j}はj列の合計

対応分析(3)

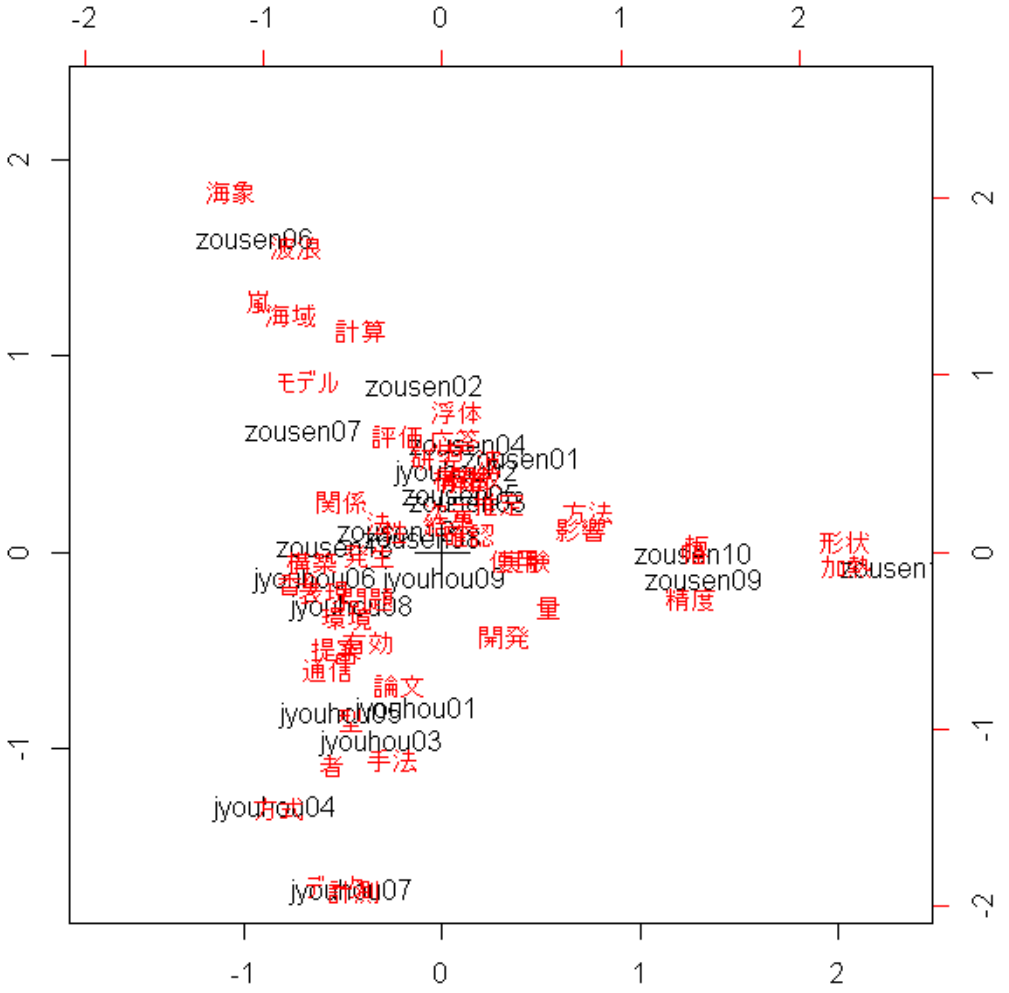
- 対応分析ではZ=[z_{ij}]を用いる、データからの変換は次の式となっている

$$z_{ij} = \frac{f_{ij} - f_{i+} f_{+j} / n}{\sqrt{f_{i+} f_{+j}}} = \frac{p_{ij} - p_{i+} p_{+j}}{eqrt\ p_{i+} p_{+j}}$$

Rで対応分析による図の作成

- パッケージ「MASS」が必要、コマンドは次のようになる

```
> gaiyou.coa <- corresp(gaiyou[,1:50],nf=2)
> biplot(gaiyou.coa)
```



潜在的意味解析（１）

- 行列の特異値分解（SVD：singular value decomposition)を用いて高次元のデータを低次元に射影する方法
- 正方行列でないn×mの行列を次のように3つの行列に分解する

$$X_{n\times m}=U_{n\times k}\Lambda_{k\times k}V_{k\times m}$$

U：左特異行列　　V：右特異行列　　Λ：特異値行列

- Λは対角行列になっており、要素が特異値で大きい順に並んでいる

潜在的意味解析（２）

- 左特異行列、右特異行列、特異値が主成分分析における主成分得点、主成分、固有値に対応している

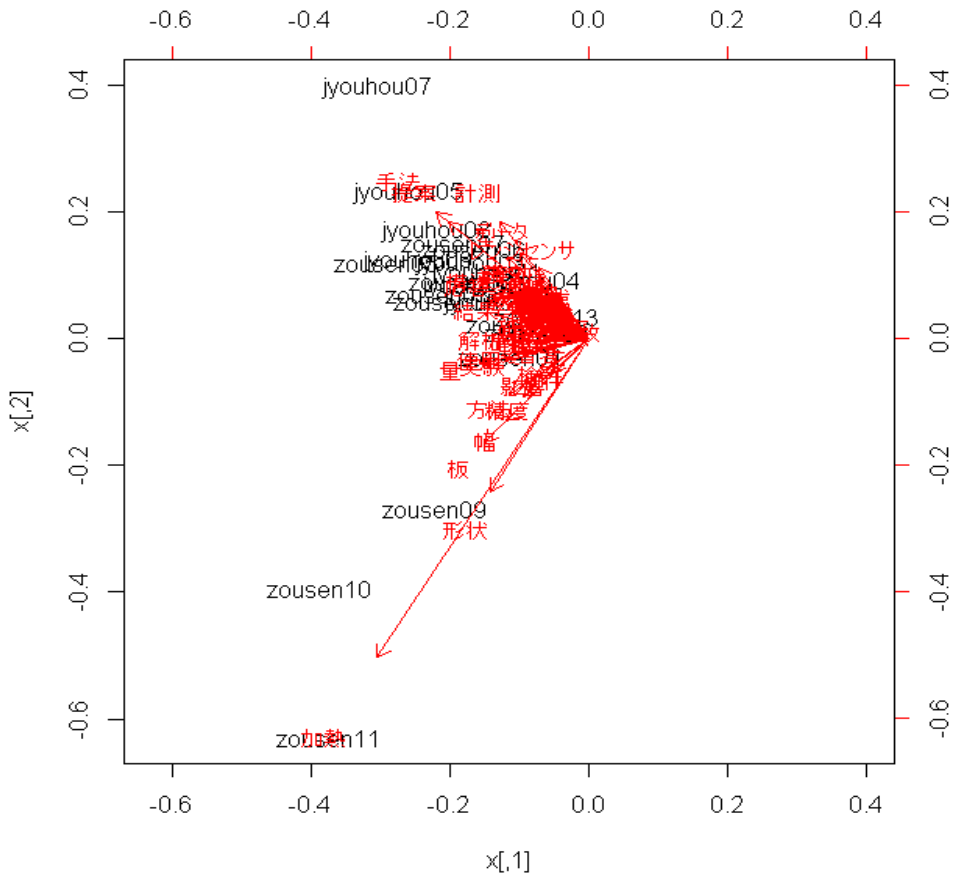
- １～３次元で分析をする場合UとVの第１～第３の値（＝ベクトル）を用いる
- 特異値は固有値の問題と同様に扱える

Rで潜在的意味解析による作図

- 特異値分解の散布図がV字型となっている

  ⇒しばしば起こることで好ましい結果ではない

- これは対応分析でもおきることがありたい処方についても研究されている




```
> gaiyou<-gaiyou/apply(gaiyou,1,sum)
> gaiyou.svd<-svd(gaiyou[,1:67])
> rownames(gaiyou.svd$u)<-rownames(gaiyou)
> rownames(gaiyou.svd$v)<-colnames(gaiyou[,1:67])
> biplot(gaiyou.svd$u[,1:2],gaiyou.svd$v[,1:2])
```

その他の方法

- 因子分析

  ⇒データ間の相関関係を用い、関連性が強い項目を一つの因子としてまとめる
- 独立成分分析

  ⇒信号分析における信号とノイズを分離する方法だがデータ解析にも用いられることがある
- カーネル法によるデータ解析