Problem_Set_2

Pierpaolo De Blasi

deadline 2025-05-13 11.59pm

Exercise 1

The data set `pulp_paper` contains measurements of properties of pulp fibers and the paper made from them. There are $n = 62$ observations on 4 paper properties: breaking length (`BL`), elastic modulus (`EM`), stress at failure (`SF`), burst strength (`BS`); and 4 pulp fiber characteristics: arithmetic fiber length (`AFL`), long fiber fraction (`LFF`), fine fiber fraction (`FFF`), zero span tensile (`ZST`).

```
pulp_paper<-read.table("data/pulp_paper.txt",header=T)
dim(pulp_paper)
```

```
## [1] 62  8
```

```
head(pulp_paper)
```

```
##       BL    EM    SF    BS    AFL    LFF    FFF   ZST
## 1 21.312 7.039 5.326 0.932 -0.030 35.239 36.991 1.057
## 2 21.206 6.979 5.237 0.871  0.015 35.713 36.851 1.064
## 3 20.709 6.779 5.060 0.742  0.025 39.220 30.586 1.053
## 4 19.542 6.601 4.479 0.513  0.030 39.756 21.072 1.050
## 5 20.449 6.795 4.912 0.577 -0.070 32.991 36.570 1.049
## 6 20.841 6.919 5.108 0.784 -0.050 31.140 38.115 1.052
```

1. Obtain the maximum likelihood solution for $m = 2$ and $m = 3$ common factors on the standardize observations and compute the proportion of total sample variance due to each factor. List the estimated communalities, specific variances, and the residual matrix $\boldsymbol{S} - (\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}})$. Compare the results. Which choice of m do you prefer? Why?
2. Give an interpretation to the common factors in the $m = 2$ solution.
3. Make a scatterplot of the factor scores for $m = 2$ obtained by the regression method. Is their correlation equal to zero? Should we expect so? Comment.
4. Suppose we have a new observation $(15.5, 5.5, 2, -0.55, 0.6, 65, -5, 1.2)$. Calculate the corresponding $m = 2$ factor scores and add this bivariate point to the plot in 3. How is it placed compared to the rest of the $n = 62$ points? Could you tell without computing the factor scores? Comment.

Exercise 2

The dataset `glass` contains data on $n = 214$ single glass fragments. Each case has a measured refractive index (`RI`) and composition (weight percent of oxides of `Na`, `Mg`, `Al`, `Si`, `K`, `Ca`, `Ba` and `Fe`). The composition sums to around 100%; what is not anything else is sand. The fragments are classified as six types (variable `type`). The classes are window float glass (`WinF`), window non float glass (`WinNF`), vehicle window glass (`Veh`), containers (`Con`), tableware (`Tabl`) and vehicle headlamps (`Head`).

```
glass<-read.table("data/glass.txt",header=T)
glass$type<-factor(glass$type)
levels(glass$type)<-c("WinF","WinNF","Veh","Con","Tabl","Head")
table(glass$type)
```

```
##
##  WinF WinNF   Veh   Con  Tabl  Head
##    70    76    17    13     9    29
```

```
dim(glass)
```

```
## [1] 214  10
```

```
head(glass)
```

```
##        RI    Na   Mg   Al    Si    K   Ca Ba   Fe type
## 1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00 WinF
## 2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00 WinF
## 3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00 WinF
## 4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00 WinF
## 5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00 WinF
## 6 1.51596 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26 WinF
```

1. Use linear discriminant analysis to predict the glass type. Look at the first two discriminant directions: what are the most important variables in separating the classes? Comment.
2. Compute the training error. Are there any groups *less homogeneous* than the others? Comment.
3. Implement a 10-fold cross validation using the partition of the observations provided by the variable `groupCV` to estimate the error rate.Comment.
4. Use the first two discriminant variables for a two-dimensional representation of the data together with centroids by using color-coding for the 6 classes of the class variable `type` (use `lookup` color vector below). Comment in view of the answer to point 2.

```
lookup<-c("black", "blue", "brown",  "gray60",
          "green3", "orange")
names(lookup)<-as.character(unique(glass$type))
```

5. Compute the training error and the 10-fold cross validation error for each reduced-rank LDA classifier. Plot both error curves against the number of discriminant directions, add full-rank LDA errors found in points 2. and 3. What classifier do you prefer? Comment.
6. (*Optional*) Find a classification rule that improves on the CV error rate estimates found in point 5. Feel free to use any classification method, even one not covered in class.