HW1: Principal Components Analysis (PCA)

Stat 154, Fall 2019

Problem 1

In lecture we mentioned that one measure of overall dispersion can be calculated as the sum of squared distances $d^2(i,l)$ between all pairs of points ($i = 1, \dots, n$; $l = 1, \dots, n$).

$$\text{Overall Dispersion} = \sum_{i=1}^{n} \sum_{l=1}^{n} d^2(i,l)$$

where:

$$d^2(i,l) = \sum_{j=1}^{p} (x_{ij} - x_{lj})^2$$

Without loss of generality, but in order to simplify notation, let's assume that there is just one variable: $p = 1$. Under this assumption, the distance between individuals i and l simplifies to:

$$d^2(i,l) = (x_i - x_l)^2$$

Let $\bar{x}$ denote the average individual G or *centroid*. The distance of any individual i to the centroid G is thus: $d^2(i,G) = (x_i - \bar{x})^2$

Show that the overall dispersion (as formulated above), can be expressed in terms of squared distances between all individuals to the centroid G as:

$$\sum_{i=1}^{n} \sum_{l=1}^{n} d^2(i,l) = 2n \sum_{i=1}^{n} d^2(i,G)$$

In other words, show that:

$$\sum_{i=1}^{n} \sum_{l=1}^{n} (x_i - x_l)^2 = 2n \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Problem 2

Consider the eigenvalue decomposition of a symmetric matrix $\mathbf{A}$. Prove that two eigenvectors $\mathbf{v_i}$ and $\mathbf{v_j}$ associated with two distinct eigenvalues λ_i and λ_j of $\mathbf{A}$ are mutually orthogonal; that is, $\mathbf{v_i}^\mathsf{T} \mathbf{v_j} = 0$

Problem 3

Consider a mean-centered data matrix $\mathbf{X}$ with n individuals (rows) and p variables (columns). Show that the variance of the j-th variable X_j can be expressed as:

$$var(X_j) = \sum_{k=1}^{p} \lambda_k v_{kj}^2$$

where:

- λ_k is the k-th eigenvalue of $\mathbf{S}$ (covariance matrix of $\mathbf{X}$)
- v_{kj} is the j-th element of the k-th eigenvector of $\mathbf{S}$

Hint: use the eigendecomposition of $\mathbf{S}$.

Problem 4

Consider a real-valued data matrix $\mathbf{X}$ with n rows (i.e. individuals) and p columns (i.e. variables). Indicate whether each of the following statements is TRUE or FALSE. No explanation needed.

a) If q eigenvalues are zero, then the rank of the covariance matrix $\mathbf{S}$ is $p - q$.

b) Any principal component with zero variance defines an exactly constant linear relationship between the variables in $\mathbf{X}$.

c) There will always be the same number of zero eigenvalues for a correlation matrix as for the corresponding covariance matrix.

d) The eigenvectors of $\frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$ and $\mathbf{X}^\mathsf{T}\mathbf{X}$ are identical.

e) The eigenvalues of $\frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$ and $\mathbf{X}^\mathsf{T}\mathbf{X}$ are identical.

f) When q zero eigenvalues do occur for a covariance or correlation matrix, then the points (i.e. row-vectors) $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ lie in a $(p-q)$-dimensional subspace of a p-dimensional space.

Problem 5

In this problem, you will investigate the results of decathlon events using the data set `decathlon` from the R package `"FactoMineR"` (by Husson, Josse, Le, and Mazet).

The dataset contains the results of decathlon events during two athletic meetings which took place one month apart in 2004: 1) the Olympic Games in Athens which took place on 23 and 24 August, and 2) the Decastar 2004 which took place on 25 and 26 September.

For both competitions, the following information is available for each athlete: performance for each of the 10 events, total number of points (for each event, an athlete earns points based on performance; here the sum of points scored) and final ranking.

The events took place in the following order: 100 meters, long jump, shot put, high jump, 400 meters (first day) and 110 meter hurdles, discus, pole vault, javelin, 1500 meters (second day).

There are 12 quantitative variables (the results for the 10 events, the ranking of the athlete, and the total number of points earned), and one categorical variable (the competition in which the athlete took part). For the sake of simplicity, we will only take into account the variables associated to the 10 decathlon events.

```r
# starting code (assuming you have it installed)
library(FactoMineR)

data(decathlon)

# decathlon events (ignore last 3 columns)
dat <- decathlon[ ,1:10]
```

5a) Applying PCA

Use the function `PCA()`—from `"FactoMineR"`—to compute the PCA results (see code below). In order to execute `PCA()`, determine whether you should perform PCA on standardized data—see argument `scale.unit` of `PCA()`. Explain your reasoning.

```r
# should you run PCA on transformed data?
pca <- PCA(X = dat, scale.unit = ???, ncp = 10, graph = FALSE)
```

5b) Eigenvalues

Consider the table of eigenvalues (see code below):

```r
# eigenvalues
pca$eig
```

i) Graph a barchart of the eigenvalues.

ii) How do you interpret the value of each eigenvalue? Explain.

iii) What can be said about the first two eigenvalues in terms of the *Proportion of Variance Explained*? Explain.

iv) Based on the different criteria discussed in lecture, how many dimensions (i.e. PCs) would you retain/use? Explain.

5c) Interpreting PCs

Examine the matrix of loadings **V**, as well as the matrix of correlations between the variables and the PCs (see code below). Provide interpretations for the first four principal components.

```r
# eigenvectors (loadings)
pca$svd$V

# correlations between variables and PCs
pca$var$cor
```

5d) Plot of individuals

Graph a scatterplot of the cloud of individuals on the first two components. Based on your answers of part 5c), provide a description of what is going on with the following athletes: *Karpov*, *Bourguignon*, *Casarsa*, and *Lorenzo*.

HW continues on next page

Problem 6

Principal Components can be computed in several ways. One approach is using the so-called PCA-NIPALS algorithm which allows you to compute PCA in an iterative way, using least-squares regressions. This is one of the algorithms that belong to a larger family of *Non-Linear Iterative Partial Least Squares* (NIPALS) algorithms.

You will have to write code to implement the PCA-NIPALS algorithm, described below.

Assume that the data is in an $n \times p$ matrix $\mathbf{X}$, with standardized variables: mean = 0, variance 1.

```
Set $\mathbf{X_0} = \mathbf{X}$
for $h = 1, 2, \dots, k$ do
    choose an arbitrary vector $\mathbf{w_h} \neq \mathbf{0}$
    repeat
        normalize weights: $\|\mathbf{w_h}\| = 1$
        $\mathbf{z_h} = \mathbf{X_{h-1}} \mathbf{w_h} / \mathbf{w_h^\mathsf{T}} \mathbf{w_h}$
        $\mathbf{w_h} = \mathbf{X_{h-1}^\mathsf{T}} \mathbf{z_h} / \mathbf{z_h^\mathsf{T}} \mathbf{z_h}$
    until convergence of $\mathbf{w_h}$
    $\mathbf{X_h} = \mathbf{X_{h-1}} - \mathbf{z_h} \mathbf{w_h^\mathsf{T}}$
end for
```

where:

- k is the number of PCs ($k \leq$ rank of $\mathbf{X}$)
- the vectors $\mathbf{w_h}$ are the *loadings* (stored in $k \times k$ matrix $\mathbf{W}$)
- the vectors $\mathbf{z_h}$ are the *PC scores* (stored in $n \times k$ matrix $\mathbf{Z}$)

Assuming that $\mathbf{X}$ is of full column-rank (i.e. $p = k$), you should be able to confirm that: $\mathbf{X} = \mathbf{Z}\mathbf{W}^\mathsf{T}$. To match the notation used in lecture, the matrix $\mathbf{W}$ actually corresponds to the matrix of eigenvectors $\mathbf{V}$.

In order to implement—and test—your algorithm, use the following data for the matrix $\mathbf{X}$ (data based on variables from the R data set `mtcars`).

```r
# starting code
M <- as.matrix(mtcars[ ,c('mpg', 'disp', 'hp', 'wt', 'qsec')])
X <- scale(M)
```

a) Write code to implement the PCA-NIPALS algorithm. The main outputs should be the matrix of PCs $\mathbf{Z}$, and the matrix of eigenvectors $\mathbf{V}$. Use good practices for writing code: indentation, blank spaces, comments, consisting naming conventions, etc. *Note: if possible, try writing a function.*

b) With the outputs obtained in part (a), how can you obtain the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$? Explain, and obtain such values.

c) With the computed PCs, graph two scatterplots: one for the first two PCs (PC1 and PC2), and another one for the 3rd and the 4th PC (PC1 and PC3).

d) Plot a *Circle of Correlations* graph using the first tow dimensions (associated to the first 2 PCs). And provide interpretations for the first two PCs.