多元统计分析

欧洲国家蛋白质营养差异分析

工 71 张岩 2017010907

2020 年 6 月 13 日

摘 要 本文基于多元统计分析的方法，对于我们得到的特定时期的欧洲 25 国不同食物类型的蛋白质消费数据进行分析。利用主成分分析和因子分析进行降维和解读，利用 fisher 判别的方法对经济状况进行分类，同时利用层次聚类方法对饮食结构进行聚类。

关键词： 蛋白质消费；数据降维；判别分析

目录

1 研究背景描述

欧洲作为海洋性气候地区，居民的膳食结构模式有着非常典型的特点，欧洲整体来说畜牧业十分发达，有着摄取肉类的饮食传统。肉、奶、蛋和鱼是欧洲饮食的重要组成部分，欧式餐饮以高蛋白食品为主的特点闻名于全世界。

针对于欧洲的饮食结构，我们对从定量的多元分析方法在更加具体的角度去对蛋白质消费与摄取的差异感兴趣。这些差异不仅来自于不同地区的差异，也来自于整体上动物性蛋白与植物性蛋白的摄取差异的关系。我们如何概况和定义一类地区的蛋白质消费的特点，在得到一批新的地区的蛋白质消费的数据后，我们又如何将其进行分类。本文针对该话题提出了以下问题：

- 不同类别国家的膳食结构的差异性主要体现在哪类食品？
- 不同类别食物的蛋白质摄入量间是否存在隐含的影响因子？
- 如何通过饮食结构对国家的经济状态进行判别？
- 如何将整个欧洲的膳食结构进行分组？
- 动物性蛋白与植物性蛋白的摄入存在着怎样的差异？

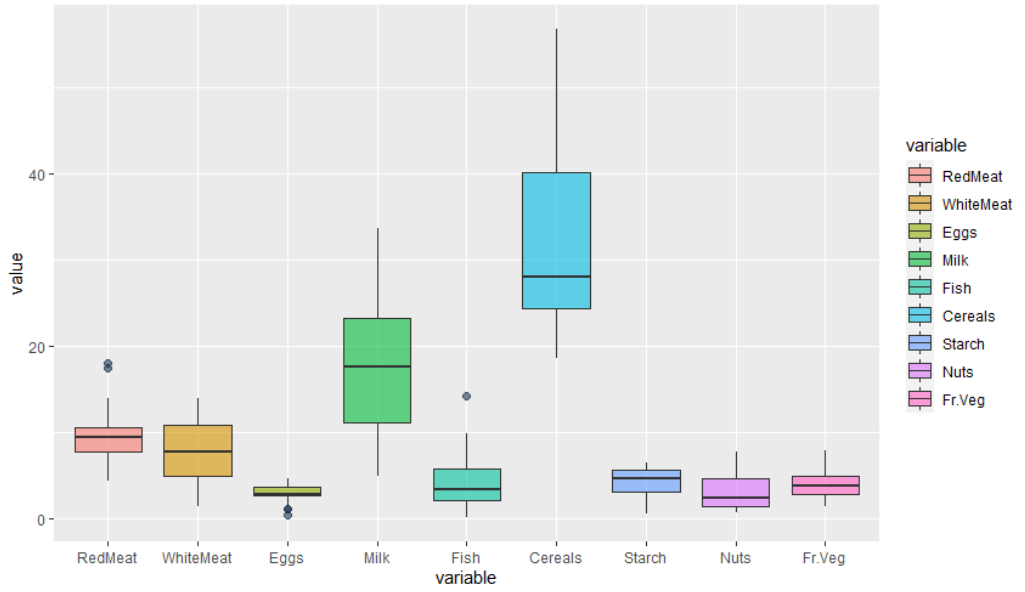
2 数据介绍

2.1 数据集描述

本项研究所用数据为课程提供的 Europe Protein consumption 数据集，数据集中有 25 条样本的记录，共有 11 个变量，其中第一列变量表示经济类型："E"表示欧盟国家，"C"表示不结盟运动国家或者经济互助委员会(苏联为首的欧洲社会主义阵营国家)，第二列为国家的名字。在这 25 个样本中有 E 类国家 16 个，C 类国家 9 个。从国家信息来看，该数据集收集于冷战时期，大约为 1970 年前后。其余 9 个变量分别 9 种食物的蛋白质摄取量：红肉、白肉、鸡蛋、牛奶、鱼、谷物、淀粉类食物、坚果类、水果蔬菜，且均为连续型变量。

2.2 单变量描述

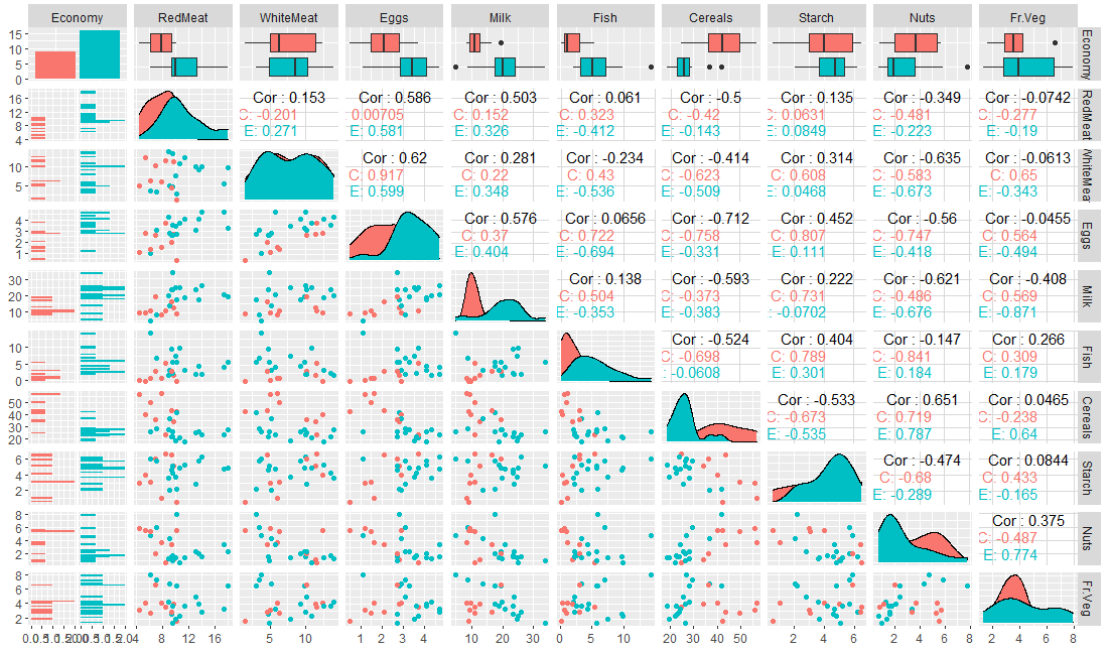
首先我们观察全体数据本身的尺度等基本信息，暂时不考虑不同国家之间的分类，画出所有 9 个连续变量的箱线图观察变量的分布：



从图中可以看到，数据中的异常值非常少，这说明数据本身的质量较高，且我们不需要对变量进行一些特殊变换。这 9 个变量中，除了 Milk 和 Cereals 的数值和方差较大之外，其他变量尺度基本相似。Milk 和 Cereals 的数值对比于其他变量也没有超出一个数量级，且这 9 个变量具有相同的含义，因此对数据不进行标准化也是可取的，但是在对数据进行降维的操作时，我们仍可以考虑标准化。

2.3 变量相关关系描述

接下来我们考虑不同变量之间的关系，同时考虑根据经济类型的分组信息，画出相关关系图，分析这些变量：



首先从单个变量的分布图中可以看到，某些变量在不同类别国家的分布差异巨大，例如蛋类、牛奶、鱼、谷类和蔬菜。且大多数变量并不服从正态分布，因

此我们不能使用依靠正态性假设的分析方法。从散点图可以看到不少变量之间存在着线性相关性，这为我们对数据进行降维提供了依据。同时我们可以从相关系数矩阵观察到，不少变量在不同的分组中具有明显的相反的相关关系。这提醒我们在对数据进行分析时，要着重考虑影响变量的潜在因素和数据可能存在的分组信息造成的影响。

3 数据降维

3.1 因子分析

我们首先考虑求得可解释这 9 种食物蛋白质消耗量的潜在因子，为了平衡数据尺度的问题，基于数据的相关系数矩阵对数据进行因子分析建模。在综合考虑之后提取 4 个公共因子，使用极大似然的估计方法，同时为了更好的解释性，对因子进行旋转。因子的载荷和方差的解释比例如下所示：

变量	Factor1	Factor2	Factor3	Factor4	共同度	特殊方差
RedMeat	0.720	0.038	0.070	-0.147	0.547	0.453
WhiteMeat	0.168	0.957	-0.065	-0.051	0.951	0.049
Eggs	0.780	0.520	0.138	0.005	0.899	0.101
Milk	0.572	0.179	0.228	-0.516	0.678	0.322
Fish	0.030	-0.177	0.967	0.136	0.986	0.014
Cereals	-0.564	-0.366	-0.610	0.137	0.844	0.156
Starch	0.227	0.339	0.474	-0.002	0.391	0.609
Nuts	-0.264	-0.610	-0.334	0.560	0.867	0.133
Fr.Veg	-0.078	-0.002	0.173	0.740	0.583	0.417
SS loadings	1.93	1.872	1.755	1.189		
Proportion Var	0.214	0.208	0.195	0.132		
Cumulative Var	0.214	0.422	0.617	0.75		

可以看到四个因子的解释占比达到了 0.75，对原数据具有较好的解释性。由因子载荷可见，Factor1 主要影响的是红肉、蛋、牛奶和谷物，Factor2 主要影响的是白肉、蛋和坚果，Factor3 主要影响的是鱼、谷物和淀粉类，Factor4 主要影响的是牛奶、坚果和蔬菜。从解释性的角度，我们可以将 Factor1 解释为日常食物，Factor2 解释为健康的蛋白质，属于更高的生活水平的人会食用的，Factor3 解释为主食，Factor4 解释为低热量的食物。

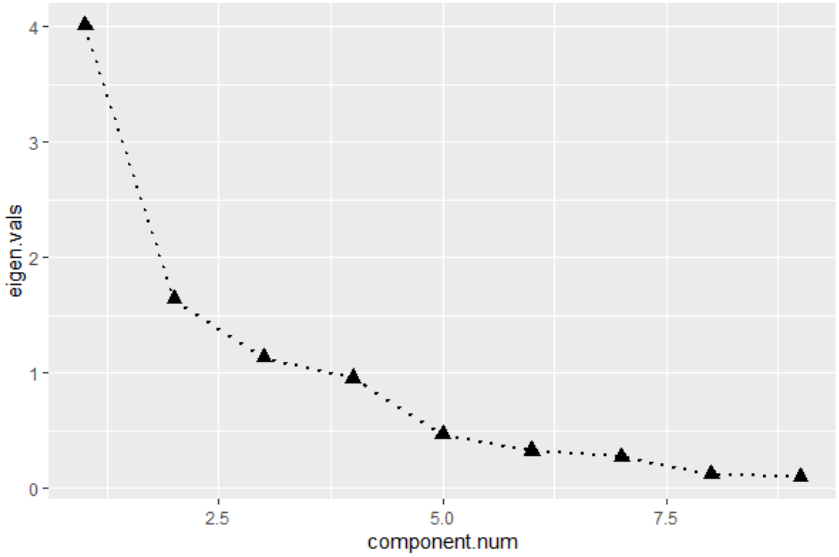
3.2 主成分分析

利用主成分分析的方法对数据进行降维和特征提取。在探索性分析时我们发现数据的尺度稍微有不同，因此为了平衡不同尺度的变量，我们对标准化后的数据进行主成分分析。列出各个主成分的方差解释占比，并画出碎石图：

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**
标准差	2.0016	1.2787	1.0620	0.9771	0.6811
方差占比	0.4452	0.1817	0.1253	0.1061	0.0515

累积占比	0.4452	0.6268	0.7521	0.8582	0.9098

	Comp.6	Comp.7	Comp.8	Comp.9
标准差	0.5702	0.5212	0.3410	0.3148
方差占比	0.0361	0.0302	0.0129	0.0110
累积占比	0.9459	0.9761	0.9890	1



从碎石图和方差比例可以清楚地看到，前四个主成分可以解释 85%的方差，这表示降维后的四维数据已经可以解释原始九维数据中 85%的变化，因此我们可以认为选取前四个主成分非常合适。与此同时，选取 3~5 个主成分都是可以接受的。此时计算得到特定个变量的得分即可作为降维后的数据。主成分不会因为保留的变量个数而改变，因此在获得主成分时只需要对得到的结果的前 n 维进行保留即可。

4 类别信息分析

4.1 对 Economy 的判别分析

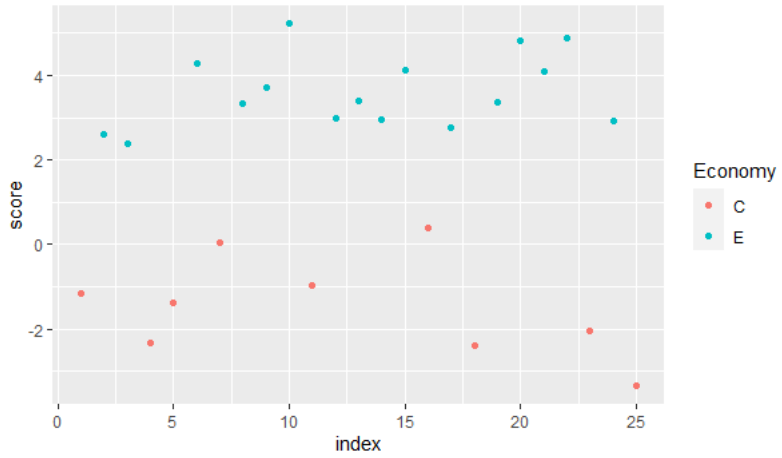
4.1.1 判别方法的介绍

我们首先查看不同 Economy 的国家之间的差距是否显著，使用双总体 Hotelling's T^2的方法进行假设检验。使用该方法需要对数据进行一些假设：1.数据服从独立多元正态分布;2.两总体方差相等。按照标准的Hotelling's T^2方法进行假设检验，具体细节在此不必展示，只需知道否认原假设的 p 值为$4.476e-05$，因此两总体间有非常大的差异。

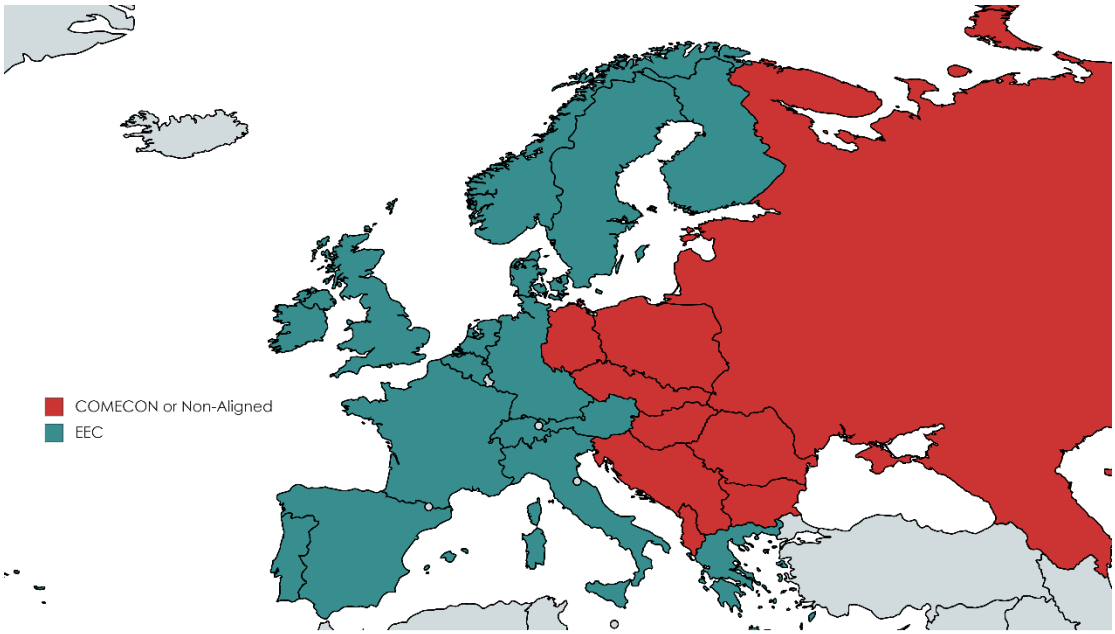
由于差异显著，我们接下来考虑使用 fisher 线性判别的方法对未知标签的数据进行判别。线性分类是一种非常简单的分类方法，在实际应用中基本不会出现过拟合的情况。由于我们要分的类别只有两类，fisher 判别的方法相当于是将样本投影到一条直线上，使得在这一条直线上两样本的区分度更大。

在进行具体的算法准确度评估前，为了对方法的应用效果有一个大致评估，我们首先使用原始的全部数据进行模型的拟合，经计算得到对原始数据进行投影的系数向量和所有样本的得分如下所示：

$$a' = [0.0279 \quad -0.1074 \quad 0.8899 \quad 0.1536 \quad 0.2463 \quad -0.1254 \quad -0.5871 \quad 0.3401 \quad 0.3812]$$



从图上可以看到两种类别间的得分的差异也十分明显，因此使用 fisher 线性判别是一个很好的分类方法。我对线性分类具有良好效果的原因很感兴趣，除了从数据对效果进行解释之外，考虑到该数据集所具有的实际意义，将样本中的每个国家在地图上标出。两种不同种类的国家分别使用两种不同的颜色，地图如下图所示：



从图中可以看出，C 和 E 两类国家在地图上的划分也非常有特点，尽管 C 和 E 的实际意义仅仅是国家的经济体制，但是在实际上他们所体现出来的是不同地区之间的地理气候因素和文化因素，因而这些地区在饮食结构上所体现出来的差异也是非常合乎我们的理解的。这说明我们的数据分析与数据的实际意义非常契合，所使用的判别方法也是非常有价值的。

4.1.2 判别方法准确度分析

为了评估分类的准确度，我采取了五折交叉验证的方法，将数据打乱顺序随机分为 5 组，这五组数据轮流做测试集，剩下四组做训练集拟合模型，进行交叉验证，将求得的五个准确率取均值，如此为一次五折交叉验证。在所使用的数据上，我们考虑使用原始数据和 PCA 方法降维后保留前 3 至 6 个主成分的这五种数据集进行验证，由于我们使用了随机分组，对每种数据分别进行 1000 次五折交叉验证，所得到的准确率求均值，结果如下：

	原始数据	3 个主成分	4 个主成分	5 个主成分	6 个主成分
准确率	0.87	0.7769	0.8889	0.9562	0.9315

综上，为了得到最高的准确度，可以选择降维至五个主成分的数据，使用 fisher 线性判别的方法进行分析。所得到的判别的投影方向与判别准则为

$$a' = [0.6857 \quad -0.7106 \quad 0.7647 \quad 0.9061 \quad -1.1192]$$

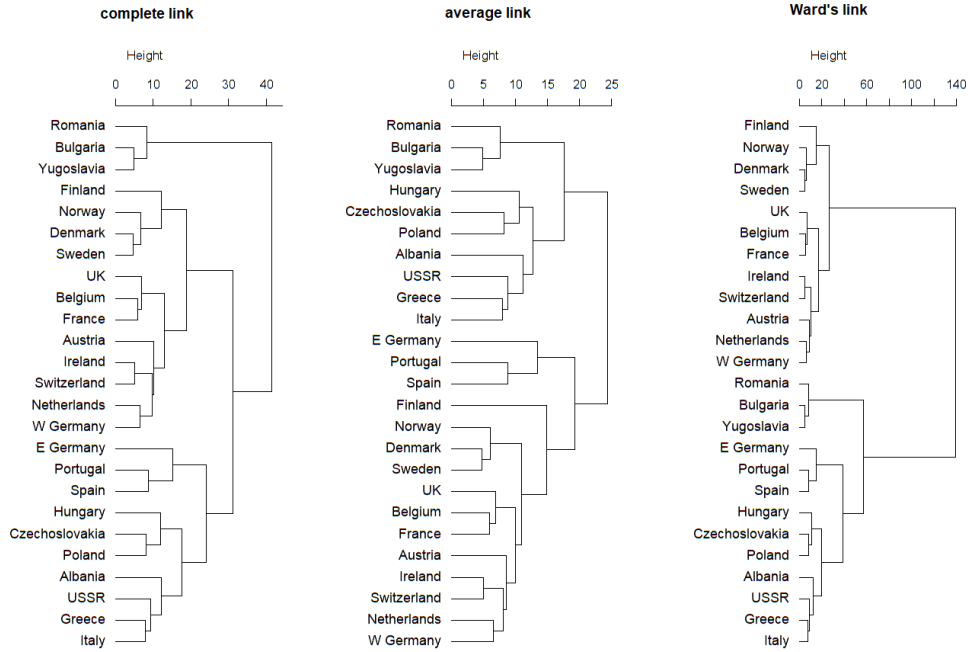
$$\begin{cases} a'x \geq a'\dfrac{\overline{x}_C + \overline{x}_E}{2} & \text{判别为 E 类国家} \\ a'x < a'\dfrac{\overline{x}_C + \overline{x}_E}{2} & \text{判别为 C 类国家} \end{cases}$$

$$\overline{x}_C = [-1.5111 \quad 0.6391 \quad -0.4745 \quad -0.4758 \quad 0.2855]$$

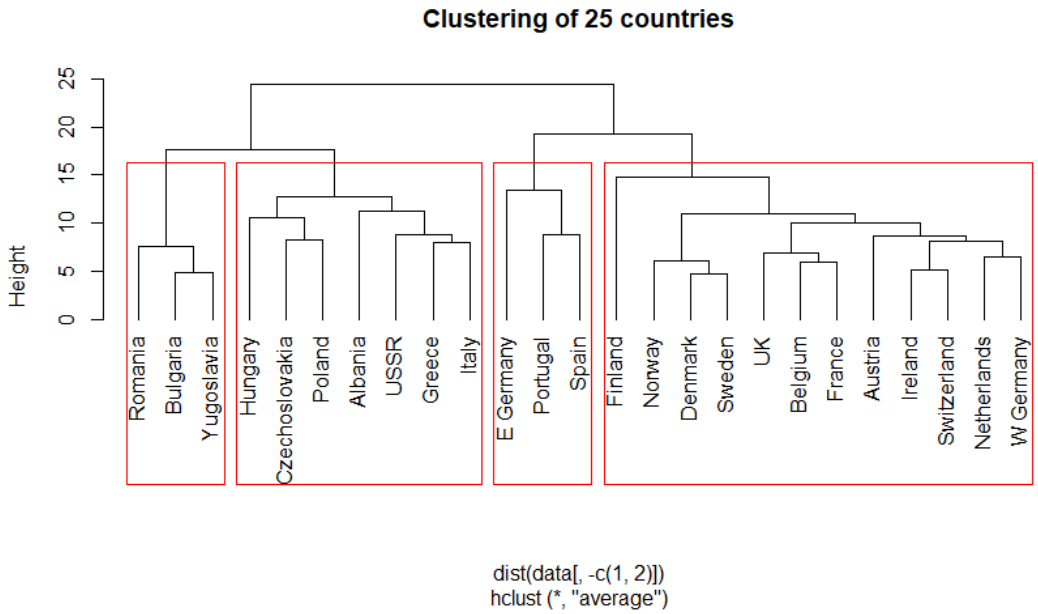
$$\overline{x}_E = [0.8500 \quad -0.3595 \quad 0.2669 \quad 0.2676 \quad -0.1606]$$

4. 2 对不同国家进行聚类分析

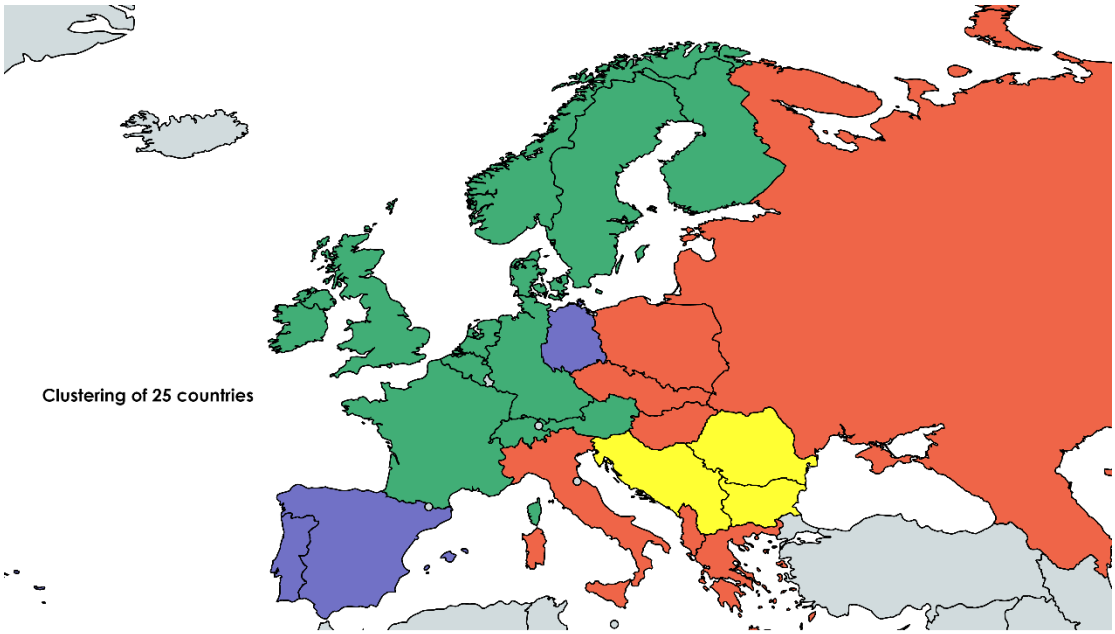
如果我们忽略掉数据中已有的 Economy 标签，仅仅从蛋白质消费的角度将这 25 个国家分为几个较为明显的类别。由于我们无法知道潜在的类别的数目，我选择了使用层次聚类的方法进行聚类分析。为了对比不同方法的聚类效果从而找到最合适的结果，我分别使用了 complete link，average link 和 Ward’s link 三种方法进行层次聚类，画出三种方法的结果的聚类树状图，如下所示：



从以上结果中，我选择了 average link 的结果，因为 average link 的结果对极端值不敏感，且该聚类结果直观上看起来较为均匀，选择聚类的类别为 4 个，将同一类别进行标记，结果如下：



为了使聚类结果能够被更好地展示，将这些国家在地图上标记，同一类别下的国家使用同样的颜色，如下图所示：

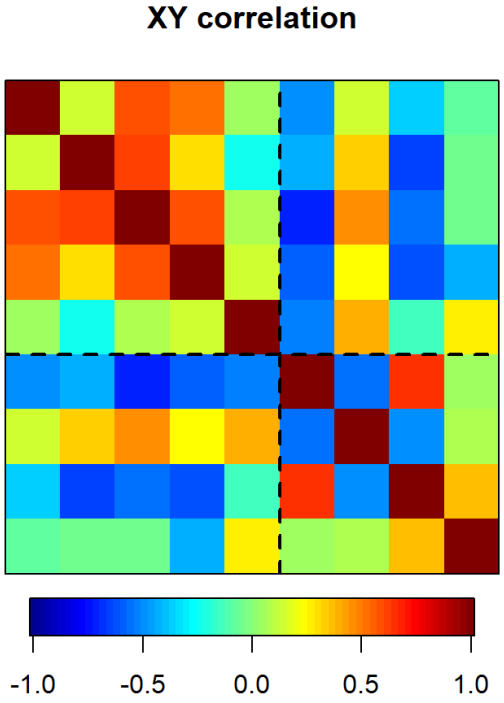


从该结果来看，同一类国家更多地趋向于位于同一地区，尽管其中可能存在一部分与该解读相悖的结果。绿色的国家可能是高纬度且受大西洋影响较大的国家，蓝色国家（尤其是西班牙和葡萄牙）受大西洋和地中海共同作用，红色国家位于内陆，受地中海和内陆的气候特点影响，而黄色国家位于亚欧交接，且位于历史较为复杂的巴尔干半岛。以上只是我们完全从蛋白质消费的数据出发对国家

进行的分组，同一类国家地区上的统一性恰好说明了地理文化等因素对饮食结构的影响，尽管这其中更加具体的原因可能是历史学和人类学的范畴，这一聚类结果仍然表明我们对数据的解读是比较好的。

5 动物性食品与植物性食品的关系

在 EDA 阶段我们仅仅考虑了两个单独变量之间的关系，接下来我们要分析两组变量之间的关系，我们按照变量的含义分为两组：动物性食品（红肉、白肉、鸡蛋、牛奶、鱼）和植物性食品（谷物、淀粉类、坚果、水果蔬菜）。我们首先将相关关系矩阵进行可视化，按照两组变量将相关关系矩阵进行分块：



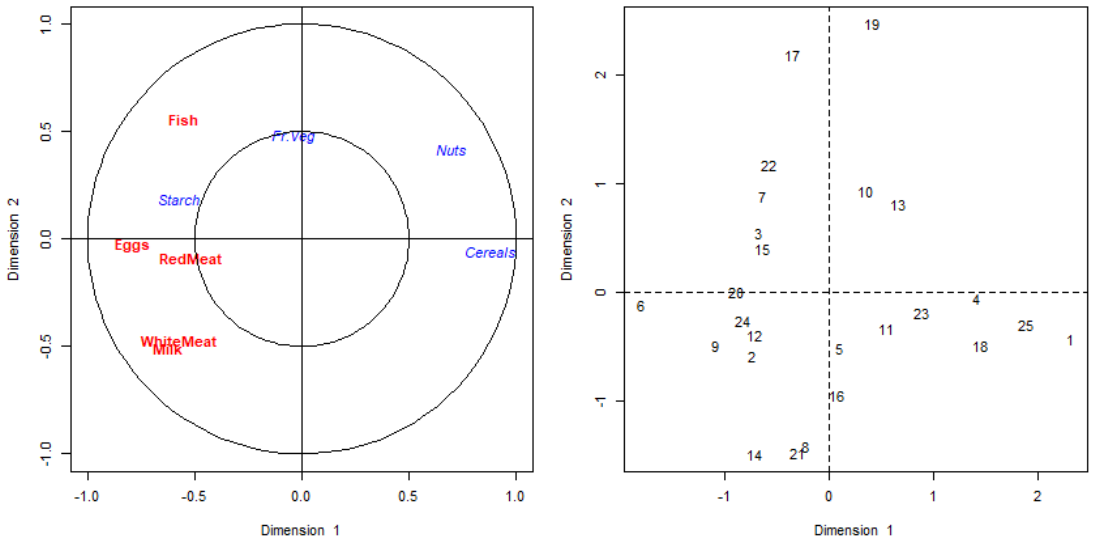
从上图可以看到，两组变量之间的相关关系体现在左下角与右上角的部分，这两组变量之间多呈现出负相关的关系。为了进行更加具体的描述，我们采用典型相关分析(CCA)的方法研究这两组变量之间的关系。CCA 的思路是将两组原始变量投影，使得两组变量的投影尽可能地重合。两组变量分别包含 5 个变量和 4 个变量，因而我们最多找到四个变量对，使用 R 中的 CCA 包中的函数进行计算，求出四个变量对之间的相关系数分别为 0.9180, 0.6955, 0.4826, 0.2317。我们最关注这最相关的一个变量对，该变量对为

$$U = -0.0516 \times RedMeat - 0.1318 \times WhiteMeat - 0.2168 \times Eggs - 0.0246 \times Milk - 0.1796 \times Fish$$

$$V = 0.065 \times Cereals - 0.0365 \times Starch + 0.175 \times Nuts - 0.11 \times Fr.Veg$$

将典型相关分析得到的在前两个维度上的变量与样本得分的关系直观地画出来，如下所示，我们可以看到动物性食物位于左下的象限中，而植物性食物多

位于右上的象限中。也即大体来说，动物性食物与植物性食物之间存在着负相关的关系，也即通过动物性食物获取蛋白质的国家通常植物性食物蛋白消费较少。



6 结论与改进

6.1 研究问题解释

首先回顾开篇提出的几个问题，本文通过多元统计的方法对这些问题进行了详细的解答。

- 不同类别国家的膳食结构的差异性主要体现蛋类、牛奶、鱼、谷类和蔬菜这几个变量中。可以通过主成分分析降维的方法对变量进行线性组合，从而找到能够体现数据的差异性的几个主成分。
- 不同类别食物的蛋白质摄入量之间存在隐含的影响因子，通过使用因子分析的方法，我们发现这些影响因子主要包括日常食物因子、健康蛋白质因子、主食因子、低热量食物因子。
- 可以通过对饮食结构进行线性判别得分的方法对国家的经济状态进行判别，通过比较我们发现主成分分析降维后的数据更适合进行判别分析。
- 可以通过聚类的方法对不同国家的膳食结构进行分组，由于我们不知道潜在的分组个数，可以用层次聚类的方法进行分析。聚类结果中，相近地区通常在同一类中。
- 动物性蛋白与植物性蛋白的摄入之间存在着显著的线性关系，我们找了最能体现线性关系的两组变量的线性组合。从结果来看，动物性蛋白与植物性蛋白之间可能存在着负相关的关系。

6.2 不足与改进

在对数据进行分析时，我并没有考虑数据的分布，仅仅是确保数据不存在异常值点之后便开始了下一步的分析，对于一些强烈依赖于对数据的假设的分析方

法，如果能够对数据的分布等信息进行更加详细推断，文章的逻辑会更加严谨。

在对 Economy 的判别中，我仅仅使用了 fisher 判别方法进行判别，由于 fisher 判别方法完全是一个线性的求解方法，本文中判别准确率较好的原因是两种的差异特别显著，如果换另一个地区的数据效果可能大打折扣。此外，该数据集中两种类别国家的数量并不均衡，使用 fisher 判别时则忽略了这一点，因而对原始数据的使用并不充分。接下来可以尝试用 ECM 方法、SVM 和逻辑回归等能够更加充分利用信息的方法进行判别。

此外，该数据较为老旧，为冷战时的数据，目前的世界格局发生了较大变化，欧洲各国也可能出现了非常大的改变，研究过去的数据对解读现状的作用不大。并且该数据的数据量较小，如果能够获得更加详细的数据，比如精确到欧洲国家的每一个城市或者按照地区面积进行收集数据，我们可以进行更加详细的解读。如果能够收集到不同时间的数据，我们也可以考察我们的研究方法的鲁棒性，从而得到更加普适的结论。

7 附录：所使用的 R 代码

```
#Europe Protein consumption
library(car)
library(GGally)
library(ggplot2)
library(corrgram)
library(psych)
library(MASS)
library(reshape)
library(CCP)
library(CCA)
library(ICSNP)
#######################################
#read data, EDA
data = read.table('Europe Protein consumption_Europrotein.dat',header = T,sep = ':')
data
summary(data)
cov(data[,-c(1,2)])
head(data)
# correlation
ggpairs(data[,-2],mapping = aes(color = Economy))
corrgram(data,order=TRUE, main="correlation graph of Europe Protein consumption",
        lower.panel=panel.pts ,upper.panel = panel.cor)
melt.data = melt(data)
ggplot(data=melt.data,aes(x=variable,y=value,fill=variable)) +
  geom_boxplot(alpha=0.6,outlier.colour="#1F3552",outlier.shape=20,outlier.size = 3.5)

#######################################
# FA : PC method
fit <- principal(data[,3:11], nfactors=4, rotate="varimax")
fit # print results
plot(fit$values,type="b") # scree plot
plot(fit$loadings)
plot(fit$loadings,type="n") # set up plot
text(fit$loadings,labels=names(data),cex=.7) # add variable names
# MLE method
```

```
fit2 <- factanal(data[,3:11],factors=4,rotation="varimax")
fit2$uniquenesses
fit2
plot(fit2$loadings)
plot(fit2$loadings,type="n") # set up plot
text(fit2$loadings,labels=names(data[,3:11]),cex=0.9) # add variable names

##########################################
# PCA
proteinpca = princomp(data[,-c(1,2)],scores=T,cor = T)
biplot(proteinpca)
evals<-data.frame(proteinpca$sdev^2)
names(evals)<-"eigen.vals"
evals$component.num<-as.integer(seq(nrow(evals)))
ggplot(evals,aes(x=component.num,y=eigen.vals))+geom_point(size=3,shape=17)+geom_line(
linetype="dotted",size=0.75)

##################################################
# LDA
HotellingsT2(data[data$Economy=='C',c(3:11)],data[data$Economy=='E',c(3:11)])
subdata = data[,-2]
L = lda(Economy~.,data=subdata)
yhat = predict(L, subdata)$class
subdata$Economy.pred = yhat
sum(subdata$Economy==subdata$Economy.pred)
ggplot(subdata,aes(x=Milk,y=Cereals,col=Economy,
                shape=Economy.pred)) + geom_point() +
  theme(legend.position='bottom')
scale = L$scaling
newdata$score = as.matrix(data[,-c(1,2)])%*%matrix(scale)
newdata$index = c(1:25)
ggplot(newdata,aes(x=index,y=score,colour = Economy))+geom_point()

# K-fold
kfold = function(inputdata,random=FALSE){
  acc = c()
  num = nrow(inputdata)
```

```
  if(random){
    inputdata = inputdata[sample(25),]
  }
  len = num/5
  for (k in 1:5) {
    testdataid = c((1+(k-1)*len):(k*len))
    testdata = inputdata[testdataid,]
    traindata = inputdata[-testdataid,]
    L = lda(Economy~.,data=traindata)
    yhat = predict(L, testdata)$class
    testdata$Economy.pred = yhat
    sampleacc = with(testdata,sum(Economy==Economy.pred)/len)
    acc = c(acc,sampleacc)
  }
  return(list(acc = acc,totalacc = mean(acc)))
}
kfold(data[,-2],random = T)
pcadata = data.frame(Economy = data$Economy)
pcadata = cbind(pcadata,as.data.frame(proteinpca$scores))
runtimes = 1000
for (i in 1:runtimes) {
  newresult = data.frame(origin = kfold(data[,-2],random = T)$totalacc,
                      usecomp.3=kfold(pcadata[,c(1:4)],random = T)$totalacc,
                      usecomp.4=kfold(pcadata[,c(1:5)],random = T)$totalacc,
                      usecomp.5=kfold(pcadata[,c(1:6)],random = T)$totalacc,
                      usecomp.6=kfold(pcadata[,c(1:7)],random = T)$totalacc)
  if(i==1){
    accresult = newresult
  }else{
    accresult = rbind(accresult,newresult)
  }
}
colMeans(accresult)
write.table(accresult,'kfoldresult.dat')

########################################
# CA
```

```
# ward.D
res <- hclust(dist(data[,-c(1,2)]),'ward.D')
plot(res,hang=-1,labels=data$Country,main = 'Ward\'s link',cex=1.1)
# average, Use this
res <- hclust(dist(data[,-c(1,2)]),'average')
plot(res,hang=-1,labels=data$Country,main = 'average link',cex=1.1)
rect.hclust(res,k=2,border = 'red')
plot(res,hang=-1,labels=data$Country,main = 'Clustering of 25 countries')
rect.hclust(res,k=4,border = 'red')
# complete
res <- hclust(dist(data[,-c(1,2)]),'complete')
plot(res,hang=-1,labels=data$Country,main = 'complete link',cex=1.1)
rect.hclust(res,k=4,border = 'red')

########################################
# CCA
ccres = cc(animal,plant)
U1 = ccres$scores$xscores[,1]
U2 = ccres$scores$yscores[,2]
xcoef = ccres$xcoef[,1]
ycoef = ccres$ycoef[,1]
scaleanimal = scale(animal,scale = F)
xscore = as.matrix(scaleanimal)%*%matrix(xcoef)
plot(U1,U2)
plt.cc(ccres,d1=1,d2=2,type="b",var.label=TRUE)
cca<-cancor(animal,plant)
rho <- cca$cor
```