Multivariate Statistical Inference HW# 6

Steven Francis

April 19, 2018

Problem 1

```
fish.corr <- array(data = c(1.0000, 0.4919, 0.2636, 0.4653, 0.4919, 1.0000,
                            0.3127, 0.3506, 0.2636, 0.3127, 1.0000, 0.4108,
                            0.4653, 0.3506, 0.4108, 1.0000), dim = c(4,4),
dimnames = list(c("x1","x2","x3","x4"),c("x1","x2","x3","x4")))

fish.corr
```

```
##        x1     x2     x3     x4
## x1 1.0000 0.4919 0.2636 0.4653
## x2 0.4919 1.0000 0.3127 0.3506
## x3 0.2636 0.3127 1.0000 0.4108
## x4 0.4653 0.3506 0.4108 1.0000
```

Part a

```
#Obtaining the eigenvalues and eigenvectors of correlation matrix
sp <- eigen(fish.corr)
sp
```

```
## eigen() decomposition
## $values
## [1] 2.1539422 0.7875151 0.6156498 0.4428929
##
## $vectors
##            [,1]       [,2]       [,3]       [,4]
## [1,] -0.5265283  0.4571532  0.2491871  0.6720749
## [2,] -0.5032995  0.4120178 -0.6142318 -0.4468223
## [3,] -0.4428007 -0.7583919 -0.3680759  0.3054332
## [4,] -0.5228624 -0.2146951  0.6520316 -0.5053471
```

```
Gam <- sp$vectors
Lamb <- diag(sp$values)

#Reproducing R to check validity of eigenvalues and eigenvectors
Gam %*% Lamb %*% t(Gam)
```

```
##        [,1]   [,2]   [,3]   [,4]
## [1,] 1.0000 0.4919 0.2636 0.4653
## [2,] 0.4919 1.0000 0.3127 0.3506
## [3,] 0.2636 0.3127 1.0000 0.4108
## [4,] 0.4653 0.3506 0.4108 1.0000
```

```r
#Fitting the single factor model
Q.hat <- (Gam %*% sqrt(Lamb))[,1]
Q.hat
```

```
## [1] -0.7727495 -0.7386583 -0.6498683 -0.7673693
```

```r
Q.hat %*% t(Q.hat)
```

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] 0.5971417 0.5707978 0.5021854 0.5929842
## [2,] 0.5707978 0.5456160 0.4800306 0.5668237
## [3,] 0.5021854 0.4800306 0.4223288 0.4986890
## [4,] 0.5929842 0.5668237 0.4986890 0.5888556
```

```r
Psi.hat <- diag(diag(fish.corr - Q.hat %*% t(Q.hat)))
Psi.hat
```

```
##           [,1]     [,2]      [,3]      [,4]
## [1,] 0.4028583 0.000000 0.0000000 0.0000000
## [2,] 0.0000000 0.454384 0.0000000 0.0000000
## [3,] 0.0000000 0.000000 0.5776712 0.0000000
## [4,] 0.0000000 0.000000 0.0000000 0.4111444
```

```r
#To check the fit of the one-factor model
round(fish.corr - (Q.hat %*% t(Q.hat) + Psi.hat), 6)
```

```
##           x1        x2        x3        x4
## x1  0.000000 -0.078898 -0.238585 -0.127684
## x2 -0.078898  0.000000 -0.167331 -0.216224
## x3 -0.238585 -0.167331  0.000000 -0.087889
## x4 -0.127684 -0.216224 -0.087889  0.000000
```

We have a single factor (k = 1) model that takes the form (y = qf + u) where: q.hat = (0.77, 0.74, 0.65, 0.77)' and psi.hat = diag(0.40, 0.45, 0.58, 0.41)

Part b

```r
factanal(covmat = fish.corr, factors = 1, rotation = "none")
```

```
##
## Call:
## factanal(factors = 1, covmat = fish.corr, rotation = "none")
##
## Uniquenesses:
##    x1    x2    x3    x4
## 0.498 0.603 0.764 0.574
##
## Loadings:
##    Factor1
## x1 0.708
## x2 0.630
## x3 0.485
## x4 0.653
##
##                Factor1
```

```
## SS loadings      1.561
## Proportion Var   0.390
##
## The degrees of freedom for the model is 2 and the fit was 0.0571
```

We have a single factor (k = 1) model that takes the form (y = qf + u) where: q.hat = (0.71, 0.63, 0.49, 0.65)' and psi.hat = diag(0.50, 0.60, 0.76, 0.57)

Part c

```
#Fitting the 2 factor model
Q.hat <- (Gam %*% sqrt(Lamb))[,1:2]
Q.hat
```

```
##            [,1]       [,2]
## [1,] -0.7727495  0.4056871
## [2,] -0.7386583  0.3656331
## [3,] -0.6498683 -0.6730125
## [4,] -0.7673693 -0.1905248
```

```
#Communalities Calculation
diag(Q.hat %*% t(Q.hat))
```

```
## [1] 0.7617238 0.6793036 0.8752746 0.6251553
```

```
Psi.hat <- diag(diag(fish.corr - Q.hat %*% t(Q.hat)))
#Uniqueness
diag(Psi.hat)
```

```
## [1] 0.2382762 0.3206964 0.1247254 0.3748447
```

```
#To check the fit of the two-factor model
fish.corr - (Q.hat %*% t(Q.hat) + Psi.hat)
```

```
##             x1         x2          x3          x4
## x1  0.00000000 -0.2272304  0.03444711 -0.05039073
## x2 -0.22723040  0.0000000  0.07874500 -0.14656149
## x3  0.03444711  0.0787450  0.00000000 -0.21611458
## x4 -0.05039073 -0.1465615 -0.21611458  0.00000000
```

```
#Variances accounted for by the two common factors are the first two eigenvalues
diag(Lamb)[1:2]
```

```
## [1] 2.1539422 0.7875151
```

The communalities range from about 0.63 to 0.88 and the specificities range between about 0.12 to 0.37

We have a two factor (k = 2) model that takes the form (y = qf + u) where: q.hat = ((0.77, 0.74, 0.65, 0.77),(0.41, 0.37, -0.67, -0.19))' and

psi.hat11 = 0.24 psi.hat22 = 0.32 psi.hat33 = 0.12 psi.hat44 = 0.37

It seems as if the first factor loads pretty equally on all four variables and the second factor loads the Smallmouth and Largemouth bass variables (x3 and x4) in the opposite direction of the Bluegill and Black Crappie variables (x1 and x2). This could be due to the peak seasons for catching the types of bass being different than the peak seasons for catching Bluegill and Black Crappie species.

```r
#Creation of orthogonal matrix G to use in factor rotation
G <- matrix(c(-1,-1,1,-1)/sqrt(2), 2,2)
G
```

```
##            [,1]       [,2]
## [1,] -0.7071068  0.7071068
## [2,] -0.7071068 -0.7071068
```

```r
#Rotation of factor loadings by 45 degrees
Q.star <- Q.hat %*% G
Q.star
```

```
##           [,1]        [,2]
## [1,] 0.2595523 -0.83328051
## [2,] 0.2637687 -0.78085188
## [3,] 0.9354180  0.01636539
## [4,] 0.6773334 -0.40789063
```

Although there are no zeros, we have a couple of smallish loadings. We might interpret factor 1 as driving much of component 3 and a good portion of component 4 (Smallmouth and Largemouth Bass respectively), and factor 2 as driving much of components 1 and 2 (Bluegill and Black Crappie respectively).

Problem 2

```r
air.pollution.data <- read.csv("AirPollution.csv", header = T)
air.pollution.data
```

```
##    x1  x2 x3 x4
## 1   8  98 12  8
## 2   7 107  9  5
## 3   7 103  5  6
## 4  10  88  8 15
## 5   6  91  8 10
## 6   8  90 12 12
## 7   9  84 12 15
## 8   5  72 21 14
## 9   7  82 11 11
## 10  8  64 13  9
## 11  6  71 10  3
## 12  6  91 12  7
## 13  7  72 18 10
## 14 10  70 11  7
## 15 10  72  8 10
## 16  9  77  9 10
## 17  8  76  7  7
## 18  8  71 16  4
## 19  9  67 13  2
## 20  9  69  9  5
## 21 10  62 14  4
## 22  9  88  7  6
## 23  8  80 13 11
## 24  5  30  5  2
## 25  6  83 10 23
```

```
## 26   8  84  7  6
## 27   6  78 11 11
## 28   8  79  7 10
## 29   6  62  9  8
## 30  10  37  7  2
## 31   8  71 10  7
## 32   7  52 12  8
## 33   5  48  8  4
## 34   6  75 10 24
## 35  10  35  6  9
## 36   8  85  9 10
## 37   5  86  6 12
## 38   5  86 13 18
## 39   7  79  9 25
## 40   7  79  8  6
## 41   6  68 11 14
## 42   8  40  6  5
```

```r
R <- cor(air.pollution.data)
R
```

```
##            x1         x2         x3         x4
## x1  1.0000000 -0.1014419 -0.1098249 -0.2535928
## x2 -0.1014419  1.0000000  0.1157320  0.3191237
## x3 -0.1098249  0.1157320  1.0000000  0.1666422
## x4 -0.2535928  0.3191237  0.1666422  1.0000000
```

Part a

```r
sp <- eigen(R)
sp
```

```
## eigen() decomposition
## $values
## [1] 1.5556393 0.9097107 0.8980289 0.6366211
##
## $vectors
##            [,1]       [,2]        [,3]        [,4]
## [1,]  0.4520666 -0.2545336  0.77712849  0.35625792
## [2,] -0.5171406 -0.5465407  0.37124600 -0.54409127
## [3,] -0.3824193  0.7708084  0.50521588 -0.06608137
## [4,] -0.6180266 -0.2058161 -0.05481445  0.75675507
```

```r
Gam <- sp$vectors
Lamb <- diag(sp$values)

Gam %*% Lamb %*% t(Gam)
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,]  1.0000000 -0.1014419 -0.1098249 -0.2535928
## [2,] -0.1014419  1.0000000  0.1157320  0.3191237
## [3,] -0.1098249  0.1157320  1.0000000  0.1666422
## [4,] -0.2535928  0.3191237  0.1666422  1.0000000
```

```r
#Fitting the single factor model
Q.hat <- (Gam %*% sqrt(Lamb))[,1]
Q.hat
```

```
## [1]  0.5638413 -0.6450049 -0.4769735 -0.7708354
```

```r
Q.hat %*% t(Q.hat)
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,]  0.3179171 -0.3636805 -0.2689374 -0.4346288
## [2,] -0.3636805  0.4160314  0.3076503  0.4971926
## [3,] -0.2689374  0.3076503  0.2275037  0.3676680
## [4,] -0.4346288  0.4971926  0.3676680  0.5941871
```

```r
Psi.hat <- diag(diag(R - Q.hat %*% t(Q.hat)))
Psi.hat
```

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] 0.6820829 0.0000000 0.0000000 0.0000000
## [2,] 0.0000000 0.5839686 0.0000000 0.0000000
## [3,] 0.0000000 0.0000000 0.7724963 0.0000000
## [4,] 0.0000000 0.0000000 0.0000000 0.4058129
```

```r
R - (Q.hat %*% t(Q.hat) + Psi.hat)
```

```
##           x1         x2         x3         x4
## x1 0.0000000  0.2622385  0.1591125  0.1810360
## x2 0.2622385  0.0000000 -0.1919183 -0.1780689
## x3 0.1591125 -0.1919183  0.0000000 -0.2010258
## x4 0.1810360 -0.1780689 -0.2010258  0.0000000
```

We have a single factor (k = 1) model that takes the form (y = qf + u) where: q.hat = (0.563, -0.645, -0.476, -0.771)' and psi.hat = diag(0.682, 0.584, 0.772, 0.406)

Proportion of total sample variance explained by single factor is: sum(Q.hat^2)/4 = 0.3889098

Part b

```r
factanal(air.pollution.data, factors = 1, rotation = "none")
```

```
##
## Call:
## factanal(x = air.pollution.data, factors = 1, rotation = "none")
##
## Uniquenesses:
##    x1    x2    x3    x4
## 0.895 0.832 0.946 0.405
##
## Loadings:
##    Factor1
## x1 -0.324
## x2  0.410
## x3  0.232
## x4  0.771
##
```

```
##                 Factor1
## SS loadings      0.921
## Proportion Var   0.230
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 0.15 on 2 degrees of freedom.
## The p-value is 0.93
```

We have a single factor (k = 1) model that takes the form (y = qf + u) where: q.hat = (-0.324, 0.410, 0.232, 0.771)' and psi.hat = diag(0.895, 0.832, 0.946, 0.405)

Part c

Both methods return Q.hat values that show wind having a converse effect when compared to Solar radiation, NO_2, and O3. The psi.hat values (covariance matrix for u) are larger in the maximum likelihood factor analysis method than the principal component solution analysis method.

Also the proportions of total variance explain was higher in the principal component solution analysis method (0.39) than the maximum likelihood factor analysis method (0.23).

Part d

```r
#Fitting the 2 factor model
Q.hat <- (Gam %*% sqrt(Lamb))[,1:2]
Q.hat
```

```
##            [,1]       [,2]
## [1,]  0.5638413 -0.2427710
## [2,] -0.6450049 -0.5212837
## [3,] -0.4769735  0.7351875
## [4,] -0.7708354 -0.1963048
```

```r
#Varimax of Q.hat
varimax(Q.hat)
```

```
## $loadings
##
## Loadings:
##      [,1]   [,2]
## [1,]  0.313 -0.528
## [2,] -0.828
## [3,]         0.875
## [4,] -0.739  0.295
##
##                 [,1]  [,2]
## SS loadings    1.332 1.133
## Proportion Var 0.333 0.283
## Cumulative Var 0.333 0.616
##
## $rotmat
##           [,1]       [,2]
## [1,] 0.8085335 -0.5884501
## [2,] 0.5884501  0.8085335
```

```r
#Communalities Calculation
diag(Q.hat %*% t(Q.hat))
```

```
## [1] 0.3768548 0.6877681 0.7680044 0.6327227
```

```r
Psi.hat <- diag(diag(R - Q.hat %*% t(Q.hat)))
#Uniqueness
diag(Psi.hat)
```

```
## [1] 0.6231452 0.3122319 0.2319956 0.3672773
```

```r
#To check the fit of the two-factor model
R - (Q.hat %*% t(Q.hat) + Psi.hat)
```

```
##           x1         x2          x3          x4
## x1 0.0000000  0.1356860  0.33759467  0.13337892
## x2 0.1356860  0.0000000  0.19132301 -0.28039938
## x3 0.3375947  0.1913230  0.00000000 -0.05670501
## x4 0.1333789 -0.2803994 -0.05670501  0.00000000
```

```r
#Variances accounted for by the two common factors are the first two eigenvalues
diag(Lamb)[1:2]
```

```
## [1] 1.5556393 0.9097107
```

After performing a varimax rotation on the Q.hat matrix, the results show that factor 1 could be interpreted as being driven by the Solar radiation and O.3 variables, and factor 2 couble be interpreted as being driven by the NO.2 variable. Within both factors, Wind has a converse effect on the other pertinent variables.