多元-08-2020270026

2020270026 王姿文

4/26/2021

1. 数据

- 数据叙述：数据为中国各省份的一些特征，一共是rows = 31, columns = 8，故共有8个维度。
- 目标： 进行主成分分析和因子分析

下表为其中几笔数据，以及数据的结构：

```
df <- read_excel("ex6.7.xls")
a <- df$...1
df <- df[,-1]
rownames(df) <- a
kbl(df) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size =
7)
```

	食品	衣着	居住	医疗	交通通讯	教育	家庭服务	耐用消费品
北京	5561.54	1571.74	1286.32	1563.10	2293.23	809.25	84.71	548.55
天津	5005.09	1153.66	1528.28	1220.92	1567.87	715.24	45.50	467.75
河北	3155.40	1137.22	1097.41	808.88	1062.31	386.60	28.84	305.70
山西	2974.76	1137.71	1250.87	769.79	931.33	570.79	35.38	259.05
内蒙古	3553.48	1616.56	1028.19	869.71	1191.70	568.35	30.49	307.92
辽宁	4378.14	1187.41	1270.95	913.13	1295.70	670.13	30.40	235.46
吉林	3307.14	1259.62	1285.28	914.47	954.96	576.17	21.25	214.28
黑龙江	3128.10	1217.04	941.25	864.89	749.05	551.73	16.11	192.87
上海	7108.62	1520.61	1646.19	755.29	3373.19	1165.96	139.86	545.30
江苏	4544.64	1166.91	1042.10	794.63	1357.96	750.97	72.09	365.56
浙江	5522.56	1546.46	1333.69	933.11	2392.63	1178.54	78.67	306.86
安徽	3905.05	1010.61	988.12	633.93	920.77	633.45	31.57	249.32
福建	5078.85	1105.31	1300.10	540.63	1777.06	686.35	78.29	320.38
江西	3633.05	969.58	851.15	483.96	872.57	388.48	27.95	229.82
山东	3699.42	1394.11	1247.04	799.79	1410.45	580.10	33.50	426.80
河南	3079.82	1141.76	963.59	790.87	915.12	464.35	23.36	332.85
湖北	3996.27	1099.16	914.26	675.32	890.12	570.99	28.21	265.01
湖南	3970.42	1090.72	960.82	790.95	971.05	543.50	38.49	254.18
广东	5866.91	975.06	1748.16	836.39	2623.08	720.58	120.04	348.66
广西	4082.99	772.28	891.33	529.36	1376.03	483.61	30.24	294.83
海南	4226.90	491.84	1106.39	536.40	1303.50	459.74	24.66	255.76
重庆	4418.34	1294.30	1096.82	878.25	1044.36	536.43	48.15	405.48
四川	4255.48	1042.45	819.28	564.93	1121.45	422.07	30.03	211.86
贵州	3597.94	851.50	836.54	471.39	871.15	436.24	25.36	186.10
云南	4272.29	1026.50	739.20	606.86	1216.46	294.29	11.89	158.97
西藏	4262.77	1011.82	634.94	317.08	966.74	205.45	3.58	39.80
陕西	3586.13	1047.61	1007.68	862.70	967.52	697.29	39.86	287.22

甘肃	3183.79	1022.62	846.26	654.82	817.17	428.40	19.30	238.04
青海	3315.94	945.14	802.73	610.02	787.63	388.96	9.74	242.21
宁夏	3352.83	1178.88	1069.15	816.87	1096.32	465.08	27.94	303.80
新疆	3235.77	1245.02	781.90	643.48	1003.89	417.06	23.89	223.22

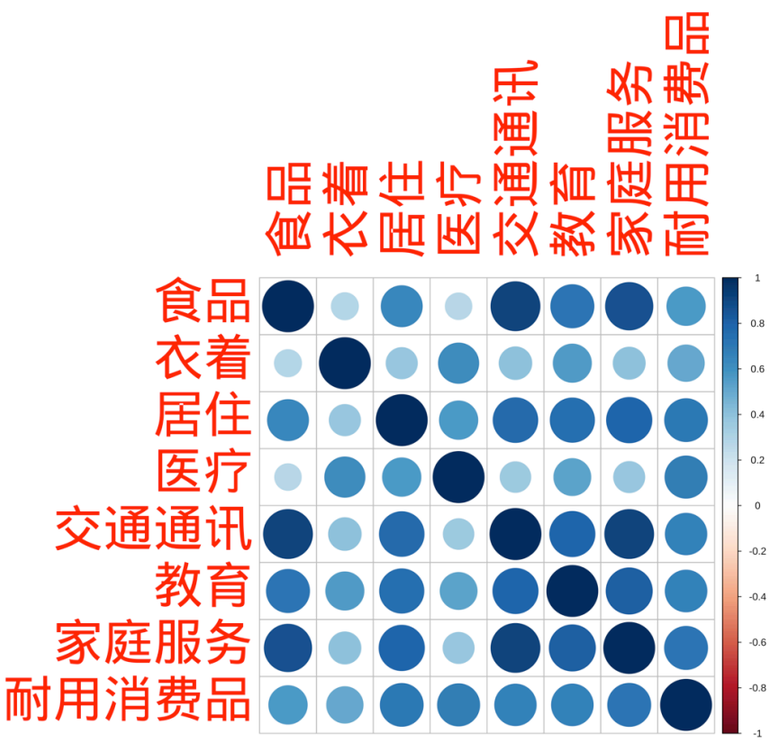
2. PCA

因为各列不可比，所以后续的pca分析需要用样本相关阵。
其中 食品 x 交通通讯 、 食品 x 教育 、 食品 x 家庭服务 、 居住 x 交通通讯 、 居住 x 教育 、 居住 x 家庭服务 、 居住 x 耐用消费品 、 交通通讯 x 教育 、 交通通讯 x 家庭服务 、 教育 x 家庭服务 、 家庭服务 x 耐用消费品 的绝对值相关性>0.7：

```
kbl(round( cor(df), 3)) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size =
7)
```

	食品	衣着	居住	医疗	交通通讯	教育	家庭服务	耐用消费品
食品	1.000	0.282	0.646	0.272	0.919	0.714	0.860	0.567
衣着	0.282	1.000	0.375	0.621	0.408	0.560	0.395	0.509
居住	0.646	0.375	1.000	0.565	0.765	0.751	0.788	0.705
医疗	0.272	0.621	0.565	1.000	0.359	0.525	0.371	0.682
交通通讯	0.919	0.408	0.765	0.359	1.000	0.783	0.917	0.662
教育	0.714	0.560	0.751	0.525	0.783	1.000	0.816	0.668
家庭服务	0.860	0.395	0.788	0.371	0.917	0.816	1.000	0.712
耐用消费品	0.567	0.509	0.705	0.682	0.662	0.668	0.712	1.000

```
corrplot(cor(df),tl.cex=4)
```



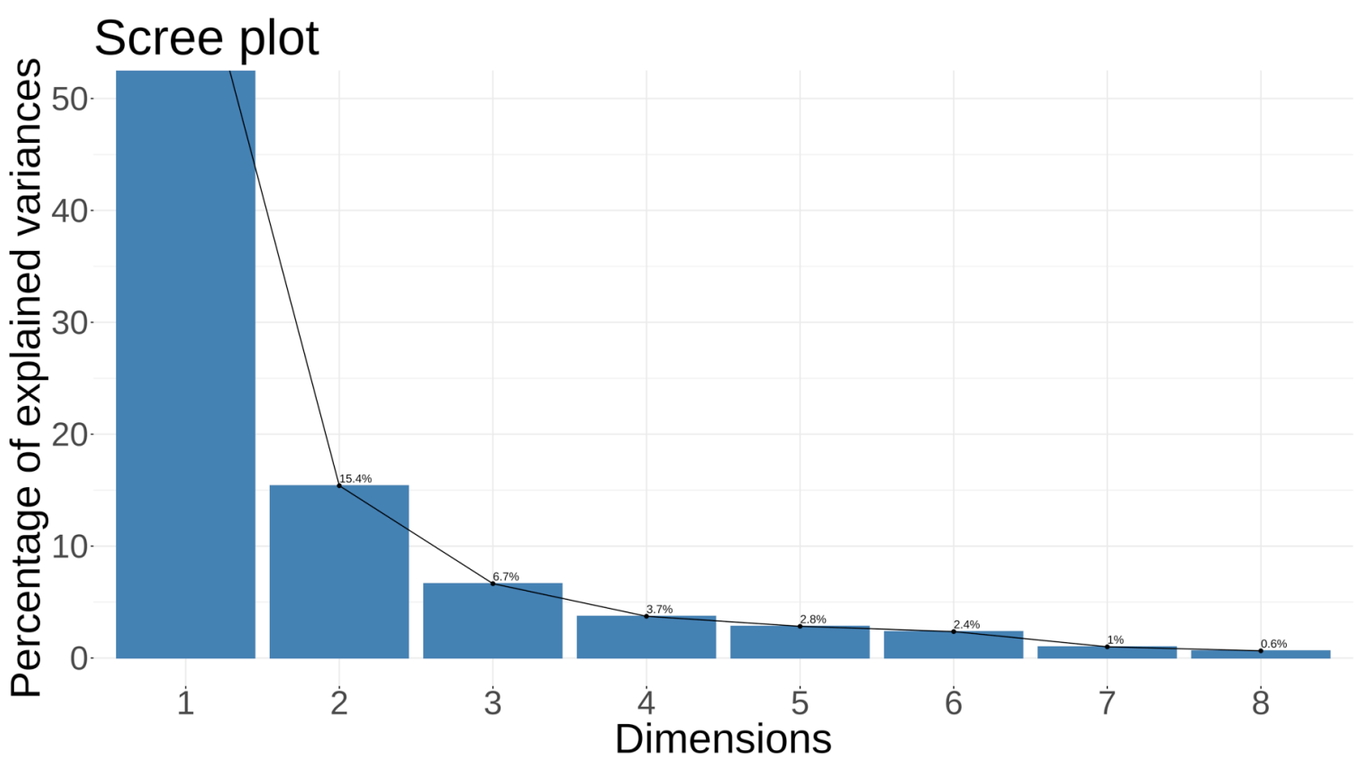
开始做主成份分析，可以看到在Comp=4时，可解释累积变异就>90%，因此选择四个维度变可解释93%的原始数据。此外，若使用两个主成分的话，其可解释累积变异为82%，已經算高了，后续可就前两维来画图看一些insight。
而我们也能从loadings来判断不同主成分的性质，例如：Comp.1主要是 食品 、 居住 、 交通通讯 、 教育 、 家庭服务 、 耐用消费品 的值较大；Comp.2主要是 衣着 、 医疗 的值较大。

```
pca1 <- princomp(df, cor=TRUE)
summary(pca1, loadings=TRUE)
```

```
## Importance of components:
##                            Comp.1    Comp.2     Comp.3    Comp.4     Comp.5
## Standard deviation     2.3213318 1.1100881 0.72943408 0.5464987 0.47643779
## Proportion of Variance 0.6735727 0.1540369 0.06650926 0.0373326 0.02837412
## Cumulative Proportion  0.6735727 0.8276096 0.89411886 0.9314515 0.95982558
##                            Comp.6     Comp.7      Comp.8
## Standard deviation     0.4351019 0.28334611 0.227588880
## Proportion of Variance 0.0236642 0.01003563 0.006474587
## Cumulative Proportion  0.9834898 0.99352541 1.000000000
##
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## 食品         0.358  0.396  0.158  0.288  0.503         0.282  0.522
## 衣着         0.257 -0.536  0.703        -0.130 -0.336         0.135
## 居住         0.374        -0.412 -0.570 -0.112 -0.512  0.224  0.198
## 医疗         0.275 -0.599 -0.336         0.600  0.148 -0.248
## 交通通讯     0.393  0.292  0.137  0.120  0.166 -0.233  0.114 -0.795
## 教育         0.386         0.195 -0.466 -0.178  0.729  0.168
## 家庭服务     0.396  0.264                -0.211        -0.837  0.152
## 耐用消费品   0.361 -0.205 -0.373  0.599 -0.503  0.114  0.251
```

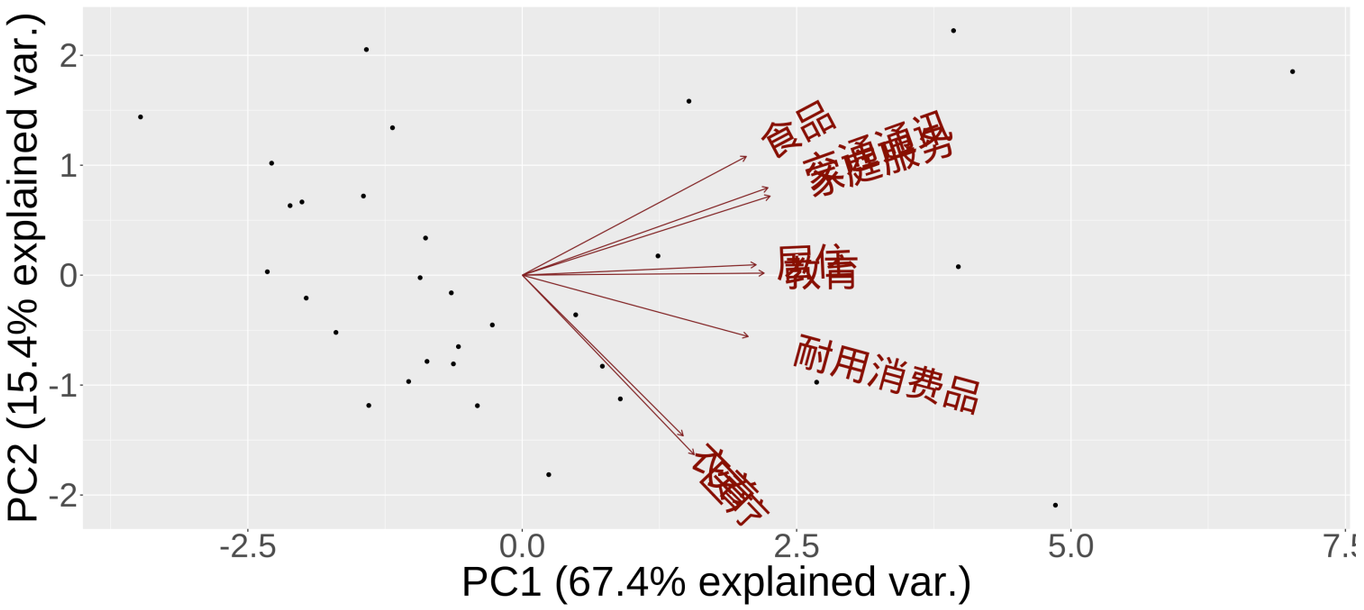
可以从**Scree plot&**简单看出每一个主成分的可解释变异占比：

```
df.pca <- PCA(df, graph = FALSE,scale.unit = TRUE)
fviz_eig(df.pca, addlabels = TRUE, ylim = c(0, 50)) +
  theme(text = element_text(size = 40))
```



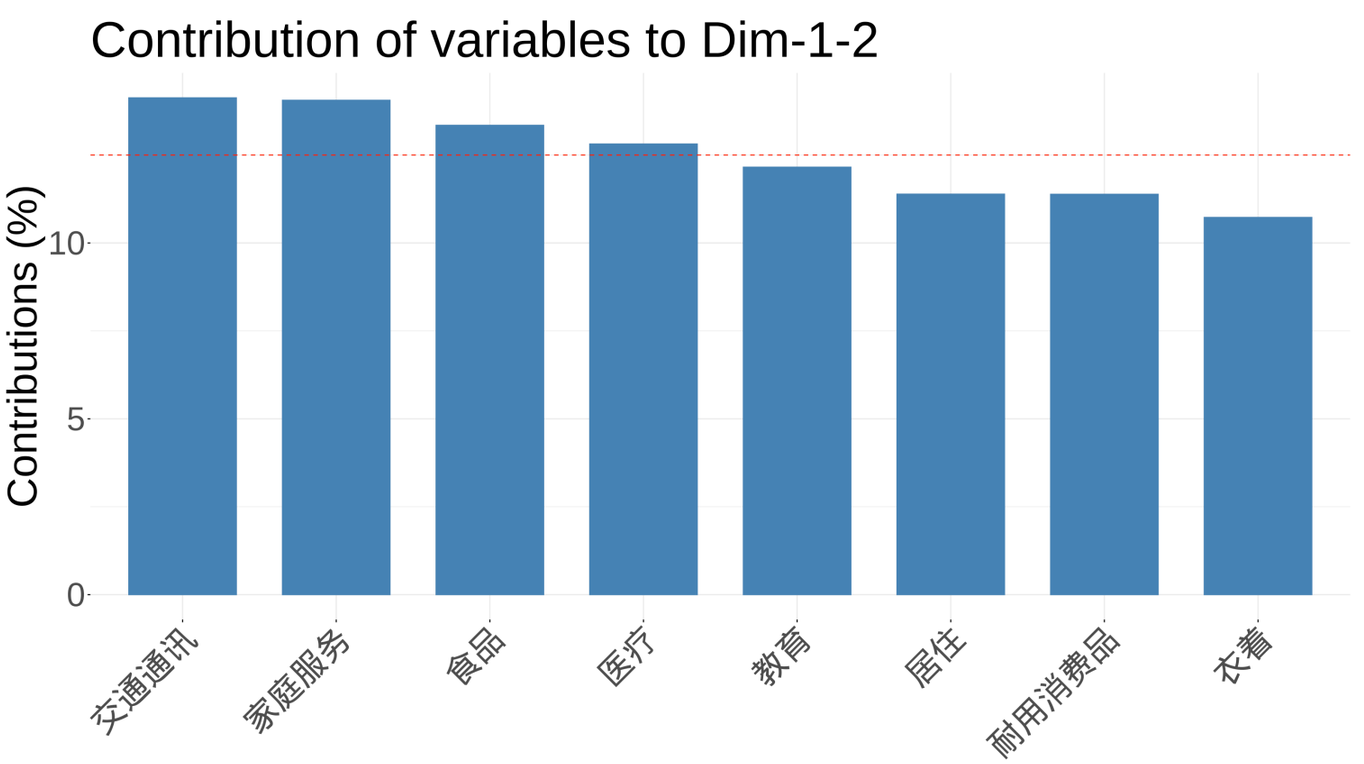
下图为前两个主成分的loadings所画出来的原始变量表现，可以看出在第一个主成分是是 食品 、 居住 、 交通通讯 、 教育 、 家庭服务 、 耐用消费品 的值较大，并且可以看出正负；在第二个主成分是 衣着 、 医疗 的值较大：

```
ggbiplot(pca1, obs.scale = 1, var.scale = 1,labels = NULL,varname.size = 13) +
  theme(text = element_text(size = 40))
```



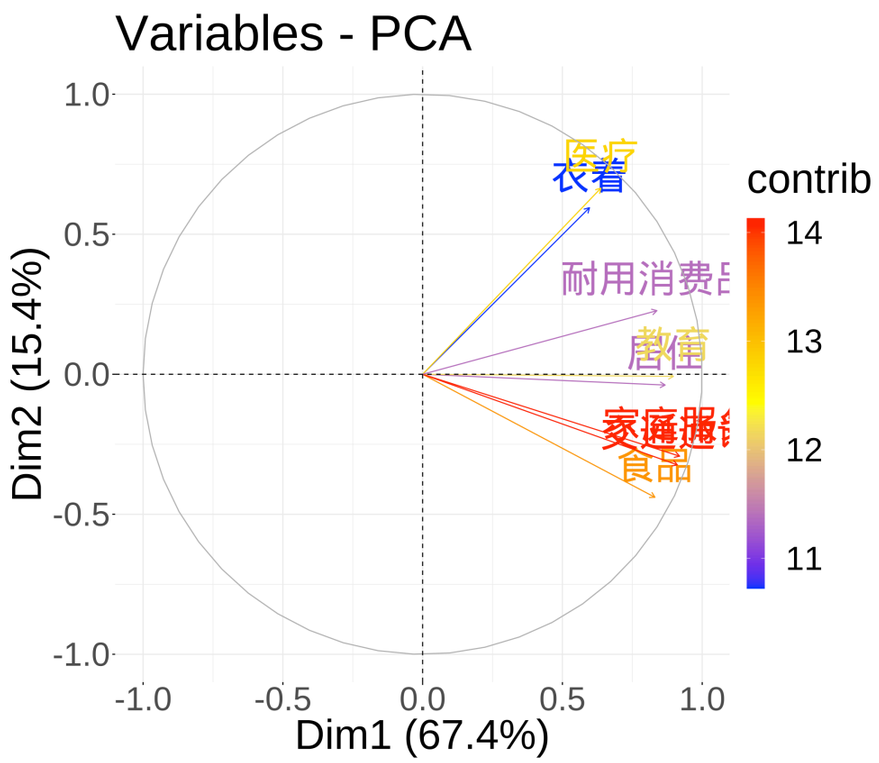
综合来看原始八个变量在第一主成分和第二主成分的贡献占比为下图，排名前几的变量与其他变量间的corr都较高：

```
fviz_contrib(df.pca, choice = "var", axes = 1:2, top = 10) +
  theme(text = element_text(size = 40))
```

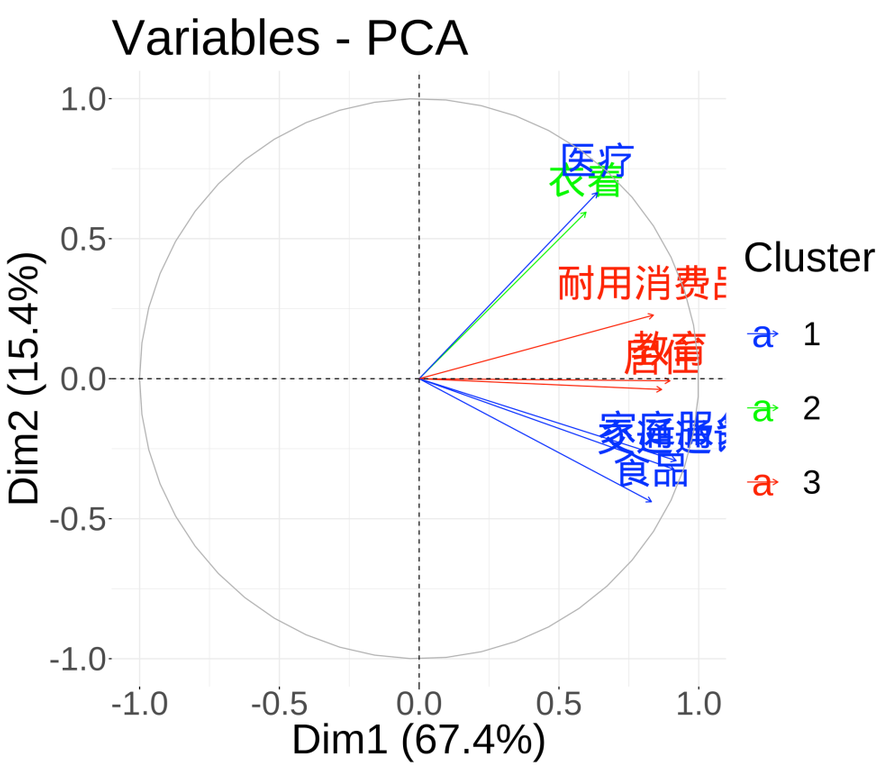


接着来看原始八个变量在第一在第一主成分和第二主成分的贡献占比和k mean分类，特别的是由两个主成分，可以用pca完成降维跟聚类的结果，此处我设定共有三个类别，结果为第一类： 医疗 、 家庭服务 、 交通通讯 、 食品 ；第二类： 衣着 ；第三类： 居住 、 教育 、 耐用消费品 ：

```
fviz_pca_var(df.pca, col.var = "contrib",labelsize = 13,
             gradient.cols = c("blue", "yellow", "red"))+
    theme(text = element_text(size = 40),
          legend.key.height = unit(2.5, "cm"))
```

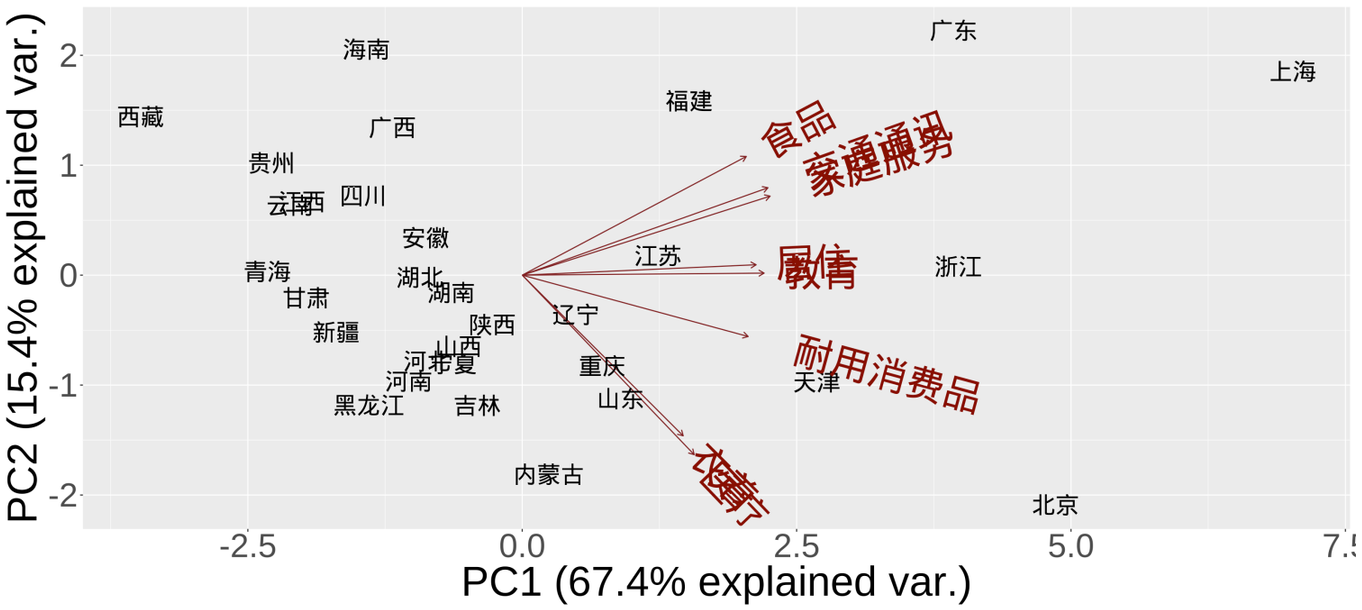



```
set.seed(123)
var <- get_pca_var(df.pca)
var.kms <- kmeans(var$contrib, centers = 3, nstart = 25)
kms.grp <- as.factor(var.kms$cluster)
fviz_pca_var(df.pca, col.var = kms.grp, palette = c("blue", "green", "red"),legend.titl
e = "Cluster",labelsize = 13)+
    theme(text = element_text(size = 40),
          legend.key.height = unit(2.5, "cm"))
```



最后将国家与前两个主成分的loadings所画出来的原始变量画在一起，可以看出不同省份在两个主成分和原始变量间的分布，例如可以看出比较主要的重点省份会是在x轴的右方：

```
ggbiplot(pca1, obs.scale = 1, var.scale = 1,labels = row.names(df)[1:31],varname.size =
13, labels.size = 8) +
  theme(text = element_text(size = 40))
```



---

3. 因子分析

以下分三种因子分析，分别为旋转及不旋转的方法，由于摆在一起看比较看得出差异，因此以下会将三种方法放在一起比较。

3.1 Loadings and Cumulative Variation

下面三种方法，不旋转和varimax旋转大致是是f1： 食品 、 居住 、 交通通讯 、 教育 、 家庭服务 、 耐用消费品 ，f2： 衣着 、 医疗 、 耐用消费品 ；而promax旋转比较不同，f1: 食品 、 居住 、 交通通讯 、 教育 、 家庭服务 ，f2: 衣着 、 医疗 、 耐用消费品

a 不旋转

用相关阵不旋转做因子分析。用极大似然估计，没有旋转

```
fac1 <- factanal(df, factors=2, rotation='none')
fac1
```

```
##
## Call:
## factanal(x = df, factors = 2, rotation = "none")
##
## Uniquenesses:
##       食品       衣着       居住       医疗   交通通讯       教育   家庭服务
##      0.123      0.544      0.281      0.191      0.043      0.256      0.106
## 耐用消费品
##      0.272
##
## Loadings:
##             Factor1 Factor2
## 食品         0.909  -0.224
## 衣着         0.460   0.495
## 居住         0.814   0.235
## 医疗         0.463   0.772
## 交通通讯     0.971  -0.119
## 教育         0.841   0.194
## 家庭服务     0.944
## 耐用消费品   0.739   0.426
##
##                Factor1 Factor2
## SS loadings      5.003   1.182
## Proportion Var   0.625   0.148
## Cumulative Var   0.625   0.773
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 14.64 on 13 degrees of freedom.
## The p-value is 0.331
```

b varimax旋转

```
## fa varimax
fac2 <- factanal(df, factors=2, rotation='varimax')
fac2
```

```
##
## Call:
## factanal(x = df, factors = 2, rotation = "varimax")
##
## Uniquenesses:
##       食品       衣着       居住       医疗   交通通讯       教育   家庭服务
##      0.123      0.544      0.281      0.191      0.043      0.256      0.106
## 耐用消费品
##      0.272
##
## Loadings:
##            Factor1 Factor2
## 食品        0.925   0.147
## 衣着        0.232   0.634
## 居住        0.659   0.533
## 医疗        0.126   0.891
## 交通通讯    0.941   0.268
## 教育        0.699   0.505
## 家庭服务    0.888   0.324
## 耐用消费品 0.515   0.680
##
##                Factor1 Factor2
## SS loadings      3.788   2.396
## Proportion Var   0.473   0.300
## Cumulative Var   0.473   0.773
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 14.64 on 13 degrees of freedom.
## The p-value is 0.331
```

c promax旋转

```
# fa promax
fac3 <- factanal(df, factors=2, rotation='promax',scores='regression')
fac3
```

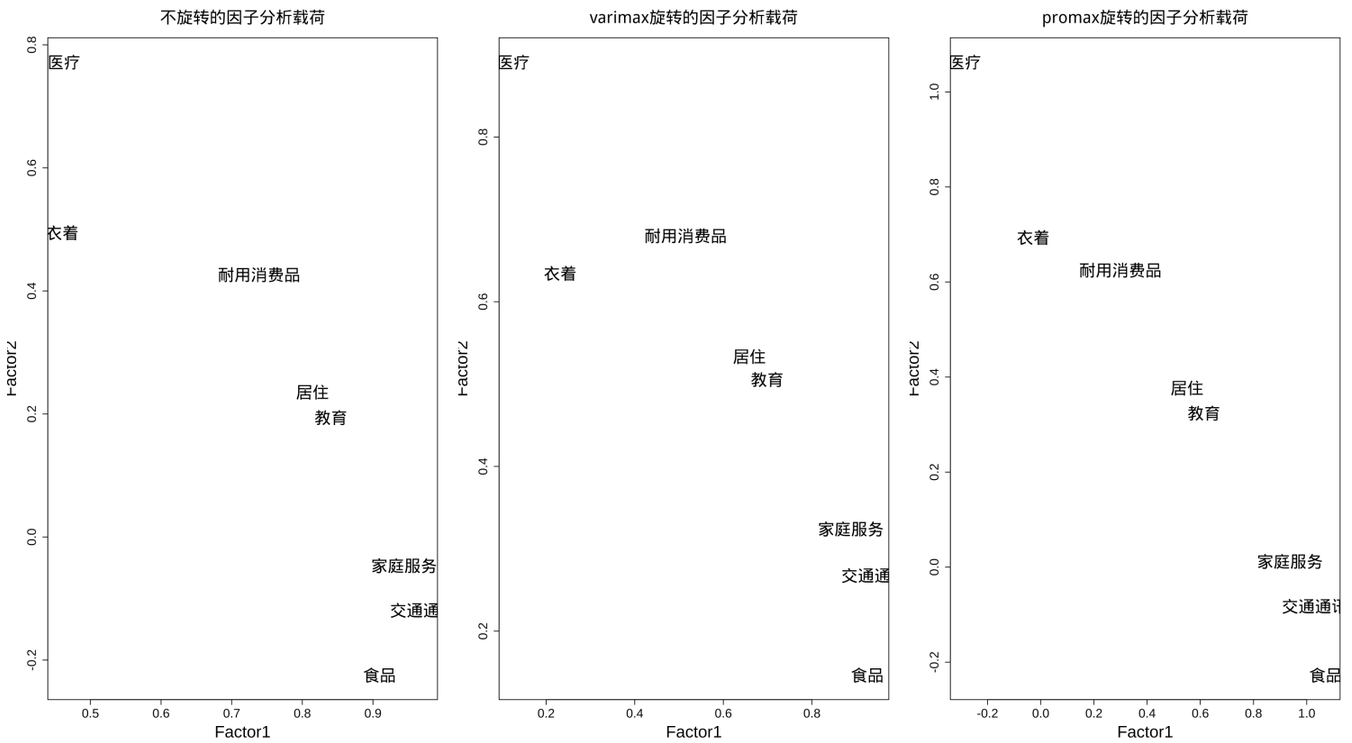
```
##
## Call:
## factanal(x = df, factors = 2, scores = "regression", rotation = "promax")
##
## Uniquenesses:
##       食品       衣着       居住       医疗   交通通讯       教育   家庭服务
##      0.123      0.544      0.281      0.191      0.043      0.256      0.106
## 耐用消费品
##      0.272
##
## Loadings:
##            Factor1 Factor2
## 食品        1.071  -0.227
## 衣着                0.694
## 居住        0.550   0.377
## 医疗       -0.286   1.062
## 交通通讯    1.031
## 教育        0.614   0.324
## 家庭服务    0.937
## 耐用消费品  0.300   0.625
##
##                Factor1 Factor2
## SS loadings      3.940   2.306
## Proportion Var   0.493   0.288
## Cumulative Var   0.493   0.781
##
## Factor Correlations:
##         Factor1 Factor2
## Factor1   1.000   0.661
## Factor2   0.661   1.000
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 14.64 on 13 degrees of freedom.
## The p-value is 0.331
```

par(mfrow=c(1,3))

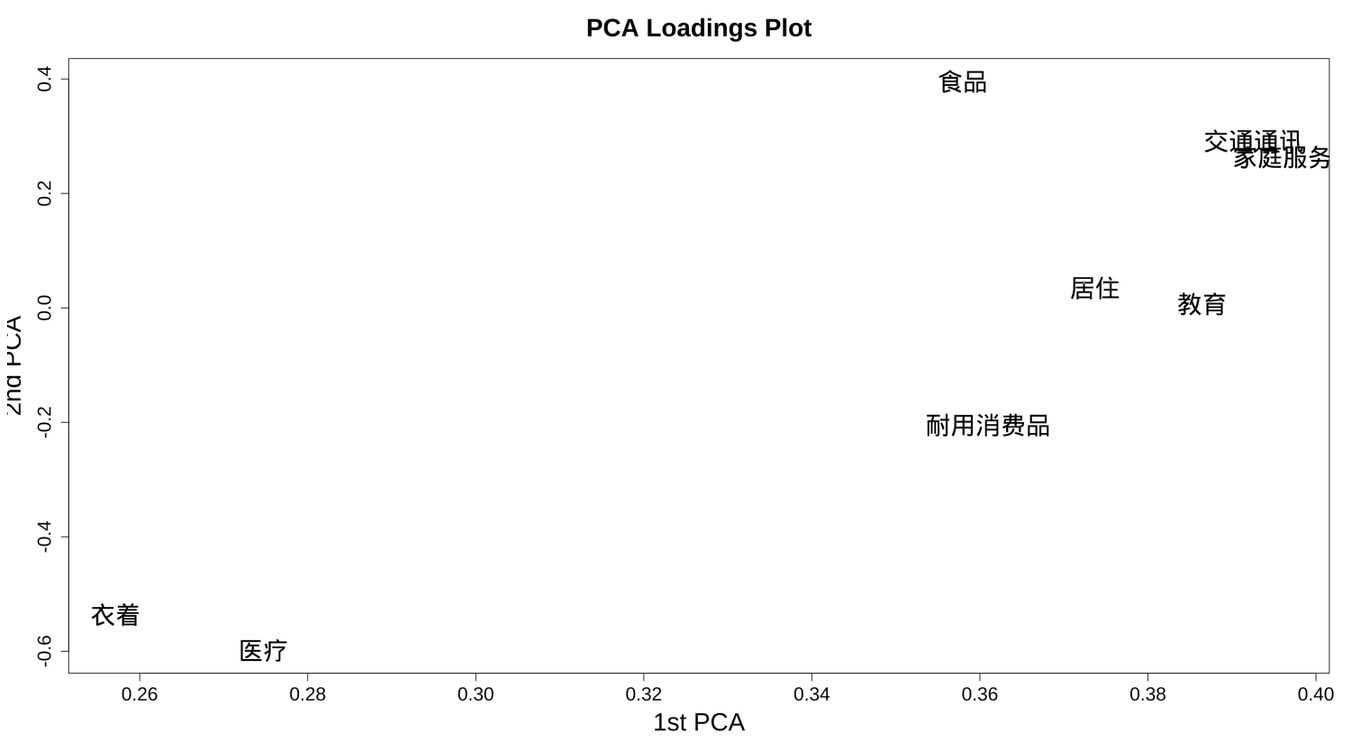
3.2 Loadings Plot

下图为factor analysis跟PCA analysis，可以看出factor analysis比较明显能将不同类型的特征区分开来

```
par(mfrow=c(1,3))
load1 <- loadings(fac1)
plot(load1, type='n', main='不旋转的因子分析载荷',cex.lab=2, cex.axis=1.5, cex.main=2)
text(load1, rownames(load1), cex = 2)
load2 <- loadings(fac2)
plot(load2, type='n', main='varimax旋转的因子分析载荷',
cex.lab=2, cex.axis=1.5, cex.main=2)
text(load2, rownames(load2), cex = 2)
load3 <- loadings(fac3)
plot(load3, type='n', main='promax旋转的因子分析载荷',
cex.lab=2, cex.axis=1.5, cex.main=2)
text(load3, rownames(load3), cex = 2)
```



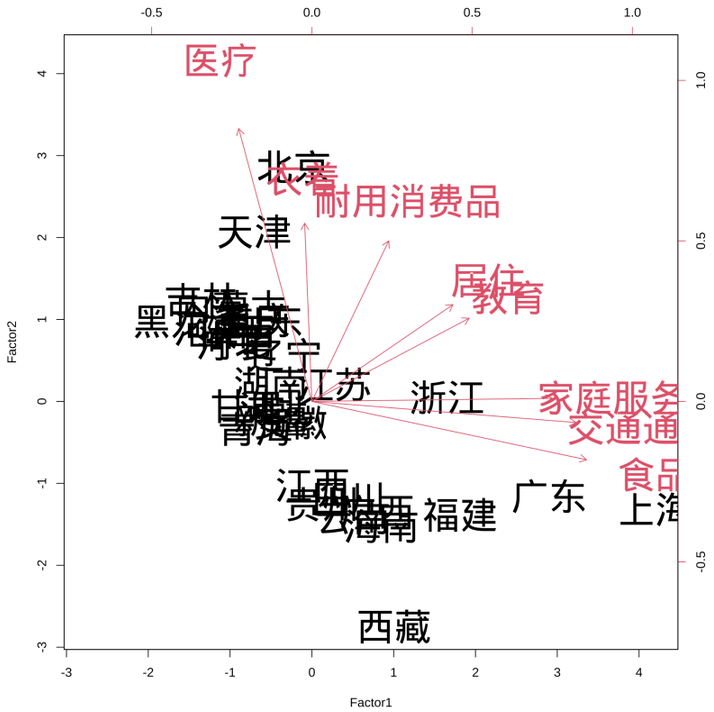

```
plot(loadings(pca1)[,1], loadings(pca1)[,2], xlab="1st PCA",
     ylab="2nd PCA", main="PCA Loadings Plot", type="n",
cex.lab=2, cex.axis=1.5, cex.main=2)
text(loadings(pca1)[,1], loadings(pca1)[,2], labels=colnames(df)[1:8]
     , cex = 2)
```



3.3 biplot

相对PCA的biplot，FA的biplot比较明显能看出聚类的效果

```
biplot(fac3$scores, fac3$loadings,cex = 3)
```



4. 小結

此次的分析结果，结合累计解释变异，PCA和FA各有好坏，像是一样两个维度，PCA的累积解释变异较高，然而以特征的分类和省份聚类来看，FA又表现较佳，因此必须依照目的和该数据的特性来选择降维方法。