Lab 7 - Principle Component Analysis (PCA)

Adam Garber

Factor Analysis ED 216B - Instructor: Karen Nylund-Gibson

February 28, 2020

Outline lab 7 - 3 examples of Principle Component Analysis:

1. 10th grade student demographics & school safety (ELS, 2002 unrestricted data)
2. California pollution burden example (from: OEHHA)
3. Combine pollution data from above with California county demographic data (from: CA census 2010)

---

install new packages for this weeks lab

```
# new packages
install.packages(c("FactoMineR",
                   "factoextra",
                   "skimr",
                   "naniar",
                   "ggfortify"))
```

load packages

```
library(FactoMineR)
library(factoextra)
library(skimr)
library(naniar)
library(ggfortify)
library(janitor)
library(tidyverse)
library(here)
```

read in ELS-2002 lab data:

```
lab_data <- read_csv(here("data", "lab_efa_cfa_data.csv"))
```

make all column names "lower_snake_case" style

```
lab_tidy <- lab_data %>%
  clean_names()
```

Prepare data for PCA

```
# remove variables that don't make sense in a PCA
lab_sub1 <- lab_tidy %>%
  select(-stu_id,    # these are random numbers
         -sch_id,
         -byrace,    # nominal (non-ordered variable)
         -byparace,  # nominal (non-ordered variable)
         -byparlng,  # nominal (non-ordered variable)
         -byfcomp,  # nominal (non-ordered variable)
         -bypared, -bymothed, -byfathed,
         -bysctrl, -byurban, -byregion)

# select columns and rename variables to have descriptive names
lab_sub2 <- lab_sub1 %>%
  select(1:9,
         bys20a, bys20h, bys20j, bys20k, bys20m, bys20n,
         bys21b, bys21d, bys22a, bys22b, bys22c, bys22d,
         bys22e, bys22g, bys22h, bys24a, bys24b) %>%
  rename("stu_exp" = "bystexp",
         "par_asp" = "byparasp",
         "mth_read" = "bytxcstd",
         "mth_test" = "bytxmstd",
         "rd_test" = "bytxrstd",
         "freelnch" = "by10flp",
         "stu_tch" = "bys20a",
         "putdownt" = "bys20h",
         "safe" = "bys20j",
         "disrupt" = "bys20k",
         "gangs" = "bys20m",
         "rac_fght" = "bys20n",
         "fair" = "bys21b",
         "strict" = "bys21d",
         "stolen" = "bys22a",
         "drugs" = "bys22b",
         "t_hurt" = "bys22c",
         "p_fight" = "bys22d",
         "hit" = "bys22e",
         "damaged" = "bys22g",
         "bullied" = "bys22h",
         "late" = "bys24a",
         "skipped" = "bys24b")

# write a CSV datafile of the new subset with renamed columns (will use for lab 8)
write_csv(lab_sub2, here("data", "lab7-8_els2002_data_subset.csv"))
```

Investigate missingness {naniar} & data summary with {skimr}

```
# Plot number of missings by variable
gg_miss_var(lab_sub2)

# Look at summary of data using skimr::skim()
skim(lab_sub2)

pca1 <- lab_sub2 %>%
  drop_na()
```

run PCA with prcomp() (function does not permit NA values)

```
pca_out1 <- prcomp(pca1, scale = TRUE)

plot(pca_out1)

#summary(pca_out1)
```

plot PCA biplot

```
jpeg(here("figures", "biplot_pca1.jpg"), res = 100) # to save the biplot

my_biplot <- autoplot(pca_out1,
                      colour = NA,
                      loadings.label = TRUE,
                      loadings.label.size = 3,
                      loadings.label.colour = "black",
                      loadings.label.repel = TRUE) +
            theme_minimal()

my_biplot

dev.off()
```

```
my_biplot
```

alternative funtion to run & plot PCA biplot

```
PCA(pca1, scale.unit = TRUE, ncp = 20, graph = TRUE)
```

This section of lab is adapted from "Lab 2" of UCSB's ESM 206 course taught by Allison Horst

These course materials are openly available @ https://allisonhorst.github.io/

A. Get the data

Data:

- California pollution burden (California Office of Environmental Health Hazard Assessment (OEHHA)'s CalEnviroScreen database, https://oehha.ca.gov/calenviroscreen/maps-data/download-data).
- See metadata, https://oehha.ca.gov/media/downloads/calenviroscreen/fact-sheet/ces30factsheetfinal.pdf.
- California county demographics (from: CA census 2010)

Read it in:

```
ca_pb <- read_csv(here("data", "ca_pollution_burden.csv"))
ca_dem <- read_csv(here("data", "ca_census_demographics_2010.csv"))
```

B. Do some cleaning

1. For the pollution burden data:

- Clean up the column headers
- Exclude any column that is a calculated percentile (contains 'percentile', 'perc', or 'pctl')

```
ca_pb_nopct <- ca_pb %>%
  clean_names() %>%
  select(-contains("pctl")) %>%
  select(-contains("perc")) %>%
  select(-latitude, - longitude)
```

2. For the demographic data:

- Clean up column names

```
ca_dem_clean <- ca_dem %>%
  clean_names()
```

C. PCA for pollution burden indicator variables

First, starting with ca_pb_nopct:

Note: The pollution burden and population characteristic variables are aggregates (averages) of existing variables in the data frame, so we won't include those. That means we'll include columns:

- From `ozone:solid_waste`, and
- From `asthma:housing_burden`

First, just selecting those:

```
ca_pb_subset <- ca_pb_nopct %>%
  select(ozone:solid_waste, asthma:housing_burden)
```

To run pca we will use the `prcomp` function:

```
pb_pca <- prcomp(ca_pb_subset, scale = TRUE) # hmmm an error
```

Look at the NA situation:

A little aside: the `naniar` package for exploring missingness! See: https://naniar.njtierney.com/

Use `naniar::gg_miss_var()` to plot the number of missings by variable:

```
# Plot number of missings by variable
gg_miss_var(ca_pb_subset)
```

Let's say our conclusion is that there are missings, but not many (compared to the actual scope of the data). We'll only keep our complete cases (census tracts without any missings).

Use `tidyr::drop_na()` with no variables specified to keep **ONLY** complete cases across all variables:

```
ca_pb_nona <- ca_pb_subset %>%
  drop_na()

# Now check for NAs:
summary(ca_pb_nona)

# Or use `skimr::skim()`!
skim(ca_pb_nona)
```

Cool. No NAs, NOW let's try PCA again:

```
my_ca_pca <- prcomp(ca_pb_nona, scale = TRUE)

plot(my_ca_pca)
```

```
jpeg(here("figures", "biplot_pca2.jpg"), res = 100) # to save the biplot

# Hmmm let's try something else (this requires ggfortify):
my_biplot <- autoplot(my_ca_pca,
```

```
                      colour = NA,
                      loadings.label = TRUE,
                      loadings.label.size = 3,
                      loadings.label.colour = "black",
                      loadings.label.repel = TRUE) +
  theme_minimal()


my_biplot

dev.off()
```

4. Join data by census tract (inner join)

```
ca_df <- ca_dem_clean %>%
  inner_join(ca_pb_nopct, by = c("census_tract_number" = "census_tract"))
```

Look at the dataframe first & then use drop_na() to get complete cases only:

```
ca_df_nona <- ca_df %>%
  drop_na()
```

5. Make a new subset for PCA, that includes % white and elderly, and some interesting pollution burden & health indicators:

Like (you can choose a different set):

- white_percent
- elderly_65_percent
- pm2_5
- pesticides
- traffic
- asthma
- cardiovascular_disease
- poverty

Make our subset:

```
my_sub <- ca_df_nona %>%
  select(white_percent,
         elderly_65_percent,
         pm2_5,
         pesticides,
         traffic,
         asthma,
         cardiovascular_disease,
         poverty)
```

Then run PCA:

```r
my_dem_pca <- prcomp(my_sub, scale = TRUE)
```

Check it out a bit:

```r
# Proportion of variance (& cumulative variance) explained by each PC
summary(my_dem_pca)

# Rotations (linear combinations for each PC):
my_dem_pca
```

Make a sweet biplot:

```r
jpeg(here("figures", "biplot_pca3.jpg"), res = 100) # to save the biplot

my_dem_biplot <- autoplot(my_dem_pca,
                      colour = NA,
                      loadings.label = TRUE,
                      loadings.label.size = 3,
                      loadings.label.colour = "black",
                      loadings.label.repel = TRUE) +
  theme_minimal() +
  scale_y_continuous(limits = c(-0.05, 0.05))

my_dem_biplot


dev.off()
```

- What are a few main things we can take out of this?
- What are the main correlations you notice?
- Are they in line with what you would expect, or is anything surprising?

End of PCA section by Allison Horst
