Assignment #3 STA437H1S/2005H1S

due Wednesday March 30, 2016

Instructions: Students in STA437H1S do problems 1 through 3; those in STA2005H1S do all 4 problems.

1. Suppose that S is a symmetric positive definite $p \times p$ matrix with $S = V \Lambda V^T$ where Λ is a diagonal matrix with elements $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and V is an orthogonal matrix with columns $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_p$.

(a) Suppose that we approximate S by $\Psi + LL^T$ where

$$L = \begin{pmatrix} \lambda_1^{1/2}\boldsymbol{v}_1 & \lambda_2^{1/2}\boldsymbol{v}_2 & \cdots & \lambda_r^{1/2}\boldsymbol{v}_r \end{pmatrix}$$

and Ψ is a diagonal matrix with the diagonal elements of $\Psi + LL^T$ equal to those of S. If ℓ_{ij} is the (i, j)-element of L, give an expression for the i-th diagonal element of Ψ, ψ_{ii}.

(b) Define $D = \Psi + LL^T - S$. If $\{d_{ij}\}$ are the elements of D, show that

$$\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 \leq \lambda_{r+1}^2 + \cdots + \lambda_p^2.$$

(Hint: Note that the diagonal elements of D are 0 so that we can consider $LL^T - S$, which can be expressed in terms of $\boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_p$ and $\lambda_{r+1}, \cdots, \lambda_p$.

(c) How might you use the result of (b) to choose the number of factors?

2. The file `marks.txt` on Blackboard contains the exam marks data considered previously in lecture and on Assignment #1. The data can be read into R as follows:

```
> exam <- scan("marks.txt",what=list(0,0,0,0,0))
> mec <- exam[[1]]
> vec <- exam[[2]]
> alg <- exam[[3]]
> ana <- exam[[4]]
> sta <- exam[[5]]
```

(a) Using the function `factanal`, carry out factor analysis on the data assuming only a single factor. Does this model seem to fit the data adequately?

(b) Which variable seems to be most important in the single factor model? Can you reconcile this with the graphical dependence model discussed in lecture?

(c) Repeat the analysis of part (a), now assuming a two factor model. Does this model seem to be an improvement over the single factor model? Try out different rotations of the factor loadings to (possibly) find an interpretable set of loadings.

3. In Assignments #1 and #2, you looked at various aspects of a data on two species of rock crabs; in this problem, you will use linear discriminant analysis to derive a rule for classifying the colour and sex of a crab.

The data can be read into R using the following code:

```
> x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
> colour <- x[[1]]
> sex <- x[[2]]
> FL <- x[[4]]
> RW <- x[[5]]
> CL <- x[[6]]
> CW <- x[[7]]
> BD <- x[[8]]
```

Define 4 groups (colour/sex combinations) as follows:

```
> group <- rep(0,200)
> group[sex=="M"&colour=="B"] <- 1
> group[sex=="F"&colour=="B"] <- 2
> group[sex=="M"&colour=="O"] <- 3
> group[sex=="F"&colour=="O"] <- 4
> group <- factor(group)
```

(a) Using the function `lda`, do a linear discriminant analysis of the data using the option `CV=T`, which does leave-one-out cross validation.

```
> library(MASS) # library with lda
>  r <- lda(group~FL+RW+CL+CW+BD, CV=T)
```

The output `r` contains a component `r$class` containing the classification (that is, the predicted group) for each observation based on the leave-one-out analysis. Give an estimate of the misclassification rate.

(b) Now do a linear discriminant analysis without the option `CV=T`:

```
> r <- lda(group~FL+RW+CL+CW+BD}
> r # this will show the results of the LDA
```

Suppose you are given the following measurements for the 5 variables: FL = 18.7, RW = 15.0, CL = 35.0, CW = 40.3, BD = 16.6. What is your prediction of the species and sex of this crab based on the linear discriminant analysis? (This can be done using the `predict` function and the output from `lda` as follows:

```
> newdata <- data.frame(FL=18.7,RW=15.0,CL=35.0,CW=40.3,BD=16.6)
> predict(r,newdata)
```

(c) Repeat parts (a) and (b), now using quadratic discriminant analysis:

```
> r1 <- qda(group~FL+RW+CL+CW+BD, CV=T)
> r2 <- qda(group~FL+RW+CL+CW+BD)
> predict(r2,newdata)
```

Comment on the results. In particular, which method gives the smaller error rate based on the leave-one-out cross validation?

4. In the case of two groups, we might consider using linear regression as a classification method. Specifically, given data $(g_1, \boldsymbol{x}_1), \cdots, (g_n, \boldsymbol{x}_n)$ (where $g_1, \cdots, g_n$ each take one of two possible values 0 or 1), we choose $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}$ to minimize

$$\sum_{i=1}^{n} (g_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

We can then use the value of $\widehat{\beta}_0 + \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$ to classify an observation $\boldsymbol{x}$; for example, we might classify $\boldsymbol{x}$ to $g = 1$ if $\widehat{\beta}_0 + \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$ is greater than some threshold.

(a) Define

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

Show that

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \right)^{-1} \left(\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) g_i \right)$$

(b) Now define

$$\begin{aligned} \bar{\boldsymbol{x}}_0 &= \frac{\sum_{i=1}^{n} (1 - g_i) \boldsymbol{x}_i}{\sum_{i=1}^{n} (1 - g_i)} \\ \bar{\boldsymbol{x}}_1 &= \frac{\sum_{i=1}^{n} g_i \boldsymbol{x}_i}{\sum_{i=1}^{n} g_i} \end{aligned}$$

to be the means of observations within the two groups and

$$S = \frac{1}{n-2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{g_i})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{g_i})^T$$

to be an estimate of a common covariance matrix for the two groups. Show that

$$\widehat{\boldsymbol{\beta}} = k S^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0)$$

for some constant k. (Hint: Note that

$$\begin{aligned} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T &= \sum_{i=1}^{n} g_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T + \sum_{i=1}^{n} (1 - g_i)(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \\ \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{g_i})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{g_i})^T &= \sum_{i=1}^{n} g_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_1)^T + \sum_{i=1}^{n} (1 - g_i)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_0)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_0)^T \end{aligned}$$

and $\bar{\boldsymbol{x}} = \lambda \bar{\boldsymbol{x}}_1 + (1 - \lambda) \bar{\boldsymbol{x}}_0$ where $\lambda = (g_1 + \cdots + g_n)/n$.)