PCA and regression

Philip Dixon

9/25/2019

make a small data set

```
test <- rbind(
  c(10, 5, 0, 0),
  c(7, 3, 1, 4),
  c(4, 2, 2, 5),
  c(2, 1, 2, 7),
  c(0, 0, 5, 12))
test
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   10    5    0    0
## [2,]    7    3    1    4
## [3,]    4    2    2    5
## [4,]    2    1    2    7
## [5,]    0    0    5   12
```

PCA on covariance matrix

```
test.pca <- princomp(test)
test.pca
```

```
## Call:
## princomp(x = test)
##
## Standard deviations:
##     Comp.1     Comp.2     Comp.3     Comp.4
## 5.73770044 0.89427411 0.33185505 0.09454978
##
##  4  variables and  5 observations.
```

```
# spp scores
test.pca$loadings[,1]
```

```
## [1]  0.6106175  0.2959878 -0.2793595 -0.6793348
```

```
# site scores
test.pca$scores[,1]
```

```
## [1]  8.48909419  3.06856743 -0.01796721 -2.89385959 -8.64583482
```

iterated regressions on centered data matrix

```
# initial values for site scores
sitec <- rnorm(5)

sppmeans <- apply(test, 2, mean)
testc <- t(t(test) - sppmeans)
```

```
for (k in 1:5) {
sppc <- rep(NA, 4)
for (i in 1:4) {
  sppc[i] <- coef(lm(testc[,i] ~ -1+sitec))[1]
  }
sppc <- sppc / sqrt(sum(sppc^2))
print(sppc)

for (i in 1:5) {
  sitec[i] <- coef(lm(testc[i,] ~  -1+sppc))[1]
  }
print(sitec)
cat('\n')
}
```

```
## [1] 0.8240652 0.4734464 0.2119895 0.2276521
## [1]  4.0767717  1.7802809 -0.7257197 -2.3919924 -2.7393406
##
## [1]  0.6434421  0.3091449 -0.2564545 -0.6516462
## [1]  8.5023208  3.0906654 -0.0569065 -2.9562280 -8.5798517
##
## [1]  0.6114334  0.2963167 -0.2788060 -0.6786846
## [1]  8.48967301  3.06919491 -0.01891258 -2.89546530 -8.64449005
##
## [1]  0.6106373  0.2959958 -0.2793461 -0.6793190
## [1]  8.48910840  3.06858271 -0.01799017 -2.89389865 -8.64580230
##
## [1]  0.6106180  0.2959880 -0.2793592 -0.6793344
## [1]  8.48909453  3.06856780 -0.01796776 -2.89386054 -8.64583403
```

```
# these are the sd's for axis 1
sd(sitec)*sqrt(4/5)
```

```
## [1] 5.7377
```

```
# spp scores on axis 1
sppc
```

```
## [1]  0.6106180  0.2959880 -0.2793592 -0.6793344
```

```
# site scores on axis 1
sitec
```

```
## [1]  8.48909453  3.06856780 -0.01796776 -2.89386054 -8.64583403
```

Use scores on axis 1 to reconstruct centered data then original data

```
cbind(testc,
  round(outer(sitec, sppc, '*'), 2))
```

```
##      [,1] [,2] [,3] [,4]  [,5]  [,6]  [,7]  [,8]
## [1,]  5.4  2.8   -2 -5.6  5.18  2.51 -2.37 -5.77
## [2,]  2.4  0.8   -1 -1.6  1.87  0.91 -0.86 -2.08
## [3,] -0.6 -0.2    0 -0.6 -0.01 -0.01  0.01  0.01
## [4,] -2.6 -1.2    0  1.4 -1.77 -0.86  0.81  1.97
## [5,] -4.6 -2.2    3  6.4 -5.28 -2.56  2.42  5.87
```

```
cbind(test,
  round(t(t(outer(sitec, sppc, '*')) + sppmeans), 2) )
```

```
##      [,1] [,2] [,3] [,4]  [,5]  [,6]  [,7]  [,8]
## [1,]   10    5    0    0  9.78  4.71 -0.37 -0.17
## [2,]    7    3    1    4  6.47  3.11  1.14  3.52
## [3,]    4    2    2    5  4.59  2.19  2.01  5.61
## [4,]    2    1    2    7  2.83  1.34  2.81  7.57
## [5,]    0    0    5   12 -0.68 -0.36  4.42 11.47
```

pca on correlation matrix

```
test.pca2 <- princomp(test, cor = T)
test.pca2
```

```
## Call:
## princomp(x = test, cor = T)
##
## Standard deviations:
##     Comp.1     Comp.2     Comp.3     Comp.4
## 1.96211876 0.36340161 0.12675817 0.04429017
##
##  4  variables and  5 observations.
```

```
# spp scores
test.pca2$loadings[,1]
```

```
## [1]  0.5011650  0.5016482 -0.4913267 -0.5057477
```

```
# site scores
test.pca2$scores[,1]
```

```
## [1]  2.88564014  1.07113342 -0.06566798 -0.89659124 -2.99451434
```

iterated regression on centered and scaled matrix

```
sppsds <- apply(test, 2, sd)

testcs <- t(t(testc)/sppsds )
round(testcs, 3)
```

```
##        [,1]   [,2]   [,3]   [,4]
## [1,]  1.359  1.456 -1.069 -1.275
## [2,]  0.604  0.416 -0.535 -0.364
## [3,] -0.151 -0.104  0.000 -0.137
## [4,] -0.654 -0.624  0.000  0.319
## [5,] -1.157 -1.144  1.604  1.457
```

```
# initial values for site scores
sitecs <- rnorm(5)

for (k in 1:5) {
sppcs <- rep(NA, 4)
for (i in 1:4) {
  sppcs[i] <- coef(lm(testcs[,i] ~ -1+sitecs))[1]
  }
```

```
sppcs <- sppcs / sqrt(sum(sppcs^2))
print(sppcs)

for (i in 1:5) {
  sitecs[i] <- coef(lm(testcs[i,] ~  -1+sppcs))[1]
  }
print(sitecs)
cat('\n')
}
```

```
## [1] -0.3094360 -0.3211903  0.6769931  0.5854626
## [1] -2.3579427080 -0.8955102869  0.0001442058  0.5893499264  2.6639588627
##
## [1] -0.4956123 -0.4960866  0.4996566  0.5085369
## [1] -2.57781623 -0.95785349  0.05693787  0.79572280  2.68300905
##
## [1] -0.5009784 -0.5014575  0.4916158  0.5058408
## [1] -2.58089157 -0.95804729  0.05867451  0.80172419  2.67854016
##
## [1] -0.5011586 -0.5016416  0.4913366  0.5057509
## [1] -2.58099146 -0.95805074  0.05873315  0.80192833  2.67838072
##
## [1] -0.5011648 -0.5016480  0.4913271  0.5057478
## [1] -2.58099488 -0.95805086  0.05873515  0.80193533  2.67837525
```

```
sd(sitecs)
```

```
## [1] 1.962119
```

```
sppcs
```

```
## [1] -0.5011648 -0.5016480  0.4913271  0.5057478
```

```
sitecs
```

```
## [1] -2.58099488 -0.95805086  0.05873515  0.80193533  2.67837525
```

Use scores on axis 1 to reconstruct original data

```
cbind(round(testcs, 2), round(outer(sitecs, sppcs, '*'), 2)  )
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
## [1,]  1.36  1.46 -1.07 -1.27  1.29  1.29 -1.27 -1.31
## [2,]  0.60  0.42 -0.53 -0.36  0.48  0.48 -0.47 -0.48
## [3,] -0.15 -0.10  0.00 -0.14 -0.03 -0.03  0.03  0.03
## [4,] -0.65 -0.62  0.00  0.32 -0.40 -0.40  0.39  0.41
## [5,] -1.16 -1.14  1.60  1.46 -1.34 -1.34  1.32  1.35
```

```
cbind(test, round(t(t(outer(sitecs, sppcs, '*'))*sppsds + sppmeans), 2)  )
```

```
##      [,1] [,2] [,3] [,4]  [,5]  [,6]  [,7]  [,8]
## [1,]   10    5    0    0  9.74  4.69 -0.37 -0.13
## [2,]    7    3    1    4  6.51  3.12  1.12  3.47
## [3,]    4    2    2    5  4.48  2.14  2.05  5.73
## [4,]    2    1    2    7  3.00  1.43  2.74  7.38
## [5,]    0    0    5   12 -0.74 -0.38  4.46 11.55
```