Stat501_Exam2

Kelby Kies

4/17/2021

Question 1:

Part 1. Our objective is to reduce the dimensions of the data to summarize the variation in the dataset.

Part 1A.): Is it appropriate to scale the dataset before performing aprincipal components analysis? [5points]

```
# Read in the places.txt file
places <- read.table('~/Desktop/stat_501/places.txt', header = T)
# Put data in log10 scale
places <- data.frame(log10(places[,1:9]), rank = places[,10])

# Is appropriate to scale the dataset before PCA?
apply(X = places[, -10], MARGIN = 2, FUN = sd)
```

```
##         Climate.and.Terrain                     Housing
##                  0.11354448                  0.10543311
## Health.Care.and.Environment                       Crime
##                  0.32051192                  0.16914698
##              Transportation                   Education
##                  0.15757179                  0.05019937
##                    The.Arts                  Recreation
##                  0.54513590                  0.18790377
##                     Economics
##                  0.08447803
```

```
# Yes!
```

We should scale the dataset to do a principal component analysis because the largest sd for 'The Arts' = 0.54513590 and it is about 10 times larger than the lowest sd for 'Education' = 0.05019937. When there is such a difference in sd between variables, it is appropriate to scale.

Part 1B.): Perform a principal components analysis using the correlation matrix. What is the minimum number of components needed to display at least 80% of the variation in the dataset? [10 points]

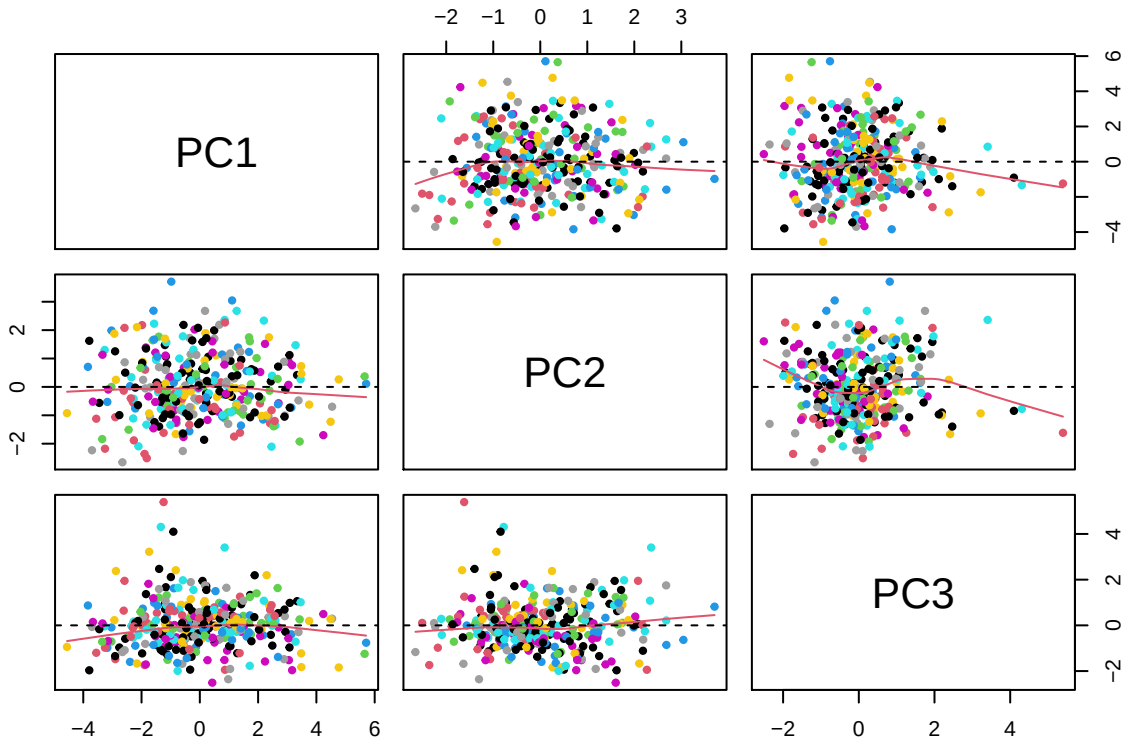
```r
# Perform PCA
place_pca <- prcomp(places[,1:9], scale = T)
summary(place_pca)$importance
```

```
##                             PC1       PC2       PC3       PC4       PC5       PC6
## Standard deviation     1.815983 1.101618 1.051442 0.9525124 0.9277008 0.7497905
## Proportion of Variance 0.366420 0.134840 0.122840 0.1008100 0.0956300 0.0624700
## Cumulative Proportion  0.366420 0.501260 0.624100 0.7249100 0.8205300 0.8830000
##                             PC7       PC8       PC9
## Standard deviation     0.6955721 0.5639789 0.5011269
## Proportion of Variance 0.0537600 0.0353400 0.0279000
## Cumulative Proportion  0.9367600 0.9721000 1.0000000
```

The number of principal components that are used to display at least 80% of the total variation in the dataset is 5. With 5 principal components at least ~82% of the total variation is accounted for.

Part 1C.): Discuss the composition of the first three principal components.[10 points]

```r
pairs(place_pca$x[,1:3],
      panel=function(x,y){panel.smooth(x,y, col = as.numeric(as.factor(colnames(places[1:9]))),
        pch = 20, cex = 1)
        abline(lsfit(x,y),lty=2) })
```



```
#legend("bottomright", fill = as.factor(colnames(places[1:9])), legend = c( levels(as.factor(colnames(p
```

I wasn't able to figure out how to add a legend to my plot, but here the points are colored by the different 9 variables. Here we can se that there isn't any real separation between the 9 variables as we might have expected. The PC1/PC2 plot seems to have more variance than the other plots which makes sense since we would expect the first PCs to contain the majority of the variance. The first component accounts for 36.6% of the total variation. The second principal component accounts for ~13.5% of the total variance and the third principal component accounts fro ~12.3% of the total variance.

Part 2: We will now do a factor analysis of the dataset.

Part 2A.) Perform a factor analysis of the dataset, using BIC to estimate the optimal number of factors needed to display the data. Use the quartimax rotation. [10 points]

```
library(fad)
```

```
## Loading required package: RSpectra
```

```
## Please cite the paper: Dai, F., Dutta, S., and, Maitra, R. (2020). A Matrix-Free Likelihood Method fo
```

```
places_fad <- fad(x=as.matrix(places[,-10]), factors=1, scores = "regression",rotation = "quartimax", me
bic <- NULL
for (q in 0:5)
  {
  res <- fad(x=as.matrix(places[,-10]), factors=q, scores = "regression",rotation = "quartimax", method
  bic <- c(bic, res$BIC)
}
```

```
## Warning in fad(x = as.matrix(places[, -10]), factors = q, scores =
## "regression", : Algorithm may not have converged. Try another starting value.
```

```
# What is the lowest BIC value? THis corresponds to the optimal # of factos needed to display the data.
which.min(bic)-1
```

```
## [1] 2
```

```
min(bic)
```

```
## [1] -8805.389
```

```
# 2 factors!
```

I ran a factor analysis for q values from 0 to 5 on the places data. The minimum BIC value, -8805.389, value was generated during the factor analysis with 2 factors and also using the quartimax rotation. This indicates that 2 factors are optimum to display the data.

Part 2B.) Interpret the first two factors. [5 points]

```
places_fda <- fad(x=as.matrix(places[,-10]), factors=2, scores = "regression",rotation = "quartimax", m
places_fda$loadings
```

```
##
## Loadings:
##       [,1]   [,2]
##  [1,]  0.192  0.105
##  [2,]  0.540  0.323
##  [3,]  0.884 -0.191
##  [4,]  0.282  0.339
##  [5,]  0.549  0.248
##  [6,]  0.491 -0.154
##  [7,]  0.823  0.258
##  [8,]  0.418  0.605
##  [9,]  0.123  0.262
##
##                  [,1]  [,2]
## SS loadings    2.599 0.854
## Proportion Var 0.289 0.095
## Cumulative Var 0.289 0.384
```

The first factor is a weighted mean of the variables Climate/Terrain, Housing, Health Care/Environment, Crime, Transportation, Education, Arts, Recreation and Economics. The second factor shows a contrast between the (Health Care/Environment & Education) variables and the (Climate/Terrain, Houseing, Crime, Transportation, Arts, Recreation and Economics) variables.

Loadings close to -1 or 1 indicate that the factor strongly influences the variable. Loadings close to 0 indicate that the factor has a weak influence on the variable. We can see the first factor has the most influence on Health Care/Environment and Arts, while the second factor has the strongest influence on Recreation.

The 2 factors explain about 38.4% of the total variance.

Part 2C.) What is the difference between factor analysis and principal components analsis? [5 points]

PCA is a linear combination of variables; Factor Analysis is a measurement model of a latent variable.

The two analyses are very similar, but differ slightly. A PCA is used to extra linear composites of observed variables. A Factor analysis explicitly assumes existence of a latent variable that underlies the observed data. The var-Cov matrix for the two analyses also differs. For PCA:

$$\Sigma_{pxp} \cong \Gamma_{pxq}\Lambda\Gamma'_{qxp}$$

But for FDA:

$$\Sigma_{pxp} = \Lambda_{pxq}\Lambda'_{qxp} + \Delta$$

Where

$$\Delta$$

is a diagonal matrix.

For FDA the Lambda matrix does not have to be orthogonal and Sigma is equal to this value rather than just an approximation.

Question 2

Part 1:Perform a multivariate mutiple regression analysis to understand which of the four pulp characteristics have linear relationships with the properties of the paper data. [10 points]

```
library(car)
```

```
## Loading required package: carData
```

```
# Read in the pulp_paper data
pulp_paper <- read.table('~/Desktop/stat_501/pulp_paper.dat', header=T)
pulp_paper_model <- lm(cbind(y1, y2, y3, y4) ~ z1 + z2 + z3 + z4, data = pulp_paper)
pulp_paper_model
```

```
##
## Call:
## lm(formula = cbind(y1, y2, y3, y4) ~ z1 + z2 + z3 + z4, data = pulp_paper)
##
## Coefficients:
##              y1          y2          y3          y4
## (Intercept)  -74.231673  -24.014741  -45.763252  -17.727292
## z1            -3.120322   -1.184892   -1.485668   -0.549977
## z2             0.097583    0.009134    0.047027    0.029166
## z3             0.049400    0.008353    0.025301    0.010951
## z4            85.076147   28.754768   45.798211   16.219950
```

```
pulp_manova <- Manova(pulp_paper_model)
summary(pulp_manova)
```

```
##
## Type II MANOVA Tests:
##
## Sum of squares and products for error:
##           y1        y2       y3        y4
## y1 127.90661 22.684786 52.09293 29.122449
## y2  22.68479  6.717486 10.99055  5.091294
## y3  52.09293 10.990554 23.87522 11.953623
## y4  29.12245  5.091294 11.95362  6.935609
##
## ----------------------------------------
##
## Term: z1
##
## Sum of squares and products for the hypothesis:
##          y1        y2        y3        y4
## y1 5.959510 2.2630275 2.8374799 1.0504020
## y2 2.263027 0.8593481 1.0774871 0.3988732
## y3 2.837480 1.0774871 1.3509991 0.5001241
## y4 1.050402 0.3988732 0.5001241 0.1851401
```

```
##
## Multivariate Tests: z1
##                  Df test stat approx F num Df den Df   Pr(>F)
## Pillai            1 0.1565833 2.506322      4     54 0.052657 .
## Wilks             1 0.8434167 2.506322      4     54 0.052657 .
## Hotelling-Lawley  1 0.1856535 2.506322      4     54 0.052657 .
## Roy               1 0.1856535 2.506322      4     54 0.052657 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ----------------------------------------
##
## Term: z2
##
## Sum of squares and products for the hypothesis:
##           y1        y2         y3        y4
## y1 20.826330 1.9494868 10.0365395 6.2245606
## y2  1.949487 0.1824853  0.9394887 0.5826614
## y3 10.036540 0.9394887  4.8367680 2.9997148
## y4  6.224561 0.5826614  2.9997148 1.8603928
##
## Multivariate Tests: z2
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            1 0.4041974 9.158512      4     54 1.0088e-05 ***
## Wilks             1 0.5958026 9.158512      4     54 1.0088e-05 ***
## Hotelling-Lawley  1 0.6784083 9.158512      4     54 1.0088e-05 ***
## Roy               1 0.6784083 9.158512      4     54 1.0088e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ----------------------------------------
##
## Term: z3
##
## Sum of squares and products for the hypothesis:
##           y1        y2       y3        y4
## y1 15.984713 2.7028275 8.186882 3.5435879
## y2  2.702828 0.4570165 1.384306 0.5991792
## y3  8.186882 1.3843058 4.193071 1.8149175
## y4  3.543588 0.5991792 1.814918 0.7855640
##
## Multivariate Tests: z3
##                  Df test stat approx F num Df den Df   Pr(>F)
## Pillai            1 0.2114291 3.619577      4     54 0.010998 *
## Wilks             1 0.7885709 3.619577      4     54 0.010998 *
## Hotelling-Lawley  1 0.2681168 3.619577      4     54 0.010998 *
## Roy               1 0.2681168 3.619577      4     54 0.010998 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ----------------------------------------
##
## Term: z4
##
```

```
## Sum of squares and products for the hypothesis:
##           y1        y2       y3        y4
## y1 103.16002 34.866913 55.53313 19.667680
## y2  34.86691 11.784619 18.76956  6.647452
## y3  55.53313 18.769565 29.89461 10.587510
## y4  19.66768  6.647452 10.58751  3.749685
##
## Multivariate Tests: z4
##                   Df test stat approx F num Df den Df    Pr(>F)
## Pillai             1 0.6952312 30.79587      4     54 2.308e-13 ***
## Wilks              1 0.3047688 30.79587      4     54 2.308e-13 ***
## Hotelling-Lawley   1 2.2811753 30.79587      4     54 2.308e-13 ***
## Roy                1 2.2811753 30.79587      4     54 2.308e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part 2:Estimate the matrix of coefficients. [10 points]

```
new_model <- lm(cbind(y1, y2, y3, y4) ~ z2 + z3 + z4, data = pulp_paper)
new_model$coefficients
```

```
##                         y1            y2           y3           y4
## (Intercept) -70.11680766 -22.452187336 -43.80405567 -17.00202079
## z2            0.05930052  -0.005402808   0.02879950   0.02241805
## z3            0.05552102   0.010677278   0.02821556   0.01203017
## z4           82.53024673  27.788003873  44.58604066  15.77121827
```

```
#pulp_paper_model$coefficients
```

Part 3: Provide an estimate of the variance-covariance matrix of the error in the four response variables after accounting for the linear effect of the four fiber characteristics. [10 points]

```
cov(new_model$residuals)
```

```
##           y1         y2        y3         y4
## y1 2.1945266 0.40898054 0.9004985 0.49463690
## y2 0.4089805 0.12421039 0.1978367 0.09000275
## y3 0.9004985 0.19783674 0.4135446 0.20415979
## y4 0.4946369 0.09000275 0.2041598 0.11673359
```

```
#cov(pulp_paper_model$residuals)
```

Part 4: Provide a detailed analysis and interpretation of your results.[10 points]

I created a multiple regression model that had paper characteristic (y1 - y4) variables as the response variables and pulp characteristics (z1 - z4) as the predictor variables. I ran a Manova to see what pulp

characteristics had a significant effect on the response variables. What I found is that long fiber fraction (z2), fine fiber fraction (z3) and zero span tensile (z4) had significant effect on the paper characteristics while arithmetic fiber length did not. Long fiber fraction (z2) and zero span tensile(z4) had the most significant effects on the paper characterisitics at the 0.0001 significance level with p-values of 1.009e-05 and 2.308e-13, respectively. The fine fiber fraction (z3) had significant effect on the paper characteristics at the 0.05 with a p-value of 0.01100.

Because Arithmetic fiber length did not have a significant effect, it was removed from the model.

For Part 2.) I have provided the matrix of coefficient values after removing the z1 variable from the model.It looks like Zero span tensile has the largest effect on the paper characteristics because if we held everything else constant and increased z4, then y1 increases by 82.53, y2 increases by 27.78, y3 increases by 44.58 and y4 will increase by 15.77. In order words zero span tensile has the largest coefficient for all of the variables.

For Part 3.) I have shown the variance-covariance matrix of the error (or residuals) for the 4 response variables after removing the arithmetic fiber length from the model. We can see that the errors associated with y3 and y1 have the strongest correlation of 0.9004985 while the errors associated with y4 and y2 have the weakest correlation, 0.09000275. Errors associated with y1 also have a very large varince of 2.1945266 .

Part 5: Provide a canonical correlation analysis of the properties of paper (y1, y2, y3, y4) and the pulp fiber characteristics (z1, z2, z3, z4). Discuss the results. [15 points]

```
paper <- pulp_paper[,1:4]
pulp <- pulp_paper[,5:8]
cancor(paper, pulp)
```

```
## $cor
## [1] 0.91732930 0.81692694 0.26538537 0.09168402
##
## $xcoef
##           [,1]        [,2]       [,3]       [,4]
## y1  0.06689193 -0.15532630 -0.2533441  0.2259421
## y2  0.03787201 -0.27574614  0.6299487  0.1048487
## y3 -0.17490257  0.09417236 -0.4125382 -0.5312127
## y4 -0.12496918  0.69613608  1.3215807  0.1267527
##
## $ycoef
##            [,1]          [,2]          [,3]          [,4]
## z1  0.081735822  0.3533117157 -0.2632020342 -1.1970693046
## z2 -0.005446751  0.0086370579  0.0006647671  0.0187756778
## z3 -0.002368852  0.0000365322 -0.0121256132 -0.0001620718
## z4 -3.550862729 -6.7807308300 -3.3801849953 -0.3870054231
##
## $xcenter
##        y1        y2        y3        y4
## 21.722823  7.266194  5.637468  1.018790
##
## $ycenter
##          z1          z2          z3          z4
## -0.02175806 39.03267742 26.67769355  1.06680645
```

The canonical correlation analysis shows that there is a very strong, positive linear relationship between the paper characteristics and pulp characterisitcs because the first canonical correlation is 0.91732930, which is

close to 1. After this is removed in the first canonical correlation,the 2nd canonical correlation continues to show a strong linear relationship with a value of 0.81692694. After this is removed, their is very little further linear relationship.

Question 3: Derive the LRT

Let $\underline{X}_1, \underline{X}_2 \cdots \underline{X}_n \sim N_p(\underline{\mu}, \underline{\Sigma})$

$H_0: \Sigma = \sigma^2 I$, $\Lambda < c_\alpha$

$H_1: \Sigma \neq \sigma^2 I$?

$$\Lambda^{2/np} = \frac{p \sum_{i=1}^{p} \hat{\lambda}_i}{\prod_{i=1}^{p} \hat{\lambda}_i}$$

The loglikelihood for μ_i & Σ_i given the Observations, $\underline{X}_i$

$$\ell(\mu_i, \Sigma_i) = -\frac{n_i}{2} \log |\Sigma_i| - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{\mu}_i)' \Sigma_i^{-1} (x_{ij} - \mu_i)$$

For the numerator (restricted) $\Sigma = \sigma^2 \cdot I$ For denominator: $\Sigma = \Sigma$

- The above function is maximized w/ $\mu_i = \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ for $i = 1 \cdots p$
- To maximize for Σ first plug in for Sigma:

$$\ell(\mu_i, \sigma_i^2 \cdot I) = -\frac{n_i}{2} \log |\sigma_i^2 I| - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)' (\sigma_i^2 \cdot I)(x_{ij} - \bar{x}_i)$$

Take derivative w/ respect to σ_i^2:

$$= -\frac{n_i}{2} (\sigma_i^2 I)^{-1} - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} -(\sigma_i^2 \cdot I)^{-1} \cdot (x_{ij} - \bar{x}) \cdot (x_{ij} - \bar{x}_i)' \cdot (\sigma_i^2 \cdot$$

$$= -\frac{n_i}{2} (\sigma_i^2 \cdot I)^{-1} + \frac{1}{2} (+(\sigma_i^2 \cdot I))^{-1} \cdot (\sigma_i^2 \cdot I)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{ij} - x_i)'$$

$$0 = (\sigma_i^2 \cdot I)^{-1} \left(-\frac{n_i}{2} + \frac{1}{2} (\sigma_i^2 \cdot I)^{-1} \right) \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{ij} - x_i)'$$

$$\frac{n_i}{2} = \frac{1}{2} (\sigma_i^2 \cdot I)^{-1} \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$$

$$\sigma_i^2 = \frac{1}{n_i} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{ij} - x_i)'$$

$$\Lambda = \frac{|\hat{\Sigma}|^{-n_i/2} + \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\hat{\mu})'\hat{\Sigma}^{-1}(x_{ij}-\hat{\mu})\right\}}{(\hat{\sigma}_i^2 \cdot I)^{-n_i/2} + \exp\left\{\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\hat{\mu})'(\hat{\sigma}_i^2 \cdot I)^{-1}(x_{ij}-\hat{\mu})\right\}}$$

Alternative

Null Model

I am not entirely sure I know what to do here. The way I understood is that for our null model $\Sigma = \sigma_i \cdot I$ so we can replace all of our Σ's w/ this value, but then we don't have an MLE for σ_i^2 so I need to find that. The LRT is the ratio of the maximized likelihood functions. I am not sure if I can reduce much further.