Homework 2. PCA. (60 Points)

Your Name

2023-09-20

Part 1. PCA vs Linear Regression (6 points).

Let's say we have two 'features': let one be x and another y. Recall that in linear regression, we are looking to get a model like:

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i$$

after the fitting, for each data point we would have:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i + r_i$$

where r_i is residual. It can be rewritten as:

$$\hat{\beta}_0 + r_i = y_i - \hat{\beta}_1 * x_i \qquad (1)$$

The first principal component z_1 calculated on (x, y) is

$$z_{i1} = \phi_{i1} y_i + \phi_{i2} x_i$$

Dividing it by ϕ_{i1}:

$$\frac{z_{i1}}{\phi_{i1}} = y_i + \frac{\phi_{i2}}{\phi_{i1}} x_i \qquad (2)$$

There is a functional resemblance between equations (1) and (2) (described linear relationship between y and x). Is the following true:

$$\hat{\beta}_0 + r_i = \frac{z_{i1}}{\phi_{i1}}$$

$$\frac{\phi_{i2}}{\phi_{i1}} = -\hat{\beta}_1$$

Answer: *(just yes or no)*

What is the difference between linear regression coefficients optimization and first PCA calculations?

Answer:

(here should be the answer. help yourself with a plot)

Part 2. PCA Exercise (27 points).

In this exercise we will study UK Smoking Data (`smoking.R`, `smoking.rda` or `smoking.csv`):

Description

Survey data on smoking habits from the UK. The data set can be used for analyzing the demographic characteristics of smokers and types of tobacco consumed.

Format

A data frame with 1691 observations on the following 12 variables.

`gender` - Gender with levels Female and Male.

`age` - Age.

`marital_status` - Marital status with levels Divorced, Married, Separated, Single and Widowed.

`highest_qualification` - Highest education level with levels A Levels, Degree, GCSE/CSE, GCSE/O Level, Higher/Sub Degree, No Qualification, ONC/BTEC and Other/Sub Degree

`nationality` - Nationality with levels British, English, Irish, Scottish, Welsh, Other, Refused and Unknown.

`ethnicity` - Ethnicity with levels Asian, Black, Chinese, Mixed, White and Refused Unknown.

`gross_income` - Gross income with levels Under 2,600, 2,600 to 5,200, 5,200 to 10,400, 10,400 to 15,600, 15,600 to 20,800, 20,800 to 28,600, 28,600 to 36,400, Above 36,400, Refused and Unknown.

`region` - Region with levels London, Midlands & East Anglia, Scotland, South East, South West, The North and Wales

`smoke` - Smoking status with levels No and Yes

`amt_weekends` - Number of cigarettes smoked per day on weekends.

`amt_weekdays` - Number of cigarettes smoked per day on weekdays.

`type` - Type of cigarettes smoked with levels Packets, Hand-Rolled, Both/Mainly Packets and Both/Mainly Hand-Rolled

Source National STEM Centre, Large Datasets from stats4schools, https://www.stem.org.uk/resources/elibrary/resource/28452/large-datasets-stats4schools.

Obtained from https://www.openintro.org/data/index.php?data=smoking

Read and Clean the Data

2.1 Read the data from smoking.R or smoking.rda (3 points) > hint: take a look at source or load functions > there is also smoking.csv file for a refference

```
# load libraries
```

```
# Load data
```

Take a look into data

```
# place holder
```

There are many fields there so for this exercise lets only concentrate on smoke, gender, age, marital_status, highest_qualification and gross_income.

Create new data.frame with only these columns.

```
# place holder
```

2.2 Omit all incomplete records.(3 points)

```
# place holder
```

2.3 For PCA feature should be numeric. Some of fields are binary (`gender` and `smoke`) and can easily be converted to numeric type (with one and zero). Other fields like `marital_status` has more than two categories, convert them to binary (e.g. is_married, is_devorced). Several features in the data set are ordinal (`gross_income` and `highest_qualification`), convert them to some king of sensible level (note that levels in factors are not in order). (3 points)

```
# place holder
```

2.4. Do PCA on all columns except smoking status. (3 points)

```
# place holder
```

2.5 Make a scree plot (3 points)

```
# place holder
```

Comment on the shape, if you need to reduce dimensions home many would you choose

`<place holder>`

2.6 Make a biplot color points by smoking field. (3 points)

```
# place holder
```

Comment on observed biplot.

`<place holder>`

Can we use first two PC to discriminate smoking?

`<place holder>`

2.7 Based on the loading vector can we name PC with some descriptive name? (3 points)

`<place holder>`

2.8 May be some of splits between categories or mapping to numerics should be revisited, if so what will you do differently? (3 points)

`<place holder>`

2.9 Follow your suggestion in 2.10 and redo PCA and biplot (3 points)

```
# place holder
```

Part 3. Freestyle. (27 points).

Get the data set from your final project (or find something suitable). The data set should have at least four variables and it shouldn't be used in class PCA examples: iris, mpg, diamonds and so on).

- Convert a columns to proper format (9 points)
- Perform PCA (3 points)
- Make a skree plot (3 points)
- Make a biplot (3 points)
- Discuss your observations (9 points)