大作业研究报告

骆炳君 2017013573 软件 71

2019/6/28

背景描述

EEC（欧洲经济共同体）和 COMECON（经济互助委员会）是冷战时期欧洲最主要的两大经济合作组织，涵盖了苏联、英国、法国、德国等大多数欧洲经济体。

本研究将利用来自 EEC 和 COMECON 的主要欧洲国家国家的蛋白质摄入量数据，分析决定蛋白质摄入结构的因素和不同种类蛋白质的摄入量间的关系，并分析两个经济合作组织在蛋白质摄入结构上的差异。

数据

介绍

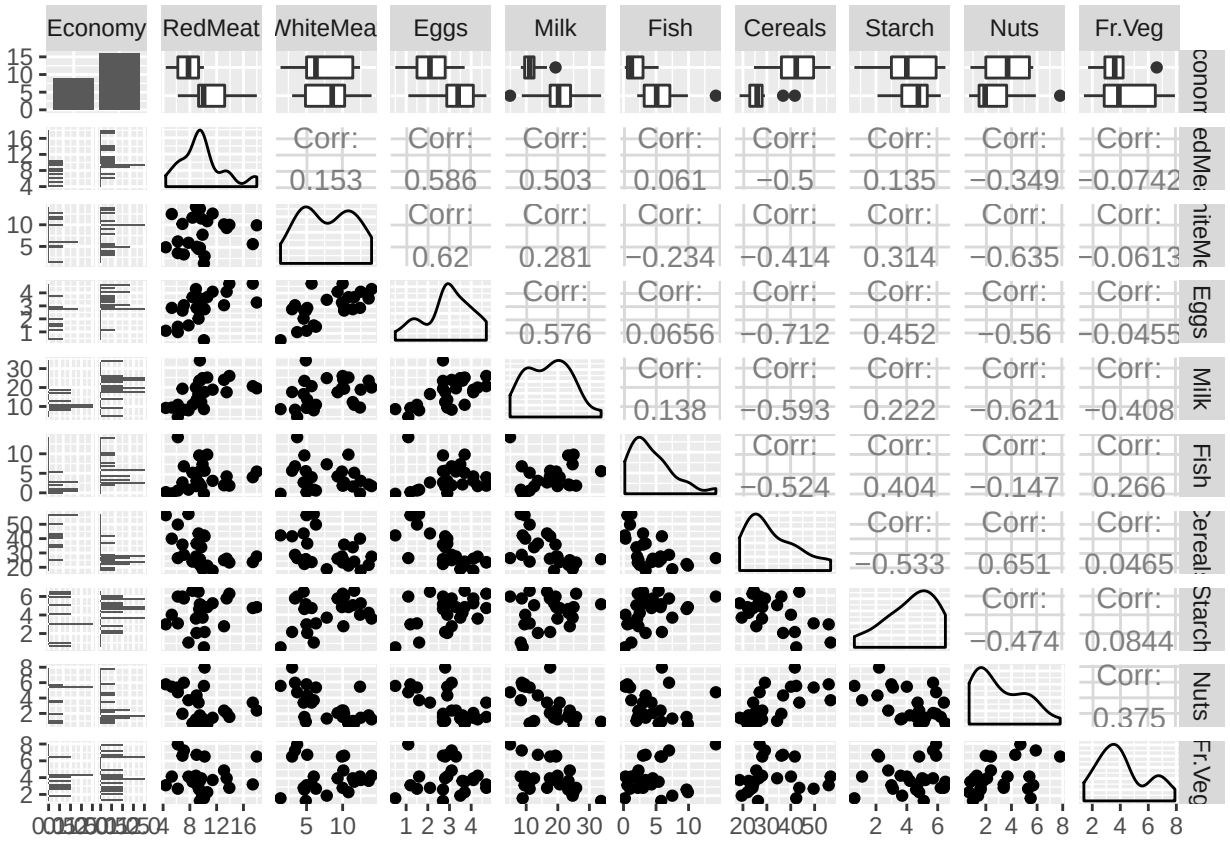
本研究使用的 Europe Protein consumption 数据集，记录了 25 个主要欧洲国家对肉、蛋、奶、鱼等 9 种类型的蛋白质摄入量情况，其中包括 16 个 EEC 国家和 9 个 COMECON 国家。由于 Country 项在本研究中没有意义，因此将此列仅作为数据标签。

表 1: 数据集概览

	Economy	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr.Veg
Albania	C	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	E	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgium	E	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgaria	C	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2

EDA

进行探索性数据分析，首先得到下面的分析图：



观察上图，可以初步得出以下结论：

- 红肉与奶，蛋与红肉、白肉、奶，坚果与谷物的摄入量存在着较强的正相关关系，谷物与蛋、奶、鱼、淀粉，坚果与白肉、蛋、奶的摄入量间存在着较强的负相关关系，蔬果与其他种类食品的摄入量都没有明显的相关关系。
- EEC 国家的红肉、蛋、奶、鱼的摄入量显著高于 COMECON 国家，COMECON 国家的谷物、坚果的摄入量显著高于 EEC 国家，EEC 国家的蔬果摄入量的方差显著大于 COMECON 国家的蔬果摄入量。

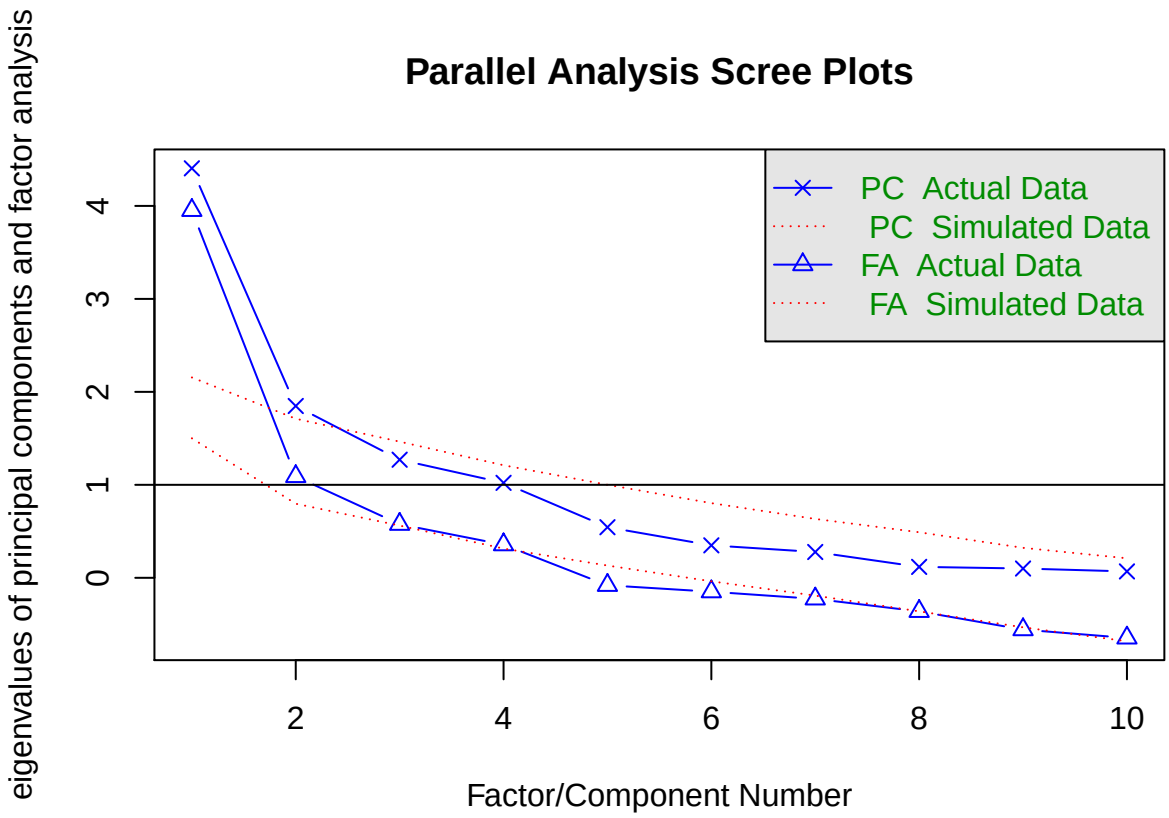
预处理

对数据进行初步处理，对 Economy 项进行数值化（C 对应 1，E 对应 2），同时求出数据集的相关系数矩阵，以方便后续的研究。

建模

因子分析

首先确定要保留的因子数量：



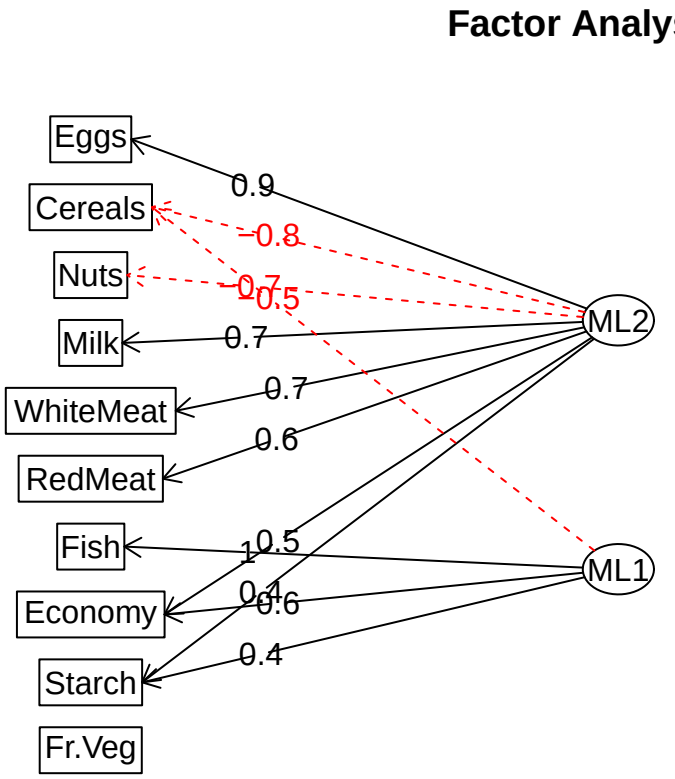

```
## Parallel analysis suggests that the number of factors =  2  and the number of components =  1
```

根据碎石图可以得出，只需保留两个因子即可。使用最大似然方法进行因子分析，得到下述结果：

```
##
## Loadings:
##           ML2    ML1
## Economy    0.505  0.562
## RedMeat    0.594
## WhiteMeat  0.675 -0.216
## Eggs       0.871
## Milk       0.677  0.157
## Fish              0.997
## Cereals   -0.773 -0.547
## Starch     0.397  0.414
## Nuts      -0.710 -0.166
## Fr.Veg    -0.222  0.261
##
##                  ML2   ML1
## SS loadings    3.591 1.962
```

```
## Proportion Var 0.359 0.196
## Cumulative Var 0.359 0.555
```

结果将数据分解为 ML2 和 ML1 两个因子，共同解释了约 55% 的方差，因子分解的效果有一定可靠性。

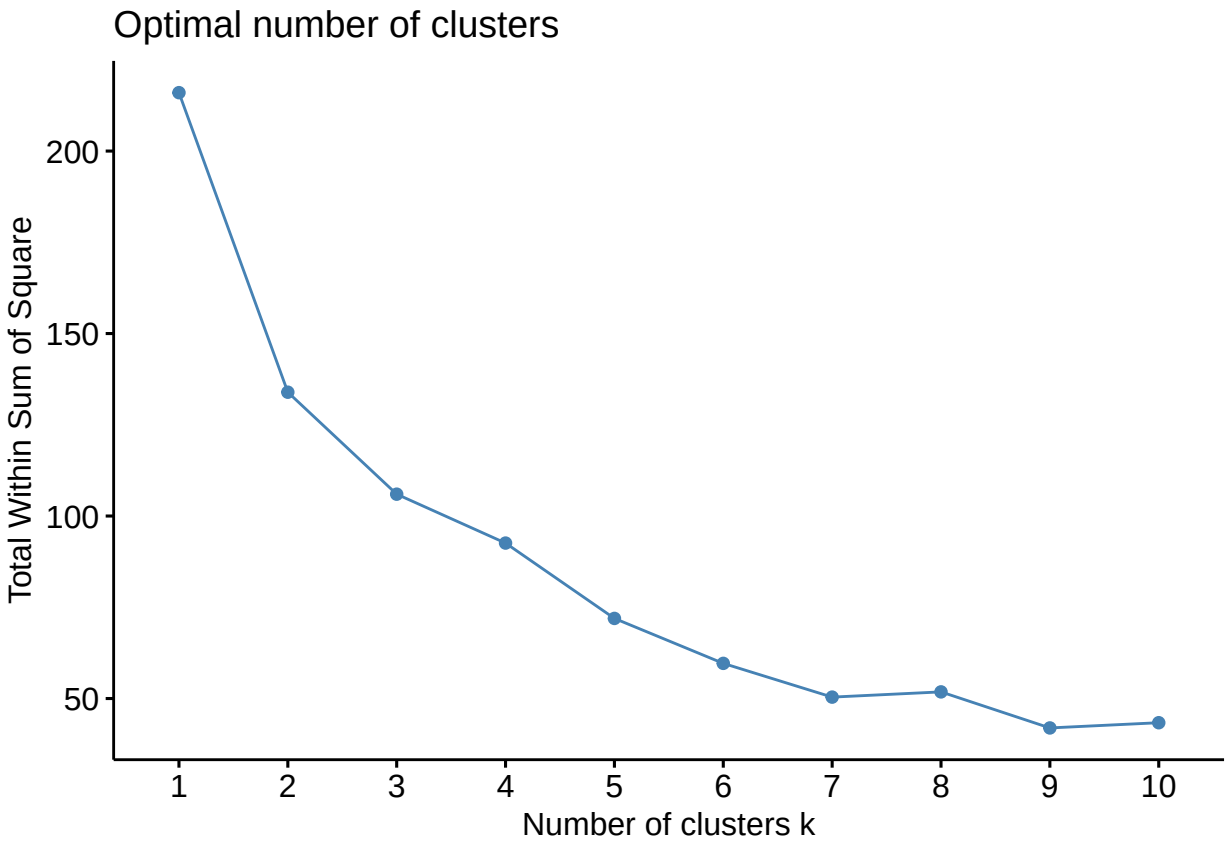


观察分解结果，可发现 ML2 是与蛋、奶、肉摄入量有关的因子，而 ML1 则是与鱼摄入量有关的因子，除此之外的谷物、经济体和淀粉是两者共有的因子，蔬果摄入量则与两个因子关系都不大。同时可以发现，两个最主要的因子都与 Economy 正相关，由此可猜测 EEC 国家的蛋白质摄入量总体上大于 COMECON 国家。

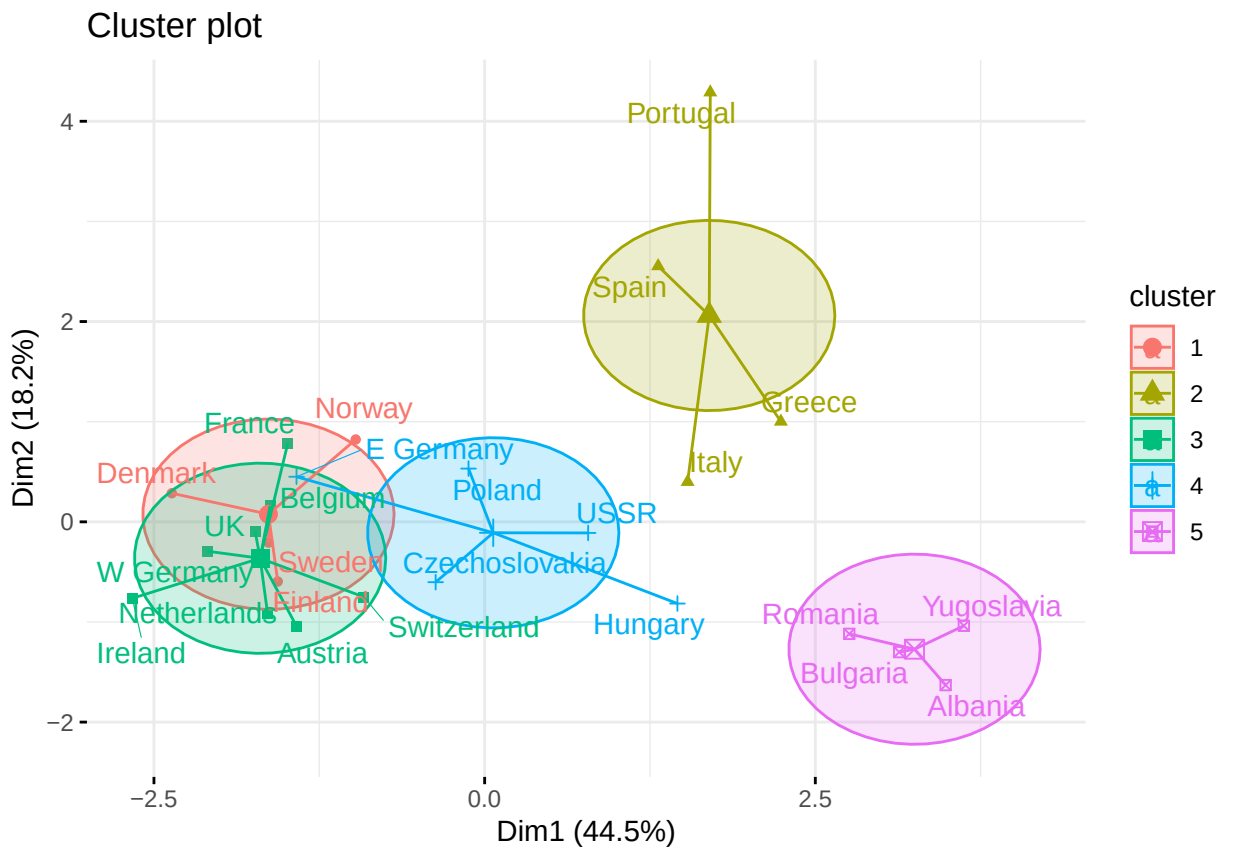
聚类分析

聚类分析试图按照各种蛋白质的摄入量对样本中的欧洲国家进行分类，同时验证 EEC 国家与 COMECON 国家的蛋白质摄入量之间是否存在显著差异，因此我们首先将数据中的 Economy 列删去，然后进行聚类，再将聚类结果与原有数据进行比较。

首先确定最优聚类数目，同样使用碎石图：



取聚类数为 5，采用 K-means 算法进行聚类操作，可得以下结果：



结合相关地理知识可知，类 1 包括了芬兰、挪威等北欧国家，类 2 包括了英国、德国、法国等主要的中西欧国家，类 3 包括了苏联、斯洛伐克、波兰等东欧国家，类 4 包括了罗马尼亚、南斯拉夫等巴尔干半岛国家，类 5 则是希腊、意大利等地中海国家。由聚类结果可知，地理环境是决定蛋白质摄入结构的主要原因，在排除 EEC 和 COMECON 国家分布上本身具有的地理因素外，可认为两种经济合作组织国家的蛋白质摄入结构没有显著差异。

结论

1. 地理环境是决定蛋白质摄入结构的主要因素，可根据不同的地理位置将欧洲国家大致分为北欧、东欧、巴尔干半岛、地中海沿岸和中西欧（或其他）5 个地理分区，不同分区间蛋白质摄入结构存在着较大差异。
2. 在蛋白质总摄入量中，可将蛋、奶、肉归为一类，它们之间存在着较强的正相关关系，并与其他种类存在着一定的负相关关系（因为比例总量一定），蔬果的摄入量与其他各类食品均无明显的相关性。
3. 在排除地理因素的差异后，不同经济合作组织的国家间不存在显著的蛋白质摄入结构差异。

附录

```
# 数据读取与预处理
rawdata <- read.table('Europrotein.dat',header = TRUE,sep = ':')
```

```
data<-rawdata
rownames(data) <- data$Country
data<-subset(data, select = -Country)
data$Economy=as.numeric(data$Economy)
data=scale(data)
R=cor(data)

# EDA
library("ggplot2")
library("GGally")
ggpairs(data)

# 因子分析
library("psych")
fa.parallel(R, n.obs = 25, fa = "both")

fa_model <- fa(R, cor=TRUE, nfactors = 2,rotate = "none", fm = "ml")
fa_model$loadings

fa.diagram(fa_model, simple = FALSE)

# 聚类分析
data_nonecon<-subset(data, select = -Economy)

library(factoextra)
fviz_nbclust(data_nonecon, kmeans, method = "wss")

km <- kmeans(data_nonecon, 5, nstart = 25)
fviz_cluster(km, data = data_nonecon,
             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)
```