Canonical Correlation Analysis & Reduced Rank Regression

Casey Tirshfield

The file `m_logret_4auto.txt` contains the monthly log returns of four automobile manufacturers (General Motors Corp., Toyota Motor Corp., Ford Motor Co., and Honda Motor Co.) from January 1994 to June 2007. The file `m_logret_4soft.txt` contains the monthly log returns of four application software companies (Adobe Systems Inc., Microsoft Corp., Oracle Corp., and SPSS Inc.) from January 1994 to June 2007.

```
auto_df <- read.table('m_logret_4auto.txt', header=TRUE)
soft_df <- read.table('m_logret_4soft.txt', header=TRUE)
auto_df<- auto_df[,-1]
soft_df <- soft_df[,-1]
```

(a) Perform a canonical correlation analysis for these two sets of returns. Give the first two estimated canonical variate pairs and the corresponding canonical correlations.

```
# this code was inspired by https://stats.idre.ucla.edu/r/dae/canonical-correlation-a
nalysis/
# determine correlations
canonical <- matcor(auto_df, soft_df)
print(canonical)
```

```
$Xcor
             GM        TM         F     Honda
GM    1.0000000 0.2258469 0.5546114 0.3388108
TM    0.2258469 1.0000000 0.2445330 0.5013922
F     0.5546114 0.2445330 1.0000000 0.2512618
Honda 0.3388108 0.5013922 0.2512618 1.0000000

$Ycor
          ADBE      msft      orcl      SPSS
ADBE 1.0000000 0.2043457 0.2842287 0.1980541
msft 0.2043457 1.0000000 0.3294099 0.1568481
orcl 0.2842287 0.3294099 1.0000000 0.3472477
SPSS 0.1980541 0.1568481 0.3472477 1.0000000

$XYcor
              GM        TM         F     Honda      ADBE      msft      orcl
GM    1.00000000 0.2258469 0.5546114 0.3388108 0.1736931 0.1531596 0.1495429
TM    0.22584690 1.0000000 0.2445330 0.5013922 0.2223710 0.4114275 0.2242932
F     0.55461142 0.2445330 1.0000000 0.2512618 0.1523644 0.2516906 0.1378794
Honda 0.33881077 0.5013922 0.2512618 1.0000000 0.2042285 0.1850978 0.1052381
ADBE  0.17369306 0.2223710 0.1523644 0.2042285 1.0000000 0.2043457 0.2842287
msft  0.15315960 0.4114275 0.2516906 0.1850978 0.2043457 1.0000000 0.3294099
orcl  0.14954288 0.2242932 0.1378794 0.1052381 0.2842287 0.3294099 1.0000000
SPSS  0.06668592 0.1762263 0.1160017 0.1544895 0.1980541 0.1568481 0.3472477
            SPSS
GM    0.06668592
TM    0.17622625
F     0.11600171
Honda 0.15448948
ADBE  0.19805413
msft  0.15684810
orcl  0.34724769
SPSS  1.00000000
```

Hide

```
cc1 <- cc(auto_df, soft_df)
# display the canonical correlations
cc1$cor
```

```
[1] 0.479160908 0.141332071 0.114717292 0.002213036
```

Hide

```
# raw canonical coefficients
cc1[3:4]
```

```
$xcoef
             [,1]       [,2]       [,3]        [,4]
GM     -0.6046899   8.136041  28.126146   0.5743063
TM    -26.6678075 -18.684736   4.678123  16.9112510
F      -8.7386900  -5.122038 -13.729627 -23.5336977
Honda   0.4058833  33.113737 -18.490662   4.1132065

$ycoef
           [,1]       [,2]       [,3]      [,4]
ADBE  -4.344641  13.200474   3.775565 -7.396720
msft -18.119942 -11.969220  -7.020909 -8.057457
orcl  -1.601346  -3.388370  15.081255 10.514573
SPSS  -3.353271   6.867945 -14.025538 11.632298
```

Hide

```
# compute canonical loadings
cc2 <- comput(auto_df, soft_df, cc1)
# display canonical loadings
canon_var <- cc2[3:6]
# tests of canonical dimensions
ev <- (1 - cc1$cor^2)
n <- dim(auto_df)[1]
p <- length(auto_df)
q <- length(soft_df)
k <- min(p, q)
m <- n - 3/2 - (p + q)/2
w <- rev(cumprod(rev(ev)))
# initialize
d1 <- d2 <- f <- vector("numeric", k)
for (i in 1:k) {
    s <- sqrt((p^2 * q^2 - 4)/(p^2 + q^2 - 5))
    si <- 1/s
    d1[i] <- p * q
    d2[i] <- m * s - p * q/2 + 1
    r <- (1 - w[i]^si)/w[i]^si
    f[i] <- r * d2[i]/d1[i]
    p <- p - 1
    q <- q - 1
}
pv <- pf(f, d1, d2, lower.tail = FALSE)
(dmat <- cbind(WilksL = w, F = f, df1 = d1, df2 = d2, p = pv))
```

```
        WilksL            F df1      df2            p
[1,] 0.7450765 3.0158715967  16 477.2255 7.764369e-05
[2,] 0.9671233 0.5874097898   9 382.2474 8.075618e-01
[3,] 0.9868351 0.5252045867   4 316.0000 7.172846e-01
[4,] 0.9999951 0.0007787107   1 159.0000 9.777726e-01
```

Hide

```
# standardized auto canonical coefficients diagonal matrix of auto sd's
s1 <- diag(sqrt(diag(cov(auto_df))))
s1 %*% cc1$xcoef
```

```
            [,1]       [,2]       [,3]        [,4]
[1,] -0.02560735  0.3445443  1.1910834  0.02432067
[2,] -0.84176785 -0.5897827  0.1476647  0.53380269
[3,] -0.36530227 -0.2141159 -0.5739377 -0.98377598
[4,]  0.01276295  1.0412574 -0.5814366  0.12933927
```

Hide

```
# standardized soft canonical coefficients diagonal matrix of soft sd's
s2 <- diag(sqrt(diag(cov(soft_df))))
s2 %*% cc1$ycoef
```

```
            [,1]       [,2]       [,3]       [,4]
[1,] -0.28361046  0.8617035  0.2464621 -0.4828447
[2,] -0.79938867 -0.5280402 -0.3097380 -0.3554669
[3,] -0.09713091 -0.2055243  0.9147655  0.6377698
[4,] -0.18219100  0.3731515 -0.7620401  0.6320098
```

Hide

```
# this code was inspired by page 202 of the textbook
# first pair
alpha_pair1 <- canonical$Xcor %*% canon_var$corr.X.xscores[1,]
beta_pair1 <- canonical$Ycor %*% canon_var$corr.Y.xscores[1,]
print(alpha_pair1)
```

```
             [,1]
GM    -0.04104459
TM     0.34633290
F      0.49874928
Honda -0.09569267
```

Hide

```
print(beta_pair1)
```

```
              [,1]
ADBE -0.2130477107
msft  0.0696821116
orcl -0.0003970646
SPSS -0.0203169636
```

Hide

```
# second pair
alpha_pair2 <- canonical$Xcor %*% canon_var$corr.X.xscores[2,]
beta_pair2 <- canonical$Ycor %*% canon_var$corr.Y.xscores[2,]
print(alpha_pair2)
```

```
             [,1]
GM    -0.82527161
TM    -0.07381735
F     -0.45023988
Honda  0.02331859
```

Hide

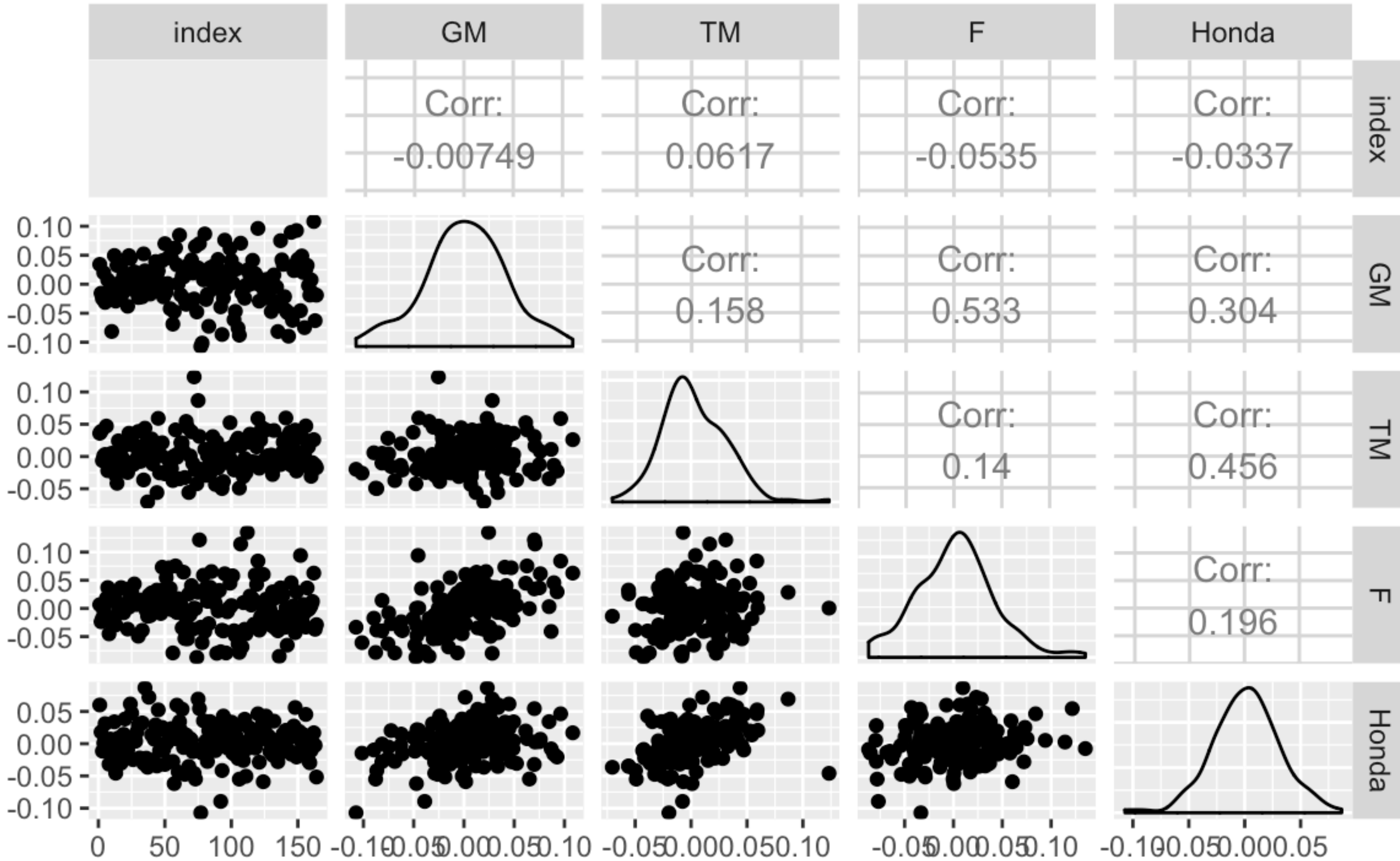
```
print(beta_pair2)
```

```
            [,1]
ADBE -0.45285145
msft -0.14389770
orcl -0.15083438
SPSS -0.09852595
```

(b) Perform reduced-rank regression of the log returns of automobile stocks on those of software company stocks, taking $rank(\mathbf{B}) = 2$ in $\mathbf{y}_k = \mathbf{B}\mathbf{x}_k + \epsilon_k, \quad k = 1, \ldots, n$ with response variable $\mathbf{y}_k = (y_{k1}, \ldots, y_{kq})^\top \in \mathbb{R}^q$, predictor variable $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^\top \in \mathbb{R}^p$, and random error $\epsilon_k \in \mathbb{R}^q$ such that $\mathbb{E}(\epsilon_k) = \mathbf{0}$ and $Cov(\epsilon_k) = \mathbf{0}$ and $Cov(\epsilon_k) = \Sigma$.

Hide

```
# calculate the regression residuals for reduced-rank regression and canonical variat
e analysis
residuals(soft_df, auto_df, type='identity', rank = 'full', k=0, plot=TRUE)
```



Hide

```
# fit the reduced-rank regression model
rrr(soft_df, auto_df, type = "identity", rank = '2', k = 0)
```

```
$mean
              [,1]
GM    -0.001215402
TM     0.001137627
F     -0.001946223
Honda  0.002988609

$A
            [,1]       [,2]
GM    -0.4056943  0.7724362
TM    -0.6420723 -0.4688881
F     -0.5357772 -0.2582809
Honda -0.3689148  0.3417269

$B
            ADBE        msft        orcl       SPSS
[1,] -0.13155359 -0.35253570 -0.04002341 -0.0759087
[2,]  0.04970285 -0.06448038  0.02444612 -0.0134459

$C
            ADBE       msft       orcl       SPSS
GM    0.09176282 0.09321473 0.03512033 0.02040962
TM    0.06116185 0.25658750 0.01423543 0.05504350
F     0.05764612 0.20553463 0.01512966 0.04414297
Honda 0.06551686 0.10802095 0.02311912 0.02340902

$eigen_values
[1] 4.709495e-04 1.649588e-05 9.183044e-06 5.713889e-09
```