DATA SKILLS
UNDERSTANDING THE WORLD

Principal Component Analysis

Alessandro Rezzani

Abstract

L'articolo descrive una delle tecniche di riduzione della dimensionalità del data set: il metodo dell'analisi delle componenti principali (Principal Component Analysis o PCA).

Oltre al metodo PCA sarà fatto un breve cenno ad altre tecniche utilizzabili in caso di variabili descrittive (Multiple Correspondence Analysis e Factor Analysis of Mixed Data).

Infine le tecniche di calcolo della PCA e FAMD saranno descritte anche attraverso l'uso di esempi in linguaggio R.

Sommario

Abstract 1

Introduzione 2

La PCA 2

PCA in R 5

Multiple Correspondence Analysis e Factor Analysis of Mixed Data in R (cenni) 8

DATA SKILLS
UNDERSTANDING THE WORLD

Introduzione

Nell'affrontare un progetto di data mining è spesso opportuno ridurre il numero di attributi utilizzati per la costruzione dei modelli. I vantaggi che si ottengono dalla riduzione del data set sono principalmente due:

- La diminuzione dei tempi di elaborazione
- Miglioramento dell'accuratezza del modello, che dovrebbe beneficiare dall'eliminazione di variabili irrilevanti o ridondanti.

Il primo intervento che possiamo operare sul data set consiste nella selezione degli attributi da mantenere nel modello, che può avvenire in maniera completamente manuale e basata sulla conoscenza del business e dei dati da parte dell'analista.

Tuttavia è spesso difficile identificare quali siano le variabili da includere o escludere da un'analisi, soprattutto quando ci si trova di front ad un gran numero variabili. Si rende perciò necessaria una tecnica che faciliti il compito dell'analista.

Nei paragrafi seguenti descriviamo quindi la PCA assieme ad un suo utilizzo per la creazione di un dataset con un minor numero di variabili rispetto all'input iniziale.

La PCA

Prima di addentrarci nei dettagli della PCA, indichiamo di seguito sulla base si quali principi essa di basa:

1) Un'elevata correlazione tra le variabili è indice di ridondanza nei dati.
2) Le variabili più importanti sono quelle che esprimono una varianza più alta.

In base a tali assunti con la PCA vogliamo trasformare i dati originali, creando un nuovo insieme di variabili che conservino l'informazione contenuta nel data set originale, ma che non presentino più la ridondanza. In termini più formali possiamo dire che si trasformano v variabili quantitative in k (con k< v) combinazioni lineari ordinate in base alla variabilità da esse spiegata (in ordine decrescente).
Si parte dalla matrice di covarianza C, costituita da tutte le covarianze tra le n variabili del data set. Prima di calcolare la matrice di covarianza occorre standardizzare le variabili, in particolare nel caso in cui esse siano espresse in unità di misura diverse. Un altro metodo consiste nell'utilizzare la matrice di correlazione R al posto della matrice di covarianza.
La prima componente principale (e, in seguito, con lo stesso procedimento, tutte le altre) si determina attraverso la seguente combinazione lineare:

$$\begin{pmatrix} P_{11} \\ .. \\ .. \\ .. \\ P_{n1} \end{pmatrix} = \begin{pmatrix} x_{11} \\ .. \\ .. \\ .. \\ x_{n1} \end{pmatrix} + a_{11+\cdots+} \begin{pmatrix} x_{1v} \\ .. \\ .. \\ .. \\ x_{nv} \end{pmatrix}$$

Il vettore a_1 è detto vettore dei pesi (*loadings*). Si intendono normalizzati, ovvero la somma dei loro quadrati è pari a 1.
I pesi sono scelti in modo da massimizzare la varianza della variabile p_i (il vettore calcolato con la combinazione lineare). Si dimostra che a_1 è l'autovettore della matrice di covarianza (o correlazione) corrispondente al più alto autovalore, che rappresenta la varianza della prima componente principale.
Come si diceva, con lo stesso procedimento è possibile calcolare l'i-esima componente principale, prendendo come riferimento l'i-esimo autovalore (in ordine decrescente) e il corrispondente autovettore.

È opportuno notare come la tecnica della PCA sia da utilizzare con le variabili quantitative o qualitative binarie, mentre non è corretto applicarla a variabili qualitative (non binarie)[1]. Per le variabili qualitative è possibile utilizzare una tecnica differente, chiamata Multiple Correspondence Analysis[2] (si veda il paragrafo successivo).

Una prima valutazione del risultato ottenuto calcolando la matrice delle componenti principali P, può essere affrontata calcolando la quota di variabilità mantenuta dai nuovi valori, rispetto ai dati originali:

$$Quota\ di\ variabilità\ mantenuta = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{v} \lambda_i}$$

[1] Le variabili **qualitative** sono generalmente espresse attraverso descrizioni in formato testuale (non numerico). Si tratta di attributi quali, per esempio, la residenza, il CAP, la professione, ecc. In realtà possiamo distinguere due tipologie di variabili qualitative: quelle nominali, per le quali non esiste un ordine e quelle ordinali, invece dove l'ordine è espresso in maniera implicita o esplicita. Spesso le variabili qualitative sono espresse sotto forma di numeri (es. nr interi ordinati), ma tale rappresentazione non le trasforma in variabili quantitative. Un ulteriore aspetto delle variabili qualitative ordinali è dato dall'impossibilità di stabilire una relazione numerica tra i valori: possiamo soltanto affermare che una certo valore è migliore o più grande di un altro, ma non possiamo determinare di quanto.
Le variabili **quantitative** sono invece l'espressione di una misurazione o di un fatto puramente numerico. Si distinguono in variabili discrete o continue a seconda che assumano o no un numero finito di valori.
Parliamo invece di variabili **binarie** quando esse possono assumere soltanto due valori (0 o 1, falso o vero) che determina l'assenza o la presenza di una certa caratteristica. Esse sono da considerarsi variabili quantitative. Si utilizza la tecnica della "binarizzazione" per trasformare variabili qualitative in variabili quantitative. La tecnica consiste nel sostituire una variabile qualitativa, per esempio "propensione al rischio" (che ipotizziamo assuma tre valori bassa, media e alta) in più variabili binarie, una per ogni valore della variabile originale (tre nel nostro caso). Nell'esempio della propensione al rischio avremmo tre variabili Prop_rischio_bassa, Prop_rischio_media, Prop_rischio_alta, ciascuna delle quali assume un valore 0 oppure 1 (in maniera esclusiva: ovvero, per ciascun elemento del dataset, solo una delle tre variabili sarà valorizzata a 1).

[2] Si veda Michael Greenacre, Jorg Blasius ,"Multiple Correspondence Analysis and Related Methods", 2006, CRC Press

Dove k è il numero di componenti principali calcolate, v è il numero di variabili originali e λ è la varianza (che nel caso della componente principale è il corrispondente autovalore).
In base alla variabilità spiegata, è possibile **scegliere le prime k (con k<n) componenti principali** da utilizzare come **nuovo dataset**, con dimensionalità ridotta rispetto a quello originale.

Un'altra misura interessante è la correlazione le componenti principali e le variabili originali. Essa permette di stabilire da cosa dipendono i valori delle componenti principali. Si noti che, in caso di mancata standardizzazione delle variabili (o della matrice di covarianza), valori molto alti di alcune variabili rispetto alle altre possono produrre, come effetto non desiderato, che alcune componenti principali dipendano esclusivamente da una variabile, vanificando così l'intenzione di ridurre la dimensionalità del dataset.
Dato che possiamo esprimere p_j come prodotto tra la matrice dei valori originali X e il vettore dei pesi a_j, si ottiene che essa è uguale al prodotto tra la matrice di correlazione delle variabili originali e il vettore dei pesi. Si arriva così ad avere che $Cov(p_j,X_i)= \lambda_j a_{ij}$.
Ricordando che la correlazione si ottiene con la formula seguente:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

la correlazione sarà quindi:

$$Corr(p_j, x_i) = \frac{\sqrt{\lambda_j} a_{ij}}{\sqrt{\sigma_X}}$$

Quindi la correlazione dipende dal fattore $a_{ij,}$ ovvero dal peso (loading), sia per quanto riguarda il segno, sia per quanto riguarda l'ampiezza.

Esiste poi una modalità visuale con cui decidere: si tratta dello screeplot, ovvero di un grafico nel quale sono rappresentate le componenti principali sull'asse delle ascisse e la varianza da esse spiegata sull'asse delle ordinate. I punti così ottenuti sono uniti da una linea. La figura seguente mostra uno screeplot.

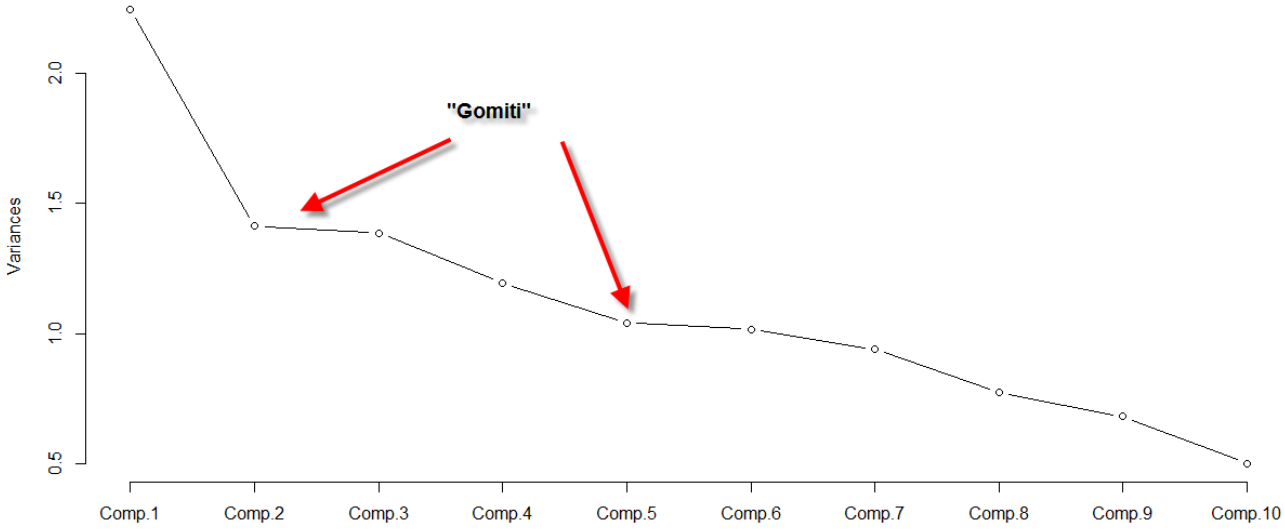

Figura 1 ScreePlot

Nella figura sono tati evidenziati i cosiddetti “gomiti”, ovvero i punti in cui la pendenza della curva si riduce significativamente. In corrispondenza di questi punti è opportuno operare la scelta del numero di componenti principali da utilizzare. Nel nostro caso una prima significativa diminuzione della pendenza si ha dopo la seconda componente. Quindi potremmo scegliere di utilizzare soltanto le prime due componenti. In alternativa, un secondo, meno evidente, “gomito” si trova dopo la quinta componente.

PCA in R

Volendo implementare in linguaggio R quanto esposto sopra, abbiamo diverse opzioni tra cui scegliere. Ne esponiamo due: il calcolo manuale ed il calcolo effettuato attraverso la funzione *princomp*.

Calcolo manuale

```
#utilizziamo il dataset iris
Xdata=t(as.matrix(iris[,-5]))

# calcoliamo gli scarti dalla media
rm=rowMeans(Xdata)
X=Xdata-matrix(rep(rm, dim(Xdata)[2]), nrow=dim(Xdata)[1])
# calcoliamo la matrice di correlazione
A=X %*% t(X)
C=A/ncol(X)

E=eigen(C,TRUE)
EV=t(E$vectors)
# calcoliamo le PC e le deviazioni standard
P = EV %*% X

#varianze
var =  E$values
```

```
# percentuale cumulativa
cumsum(var)/sum(var)
```

```
> cumsum(var)/sum(var)
[1] 0.9246187 0.9776852 0.9947878 1.0000000
```

```
# sceeeplot
plot(var,type="lines")

#prime 2 PC
P[,c(1,2)]
```

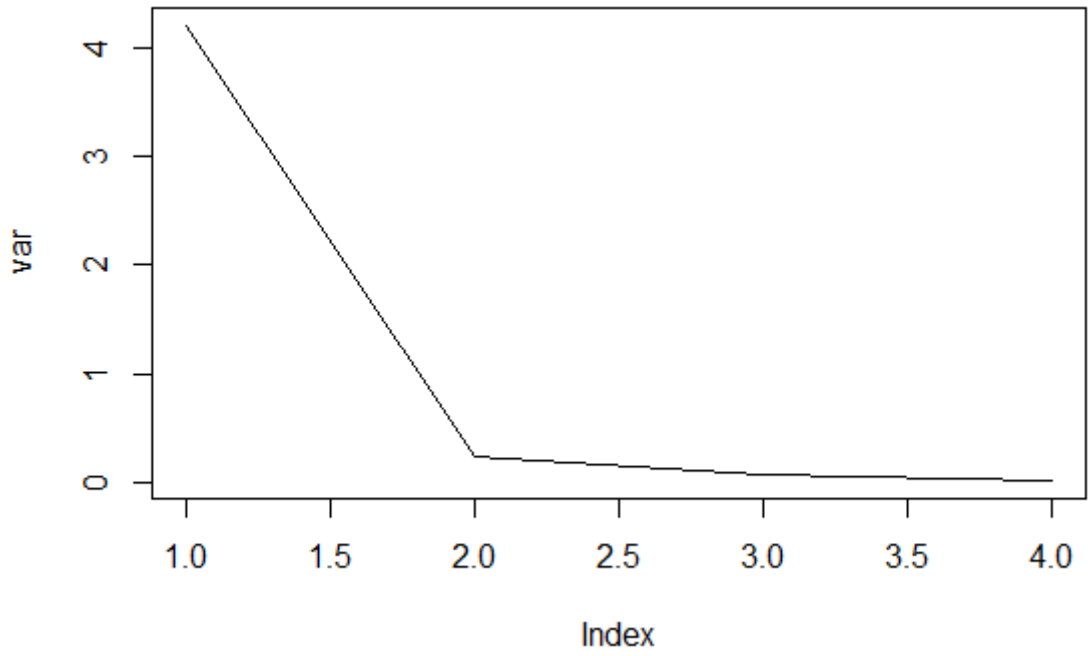

Figura 2 Screeplot delle PC per il dataset iris.

La Figura precedente mostra lo screeplot ottenuto con il codice R. Come si può vedere la pendenza della curva cambia radicalmente dopo la seconda componente principale.

Calcolo con la funzione princomp

```
Xdata=t(as.matrix(iris[,-5]))

# calcoliamo gli scarti dalla media
rm=rowMeans(Xdata)
X=Xdata-matrix(rep(rm, dim(Xdata)[2]), nrow=dim(Xdata)[1])

#utilizziamo la funzione princomp per il calcolo delle PC
fit <- princomp( data.frame(t(X)), cor=FALSE)

# deviazioni standard e %della varianza
summary(fit)
```

```
> summary(fit)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
```

```
# scree plot
plot(fit,type="lines")

# prime 2 PC
fit$scores[,c(1,2)]
```

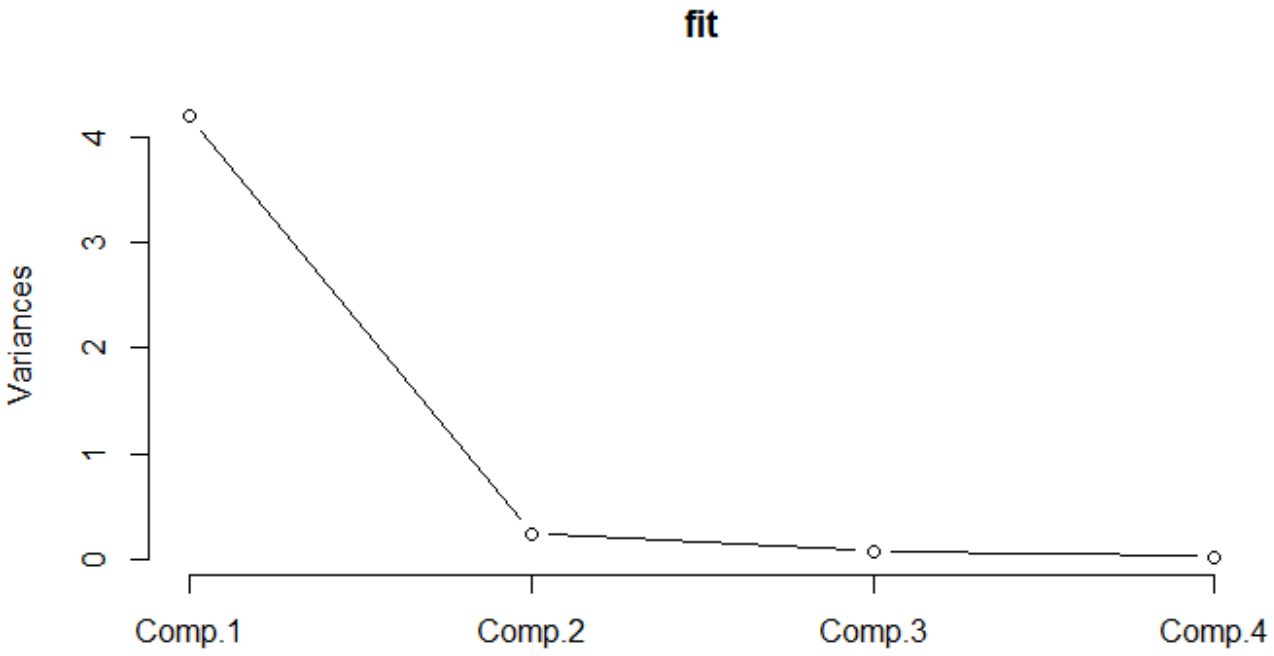

Figura 3 Scree plot ottenuto dal risultato di princomp.

Multiple Correspondence Analysis e Factor Analysis of Mixed Data in R (cenni)

Come accennato, la MCA è una tecnica di riduzione della dimensionalità del dataset, applicabile alle variabili qualitative. In questo documento facciamo soltanto un cenno al procedimento e, par approfondimenti rimandiamo alla bibliografia.

Un procedimento per realizzare la MCA si basa sulla *complete disjunctive table*, ovvero una matrice dove le righe sono gli individui (osservazioni) e le colonne sono le variabili, o meglio, indicatori che esprimono le diverse categorie presenti nelle variabili (in pratica si trasformano le variabili qualitative in una serie di variabili binarie). A questo punto trasformando la matrice i dati così ottenuta secondo la formula seguente, si può applicare la PCA:

$$x_{ik} = \frac{y_{ik}}{pk} - 1$$

Dove:
y_{ik} rappresenta il valore della k-esima variabile binaria per la riga (individuo) i
p_k rappresenta la percentuale di individui per cui $y_{ik}=1$

La FactorAnalysis of Mixed Data (FAMD) è un metodo di analisi di data set contenti sia variabili quantitative, sia variabili qualitative, sviluppato da Jean-Paul Benzécri. Semplificando molto, possiamo dire che FAMD applica la PCA alle variabili quantitative e la MCA alle variabili qualitative.

Esempio R di FAMD

Vediamo quindi un esempio di FAMD in R.
Partendo dal dataset Clienti che contiene variabili quantitative e qualitative, applichiamo la funzione FAMD del package FactoMineR. Essa utilizza le tecniche di PCA per le variabili quantitative e MCA per le variabili qualitative.

```
Clienti                 10000 obs. of 13 variables
  ID : int 22711 13555 28907 2 25410 4 15756 11085 17974 21008 ...
  Marital.Status : Factor w/ 2 levels "Married","Single": 2 1 1 2 2 
  Gender : Factor w/ 2 levels "Female","Male": 2 1 2 2 1 1 1 1 1 1 ..
  Yearly.Income : int 30000 40000 160000 160000 70000 120000 70000 6
  Children : int 0 0 5 0 2 2 0 0 1 1 ...
  Education : Factor w/ 5 levels "Bachelors","Graduate Degree",..: 4
  Occupation : Factor w/ 5 levels "Clerical","Management",..: 1 1 4 
  Home.Owner : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 2 1 2 1 ..
  Cars : int 1 0 3 2 1 3 2 2 0 0 ...
  Commute.Distance: Factor w/ 5 levels "0-1 Miles","1-2 Miles",..: 1
  Region : Factor w/ 3 levels "Europe","North America",..: 1 1 1 3 2
  Age : int 33 37 55 47 38 40 49 49 38 46 ...
  BikeBuyer : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 2 1 2 2 ...
```

Figura 4 Il dataset clienti.

#utilizziamo la libreria FactoMineR

```
library("FactoMineR")
```

#utilizziamo la funzione FAMD su tutte le colonne tranne la prima (l'ID)

```
res<-FAMD(Clienti[,2:12])
```

#autovalori /autovettori

```
res$eig
```

```
> res$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  3.139329e+00           1.494919e+01                          14.94919
comp 2  2.002544e+00           9.535922e+00                          24.48511
comp 3  1.656321e+00           7.887241e+00                          32.37235
comp 4  1.393632e+00           6.636343e+00                          39.00869
comp 5  1.279253e+00           6.091681e+00                          45.10038
comp 6  1.227797e+00           5.846654e+00                          50.94703
comp 7  1.177604e+00           5.607638e+00                          56.55467
comp 8  1.137593e+00           5.417109e+00                          61.97178
comp 9  1.079405e+00           5.140024e+00                          67.11180
comp 10 9.599085e-01           4.570993e+00                          71.68279
comp 11 9.189665e-01           4.376031e+00                          76.05882
comp 12 8.534527e-01           4.064060e+00                          80.12288
comp 13 7.821944e-01           3.724735e+00                          83.84762
comp 14 7.311092e-01           3.481472e+00                          87.32909
comp 15 6.454562e-01           3.073601e+00                          90.40269
```

#principal components

```
res$ind$coord
```

```
> res$ind$coord
          Dim.1        Dim.2        Dim.3         Dim.4        Dim.5
1   -2.83151295  0.221403406  0.867405267  0.1740195975 -0.038252426
2   -2.33081116 -2.508137980  0.664103517 -0.0601373858  0.492522530
3    3.26159963  1.755925031  2.909384886 -1.0527336426 -0.455556126
4    1.74593080 -1.096838852  0.109956205  2.0519681634  1.950241149
5   -0.52401969  0.696836607 -1.329816343  1.4375316237 -0.180901361
6    2.56561427 -0.690340823 -0.236040827  0.5931205297  0.299771930
7    0.75204184  1.254390916 -0.220802616  1.6826405462 -1.197719931
8    0.52236873  1.844485582 -1.372261033  0.2016601105 -0.680746055
9   -1.82532417 -1.944891213  0.752959349  0.0784681554 -0.919144139
10  -3.02924520  1.299631174  1.293151429 -0.4982965270  0.781266390
11   0.72293844 -0.087512996 -1.755666672  0.0445415269  2.071707981
12  -0.05213744  0.107667290  0.123399910 -0.5813309272 -1.333764212
13  -0.71447841  2.049052047 -2.437025070 -0.4802448603 -0.042562924
14  -0.92615027 -1.051437765 -0.823717469 -1.2505343875 -0.488935337
```

#correlazioni tra le componenti e le variabili

resvarcorrelation

plot(res$eig[1:10,"eigenvalue"],type="lines")

```
> res$var$correlation
                     Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
Marital.Status  0.21538441 0.44150170 0.16398479 0.21967161 0.14976133
Gender          0.01211726 0.05079614 0.16898939 0.18208228 0.05208527
Yearly.Income   0.82021749 0.02069510 0.03989762 0.13430429 0.07513654
Children        0.50424571 0.27367445 0.45571278 0.06090340 0.01209807
Education       0.12308733 0.76935774 0.31965836 0.54353757 0.59876450
Occupation      0.88338162 0.65291589 0.61666435 0.28448702 0.75886585
Home.Owner      0.19091533 0.50837800 0.27601740 0.14824405 0.18876761
Cars            0.68102967 0.49009585 0.12777734 0.06736123 0.18500223
Commute.Distance 0.62962617 0.28832814 0.61357823 0.38545822 0.36187703
Region          0.51035811 0.10804953 0.65372223 0.78311979 0.26281206
Age             0.46164223 0.34358876 0.11316402 0.35460011 0.20985541
```

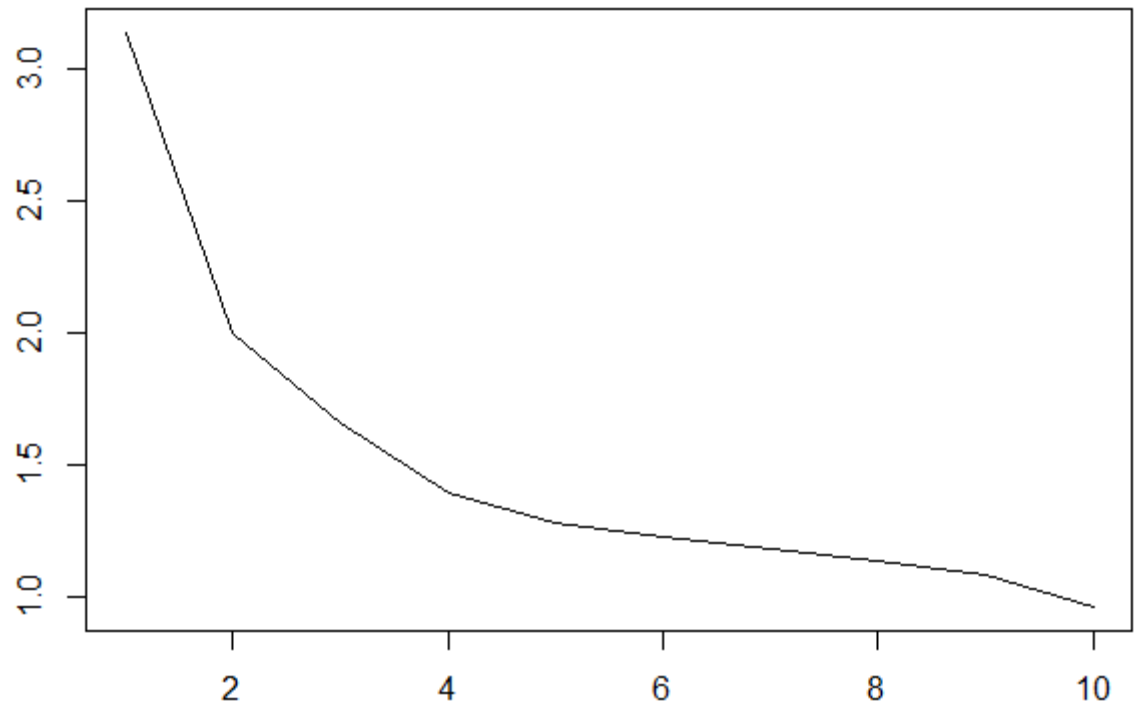

Figura 5 Scree plot.

Come si nota i risultati sono interpretabili nello stesso modo della PCA, per esempio in termini di correlazione e scree plot.

Bibliografia

P.Giudici, *Data Mining. Metodi informatici, statistici e applicazioni* (seconda edizione), McGraw-Hill, 2005

Mehmed Kantardzic *Data Mining: Concepts, Models, Methods, and Algorithms,* John Wiley & Sons, 2003

Jiawei Han, Micheline Kamber**,** *Data Mining:Concepts and Techniques Second Edition,* Morgan Kaufmann Publishers, 2006

FactoMineR (http://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf)

Sebastien Le, Julie Josse, Francois Husson, *FactoMineR: An R Package for Multivariate Analysis*, Journal of Statistical Software, 2008

Michael Greenacre, Jorg Blasius ,*Multiple Correspondence Analysis and Related Methods*, CRC Press, 2006

Pagès Jérôme, *Multiple Factor Analysis by Example Using R,* Chapman & Hall, 2014