Chapter8 - Canonical Correlation Analysis

张笑竹 / 201618070114

2018年12月10日

第八章实验采用课本例8-1，例8-2和例8-3中给出的数据进行典型相关分析。例8-1是SAS中一个生理指标与运动指标关系的的数据；例8-2是城市竞争力与城市设施关系研究；例8-3是生猪生产的主要指标。
将本次的实验任务拆分如下：
1）调用cancor()函数，对例8-1进行典型相关分析；
2）将上述函数进行封装，从而输出更加清晰的结果；
3）调用封装的函数，对例8-2进行典型相关分析；
4）继续调用封装的函数，对例8-3进行典型相关分析。

1 例8-1的分析

1.1 变量间的相关关系

首先，在R中输入例8-1的数据集。其中，第一组变量表示生理指标(physiological measurements)，X_1表示体重(weights)，X_2表示腰围(waist)，X_3表示脉搏(pulse)；第二组变量表示运动指标(exercise)，Y_1表示引体向上(chins)，Y_2表示仰卧起坐(situps)，Y_3表示跳跃次数(jumps).

```
X<-data.frame(
  X1=c(191, 193, 189, 211, 176, 169, 154, 193, 176, 156,
       189, 162, 182, 167, 154, 166, 247, 202, 157, 138),
  X2=c(36, 38, 35, 38, 31, 34, 34, 36, 37, 33,
       37, 35, 36, 34, 33, 33, 46, 37, 32, 33),
  X3=c(50, 58, 46, 56, 74, 50, 64, 46, 54, 54,
       52, 62, 56, 60, 56, 52, 50, 62, 52, 68),
  Y1=c( 5, 12, 13, 8, 15, 17, 14, 6, 4, 15,
        2, 12, 4, 6, 17, 13, 1, 12, 11, 2),
  Y2=c(162, 101, 155, 101, 200, 120, 215, 70, 60, 225,
       110, 105, 101, 125, 251, 210, 50, 210, 230, 110),
  Y3=c(60, 101, 58, 38, 40, 38, 105, 31, 25, 73,
       60, 37, 42, 40, 250, 115, 50, 120, 80, 43)
)
```

然后，将数据进行标准化，并分别计算：
①第一组变量间的相关系数；
②第二组变量间的相关系数；
③第一组和第二组变量之间的相关系数；

```
X=scale(X,center=T,scale = T)
#第一组变量间的相关系数
cor(X[,c(1,2,3)])
```

```
##            X1         X2         X3
## X1  1.0000000  0.8702435 -0.3657620
## X2  0.8702435  1.0000000 -0.3528921
## X3 -0.3657620 -0.3528921  1.0000000
```

```
#第二组变量间的相关系数
cor(X[,c(4,5,6)])
```

```
##           Y1        Y2        Y3
## Y1 1.0000000 0.6957274 0.4957602
## Y2 0.6957274 1.0000000 0.6692061
## Y3 0.4957602 0.6692061 1.0000000
```

```
#第一、二组变量间的相关系数
cor(X[,c(1,2,3)],X[,c(4,5,6)])
```

```
##            Y1         Y2          Y3
## X1 -0.3896937 -0.4930836 -0.22629556
## X2 -0.5522321 -0.6455980 -0.19149937
## X3  0.1506480  0.2250381  0.03493306
```

观察输出结果，体重和腰围有较强的正相关关系，引体向上和仰卧起坐有较强的正相关关系；而体重和腰围与三个运动指标的相关系数为负数，说明体重和腰围较大对运动能力有负面影响。

1.2 典型相关系数的检验

下面，调用cancor()函数，对这两组数据进行典型相关系数的检验。

```
can.x<-cancor(X[,1:3],X[,4:6])
```

可以直接输出典型相关系数。

```
can.x$cor
```

```
## [1] 0.79560815 0.20055604 0.07257029
```

那么，三对典型变量的相关系数是否显著？这里采用序贯检验进行说明。首先检验第一个相关系数λ_1.

$$H_0:\lambda_1=0 \qquad H_1:\lambda_1\neq 0$$

计算统计量：

$$\Lambda_1=(1-\lambda_1^2)(1-\lambda_2^2)\dots(1-\lambda_p^2)=\prod_{i=1}^{p}(1-\lambda_i^2)$$

$$Q_1=-[n-1-\frac{1}{2}(p+q-1)]\ln\Lambda_1$$

这里Q_1近似服从$\chi^2(pq)$. 因此，若$Q_1>\chi_\alpha^2(pq)$，则拒绝原假设H_0，说明第一对典型变量显著相关；否则，所有的典型变量都不显著相关。
在去掉第一典型相关系数后，继续检验余下的$p-1$个典型相关系数的显著性。更一般地，若前$j-1$个典型相关系数在水平α下是显著的，则当检验第j个典型相关系数的显著性时，计算统计量

$$\Lambda_j=(1-\lambda_j^2)(1-\lambda_{j+1}^2)\dots(1-\lambda_p^2)=\prod_{i=j}^{p}(1-\lambda_i^2)$$

$$Q_j = -[n - j - \frac{1}{2}(p + q - 1)] \ln \Lambda_j$$

这里Q_j近似服从$\chi^2[(p - j + 1)(q - j + 1)]$.

编写检验函数如下：

```
#编写函数进行检验
cor.test<-function(X,Y,n,p,q,alpha){
  can<-cancor(X,Y)
  j=min(p,q)
  count=0
  QSTA=c()
  PVAL=c()
  for(i in 1:j){
    Lambda=prod(1-can$cor[i:j]^2)
    Q=-(n-i-1/2*(p+q+1))*log(Lambda)
    pvalue=pchisq(Q,(p-i+1)*(q-i+1),lower.tail = F)
    if(pvalue<alpha){
      count=count+1
    }
    QSTA=c(QSTA,Q)
    PVAL=c(PVAL,pvalue)
  }
  return(list("count"=count,"Q-statistics"=QSTA,"pvalue"=PVAL))
}
```

函数的输入包括：第1组变量，第2组变量，样本容量，第1组变量个数，第2组变量个数，显著性水平。函数的输出包括：显著的典型变量对数，Q统计量值，以及p-value.
下面，在0.05的显著性水平下，对本案列的典型相关系数进行检验。

```
cor.test(X[,1:3],X[,4:6],20,3,3,0.05)
```

```
## $count
## [1] 0
##
## $`Q-statistics`
## [1] 16.2549575  0.6718487  0.0712849
##
## $pvalue
## [1] 0.06174456 0.95475464 0.78947507
```

然而，在该显著性水平下显著的相关系数为0个，无法进行后续分析。因此，放宽显著性水平到0.10，再次调用函数进行检验。

```
cor.test(X[,1:3],X[,4:6],20,3,3,0.10)
```

```
## $count
## [1] 1
##
## $`Q-statistics`
## [1] 16.2549575  0.6718487  0.0712849
##
## $pvalue
## [1] 0.06174456 0.95475464 0.78947507
```

在该显著性水平下，只有第1个相关系数，也就是第一对典型相关变量是显著的。输出的p-value为0.06.

1.3 典型权重、典型载荷和交叉载荷

1.3.1 典型权重

所谓典型权重，就是标准化系数，也就是

$$\begin{cases} b_i = V_{yy}^{-1/2} r_i \\ a_i = \frac{1}{\lambda_i} V_{xx}^{-1} V_{xy} b_i \end{cases}$$

计算结果为：

```
#第1组变量
can.x$xcoef
```

```
##           [,1]        [,2]        [,3]
## X1 -0.17788841 -0.43230348 -0.04381432
## X2  0.36232695  0.27085764  0.11608883
## X3 -0.01356309 -0.05301954  0.24106633
```

```
#第2组变量
can.x$ycoef
```

```
##           [,1]        [,2]        [,3]
## Y1 -0.08018009 -0.08615561 -0.29745900
## Y2 -0.24180670  0.02833066  0.28373986
## Y3  0.16435956  0.24367781 -0.09608099
```

由此，可以计算出各个样本的典型变量值。

$$\begin{cases} U = a' X \\ V = b' Y \end{cases}$$

```
U<-X[,1:3]%*%can.x$xcoef
V<-X[,4:6]%*%can.x$ycoef
colnames(U)<-c("U1","U2","U3")
colnames(V)<-c("V1","V2","V3")
Score<-data.frame(U,V)
Score
```

```
##              U1           U2          U3          V1           V2
## 1  -0.009969788 -0.121501078 -0.20419401 -0.02909460  0.031027608
## 2   0.186887139 -0.046163013  0.13223387  0.23190170  0.084158321
## 3  -0.101193522 -0.141661215 -0.37063341 -0.12979237 -0.112030106
## 4   0.060964112 -0.346616669  0.03342558  0.09063830 -0.150034732
## 5  -0.512831098 -0.458299483  0.44354554 -0.39173848 -0.209788233
## 6  -0.077780541  0.094512914 -0.23766491 -0.11930102 -0.288113119
## 7   0.003955674  0.254201102  0.25701898 -0.22619839  0.122191086
## 8  -0.016855040 -0.127105942 -0.34147617  0.21834490 -0.164740837
## 9   0.203734347  0.196310283 -0.00758741  0.26809619 -0.165185828
## 10 -0.104800666  0.208124774 -0.11711820 -0.38258341 -0.041647344
## 11  0.113834968 -0.016598895 -0.09752299  0.21737727  0.056375494
## 12  0.063237343  0.213427257  0.21221151  0.01130349 -0.218167527
## 13  0.043586465 -0.008040409  0.01237648  0.16412985 -0.065834278
## 14 -0.082181602  0.055998387  0.10021686  0.03462910 -0.097067107
## 15 -0.094153311  0.228436101 -0.04670258  0.05393456  0.778658842
## 16 -0.173085857  0.047742282 -0.20173015 -0.15965387  0.183746435
## 17  0.718139369 -0.256090676  0.05898572  0.43237930 -0.002016448
## 18  0.001362964 -0.317746855  0.21374067 -0.12845980  0.223805116
## 19 -0.221400693  0.120731486 -0.22201469 -0.31879992  0.059073577
## 20 -0.001450263  0.420339649  0.38288931  0.16288721 -0.024410919
##              V3
## 1   0.344302062
## 2  -0.403047146
## 3  -0.133855684
## 4  -0.059921010
## 5  -0.008592976
## 6  -0.480186091
## 7  -0.006091381
## 8  -0.074849953
## 9   0.003582504
## 10  0.042948559
## 11  0.277291769
## 12 -0.264987331
## 13  0.157664098
## 14  0.157711712
## 15 -0.283334468
## 16  0.008766141
## 17  0.080198839
## 18  0.055667431
## 19  0.277587462
## 20  0.309145462
```

观察典型权重的符号和大小，是传统的解释解释典型函数的方法。典型权重较大，说明原始变量对它的典型变量贡献较大；典型权重符号为负，说明变量之间存在一种反向关系。反之亦然。然而，这种解释遭到了很多批评；因此，在解释典型相关的时候应慎用典型权重。

1.3.2 典型载荷

典型载荷，也成为典型结构相关系数，是原始变量（自变量或者因变量）与它的典型变量间的简单线性相关系数。即：

$$\begin{cases} cor(U, X) \\ cor(V, Y) \end{cases}$$

计算结果如下：

```
#第1组变量及其典型变量的相关系数
cor(X[,1:3],U)
```

```
##            U1         U2          U3
## X1  0.6206424 -0.7723919 -0.13495886
## X2  0.9254249 -0.3776614 -0.03099486
## X3 -0.3328481  0.0414842  0.94206752
```

```
#第2组变量及其典型变量的相关系数
cor(X[,4:6],V)
```

```
##            V1        V2          V3
## Y1 -0.7276254 0.2369522 -0.64375064
## Y2 -0.8177285 0.5730231  0.05444915
## Y3 -0.1621905 0.9586280 -0.23393722
```

典型载荷的解释类似于因子载荷，反映原始变量与典型变量的共同方差。根据结果，生理测量的第一个典型变量U_1与腰围X_2的相关系数最大，说明这个典型变量主要反映人的体型肥胖程度。运动因素的第一个典型变量V_1与仰卧起坐Y_2和引体向上Y_1有较强的负相关关系，说明这个典型变量主要反映人不适合运动的程度。

1.3.3 典型交叉载荷

所谓典型交叉载荷，就是原始变量与另一组变量的典型变量的相关系数。即：

$$\begin{cases} cor(X,V) \\ cor(Y,U) \end{cases}$$

计算结果如下：

```
#交叉载荷
#第1组变量与第2组典型变量的相关系数
cor(X[,1:3],V)
```

```
##            V1           V2           V3
## X1  0.4937881 -0.154907853 -0.009794003
## X2  0.7362756 -0.075742277 -0.002249306
## X3 -0.2648166  0.008319907  0.068366110
```

```
#第2组变量与第1组典型变量的相关系数
cor(X[,4:6],U)
```

```
##            U1         U2          U3
## Y1 -0.5789047 0.0475222 -0.04671717
## Y2 -0.6505914 0.1149232  0.00395139
## Y3 -0.1290401 0.1922586 -0.01697689
```

根据结果看出，腰围X_2与运动的第一典型变量V_1的相关性较强，说明了腰围大（体型较胖）则运动能力差；仰卧起坐Y_1和引体向上Y_2与生理测量的第一典型变量U_1呈一定的负相关关系，说明人的体型肥胖程度对这两

种运动能力有负面影响。

1.4 冗余分析

事实上，典型相关系数的平方可以提供典型变量间共同方差的一个估计，但它表示的是由因变量组和自变量组的线性组合所共享的方差，而不是来自两组变量的方差。这样，即使两个典型变量可能并没有从它们各自的变量组中提取显著方差，但这两个典型变量间仍可能得到一个相对较强的典型相关系数。为了克服这种有偏性和不稳定性，提出了冗余指数，它计算的是一组变量的方差能被另一组变量的方差解释的比例。

计算冗余指数分三步（首先利用第二组变量进行计算）：

①共同方差的比例。

通过对1.3.2节中得到的“典型载荷”进行平方，可以得到每个原始变量通过本组典型变量解释的方差比例。然后，再计算每个典型变量载荷平方的简单平均，得到所有原始变量通过本组的每个典型变量，平均能够解释的共同方差的比例。

```
#(1)共同方差的比例
L1=cor(X[,4:6],V)
L1^2   #典型载荷的平方
```

```
##            V1         V2          V3
## Y1 0.52943876 0.05614635 0.414414891
## Y2 0.66867982 0.32835547 0.002964709
## Y3 0.02630576 0.91896762 0.054726623
```

输出的结果为典型载荷的平方。可以验证，该矩阵按行求和，每行为1.

```
apply(L1^2,1,sum)   #验证，按行求和为1
```

```
## Y1 Y2 Y3
##  1  1  1
```

这说明，类似于主成分分析，典型载荷的平方，就是一种比例（某个原始变量通过该组每个典型变量所解释的方差比例）。

然后，按列求均值，得到该组所有原始变量，通过每个典型变量，平均能够解释的共同方差的比例。

```
R2=colMeans(L1^2)   #计算典型变量所解释的共同方差的比例
R2
```

```
##        V1        V2        V3
## 0.4081414 0.4344898 0.1573687
```

此外，将各个比例累计求和，可以计算出累计比例。

```
cumsum(R2)
```

```
##        V1        V2        V3
## 0.4081414 0.8426313 1.0000000
```

可见，第一个典型变量仅仅解释了共同方差的40.8%，而前两个典型变量一共解释了共同方差的84.3%.

②解释的方差比例。
解释的方差比例，就是典型相关系数的平方。

```
#(2)解释的方差比例
P=can.x$cor^2
```

它反映了，通过每个自变量典型变量能够解释的每个因变量典型变量的方差比例。

③冗余指数。
冗余指数 = 共同方差的比例 × 典型相关系数^2
由此，求得冗余指数为

```
R2*P
```

```
##          V1          V2          V3
## 0.258350408 0.017476364 0.000828774
```

将各个比例累计求和，可以计算出累计比例。

```
cumsum(R2*P)
```

```
##        V1        V2        V3
## 0.2583504 0.2758268 0.2766555
```

然后，将各个指标放入一个数据框进行汇总。

```
data.frame("R2"=R2,"Cumulative R2"=cumsum(R2),
           "Cor2"=P,"Redundancy"=R2*P,"Cumulative Redundancy"=cumsum(R2*P))
```

```
##           R2 Cumulative.R2        Cor2  Redundancy Cumulative.Redundancy
## V1 0.4081414     0.4081414 0.632992335 0.258350408             0.2583504
## V2 0.4344898     0.8426313 0.040222726 0.017476364             0.2758268
## V3 0.1573687     1.0000000 0.005266446 0.000828774             0.2766555
```

因此，运动指标的各个变量，通过其组内第一个典型变量，可以解释生理指标的方差比例是25.8；而通过第二、第三个典型变量可以解释的方差比例则很低。

最后，重复上述过程，输出第一组变量的冗余分析结果。

```
R2=colMeans(cor(X[,1:3],U)^2)
R2*P
```

```
##          U1          U2          U3
## 0.285352091 0.009934185 0.001591636
```

```
data.frame("R2"=R2,"Cumulative R2"=cumsum(R2),
           "Cor2"=P,"Redundancy"=R2*P,"Cumulative Redundancy"=cumsum(R2*P))
```

```
##           R2 Cumulative.R2        Cor2  Redundancy Cumulative.Redundancy
## U1 0.4507987     0.4507987 0.632992335 0.285352091             0.2853521
## U2 0.2469794     0.6977781 0.040222726 0.009934185             0.2952863
## U3 0.3022219     1.0000000 0.005266446 0.001591636             0.2968779
```

可见，生理指标的各个变量，通过其组内的第一个典型变量，可以解释运动指标的方差比例是28.5%；而通过第二、第三个典型变量可以解释的方差比例则很低。

2 例8—2的分析

2.1 函数的封装

实际上，利用R语言自带的函数cancor()，能够输出的结果少之又少。如果想要方便地得到典型相关分析结果，就需要对上述指令进行封装，从而减少重复，提高效率。

```
#函数的封装
cca<-function(X,p,q,alpha){
  X=scale(X,center = T,scale = T)
  #变量间的相关性
  X1=X[,1:p]
  X2=X[,(p+1):(p+q)]
  cor1=cor(X1)
  cor2=cor(X2)
  cor12=cor(X1,X2)

  #典型相关系数及其检验
  can.x<-cancor(X1,X2)    #进行典型相关分析
  test<-cor.test(X1,X2,nrow(X),p,q,alpha)   #进行检验

  U<-as.matrix(X1)%*%as.matrix(can.x$xcoef)
  V<-as.matrix(X2)%*%as.matrix(can.x$ycoef)
  colnames(U)<-paste("U",c(1:p),sep = "",collapse = NULL)
  colnames(V)<-paste("V",c(1:q),sep = "",collapse = NULL)
  U=U[,1:min(p,q)]
  V=V[,1:min(p,q)]
  #典型载荷
  L11=cor(X1,U)
  L22=cor(X2,V)
  #交叉载荷
  L12=cor(X1,V)
  L21=cor(X2,U)

  #冗余分析
  #第一组变量
  R2_1=colMeans(L11^2)
  P=can.x$cor^2
  redun1<-data.frame("Proportion"=R2_1,"Cumulative Proportion"=cumsum(R2_1),
             "R-Squared"=P,"Canonical Proportion"=R2_1*P,"Cumulative Proportion"=cumsum(R2_1*P))
  #第二组变量
  R2_2=colMeans(L22^2)
  redun2<-data.frame("R2"=R2_2,"Cumulative R2"=cumsum(R2_2),
                     "Cor2"=P,"Redundancy"=R2_2*P,"Redundancy"=cumsum(R2_2*P))

  return(list("cor1"=cor1,"cor2"=cor2,"cor12"=cor12,"CanonicalCor"=can.x$cor,"cor.test"=test,
              "xcoef"=can.x$xcoef,"ycoef"=can.x$ycoef,
              "Loadings_xU"=L11,"Loadings_yV"=L22,"LoadingsxV"=L12,"LoadingsyU"=L21,
              "RedunAna_x"=redun1,"RedunAna_y"=redun2,"scores_U"=U,"scores_V"=V))
}
```

函数的输入为：待分析的数据集，第一组变量个数，第二组变量个数，以及检验的显著性水平。
函数的输出为：第一组变量的相关系数，第二组变量的相关系数，第一和第二组变量的相关系数，典型相关系数，典型相关系数的检验，第一组典型权重，第二组典型权重，第一组典型载荷，第二组典型载荷，第一组典型交叉载荷，第二组典型交叉载荷，第一组冗余分析结果，第二组冗余分析结果，第一组典型变量值，以及第二组典型变量值。

2.2 利用封装函数分析例8-2

首先读取数据，并对数据集进行微调。

```
Y=read.csv("eg8-2.csv")
rownames(Y)=Y[,1]
Y=Y[,-1]
head(Y)
```

```
##      劳动生产率 市场占有率 居民人均收入 经济增长率 对外设施指数
## 上海   45623.05       2.50         8439      16.27         1.03
## 深圳   52256.67       1.30        18579      21.50         1.34
## 广州   46551.87       1.13        10445      11.92         1.07
## 北京   28146.76       1.38         7813      15.00        -0.43
## 厦门   38670.43       0.12         8980      26.71        -0.53
## 天津   26316.96       1.37         6609      11.07        -0.11
##      对内设施指数 百人电话数 技术设施指数 文化设施指数 卫生设施指数
## 上海         0.42         50         2.15         1.23         1.64
## 深圳         0.13        131         0.33        -0.27        -0.64
## 广州         0.40         48         1.31         0.49         0.09
## 北京         0.19         20         0.87         3.57         1.80
## 厦门         0.25         32        -0.09        -0.33        -0.84
## 天津         0.07         27         0.68        -0.12         0.87
```

直接调用函数进行分析。

```
cca.Y=cca(Y,4,6,0.05)
cca.Y[c(1:13)]
```

```
## $cor1
##                 劳动生产率   市场占有率 居民人均收入   经济增长率
## 劳动生产率     1.0000000  0.25093499    0.7106959  0.20989250
## 市场占有率     0.2509350  1.00000000    0.2252807 -0.07546211
## 居民人均收入   0.7106959  0.22528068    1.0000000  0.38802437
## 经济增长率     0.2098925 -0.07546211    0.3880244  1.00000000
##
## $cor2
##                 对外设施指数 对内设施指数 百人电话数 技术设施指数
## 对外设施指数    1.00000000    0.1599611  0.5556427    0.6502246
## 对内设施指数    0.15996109    1.0000000  0.3898509    0.2398817
## 百人电话数      0.55564273    0.3898509  1.0000000    0.1622243
## 技术设施指数    0.65022457    0.2398817  0.1622243    1.0000000
## 文化设施指数    0.06026078    0.1715953 -0.1032157    0.5563953
## 卫生设施指数    0.04992627   -0.1531462 -0.2385656    0.4261477
##                 文化设施指数 卫生设施指数
## 对外设施指数    0.06026078   0.04992627
## 对内设施指数    0.17159529  -0.15314624
## 百人电话数     -0.10321568  -0.23856564
## 技术设施指数    0.55639529   0.42614768
## 文化设施指数    1.00000000   0.63681818
## 卫生设施指数    0.63681818   1.00000000
##
## $cor12
##                 对外设施指数 对内设施指数 百人电话数 技术设施指数
## 劳动生产率      0.33743410    0.7531945  0.5938129   0.12353209
## 市场占有率      0.63496881    0.3353388  0.3160558   0.90657406
## 居民人均收入    0.43189976    0.6132107  0.8642810   0.06624152
## 经济增长率     -0.05954136    0.1407383  0.3989202  -0.08710869
##                 文化设施指数 卫生设施指数
## 劳动生产率     -0.05608530   -0.2285757
## 市场占有率      0.57224039    0.5265033
## 居民人均收入   -0.02154236   -0.3489535
## 经济增长率     -0.01027332   -0.2955984
##
## $CanonicalCor
## [1] 0.9601027 0.9499374 0.6469903 0.3571362
##
## $cor.test
## $cor.test$count
## [1] 2
##
## $cor.test$`Q-statistics`
## [1] 74.977516 37.568007  7.805693  1.432680
##
## $cor.test$pvalue
## [1] 3.760855e-07 1.044780e-03 4.526773e-01 6.978926e-01
##
##
## $xcoef
##                       [,1]         [,2]        [,3]        [,4]
## 劳动生产率    -0.032005929 -0.030335607  0.26021162  0.19819538
## 市场占有率    -0.164825128  0.168873541 -0.04721476  0.01440014
## 居民人均收入 -0.097957189 -0.177100236 -0.14362860 -0.24749073
## 经济增长率    -0.006547108 -0.001346956 -0.13773433  0.21321648
##
```

```
## $ycoef
##                      [,1]        [,2]       [,3]        [,4]        [,5]
## 对外设施指数 -0.03521614 -0.04895944  0.1598140 -0.37421779  0.04515704
## 对内设施指数 -0.07853279 -0.06049260  0.2426494  0.05197925  0.07628689
## 百人电话数   -0.11270945 -0.09067860 -0.2097401  0.09587088 -0.04508868
## 技术设施指数 -0.07735448  0.19936455 -0.1129037  0.30766704 -0.21681184
## 文化设施指数 -0.02636916 -0.05573211 -0.1534132 -0.11658024  0.25384427
## 卫生设施指数 -0.03254930  0.08845328  0.1349969 -0.04763363  0.02144587
##                      [,6]
## 对外设施指数  0.126595723
## 对内设施指数  0.007494276
## 百人电话数   -0.178352909
## 技术设施指数  0.028211472
## 文化设施指数  0.163457971
## 卫生设施指数 -0.276619396
##
## $Loadings_xU
##                        U1         U2          U3          U4
## 劳动生产率   -0.6292432 -0.4973785  0.51163936  0.30804671
## 市场占有率   -0.8475022  0.5294565 -0.01691942 -0.03360866
## 居民人均收入 -0.6990629 -0.7023857 -0.09928747 -0.09004096
## 经济增长率   -0.1692850 -0.3887132 -0.58969946  0.68738574
##
## $Loadings_yV
##                        V1          V2          V3           V4
## 对外设施指数 -0.7144950  0.09445123  0.02689560 -0.531715458
## 对内设施指数 -0.6372790 -0.34418336  0.48978011  0.394865321
## 百人电话数   -0.7190224 -0.54256684 -0.26602448 -0.080590102
## 技术设施指数 -0.7232205  0.63201266 -0.05507999  0.031380756
## 文化设施指数 -0.4101787  0.46880356 -0.24997109  0.003241507
## 卫生设施指数 -0.1967986  0.72520503  0.04377133 -0.175566310
##
## $LoadingsxV
##                        V1         V2          V3          V4
## 劳动生产率   -0.6041381 -0.4724785  0.33102572  0.11001464
## 市场占有率   -0.8136891  0.5029506 -0.01094670 -0.01200287
## 居民人均收入 -0.6711722 -0.6672225 -0.06423803 -0.03215689
## 经济增长率   -0.1625310 -0.3692533 -0.38152985  0.24549035
##
## $LoadingsyU
##                        U1          U2          U3           U4
## 对外设施指数 -0.6859886  0.08972276  0.01740119 -0.189894855
## 对内设施指数 -0.6118533 -0.32695266  0.31688300  0.141020713
## 百人电话数   -0.6903353 -0.51540455 -0.17211527 -0.028781645
## 技术设施指数 -0.6943659  0.60037249 -0.03563622  0.011207205
## 文化设施指数 -0.3938136  0.44533405 -0.16172888  0.001157659
## 卫生设施指数 -0.1889469  0.68889941  0.02831963 -0.062701091
##
## $RedunAna_x
##    Proportion Cumulative.Proportion R.Squared Canonical.Proportion
## U1  0.4078883             0.4078883 0.9217972           0.37599031
## U2  0.2930383             0.7009266 0.9023811           0.26443226
## U3  0.1549161             0.8558428 0.4185965           0.06484735
## U4  0.1441572             1.0000000 0.1275463           0.01838672
##    Cumulative.Proportion.1
## U1               0.3759903
```

```
## U2               0.6404226
## U3               0.7052699
## U4               0.7236566
##
## $RedunAna_y
##            R2 Cumulative.R2      Cor2 Redundancy Redundancy.1
## V1 0.36060747     0.3606075 0.9217972 0.33240695    0.3324070
## V2 0.26115018     0.6217577 0.9023811 0.23565700    0.5680640
## V3 0.06313537     0.6848930 0.4185965 0.02642825    0.5944922
## V4 0.07949225     0.7643853 0.1275463 0.01013894    0.6046311
```

根据输出结果，共有2个典型相关系数显著。前两对典型相关变量为

$$\begin{cases} U_1 = -0.032X_1 - 0.165X_2 - 0.098X_3 - 0.007X_4 \\ V_1 = -0.035Y_1 - 0.079Y_2 - 0.113Y_3 - 0.077Y_4 - 0.026Y_5 - 0.033Y_6 \end{cases}$$

$$\begin{cases} U_2 = -0.030X_1 + 0.169X_2 - 1.177X_3 - 0.001X_4 \\ V_2 = -0.049Y_1 - 0.060Y_2 - 0.091Y_3 + 0.199Y_4 - 0.056Y_5 + 0.088Y_6 \end{cases}$$

根据典型载荷，X_1至X_3与竞争力组第一典型变量U_1高度相关，说明劳动生产率、市场占有率、居民人均收入在反映城市竞争力水平方面占主导地位。同时，X_1至X_3与城市基础设施变量的第一典型变量V_1高度相关。此外，Y_1至Y_4与基础设施组的第一典型变量V_1高度相关，说明对外基础设施、对内基础设施、百人电话数、技术设施指数在反映城市基础设施水平方面占主导地位。同时，Y_1至Y_4与城市竞争力变量的第一典型变量U_1也高度相关。

最后，进行冗余分析。通过输出的结果，城市竞争力变量组，通过它的前2个典型相关变量，解释了基础设置方差比例的64.04%；而基础设施变量组，通过它的前2个典型相关变量，解释了城市竞争力方差的56.81%.

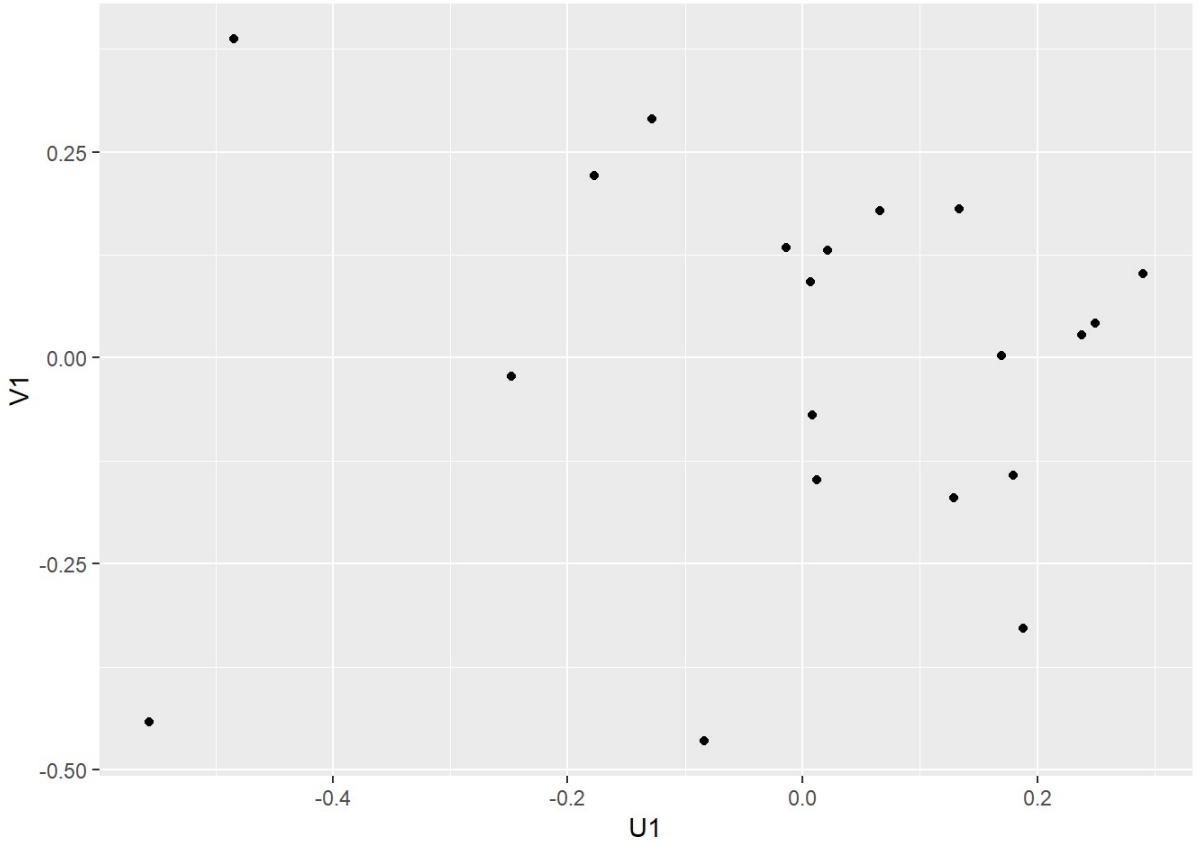
2.3 可视化：散点图

分别以第一对典型相关变量值为横轴、纵轴，做出各个样本的散点图。

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
Yplot=data.frame(cca.Y$scores_U[,1],cca.Y$scores_U[,2])
ggplot(Yplot,aes(x=Yplot[,1], y = Yplot[,2]))+
  geom_point()+
  labs(x = "U1", y = "V1")
```



由于U_1和V_1分别反映了城市竞争力和城市基础设施。因此，从散点的分布位置来看，城市竞争力较强的城市，基础设施也往往更好一些。反之亦然。

3 例8—3的分析

3.1 调用封装函数分析例8-3

首先，导入例8-3的数据，并对行名等进行处理。

```
Z=read.csv("eg8-3.csv")
rownames(Z)=Z[,1]
Z=Z[,-1]
head(Z)
```

```
##          x1      x2     x3    x4    x5    x6    y1    y2      y3     y4
## 1978 100.30 30476.5 101.30 133.6 79014 15.34 16110 30129  975.30 246.28
## 1979 124.77 33211.5 105.05 160.2 79047 19.32 18768 31971 1001.35 242.21
## 1980 127.77 32055.5 108.72 191.3 79565 22.31 19861 30543 1134.10 246.28
## 1981 128.02 32502.0 112.97 223.4 79901 22.15 19495 29370 1188.42 257.42
## 1982 128.15 35450.0 113.19 270.1 80174 23.51 20063 30078 1271.80 264.96
## 1983 127.77 38727.5 113.08 309.8 80734 24.69 20661 29854 1316.10 262.06
```

与上节完全类似，这里调用封装的函数，对例8—3进行典型相关分析。

```
cca.Z=cca(Z,6,4,0.05)
cca.Z[c(1:13)]
```

```
## $cor1
##             x1        x2         x3         x4          x5         x6
## x1  1.00000000 0.8805320  0.9632730  0.8756268 -0.09821871  0.5379226
## x2  0.88053199 1.0000000  0.8612600  0.7915370  0.03004910  0.6921372
## x3  0.96327299 0.8612600  1.0000000  0.9390244 -0.26394328  0.5256345
## x4  0.87562676 0.7915370  0.9390244  1.0000000 -0.52620024  0.6035812
## x5 -0.09821871 0.0300491 -0.2639433 -0.5262002  1.00000000 -0.1326758
## x6  0.53792257 0.6921372  0.5256345  0.6035812 -0.13267584  1.0000000
##
## $cor2
##            y1         y2         y3         y4
## y1  1.0000000  0.9515612  0.9982374 -0.7664580
## y2  0.9515612  1.0000000  0.9579827 -0.6202074
## y3  0.9982374  0.9579827  1.0000000 -0.7368713
## y4 -0.7664580 -0.6202074 -0.7368713  1.0000000
##
## $cor12
##            y1         y2         y3         y4
## x1  0.9064604  0.9056953  0.9156059 -0.6422717
## x2  0.8588213  0.8642073  0.8779913 -0.4928476
## x3  0.9477899  0.8971807  0.9466380 -0.7744657
## x4  0.9511073  0.8544873  0.9408628 -0.8579433
## x5 -0.3191810 -0.1169779 -0.2748066  0.6768307
## x6  0.6887394  0.6598648  0.7010530 -0.2556218
##
## $CanonicalCor
## [1] 0.9879045 0.9164870 0.7621484 0.2440953
##
## $cor.test
## $cor.test$count
## [1] 3
##
## $cor.test$`Q-statistics`
## [1] 152.540910  62.173461  20.016626   1.259343
##
## $cor.test$pvalue
## [1] 1.115537e-20 1.062788e-07 1.027333e-02 7.388093e-01
##
##
## $xcoef
##           [,1]        [,2]        [,3]         [,4]         [,5]
## x1  0.10435092  0.16656151 -0.08865784  0.279860002 -0.707278618
## x2  0.05352343  0.03765392 -0.05361523 -0.510174598 -0.479539222
## x3 -0.06917091 -0.27966924 -0.52838973  0.153924411 -0.346548993
## x4 -0.29011307  0.11552311  0.99062879  0.001642154  1.695803558
## x5 -0.10138241  0.16770647  0.36392511  0.114867367  0.734276487
## x6 -0.02067250  0.04357078 -0.29670795  0.111490365  0.004196204
##          [,6]
## x1 -0.7170273
## x2 -0.0925953
## x3  0.9450994
## x4 -0.1287096
## x5  0.1774485
## x6  0.0185541
##
## $ycoef
```

```
##           [,1]        [,2]       [,3]       [,4]
## y1  0.09294407 -1.83547064 -3.7219371  2.6626695
## y2  0.02576189  0.15013763  0.4333519  0.5622526
## y3 -0.28282160  1.77553740  2.9679604 -3.1201990
## y4  0.02633280  0.09756022 -0.4430636  0.0916654
##
## $Loadings_xU
##            U1           U2          U3          U4
## x1 -0.9173294  0.207006176  0.14658208  0.15635904
## x2 -0.8678531  0.377424215  0.01577905 -0.26528321
## x3 -0.9644030  0.001729207  0.09823101  0.07506778
## x4 -0.9717966 -0.179819481  0.02048245 -0.03089989
## x5  0.3426807  0.860152024  0.15392774  0.08490304
## x6 -0.6758848  0.321509614 -0.59025237 -0.13149716
##
## $Loadings_yV
##            V1         V2          V3          V4
## y1 -0.9965140 0.02706898 -0.03908272 0.068553865
## y2 -0.9319897 0.23695543  0.05248881 0.269243962
## y3 -0.9949942 0.08207986 -0.03119119 0.047712399
## y4  0.7944221 0.55419993 -0.24840675 0.007070566
##
## $LoadingsxV
##            V1           V2          V3           V4
## x1 -0.9062338  0.189718467  0.11171730  0.038166507
## x2 -0.8573560  0.345904383  0.01202598 -0.064754384
## x3 -0.9527380  0.001584796  0.07486660  0.018323692
## x4 -0.9600422 -0.164802215  0.01561066 -0.007542518
## x5  0.3385358  0.788318140  0.11731578  0.020724432
## x6 -0.6677096  0.294659379 -0.44985989 -0.032097839
##
## $LoadingsyU
##            U1         U2          U3          U4
## y1 -0.9844606 0.02480837 -0.02978683 0.016733676
## y2 -0.9207168 0.21716657  0.04000426 0.065721185
## y3 -0.9829592 0.07522512 -0.02377231 0.011646372
## y4  0.7848132 0.50791702 -0.18932280 0.001725892
##
## $RedunAna_x
##    Proportion Cumulative.Proportion  R.Squared Canonical.Proportion
## U1 0.67389572             0.6738957 0.97595523          0.657692053
## U2 0.17681143             0.8507072 0.83994841          0.148512477
## U3 0.06731596             0.9180231 0.58087015          0.039101832
## U4 0.02098556             0.9390087 0.05958252          0.001250372
##    Cumulative.Proportion.1
## U1               0.6576921
## U2               0.8062045
## U3               0.8453064
## U4               0.8465567
##
## $RedunAna_y
##            R2 Cumulative.R2       Cor2  Redundancy Redundancy.1
## V1 0.87069125     0.8706912 0.97595523 0.849755673    0.8497557
## V2 0.09268882     0.9633801 0.83994841 0.077853824    0.9276095
## V3 0.01674033     0.9801204 0.58087015 0.009723960    0.9373335
## V4 0.01987960     1.0000000 0.05958252 0.001184477    0.9385179
```

根据输出的结果，典型相关系数显著的典型变量，共有3对。因此，两组变量的相关性，可以转化为前3对典型相关变量的相关性。线性组合为：

$$\begin{cases} U_1 = 0.104X_1 + 0.054X_2 - 0.069X_3 - 0.290X_4 - 0.101X_5 - 0.021X_6 \\ V_1 = 0.093Y_1 + 0.026Y_2 - 0.283Y_3 + 0.026Y_4 \end{cases}$$

$$\begin{cases} U_2 = 0.167X_1 + 0.038X_2 - 0.280X_3 + 0.116X_4 + 0.168X_5 + 0.044X_6 \\ V_2 = -1.835Y_1 + 0.150Y_2 + 1.776Y_3 + 0.098Y_4 \end{cases}$$

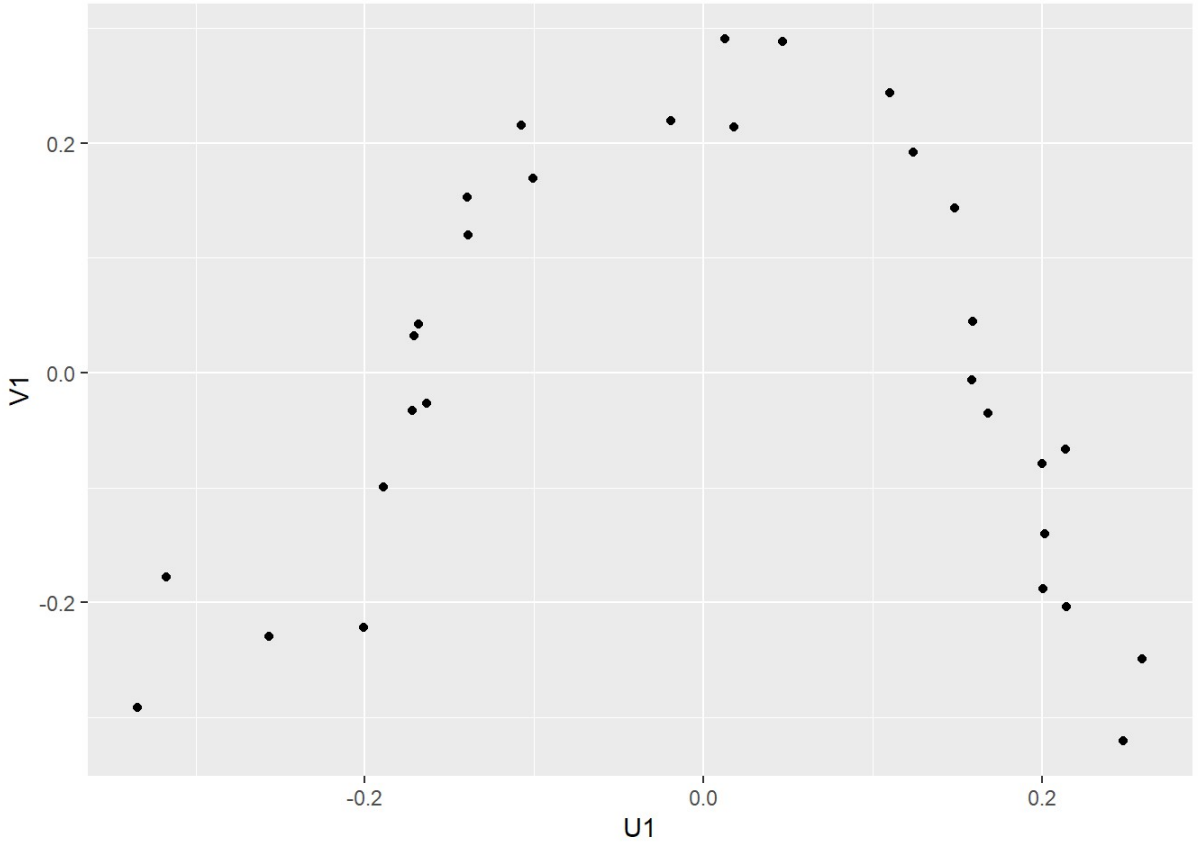
$$\begin{cases} U_3 = -0.089X_1 - 0.054X_2 - 0.528X_3 + 0.991X_4 + 0.364X_5 - 0.297X_6 \\ V_3 = -3.722Y_1 + 0.433Y_2 - 2.968Y_3 - 0.443Y_4 \end{cases}$$

根据典型载荷看出，U_1与X_1至X_4关系密切，说明猪（毛重）生产价格指数、粮食产量、粮食零售价格指数、农村居民人均纯收入都对生猪的生产起到了重要作用。而V_1与Y_1至Y_3关系密切，说明生猪生产量可主要由肉猪出栏头数、生猪年底存栏头数、猪肉产量进行解释。
此外，X_1至X_4与V_1的联系较为密切，Y_1至Y_3与U_1的联系也较为密切。基本可以说明，猪（毛重）价格指数、粮食产量等，直接影响了肉猪出栏头数、生猪年底存栏头数等指标，这也是与实际情况相符的。
最后，进行冗余度分析。生猪生产“影响指标”，可通过它的前2个典型变量，解释生猪生产“代表指标”方差的80.62%；而生猪生产“代表指标”，则通过它的第1个典型变量，就能够解释生猪生产“影响指标”的84.98%.

3.2 可视化：散点图

分别以第一对典型相关变量值为横轴、纵轴，做出各个样本的散点图。

```
Zplot=data.frame(cca.Z$scores_U[,1],cca.Z$scores_U[,2])
ggplot(Zplot,aes(x=Zplot[,1], y = Zplot[,2]))+
  geom_point()+
  labs(x = "U1", y = "V1")
```



由于U_1和V_1分别反映了生猪生产“影响指标”和生猪生产“代表指标”。因此，从散点的分布位置来看，随着生猪生产“各种投入”（影响因素）的增加，“产量指标”往往会出现先上升再下降的趋势。