Rによる主成分分析

**主成分分析は、データの分散共分散、あるいは相関の情報に基づいて多くの変数のデータの情報の損失を抑えながら少ない変数に集約して分析する方法（金　2009, *Estrela* 179, p.46）**

**列に（高頻度）語、行にテキストからなる頻度表データを作成し、Rで主成分分析を行う。**

> nnfreq.df <- read.table( “F:\\CL\_work\\data\\nnfreq.dat”, header=T)

> head(nnfreq.df,3)

author textname man time thing way hand eye day face nothing head

1 carrol alicewonderland 5 68 79 54 31 36 27 20 33 60

2 carrol huntingsnark 12 6 11 5 6 5 15 4 3 5

3 carrol lookingglass 18 80 84 72 68 45 22 22 28 66

moment something word life house year place sir night door woman room girl

1 31 18 31 13 19 3 10 6 3 32 2 13 7

2 6 18 17 8 19 4 5 6 3 2 2 13 7

3 43 23 28 7 9 6 12 6 9 16 2 13 3

friend side name course voice people father sort mind boy anything morning

1 5 21 12 25 51 13 2 22 9 3 20 5

2 6 21 4 2 1 1 1 22 4 3 20 5

3 4 33 33 35 45 14 1 18 10 2 17 4

world hour child mother lady matter fact round end foot home part air

1 7 6 20 0 3 3 8 28 16 28 4 3 14

2 7 1 20 1 3 2 2 28 2 1 1 1 1

3 5 6 18 1 4 3 7 33 17 21 5 7 15

person gentleman minute heart light case window kind idea business arm

1 5 0 32 3 0 5 8 5 15 7 18

2 5 0 2 2 2 3 8 5 1 7 18

3 4 3 30 3 3 3 6 12 13 3 18

manner point fellow table question thought water paper right evening money

1 3 0 4 17 21 6 5 4 7 5 0

2 1 1 1 17 2 1 5 2 2 2 0

3 6 1 1 21 13 16 10 5 4 2 2

work ship country sea letter reason wife sight sense everything book hair

1 1 0 1 13 4 10 1 10 3 12 8 7

2 2 8 1 1 4 1 1 1 1 12 1 4

3 3 1 7 5 2 5 1 7 2 5 11 20

bed week street look bit fire tone rest family back silence brother wall

1 2 4 0 6 15 3 42 8 1 5 12 1 0

2 2 6 0 1 15 3 4 2 1 5 3 1 0

3 4 4 1 8 11 8 41 6 1 9 16 8 5

doubt chair love companion line

1 3 0 0 0 3

2 2 1 0 0 1

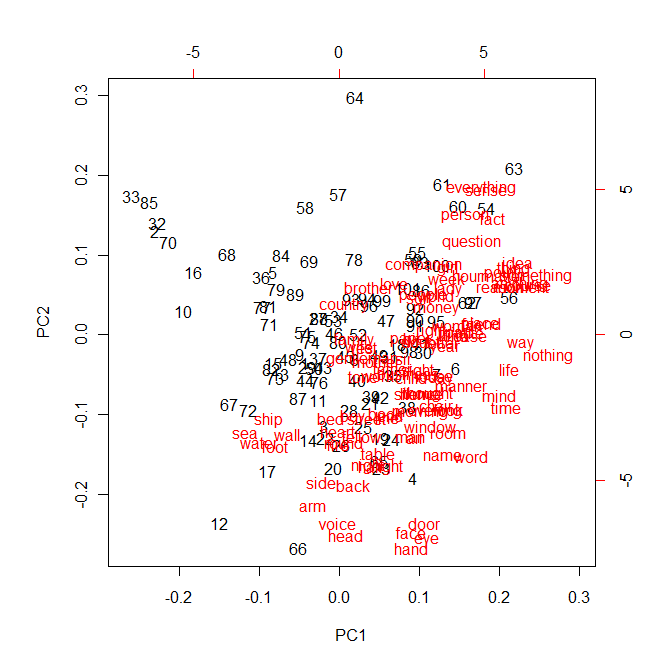
3 2 6 1 1 1

**主成分分析には、分散共分散または相関係数を用いる方法がある。相関情報を用いる場合には、相関が最大になる方向の軸を第１主成分、これと直交し、かつ次に相関が大きい方向の軸を第２主成分とする。この方法は、データの相関係数行列の固有値と固有ベクトルを求める問題に帰する。**

**#関数prcomp() で主成分分析を行い、関数biplot()で第１成分と第２成分を軸にして作図を行う。**

**> nn.pr <- prcomp(nnfreq.df[,3:102], scale=TRUE)　　# 標準化を行い、相関情報を使うときはscale=T**

**> biplot(nn.pr, var.axes=F )**



**２変数に集約した場合、もともとの個数の変数のデータの何パーセントの情報が、この２変数に含まれているか**

**⇒寄与率、累積寄与率から読み取る**

**任意の固有値λ***i* **が、固有値の合計（ λ***i* **）の中に占める割合（λ***j* **/ λ***i* **、あるいは百分率）を寄与率と呼ぶ。**

**通常、固有値は大きいものから降順（λ***i* **＞ λ***i* **＋ 1）に並べられている。第１固有値の寄与率から順番に累積したものを累積寄与率と呼ぶ。**

**主成分分析では、固有ベクトルを主成分と呼ぶ。第１固有ベクトルを第１主成分、第２固有ベクトルを第２主成分と呼ぶ。**

**> summary(nn.pr)**

**Importance of components:**

**PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9**

**Standard deviation 3.503 3.0104 2.5598 2.2565 2.1277 2.0224 1.9107 1.7021 1.6467　固有値＝SD^2**

**Proportion of Variance 0.123 0.0906 0.0655 0.0509 0.0453 0.0409 0.0365 0.0290 0.0271　寄与率**

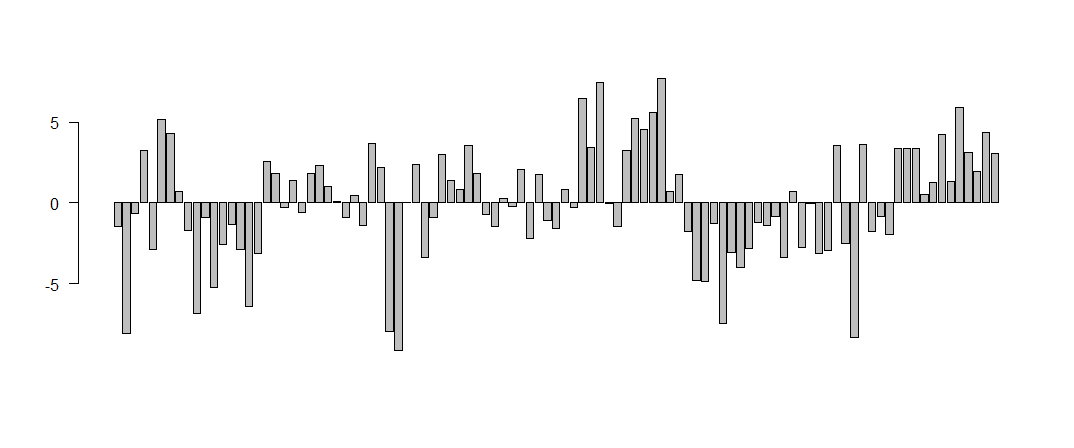
**Cumulative Proportion 0.123 0.2134 0.2789 0.3298 0.3751 0.4160 0.4525 0.4814 0.5086　累積寄与率**

**主成分は、データ行列の列のスコアである。　xxx.pc$rotation**

**データ行列の行のスコアを主成分得点と呼ぶ。xxx.pc$x**

**テキストの数と変数の数が多い場合は、バイプロットでは考察しにくい。その際は、個別の主成分と主成分得点の棒グラフなどで考察する**

**> barplot(nn.pr$x[,1], las=2) # 第１主成分得点の棒グラフ**



**どのような名詞が多く使用されているかについては、主成分の棒グラフで考察する。用いた名詞が多い場合、主な名詞を抽出するために主成分の値をソートし、値が大きい上位10個と値が小さい下位10個を抽出する。**

**> y<-nn.pr$rotation[,1] ♯第１主成分**

**> y1<-y[sort.list(y, dec=T)]　#降順にソート**

**> barplot(y1[c(1:10, 91:100)], las=2)♯上位10語と下位10語の主成分得点を棒グラフに**

