Computer Lab 4

Thomas Zhang

2015-12-09

Assignment 1

Given the Andrew and Herzberg diabetic and non-diabetic data set we want to test whether there is any correlation between the primary and secondary variables at the 5% significance level. In order to do this we set null hypothesis to $H_0 = \mathbf{\Sigma_{12}} = 0$ and perform a Chi squared test on the likelihood ratio test statistic with Bartlett correction at pq degrees of freedom, as described in section 10.6 of the Johnson & Wichern course textbook.

```
## [1] "Bartlett likelihood ratio test statistic for all canonical correlations zero: 13.749"
```

```
## [1] "Chi square statistic value at 5% significance level, 6 degrees of freedom: 12.592"
```

We see that the liklihood ratio test statistic is greater than the Chi square test value at 5% significance level, so we reject the null hypothesis H_0 that $\mathbf{\Sigma_{12}}$ is zero.

We can extend our testing by examining the "significance" of the individual canonical correlations. Since we in this case only have two canonical correlations we set as null hypothesis H_0 that the largest canonical correlation is non-zero and the second largest canonical correlation is zero. Now we test this new null hypothesis as before.

```
## [1] "Bartlett likelihood ratio test statistic for second canonical correlation zero: 0.667"
```

```
## [1] "Chi square statistic value at 5% significance level, 2 degrees of freedom: 5.991"
```

It is observed that the liklihood ratio test statistic is smaller than the Chi square test value at 5% significance level, so we cannot reject the null hypothesis H_0. In summary, there exists an association between primary and secondary variables, but this association is only significant at the 5% significance level for the first pair out of two pairs of canonical variates.

Now, the squared canonical correlations indicate how large a fraction of the variance in the primary variables can be explained by the variance in the secondary variables, and vice versa. For our first canonical correlation:

```
## [1] "First canonical correlation: 0.517"
```

```
## [1] "First canonical correlation squared: 0.268"
```

The canonical variates themselves, in the first canonical correlation, seem to mainly represent a negative correlation between insulin resistance and relative weight, and a positive correlation between insulin response to oral glucose and relative weight. Further, it can be seen that V1 is mostly correlates with the negative of relative weight, while U1 mainly correlates with positive of insulin resistance.

```
## [1] "Coefficients for canonical variable U for standardized X:"
```

```
##        Glucose intolerance Insulin response to oral glucose
## First            0.4356829                       -0.7046696
## Second           0.8231828                       -0.4547487
##        Insulin resistance
## First           1.0814622
## Second         -0.4005717
```

```
## [1] "Coefficients for canonical variable V for standardized X:"
```

```
##        Relative weight Fasting plasma glucose
## First      -1.02022352              0.160936
## Second      0.04745539             -1.008567
```

```
## [1] "correlations between Ui and its standardized components:"
```

```
##    Glucose intolerance Insulin response to oral glucose Insulin resistance
## U1           0.3397282                      -0.0501787          0.7551136
## U2           0.6837882                      -0.4565378         -0.5729495
```

```
## [1] "correlations between Vi and its standardized components:"
```

```
##    Relative weight Fasting plasma glucose
## V1      -0.9875069            -0.04646446
## V2      -0.1575755            -0.99891994
```

Since U1 and V1 only correlates strongly with one of the standardized variables each, I can not say that they are good summary measures of their respective data sets. In particular, glucose intolerance and fasting plasma glucose variables have little to no influence on the first, and only significant canonical correlation. Furthermore, only 26% or so of the variance in U1 can be explained by the variance in V1. There is also the questionable correlation that relative weight is negatively correlated with insulin resistance. (should it not be positive?). In conclusion, this canonical correlation analysis is not very good.

Appendix - R-Code

```
smpcovmatr <- matrix(0,5,5)
diag(smpcovmatr) <- c( 1106,2382,2136,.016,70.56)
diag2 <- c(396.7,1143,2.189,.216)
for(i in 1:4){
smpcovmatr[i,i+1] <- diag2[i]
}
diag3 <- c(108.4,-.214,-20.84)
for(i in 1:3){
  smpcovmatr[i,i+2] <- diag3[i]
}
diag4 <-c(.787,-23.96)
for(i in 1:2){
  smpcovmatr[i,i+3] <- diag4[i]
}
smpcovmatr[1,5] <- c(26.23)
smpcovmatr[lower.tri(smpcovmatr)] = t(smpcovmatr)[lower.tri(smpcovmatr)]
```

```r
cormatr <- cov2cor(smpcovmatr)
s11 <- smpcovmatr[1:3,1:3]
s12 <- smpcovmatr[1:3,4:5]
s22 <- smpcovmatr[4:5,4:5]

matrsqrtinv <- function(A){
  eig <- eigen(A)
  vects <- eig$vectors
  vals <- eig$values
  result <- matrix(0,nrow=ncol(vects),ncol=ncol(vects))
  for(i in 1:length(vals)){
    result <- result + 1/sqrt(vals[i]) * crossprod(t(vects[,i]),t(vects[,i]))
  }
  return(result)
}

thatmatr <- matrsqrtinv(s22) %*% t(s12) %*% solve(s11) %*% s12 %*% matrsqrtinv(s22)
ees <- eigen(thatmatr)


ffs <- eigen(matrsqrtinv(s11) %*% (s12) %*% solve(s22) %*% t(s12) %*% matrsqrtinv(s11))
thosevects <- ffs$vectors
vcoefs <- t(thosevects) %*% matrsqrtinv(s11)
ucoefs <- t(ees$vectors) %*% matrsqrtinv(s22)

dmat <- diag(x=sqrt(c( 1106,2382,2136,.016,70.56)))

# Self-correlations
selfcorrs1 <- vcoefs %*% s11 %*% solve(dmat[1:3,1:3])
selfcorrs2 <- ucoefs %*% s22 %*% solve(dmat[4:5,4:5])
colnames(selfcorrs1) <- c("Glucose intolerance","Insulin response to oral glucose",
                          "Insulin resistance")
colnames(selfcorrs2) <- c("Relative weight","Fasting plasma glucose")
selfcorrs1 <- selfcorrs1[1:2,]
rownames(selfcorrs1) <- c("U1","U2")
rownames(selfcorrs2) <- c("V1","V2")

#Let us say obsvs are standardized
vcoefs <- vcoefs %*% dmat[1:3,1:3]
vcoefs <- vcoefs[1:2,]
ucoefs <- ucoefs %*% dmat[4:5,4:5]
colnames(vcoefs) <- c("Glucose intolerance","Insulin response to oral glucose",
                      "Insulin resistance")
rownames(vcoefs) <- c("First","Second")
colnames(ucoefs) <- c("Relative weight","Fasting plasma glucose")
rownames(ucoefs) <- c("First","Second")

#reject criterion -(n-1 - 1/2 (p + q + 1))ln PI(1- rho^2) > chisq_pq(alpha)
p <- 2
q <- 3
n <- 46
const <- -(n - 1 - 1/2 * (p + q + 1))
logfactor <- log((1 - ees$values[1] ) * (1- ees$values[2]))
```

```
tryingtobeat <- const * logfactor
beatthis <- qchisq(0.95,6)
logfactor2 <- log((1- ees$values[2]))
tryingtobeat2 <- const * logfactor2
nextobeat <- qchisq(0.95,2)
paste("Bartlett likelihood ratio test statistic for all canonical correlations zero:",
      round(tryingtobeat,3))
paste("Chi square statistic value at 5% significance level, 6 degrees of freedom:",round(beatthis,3))
paste("Bartlett likelihood ratio test statistic for second canonical correlation zero:",
      round(tryingtobeat2,3))
paste("Chi square statistic value at 5% significance level, 2 degrees of freedom:",round(nextobeat,3))
paste("First canonical correlation:",signif(sqrt(ees$values[1]),3))
paste("First canonical correlation squared:",signif(ees$values[1],3))

paste("Coefficients for canonical variable U for standardized X:")
vcoefs
paste("Coefficients for canonical variable V for standardized X:")
ucoefs
paste("correlations between Ui and its standardized components:")
selfcorrs1
paste("correlations between Vi and its standardized components:")
selfcorrs2
## NA
```