地理建模实验 6 实验报告

42109232　吕文博　地信 2101 班

2024-06-11

主成分分析

读取和清洗数据

```
library(tidyverse)
library(broom)
library(ggfortify)
urban_econindic = readxl::read_xlsx('../data/exp6/6.xlsx',sheet = 'Chp 第 16 题')
names(urban_econindic) = str_extract(names(urban_econindic),"^[^/]+")
head(urban_econindic) %>% knitr::kable()
```

城市编号	总人口	非农业人口比例	农业总产值	工业总产值	地方财政预算内收入	城乡居民年底储蓄余额	在岗职工工资总额
1	1249.90	0.60	184.34	1999.97	279.09	2680.66	577.33
2	910.17	0.58	150.11	2264.55	112.81	1130.19	225.43
3	875.40	0.23	291.87	688.58	35.23	709.59	75.89
4	299.92	0.66	23.60	273.78	20.33	394.31	65.40
5	207.78	0.44	36.53	81.65	10.58	139.66	30.93
6	677.08	0.63	129.54	582.67	56.79	901.70	115.28

使用 baseR 的 prcomp 函数联同 tidyverse 执行主成分分析

```
urban_econpca = urban_econindic %>%
  nest() %>%
  mutate(pca = map(data, \(x) prcomp(select(x,-`城市编号`),
                                     center = TRUE, scale = TRUE)),
         pca_aug = map2(pca, data, \(x,y) augment(x, data = y)))

urban_econpca
```

```
## # A tibble: 1 x 3
##   data              pca      pca_aug
##   <list>            <list>   <list>
## 1 <tibble [35 x 8]> <prcomp> <tibble [35 x 16]>
```

```
var_exp = urban_econpca %>%
  unnest(pca_aug) %>%
  summarise(across(contains(".fittedPC"),\(x) stats::var(x))) %>%
  gather(key = pc, value = variance) %>%
  mutate(var_exp = variance / sum(variance),
         cum_var_exp = cumsum(var_exp),
         pc = str_replace(pc, ".fitted", ""))

var_exp
```

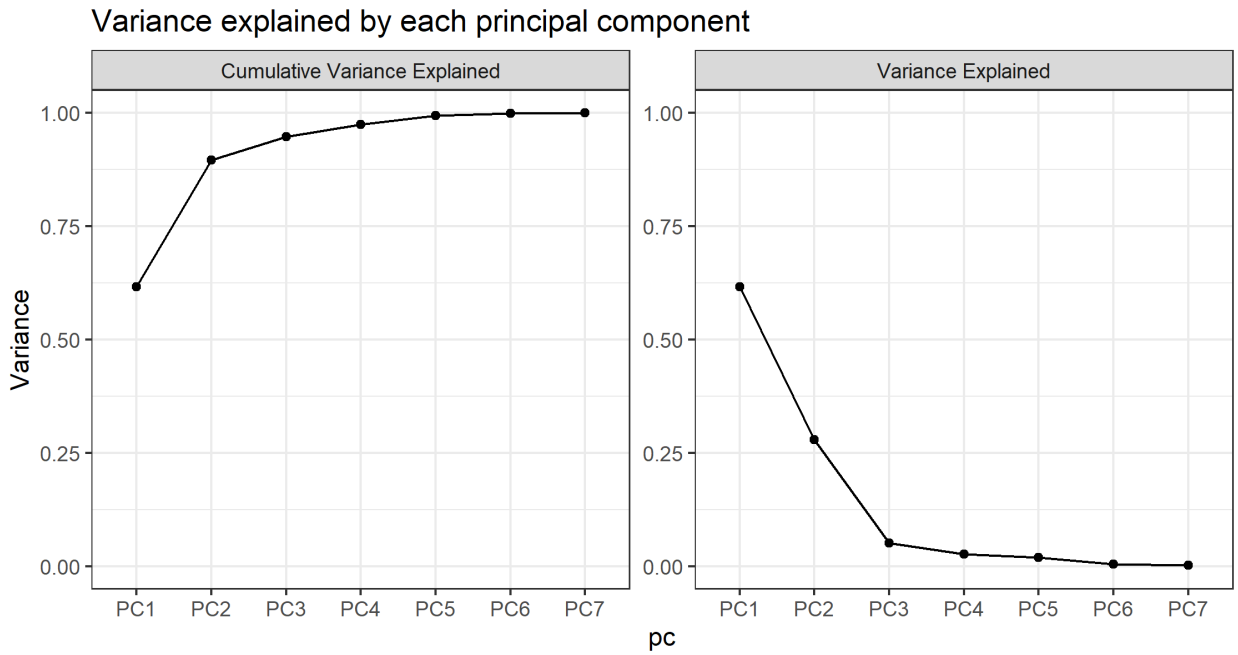
```
## # A tibble: 7 x 4
##   pc    variance var_exp cum_var_exp
##   <chr>    <dbl>   <dbl>       <dbl>
## 1 PC1     4.31   0.616         0.616
## 2 PC2     1.95   0.279         0.896
## 3 PC3     0.360  0.0514        0.947
## 4 PC4     0.185  0.0264        0.973
## 5 PC5     0.138  0.0198        0.993
## 6 PC6     0.0331 0.00473       0.998
## 7 PC7     0.0150 0.00214       1
```

按照特征值大于 1 的原则, 第 1 主成分的初始特征值为 4.31, 第 2 主成分的初始特征值为 1.95. 从第 3 主成分开始, 其初始特征值均小于 1. 因此, 选择前 2 个主成分可以得到 89.6% 的累计贡献率, 即表示前 2 个主成分可以解释 89.6% 的总方差.

碎石图

```
var_exp %>%
  rename(`Variance Explained` = var_exp,
         `Cumulative Variance Explained` = cum_var_exp) %>%
  gather(key = key, value = value, `Variance Explained`:`Cumulative Variance Explained`) %>%
  ggplot(aes(pc, value, group = key)) +
  geom_point() +
  geom_line() +
  facet_wrap(~key, scales = "free_y") +
  theme_bw() +
```

```
  lims(y = c(0, 1)) +
  labs(y = "Variance",
       title = "Variance explained by each principal component")
```

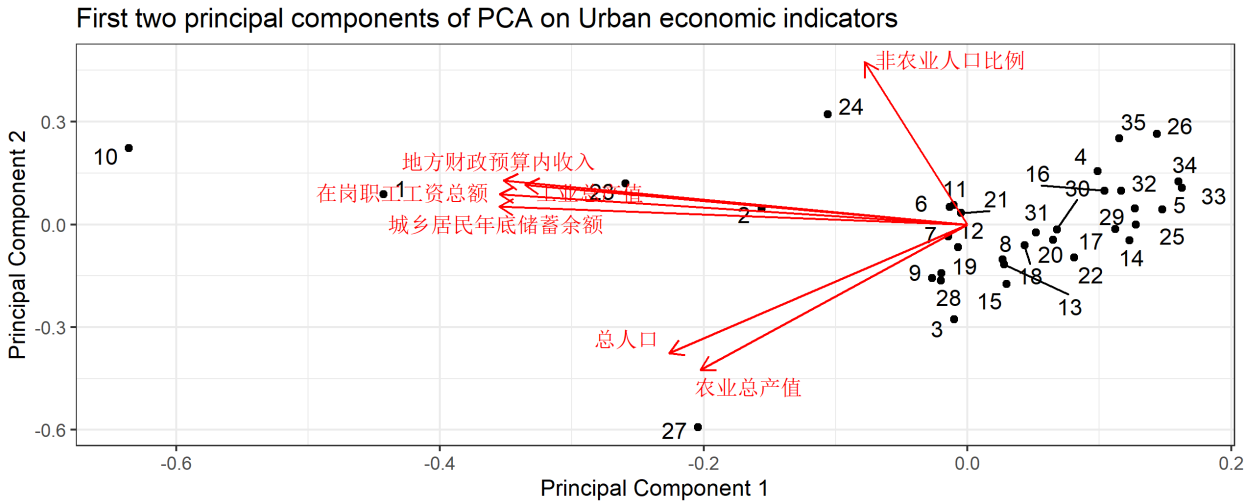


从碎石图中可以看出, 从第 3 个组件开始, 特征值就处于一个较低的水平. 因此选择前 2 个主成分是科学的。

将前两个主成分为 x 和 y 轴展示数据

```
urban_econpca %>%
  mutate(
    pca_graph = map2(
      .x = pca,
      .y = data,
      ~ autoplot(.x, loadings = TRUE, loadings.label = TRUE,
                 loadings.label.repel = TRUE,
                 data = .y, label = TRUE,
                 label.label = " 城市编号",
                 label.repel = TRUE) +
        theme_bw() +
        theme(pan)
        labs(x = "Principal Component 1",
             y = "Principal Component 2",
```

```
            title = "First two principal components of PCA on Urban economic indicators")
    )
  ) %>%
  pull(pca_graph)
```



计算综合评价得分

R 中的特征向量默认指向负方向，因此我们将乘以 -1 来反转主成分得分的符号.

```
weight_pca = var_exp %>%
  dplyr::filter(pc %in% c("PC1","PC2")) %>%
  pull(variance) %>%
  {. / sum(.)}
weight_pca
```

```
## [1] 0.6882711 0.3117289
```

```
urban_rank = urban_econpca %>%
  unnest(pca_aug) %>%
  select(城市编号,num_range(".fittedPC",1:2)) %>%
  mutate(across(-城市编号,\(x) -x)) %>%
  mutate(综合得分 = weight_pca[1] * .fittedPC1 + weight_pca[2] * .fittedPC2) %>%
  mutate(综合排名 = min_rank(desc(综合得分))) %>%
  select(城市编号, 综合得分, 综合排名) %>%
  arrange(综合排名)
```

注：与 SPSS 结果计算有出入,SPSS 中通过因子分析和综合得分两步得出结果,R 直接可以运行主成分分析，R 计算结果较准确

```
knitr::kable(urban_rank)
```

城市综合排名

城市编号	综合得分	综合排名
10	4.8032455	1
1	3.5177835	2
27	3.2554747	3
23	1.8837192	4
2	1.2007658	5
3	0.7998288	6
9	0.6319004	7
28	0.5925233	8
19	0.5331693	9
12	0.2330736	10
7	0.2131860	11
15	0.1980604	12
13	0.0661563	13
24	0.0650758	14
8	0.0396344	15
6	-0.0191400	16
21	-0.0454772	17
11	-0.0561301	18
18	-0.2079732	19
31	-0.3791397	20
20	-0.4330530	21
22	-0.4339440	22
30	-0.5374285	23
17	-0.9126734	24
14	-0.9188985	25
25	-1.0809321	26
16	-1.1324317	27
29	-1.1946543	28
4	-1.2361824	29
32	-1.2370102	30
5	-1.3626954	31
35	-1.6209127	32
33	-1.6502416	33

城市编号	综合得分	综合排名
34	-1.6779540	34
26	-1.8967252	35

```
urban_rank %>%
  head(5) %>%
  knitr::kable()
```

综合实力前五城市

城市编号	综合得分	综合排名
10	4.803246	1
1	3.517784	2
27	3.255475	3
23	1.883719	4
2	1.200766	5

```
urban_rank %>%
  tail(5) %>%
  knitr::kable()
```

综合实力后五城市

城市编号	综合得分	综合排名
5	-1.362695	31
35	-1.620913	32
33	-1.650242	33
34	-1.677954	34
26	-1.896725	35