《系统工程导论》主成分分析作业

1. 编程实现 PCA 算法

按照 PCA 的规则算法以及重建数据的方法，即可编写相关算。**具体实验过程和结果在第二题当中阐述。**

在特征值选取时，以此删去较小的特征值，直到满足误差要求，进而得到所需主成分的量。

2. 利用上面编写的函数，以及线性回归章节作业中编写的函数，对附件的数据进行建模。附件的数据为美国 1992 年总统竞选各个 county 的投票情况。

2.1 算法设计

【1】用之前的线性回归方法可检验前 14 维的数据是否线性相关。

【2】检验得到之前的数据线性相关（相关即病态），则可以用 PCA 对数据进行压缩，这样便会消除数据的线性相关。

【3】使用压缩后的数据，将 turnout 作为 label，进行回归。由于压缩后的数据是降维数据，即消除了线性相关，没有了病态。

【4】利用 PCA 的映射关系，即可反向映射得到回归模型。

2.2 实验

【1】直接对原始数据进行线性回归：

```
[1] 经检验，该线性回归问题为病态线性回归问题
[2] 线性回归方程为： Z = (-0.000379)*Y1 + (-2e-06)*Y2 + (-0.001011)*Y3 + 0.607045*Y4 + 0.679938*Y5 +
(-0.000415)*Y6 + 0.330514*Y7 + 0.000252*Y8 + 0.163893*Y9 + 0.000438*Y10 + (-0.095966)*Y11 + 0.153984*Y12
+ 0.055389*Y13 + (-0.030923)*Y14 + 19.563163
[3] F检验的值： 208.44718328855717      F检验的临界值： 1.6949614976166913
[4] 在显著水平取： 0.05 时，经检验，线性相关
[5] 置信区间（误差范围）：[ -10.723661591868424 , 10.723661591868424 ]
```

发现原问题数据线性相关，回归问题变为病态问题，需要进行压缩。

【2】对原始数据进行压缩

取相对误差界为 0.10，比较特征值后，只需取前 9 个特征值即可，即取前 9 个主要成分，舍弃后面的 5 个特征值。

```
特征值： [12010.55935341  9431.41941791  4973.5835387   3667.36182353
  2797.48799589  2304.97386666  1978.51187178  1902.94538737
  1747.02623643  1206.01223884   845.70430037   366.86767029
   349.96907789    13.57722092]
```

得各个主成分：14*9 维，即使用的前九个主成分。

```
各个主成分（每一列一个主成分）：
[[-0.16517342 -0.02873452 -0.39146795 -0.44474069  0.34457098 -0.31285937
   0.01562619 -0.43058939 -0.46550411]
 [-0.22891924  0.06638947 -0.38287661 -0.3623429   0.03698412 -0.1233878
  -0.4985621   0.39256142  0.47169695]
 [-0.2558878   0.23032151  0.1375672   0.19646095 -0.28832833 -0.46375934
  -0.12161679 -0.51482051  0.33955965]
 [ 0.341662   -0.12593406 -0.27353384 -0.14406795 -0.51331412 -0.23598956
   0.18590729 -0.08099073  0.10624612]
 [ 0.39061284 -0.12354869 -0.28110774 -0.20785886 -0.31703052  0.02594911
   0.20842256 -0.01567969  0.07856136]
 [-0.32470991  0.08382458 -0.14386755 -0.01281911 -0.51349293  0.17914556
  -0.18195877  0.25318686 -0.4869421 ]
 [-0.2704543   0.29810229 -0.22049992 -0.08414874 -0.03232797  0.36030546
   0.48477781 -0.05585894  0.10192293]
 [-0.25496275  0.37772636 -0.14762589 -0.04472661  0.04558878  0.13040247
   0.35566006 -0.02469943  0.25770246]
 [ 0.36741116  0.01981059 -0.03173534 -0.21217342  0.25598619  0.38672937
  -0.16653082 -0.23118366  0.23966525]
 [-0.17105344 -0.4356875  -0.28290495  0.30520488  0.13864519 -0.08019064
   0.17679358  0.0827443   0.12391585]
 [ 0.05885274  0.24235619  0.50991029 -0.53985149 -0.07528481 -0.15384389
   0.11276357  0.15967633 -0.04788401]
 [ 0.19805113  0.37380052 -0.20375726  0.19550339 -0.13053486  0.30966491
  -0.40471057 -0.31830304 -0.12532209]
 [ 0.27736633  0.36745331 -0.15063732  0.18838411  0.14534822 -0.29273252
   0.16630188  0.27746623 -0.07639804]
 [-0.24096799 -0.38512987  0.17612587 -0.23620412 -0.19335706  0.27440222
   0.02521636 -0.2381503   0.13230964]]
```

对数据进行压缩，得到压缩数据：压缩后的数据大小：(3114, 9)：

```
压缩后的数据（每一行对应一个数据）：
[[-1.84082979  0.56231471  2.19053966 ...  0.26436835  0.54409464
  -0.47094261]
 [-0.67743835  1.26159402  1.53805662 ...  0.41504255 -0.27380622
   0.34471203]
 [-1.46645934 -2.79270069  1.35221828 ...  0.27945332 -0.31264393
   0.24206901]
 ...
 [-1.98621852  2.95327645  1.11730156 ... -1.12584929 -0.94179065
  -0.13442088]
 [ 0.47466912  1.41599876  0.17736583 ...  0.07521965  0.3093336
  -0.72130675]
 [ 0.82779611  1.79879119  0.87784009 ... -0.3003406   0.23417773
  -0.55150028]]
```

同时获得原始数据的归一化参数：

```
归一化参数：
[均值，方差]
[[224.75369299935775, 2057130.8234047948],
 [79691.60179833013, 70131512457.83244],
 [6.255459209749752, 412.90761999274],
 [8.326204238025209, 4.681355709963249],
 [6.61859345009255, 5.567595830038497],
 [3005.7029543994863, 5352164.166420302],
 [13.488310858616863, 43.19651088240377],
 [28365.796403339755, 49463875.24885027],
 [6.490269759678274, 53.96134758325836],
 [39.72594732095894, 116.0760773923382],
 [39.78782915594337, 73.7143443993207],
 [19.813680173650948, 47.34351239435449],
 [87.36913957332447, 239.55523559356843],
 [8.648582734371844, 206.7565282478633]]
```

【3】对降维得到的 9 维数据进行线性回归：

```
对压缩数据进行线性回归：

[1] 经检验，该线性回归问题没有病态
[2] 线性回归方程为：
                Z = 0.269472*Y1 + 0.18102*Y2 + (-0.328584)*Y3 + 0.140607*Y4 + (-0.200781)*Y5 +
                    0.122156*Y6 + 0.226103*Y7 + (-0.164467)*Y8 + 0.310139*Y9 + 0.0
[3] F检验的值：  424.8304776251011      F检验的临界值：  1.8828917680653123
[4] 在显著水平取：  0.05 时，经检验，线性相关
[5] 置信区间（误差范围）：[ -1.5063684207147081 , 1.5063684207147081 ]
```

可见，此时得到的回归模型已经没有了病态，各个变量之间在显著水平下线性相关。同时由于对各维度数据都进行了归一化处理，使得最后的回归模型偏执为 0。

置信区间：[-11.456878752789025 , 11.456878752789025]

【4】去归一化。使用之前得到的数据，去归一化。即可得到最终的回归模型。

$$z(t) \approx a^T x(t)$$
$$x(t) \approx \hat{L}\hat{y}(t)$$
$$\hat{y}(t) = \hat{L}^T x(t)$$
$$\hat{z}(t) = \hat{b}^T \hat{y}(t) = \hat{b}^T \hat{L}^T x(t)$$

```
[*] 线性回归方程为： Turnout = (-0.000853941)*X1 + (-8.13e-07)*X2 + 0.044496894*X3 +
                          1.059604556*X4 + 0.924855926*X5 + (-0.000446036)*X6 +
                          0.280755291*X7 + 0.000230413*X8 + 0.159167211*X9 +
                          0.026876745*X10 + (-0.179843049)*X11 + 0.222717677*X12 +
                          0.059250276*X13 + (-0.035697272)*X14 + 15.91985202
```

2.3 遇到的问题及解决方案

（一）数据读入

之前没有处理过 xlsx 数据，调研后发现 panda 擅长对表格数据数 据，故用 panda 对数据进行了载入。

（二）数据恢复

上课主要讲解了数据主成分的提取，即数据压缩的过程，对数据复原不是很了解。因此也是顺便学习了一下。

（三）病态问题降维回归结果的复原

对降维之后的数据进行回归之后，然后对其进行复原到原始数据。整个过程的矩阵操作比较繁琐，但经过自己的分析和理解其中的本质原理，对其进行了复原。