Principal Components Analysis

Claire Le Barbenchon and Federico Ferrari

Data Expeditions

Welcome to Data Expeditions!

- Data Expeditions is a program funded by iiD that aims to introduce undergraduate students to exploratory data analysis.
- Pairs of graduate students, often from different disciplines, work with the course instructor to formulate a question that will engage the students, and a pathway through a dataset that will provide insight.

Introductions

- Claire Le Barbenchon: 2^{nd} year Ph.D. student in Public Policy and Sociology. I work on demography, particularly migrant social networks and labour markets.
- Federico Ferrari: 2^{nd} year Ph.D. student in Statistics advised by David B. Dunson. Currently working on latent factor regression with application to chemical exposures.

The World Bank Living Standards Measurement Study

- International survey that collects information about health, education, poverty, and employment
- Collected data from dozens of countries since 1980
- For more information visit: World Bank LSMS

Our data

- A subset of four countries: Bulgaria (2007), Tajikistan (2009), Tanzania (2010-2011) and Panama (2008)
- We selected countries to represent different continents with comparable and recent survey data
- A random sample of participants from each country
- For each participant we have the following information: age, gender, marital status, relationship to household head, education, a health proxy variable (hospitalization in past 12 months), water access, and household assets (10 assets from TV to car)

ae-09-household-explore

- Go to the course GitHub organization
- Clone your application exercise repo: `ae-09-household-explore-TEAMNAME`
- Knit the R Markdown document

```r
household <- read_csv("data/household-survey.csv")
```

A quick peek

```r
names(household)
```

```
##  [1] "household_id" "person_id"    "country"      "sex"
##  [5] "age"          "relhh"        "marstat"      "educ"
##  [9] "hosp12"       "wat_source"   "stove"        "refrigerat
## [13] "tv"           "bike"         "motorbike"    "computer"
## [17] "car"          "video"        "stereo"       "sew"
## [21] "merge_index"
```

Your turn: A quick peek

How many observations are there from each country?

Your turn: A quick peek

How many observations are there from each country?

```
household %>%
   count(country)
```

```
## # A tibble: 4 x 2
##   country        n
##   <chr>      <int>
## 1 Bulgaria    2500
## 2 Panama      2500
## 3 Tajikistan  2500
## 4 Tanzania    2500
```

Your turn: Stoves

What percent of households in each country have stoves?

Your turn: Stoves

What percent of households in each country have stoves?

```
household %>%
   group_by(country) %>%
   summarise(mean(stove))
```

```
## # A tibble: 4 x 2
##   country    `mean(stove)`
##   <chr>              <dbl>
## 1 Bulgaria           0.830
## 2 Panama             0.859
## 3 Tajikistan         0.716
## 4 Tanzania           0.592
```

Your turn: All assets

What percent of households in each country have each of the ten assets?

Hint: Use the `summarise_at()` function for summarizing multiple variables at once. See the help for examples for use.

Answer the following questions looking at the table of percentages you calculate:

- Which country has the highest level of asset-holdings?
- Which country has the lowest?
- Do households in these countries tend to have the same asset levels, or is there lots of variability across countries?

Your turn: All assets

```
assets <- c("stove", "refrigerator", "tv", "bike", "motorbike",
            "computer", "car", "video", "stereo", "sew")

asset_means <- household %>%
   group_by(country) %>%
   summarise_at(assets, mean)
```

Your turn: All assets

This is a bit difficult to view...

```
asset_means
```

```
## # A tibble: 4 x 11
##   country stove refrigerator    tv  bike motorbike compu
##   <chr>   <dbl>        <dbl> <dbl> <dbl>     <dbl>    <c
## 1 Bulgar~ 0.830        0.927 0.964 0.217    0.0288   0.2
## 2 Panama  0.859        0.62  0.786 0.260    0.01     0.1
## 3 Tajiki~ 0.716        0.339 0.34  0.166    0.0172   0.0
## 4 Tanzan~ 0.592        0.171 0.279 0.486    0.0632   0.0
## # ... with 2 more variables: stereo <dbl>, sew <dbl>
```

Your turn: All assets

```r
asset_means %>%
    t() %>%  # transpose
    kable()  # pretty print
```

country	Bulgaria	Panama	Tajikistan	Tanzania
stove	0.8304	0.8588	0.7160	0.5924
refrigerator	0.9268	0.6200	0.3388	0.1708
tv	0.9640	0.7856	0.3400	0.2792
bike	0.2172	0.2596	0.1656	0.4856
motorbike	0.0288	0.0100	0.0172	0.0632
computer	0.2468	0.1596	0.0148	0.0392
car	0.4120	0.2088	0.1856	0.0436
video	0.3360	0.4204	0.1628	0.2036
stereo	0.1828	0.3520	0.0864	0.7416
sew	0.2368	0.1732	0.2136	0.1400

Constructing a Poverty Index

In developing countries, income is not always a good measure of well-being.

- Sensitive to seasons
- Does not capture in-kind revenue
- Not applicable to subsistence households

Household asset ownership does **not vary seasonally** and is **not tied to payment**.The descriptive statistics of household assets showed variability in asset holdings across countries.

How can we turn this into an index to determine the poverty level of households in our dataset?

Principal Component Analysis

We have variables that display strong pairwise correlation $\rightarrow$ we can reasonably think that we can reduce dimensionality of the data without losing too much information.

Key Ideas:

- **Reduce dimesionality** of the data
- **Avoid loss** of relevant information

Principal Component Analysis

So our goal, starting from p variables $x_1, ..., x_p$, will be to find new variables $y_1, ..., y_p$, such that:

- They are linear combinations of original variables

$$y_k = \mathbf{a_k}^T \mathbf{x} = \sum_{j=1}^{p} a_{k,j} x_j$$

- They are uncorrelated

$$Cov(y_j, y_k) = 0$$

- They are arranged in order of decreasing variance

$$var(y_1) > var(y_2) > \cdots > var(y_p)$$

Principal Component Analysis

So why not the **Mean**?

- We want to allow flexibility and estimates the coefficients of the linear combination from the data
- The mean does not always maximize the variability of the data: Take x_1 and $x_2 = -x_1$, the mean is always $0 \rightarrow$ no variability at all

A brief diversion: turtles

We'll demonstrate PCA in R with data on turtles.

```
turtle <- read_csv("data/turtle.csv")
turtle
```

```
## # A tibble: 48 x 4
##    length width height sex
##     <int> <int>  <int> <chr>
##  1     98    81     38 female
##  2    103    84     38 female
##  3    103    86     42 female
##  4    105    86     42 female
##  5    109    88     44 female
##  6    123    92     50 female
##  7    123    95     46 female
##  8    133    99     51 female
##  9    133   102     51 female
## 10    133   102     51 female
## # ... with 38 more rows
```

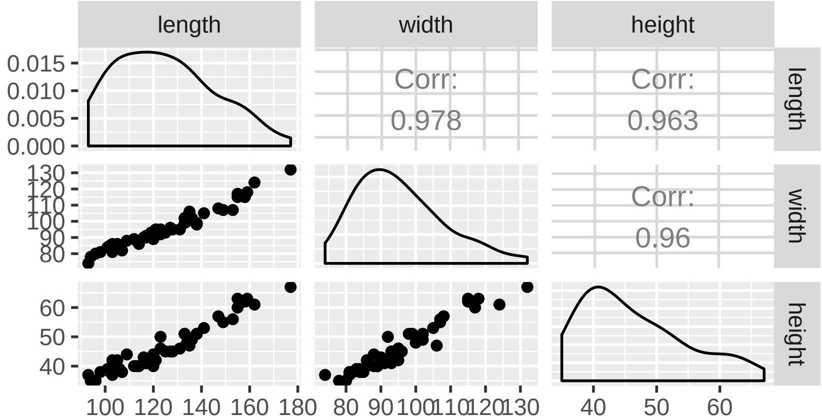
Exploring turtles

The `ggpairs()` function from the **GGally** package is useful for plotting the relationships between many variables as once.

Exploring turtles

The ggpairs() function from the **GGally** package is useful for plotting the relationships between many variables as once.

```
turtle %>%
  select(length, width, height) %>%
  ggpairs()
```



Turtles on a log: create

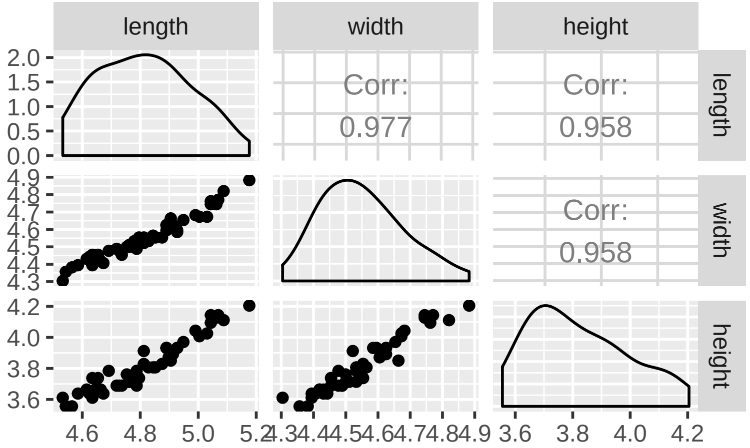
We do a variance stabilizing transformation using the **logarithm**:

```
turtle_log <- turtle %>%
   select(length, width, height) %>%
   mutate_all(log)
turtle_log
```

```
## # A tibble: 48 x 3
##    length width height
##     <dbl> <dbl>  <dbl>
##  1   4.58  4.39   3.64
##  2   4.63  4.43   3.64
##  3   4.63  4.45   3.74
##  4   4.65  4.45   3.74
##  5   4.69  4.48   3.78
##  6   4.81  4.52   3.91
##  7   4.81  4.55   3.83
##  8   4.89  4.60   3.93
##  9   4.89  4.62   3.93
```

Turtles on a log: visualize

```
ggpairs(turtle_log)
```



Correlation pattern

The `cor()` function will compute the correlations between the variables in a data frame, and return a matrix as a result:

```r
cor(turtle_log)
```

```
##           length     width    height
## length 1.0000000 0.9765071 0.9581337
## width  0.9765071 1.0000000 0.9580907
## height 0.9581337 0.9580907 1.0000000
```

Principal Components

```
turtle_pca <- prcomp(turtle_log)
turtle_pca$rotation
```

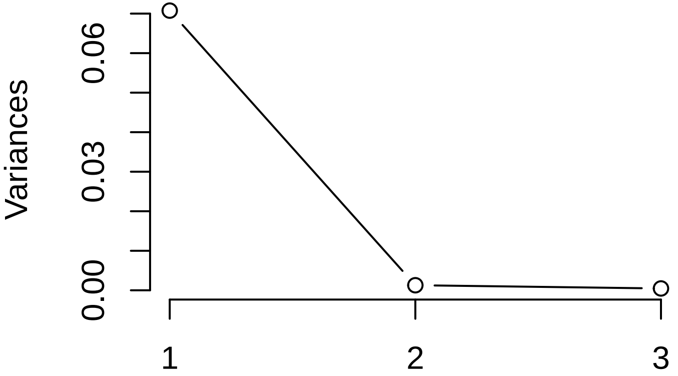
```
##              PC1        PC2         PC3
## length 0.6018462 -0.5549734 -0.57426963
## width  0.4756146 -0.3285717  0.81598495
## height 0.6415387  0.7642285 -0.06620384
```

We can intepret the first principal component as a weighted average
What is the interpretation in the original scale?

Screeplot

How many principal components ?

```
screeplot(turtle_pca, type = "lines", main = "")
```



It forms a steep curve followed by a bend and then a straight-line trend

Recap

- We explored the *World Bank Data*
- We used PCA in order to summarize the variability in the data
- PCA performs best when the correlation is high
- Using the first PC we constructed a *size* index for the turtles

Homework: Replicate the Principal Component Analysis using the *World Bank Data*

Bibliography

- The World Bank, Living Standards Measurement Study LSMS (2007). Bulgaria Multitopic Household Survey 2007 [BGR_2007_MTHS_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/2273/study-description
- The World Bank, Living Standards Measurement Study - Integrated Surveys on Agriculture (2010-2011). Tanzania - National Panel Survey 2010-2011, Wave 2 [TZA_2010_NPS-R2_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/1050
- The World Bank, Living Standards Measurement Study LSMS (2008). Panama - Encuesta de Niveles de Vida 2008 [PAN_2008_ENV_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/70
- Tajikistan Statistical Agency, Living Standards Measurement Study LSMS (2009). Tajikistan - Living Standards Survey 2009 [TJK_2009_TLSS_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/73%5Bc1%5D