LAB: Principal Component Analysis

2025-03-11

```
# We will use the following packages.
# If needed, install them : pak::pkg_install().
stopifnot(
  require("magrittr"),
  require("lobstr"),
 require("ggforce"),
  require("patchwork"),
  require("gt"),
  require("glue"),
  require("skimr"),
  require("corrr"),
  require("GGally"),
  require("broom"),
  require("tidyverse"),
  require("ggfortify"),
  require("autoplotly")
old_theme <- theme_set(theme_minimal())</pre>
```

 $\rm M1~MIDS/MFA/LOGOS$

Université Paris Cité Année 2024 Course Homepage

Moodle



Swiss fertility data

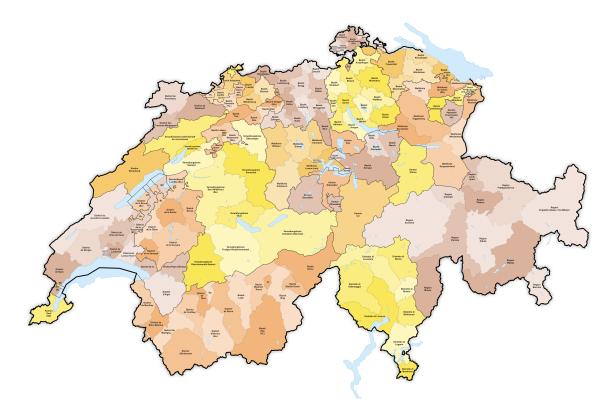
Dataset swiss from datasets::swiss connect fertility and social, economic data within 47 French-speaking districts in Switzerland.

- Fertility: fertility index
- Agriculture : jobs in agricultural sector
- Examination: literacy index (military examination)
- Education: proportion of people with successful secondary education
- Catholic: proportion of Catholics
- Infant.Mortality: mortality quotient at age 0

Fertility index (Fertility) is considered as the response variable

The social and economic variables are *covariates* (*explanatory* variables).

See European Fertility Project for more on this dataset.



PCA (Principal Component Analysis) is concerned with covariates.

```
data("swiss")

swiss %>%
  glimpse(50)
```

Have a look at the documentation of the dataset

Describe the dataset

i Question

Compute summary for each variable

g solution

It is enough to call summary() on each column of swiss. This can be done in a functional programming style using package purr. The collections of summaries can be rearranged so as to build a dataframe that is fit for reporting.

```
tt <- map_dfr(swiss, summary, .id = "var")</pre>
```

tt %>%
 knitr::kable()

var	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Fertility	35.00	64.700	70.40	70.14255	78.450	92.5
Agriculture	1.20	35.900	54.10	50.65957	67.650	89.7
Examination	3.00	12.000	16.00	16.48936	22.000	37.0
Education	1.00	6.000	8.00	10.97872	12.000	53.0
Catholic	2.15	5.195	15.14	41.14383	93.125	100.0
Infant.Mortality	10.80	18.150	20.00	19.94255	21.700	26.6

Function skim from skimr delivers all univariate summaries in proper form.

```
foo <- swiss %>%
  select(-Fertility) %>%
  skim()
```

```
foobar <- foo %>%
  filter(skim_type=="numeric") %>%
  rename(variable=skim_variable) %>%
  mutate(across(where(is.numeric), ~ round(.x, digits=1)))
```

```
foobar %>%
knitr::kable()
```

skim_typieble n_missingpletaumeteianmeeniauscheriaupfleriaupfleriaupfleriaupff

```
numeria griculture 0
                         1
                              50.7
                                    22.7
                                            1.2
                                                  35.9
                                                                67.7
                                                                        89.7
                                                         54.1
numer Examinatio 1
                         1
                              16.5
                                      8.0
                                            3.0
                                                  12.0
                                                         16.0
                                                                22.0
                                                                        37.0
numerEducation 0
                         1
                                     9.6
                                            1.0
                                                   6.0
                                                          8.0
                                                                12.0
                                                                        53.0
                              11.0
numer@atholic
                         1
                              41.1
                                    41.7
                                            2.1
                                                   5.2
                                                         15.1
                                                                93.1
                                                                       100.0
numerIc fant.Mortality
                         1
                              19.9
                                     2.9
                                          10.8
                                                  18.1
                                                         20.0
                                                                21.7
                                                                        26.6
```

Question

Display graphic summary for each variable.

We have to pick some graphical summary of the data. Boxplots and violine plots could be used if we look for concision.

We use histograms to get more details about each column.

Not that covariates have different meanings: Agriculture, Catholic, Examination, and Education are percentages with values between 0 and 100.

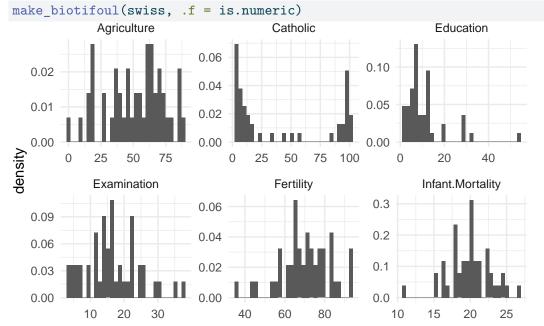
We have no details about the standardized fertility index ${\tt Fertility}$

Infant.Mortality is also a rate:

Infant mortality is the death of an infant before his or her first birthday. The infant mortality rate is the number of infant deaths for every 1,000 live births. In addition to giving us key information about maternal and infant health, the infant mortality rate is an important marker of the overall health of a society.

see Center for Desease Control

We reuse the function we have already developed during previous sessions.



Histograms reveal that our covariates have very different distributions.

Religious affiliation (Catholic) tells us that there two types of districts, which is reminiscent of the old principle *Cujus regio*, *ejus religio*, see Old Swiss Confederacy. Agriculture shows that in most districts, agriculture was still a very important activity.

Education reveals that in all but a few districts, most children did not receive secondary education. Examination shows that some districts lag behind the bulk of districts. Even less exhibit a superior performance.

The two demographic variables Fertility and Infant.Mortality look roughly unimodal with a few extreme districts.

Investigate correlations

i Question

- Compute, display and comment the sample correlation matrix
- Display jointplots for each pair of variables

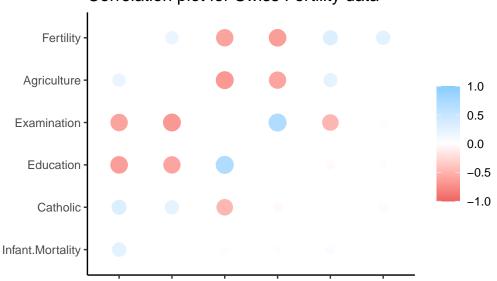
Package corrr, functions correlate and rplot provide a convenient tool. Note that rplot() creates a graphical object of class ggplot. We can endow it with more layers.

```
corrr::correlate(swiss) %>%
  corrr::rplot() %>% +
  ggtitle("Correlation plot for Swiss Fertility data")
```

Correlation computed with

- * Method: 'pearson'
- * Missing treated using: 'pairwise.complete.obs'

 Correlation plot for Swiss Fertility data



Fertility Agriculturexamination Catholic fant. Mortality

The high positive linear correlation between Education and Examination is moderately surprising. The negative correlation between the proportion of people involved in Agriculture and Education and Examinationis also not too surprising. Secondary schooling required pupils from rural areas to move to cities.

A more intriguing observation concerns the pairs Catholic and Examination (negative correlation) and Catholic and Education (little correlation).

The response variable Fertility looks negatively correlated with Examination an Education. These correlations are worth being further explored. In Demography, the decline of Fertility is often associated with the trise of women education. Note that Examination is about males, and that Education does not give details about the way women complete primary education.

Perform PCA on covariates

i Question

Pairwise analysis did not provide us with a clear and simple picture of the French-speaking districts.

Play with centering and scaling

We first call prcomp() with the default arguments for centering and scaling, that is, we center columns and do not attempt to standardize columns.

```
pco <- swiss |>
  select(-Fertility) |>
  prcomp()
```

i Question

Center the data. Compare the results with the output of scale()

solution

Hand-made centering of the dataframe

```
X <- select(swiss, -Fertility)
n <- nrow(X)

Y <- (X - matrix(1, nrow = n, ncol=1) %*% rep(1/n,n) %*% as.matrix(X))

Y <- as.matrix(Y)

tibble(var=names(X), mX=colMeans(X), mY=colMeans(Y)) %>%
  mutate(across(where(is.numeric), ~ round(.x, digits=2))) %>%
  knitr::kable()
```

var	mX	mY
Agriculture	50.66	0
Examination	16.49	0
Education	10.98	0
Catholic	41.14	0
Infant.Mortality	19.94	0

Function scale(X, scale=F) from base R does the job.

Question

Check that the ouput of svd() actually defines a Singular Value Decomposition.

```
svd_Y <- svd(Y)

svd_Y %$%
  (as.matrix(Y) - u %*% diag(d) %*% t(v)) %>%
  norm(type = "F")

[1] 2.445535e-13

norm( diag(1, ncol(Y)) -
  (svd_Y %$% (t(v) %*% v)), 'F')

[2] 9.589856e-16
```

Note that we used the exposing pipe %\$% from magrittr to unpack svd_Y which is a list with class svd and members named u, d and v.
We could have used with(,) from base R.

i Question

Relate the SVD of Y and the eigen decomposition of $Y^{\top} \times Y$

g solution

The matrix $1/nY^{\top} \times Y$ is the covariance matrix of the covariates. The spectral decomposition of the symmetric Semi Definite Positive (SDP) matrix $1/nY^{\top} \times Y$ is related with the SVD factorization of Y.

The spectral decomposition of $Y^{\top} \times Y$ can be obtained using eigen.

```
(t(eigen(t(Y) %*% Y )$vectors) %*% svd_Y$v ) %>%
 round(digits=2)
     [,1] [,2] [,3] [,4] [,5]
[1,]
        1
             0
                   0
[2,]
        0
            -1
                   0
                         0
                              0
             0
                         0
                              0
[3,]
        0
                   1
[4,]
        0
             0
                   0
                         1
                              0
[5,]
              0
                   0
                              1
```

lacktriangle Here, the eigenvectors of $Y^{\top} \times Y$ coincide with the right singular vectors of Y corresponding to non-zero singular values. Up to sign changes, it is always true when the non-zero singular values are pairwise distinct.

Question

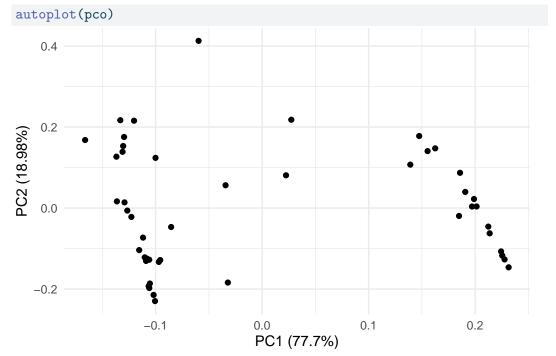
Check that prcomp is indeed a wrapper for svd().

```
solution
(Y - pco$x %*% t(pco$rotation )) %>%
  round(digits = 2) %>%
  head()
              Agriculture Examination Education Catholic Infant. Mortality
Courtelary
                        0
                                     0
                                               0
                                                         0
                        0
                                     0
                                               0
                                                         0
Delemont
                        0
                                     0
                                               0
                                                         0
Franches-Mnt
Moutier
                        0
                                     0
                                               0
                                                         0
Neuveville
                        0
                                     0
                                               0
                                                         0
                        0
                                               0
                                                         0
Porrentruy
(svd_Y$v %*% t(pco$rotation )) %>%
round(2)
     Agriculture Examination Education Catholic Infant.Mortality
[1,]
                            0
                                       0
                                                0
                                                                  0
[2,]
                0
                            1
                                       0
                                                0
                                                                  0
[3,]
                0
                            0
                                       1
                                                0
                                                                  0
[4,]
                            0
                                       0
                                                                  0
                0
                                                1
[5,]
                            0
                                       0
                0
                                                0
                                                                  1
(t(pco$x) %*% pco$x) %>%
round(2)
         PC1
                   PC2
                           PC3
                                  PC4
                                          PC5
PC1 86484.49
                  0.00
                          0.00
                                 0.00
                                         0.00
PC2
        0.00 21127.44
                          0.00
                                 0.00
                                         0.00
PC3
        0.00
                  0.00 2706.14
                                  0.00
                                         0.00
PC4
        0.00
                  0.00
                          0.00 639.22
                                         0.00
PC5
        0.00
                  0.00
                          0.00
                                 0.00 348.01
```

i Question



Objects of class pca can be handled by generic functions like plot.



The displayed plot is the so-called screeplot, it provides information about the approximation of the decomposed matrix by its truncated SVDs.

i Question

```
p_screeplot <- . %>%
    tidy(matrix="pcs") %>% {
    ggplot(.) +
    aes(x=PC, y=percent, label=format(1.-cumulative,2)) +
    geom_text(angle=45, vjust=-1, hjust=-0.1) +
    geom_col(fill=NA, colour="black")
}
```

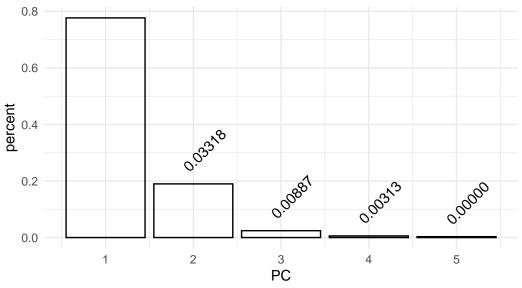
- 1 Define a pipeline for building a screeplot
- ② Mind the braces on the right side of the first pipe
- (3) 1- percent tell the reader about the relative Frobenious error achieved by keeping the first components of the SVD expansion.

```
pco %>%

p_screeplot() +
labs(
   title="Screeplot for swiss fertility data",
   subtitle="Keeping the first two components is enough to achieve relative Froebenium
```

Screeplot for swiss fertility data

Keeping the first two components is enough to achieve relative Froebenius



Question

Project the dataset on the first two principal components (perform dimension reduction) and build a scatterplot. Colour the points according to the value of original covariates.

```
solution
p <- pco %>%
                                                                                1
  augment(swiss) %>%
  ggplot() +
  aes(x=.fittedPC1, y=.fittedPC2, label=.rownames) +
  geom_point() +
  coord_fixed() +
                                                                                (2)
  ggrepel::geom_text_repel()
(p +
  aes(color=Infant.Mortality)) +
   aes(color=Education)) +
(p +
   aes(color=Examination)) +
(p +
   aes(color=Catholic)) +
(p +
   aes(color=Agriculture)) +
(p +
   aes(color=Fertility)) +
plot_layout(ncol = 2) +
plot_annotation(title="Swiss data on first two PCs" ,
                   subtitle = "centered, unscaled")
Swiss data on first two Praise
                                                       Education
centered, unscaled
                                                            50
                                     fittedPC2
                                                            40
                       20
                                                            30
                                                         Catholic
      -502502550
                    Examination
      .fittedPC1
                                           .fittedPC1
                                                            1000
fittedPC2
                                    fittedPC2
                        30
                                                             75
                        20
                                                             50
                     Agriculture
      -502502550
                                           -502502550
                                                         -erti<u>lit</u>y
                         10
      .fittedPC1
                                           .fittedPC1
                                                             90
                         80
                                    fittedPC2
fittedPC2
                                                             80
                         60
                                                             70
                         40
                                                             60
      -502502550
                                           -502502550
                                                             50
                         20
      .fittedPC1
                                           .fittedPC1
                                                             40
```

Question

We can extract factor V from the SVD factorization using generic function tidy from package ${\tt broom}$

```
pco %>%
        tidy(matrix="v") %>%
        glimpse()
Rows: 25
Columns: 3
$ column <chr> "Agriculture", "Agric
                                        <dbl> 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, ~
$ value <dbl> 0.28151505, -0.88377692, -0.36961938, -0.02652821, -0.04863543,~
The result is a tibble in long form. It is worth pivoting the dataframe
pco %>%
        tidy(matrix="v") %>%
        pivot_wider(id_cols =column,
                                                              names_from = PC,
                                                              values_from = value)
# A tibble: 5 x 6
                                                                                                        `1`
                                                                                                                                             `2`
                                                                                                                                                                                      `3`
                                                                                                                                                                                                                          `4`
                                                                                                                                                                                                                                                               `5`
         column
         <chr>>
                                                                                               <dbl>
                                                                                                                                   <dbl>
                                                                                                                                                                            <dbl>
                                                                                                                                                                                                                <dbl>
                                                                                                                                                                                                                                                    <dbl>
1 Agriculture
                                                                                         0.282 -0.884 -0.370
                                                                                                                                                                                                       -0.0265 -0.0486
```

5 Infant.Mortality 0.0105 0.0193 0.0985 -0.0867 -0.991 Think of the rows of swiss as vectors. Then matrix v In wide form, we readily access to the decomposition of the or

0.174 -0.450

-0.0584 0.311 -0.807

0.303

-0.867

0.00166 -0.0715 0.0223

0.485 -0.117

0.0332

-0.121

0.950

Question

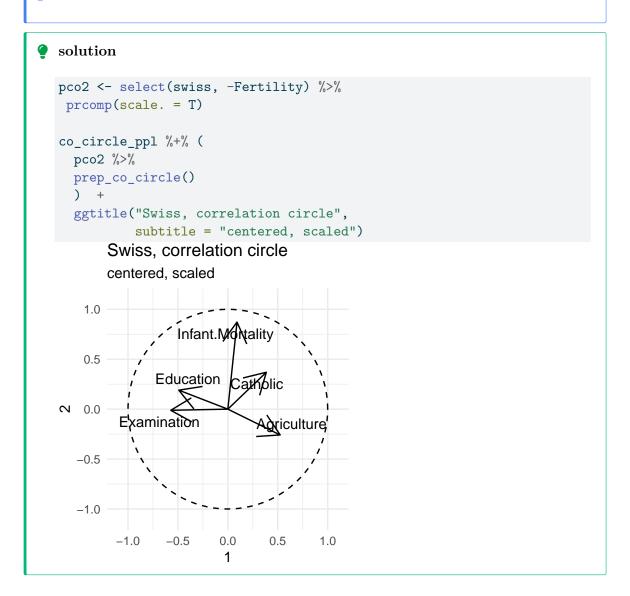
2 Examination

3 Education

4 Catholic

solution prep_co_circle <- . %>% tidy(matrix="v") %>% pivot_wider(id_cols =column, names_from = PC, values_from = value) co_circle_ppl <- (</pre> pco %>% prep_co_circle() %>% filter(F)) %>% ggplot() + aes(x=1), y=2, label=column) + geom_segment(aes(xend=0, yend=0), arrow = grid::arrow(ends = "first")) + ggrepel::geom_text_repel() + coord_fixed() + xlim(c(-1.1, 1.1)) + ylim(c(-1.1, 1.1)) +annotate ("path", x=0+1*cos(seq(0,2*pi,length.out=100)),y=0+1*sin(seq(0,2*pi,length.out=100)), linetype="dashed") co_circle_ppl %+% (pco %>% prep_co_circle()) + ggtitle("Swiss, correlation circle", subtitle = "centered, unscaled") Swiss, correlation circle centered, unscaled 1.0 0.5 Education Catholic Examination ($^{\circ}$ 0.0 Infant.Mortality -0.5-1.0Agriculture -1.0-0.50.0 0.5 1.0 # pco %\$% { ifelse(!is.null(center), "centered", "not centered"); ifelse(!is.null(scale), "scaled", "not scaled") # }

Question



Sanity checks

• X : data matrix after column centering (use scale(., center=T, scale-F))

X

```
solution
  X <- as.matrix(select(swiss, -Fertility)) |>
    scale(center = T, scale=F)
  # check centering, spot the difference in variances
  X |>
    as_tibble() |>
    summarise(across(everything(), c(var, mean)))
  # A tibble: 1 x 10
    Agriculture 1 Agriculture 2 Examination 1 Examination 2 Education 1
                                        <dbl>
                                                      <dbl>
            <dbl>
                          <dbl>
                                                                   <dbl>
             516.
                       2.64e-15
                                         63.6
                                                  -1.51e-16
                                                                    92.5
  # i 5 more variables: Education_2 <dbl>, Catholic_1 <dbl>, Catholic_2 <dbl>,
      Infant.Mortality_1 <dbl>, Infant.Mortality_2 <dbl>
  # should be 0
  norm(X %*% pco$rotation - pco$x)
  [1] 0
  # check the left singular vectors
  pco$x %*% diag((pco$sdev)^(-1)) |>
    as_tibble() |>
    summarise(across(everything(), c(mean, var)))
  Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if
  `.name_repair` is omitted as of tibble 2.0.0.
  i Using compatibility `.name_repair`.
  # A tibble: 1 x 10
        V1_1 V1_2
                        V2_1 V2_2
                                        V3_1 V3_2
                                                        V4_1 V4_2
                                                                       V5_1 V5_2
       <dbl> <dbl>
                       <dbl> <dbl>
                                       <dbl> <dbl>
                                                       <dbl> <dbl>
                                                                       <dbl> <dbl>
  1 9.45e-17
                 1 -1.36e-16
                                 1 -9.32e-17 1.00 -1.71e-17
                                                                 1 5.38e-16
  pco$rotation %*% (diag((pco$sdev)^(-2)) %*% t(pco$x) %*% X)
                     Agriculture
                                  Examination
                                                   Education
                                                                  Catholic
  Agriculture
                    4.600000e+01 6.272760e-15 4.884981e-15 -2.522288e-14
                    1.002115e-13 4.600000e+01 1.140754e-14 4.335421e-14
  Examination
  Education
                    1.257328e-13 1.104672e-14 4.600000e+01 -1.309508e-13
  Catholic
                   -1.120631e-14 -2.477185e-15 -4.218847e-15 4.600000e+01
  Infant.Mortality 1.318945e-13 -3.708145e-14 -5.773160e-14 3.819167e-14
                   Infant.Mortality
                      -5.329071e-15
  Agriculture
  Examination
                      -4.218847e-15
  Education
                      -1.065814e-14
  Catholic
                       8.881784e-16
  Infant.Mortality
                       4.600000e+01
```

Question

```
pco |>
  tidy(matrix="v") |>
  pivot_wider(id_cols =column,
              names_from = PC,
              values_from = value) |>
  rowwise() |>
  summarise(column, 12=sum((c_across(where(is.numeric)))^2))
# A tibble: 5 x 2
  column
                      12
  <chr>>
                    <dbl>
                    1.00
1 Agriculture
2 Examination
                    1
3 Education
4 Catholic
5 Infant.Mortality 1
```

i Question

Checking Orthogonality of V

solution

```
# checking that pco$rotation is an orthogonal matrix
t(pco$rotation) %*% pco$rotation
             PC1
                           PC2
                                         PC3
                                                       PC4
    1.000000e+00 -4.340061e-16 -1.825796e-16 0.000000e+00 -2.255141e-17
PC2 -4.340061e-16 1.000000e+00 3.384879e-16 -1.951564e-18 2.775558e-17
PC3 -1.825796e-16 3.384879e-16 1.000000e+00 1.214306e-17 -5.551115e-17
PC4 0.000000e+00 -1.951564e-18 1.214306e-17 1.000000e+00 4.024558e-16
PC5 -2.255141e-17 2.775558e-17 -5.551115e-17 4.024558e-16 1.000000e+00
pco$rotation %*% t(pco$rotation)
                  Agriculture
                               Examination
                                                Education
                                                               Catholic
Agriculture
                 1.000000e+00 -2.190088e-17 1.040834e-17 2.671474e-16
Examination
                -2.190088e-17 1.000000e+00 -2.129373e-16 -4.011548e-18
Education
                 1.040834e-17 -2.129373e-16 1.000000e+00 -3.165870e-17
                 2.671474e-16 -4.011548e-18 -3.165870e-17 1.000000e+00
Catholic
Infant.Mortality 2.081668e-17 3.469447e-16 -1.387779e-16 2.081668e-17
                Infant.Mortality
Agriculture
                    2.081668e-17
Examination
                    3.469447e-16
Education
                   -1.387779e-16
Catholic
                    2.081668e-17
Infant.Mortality
                    1.000000e+00
```

Compare standardized and non-standardized PCA

i Question

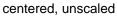
Pay attention to the correlation circles.

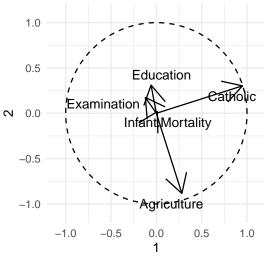
- 1. How well are variables represented?
- 2. Which variables contribute to the first axis?

```
pco_c <- swiss %>%
   select(-Fertility) %>%
   prcomp()

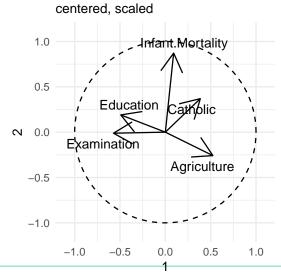
pco_cs <- swiss %>%
   select(-Fertility) %>%
   prcomp(scale.=T, center=T)
```

Swiss, correlation circle





Swiss, correlation circle



i Question

Explain the contrast between the two correlation circles. In the sequel we focus on standardized PCA.

```
solution
q <- p %+% (pco_cs %>%
  augment(swiss)) +
  ggtitle("Swiss data on first two PCs", subtitle = "centered, scaled")
(q +
  aes(color=Infant.Mortality)) +
   aes(color=Education)) +
(q +
   aes(color=Examination)) +
(q +
   aes(color=Catholic)) +
(q +
   aes(color=Agriculture)) +
(q +
   aes(color=Fertility)) +
plot_layout(ncol = 2)
Warning: ggrepel: 47 unlabeled data points (too many overlaps). Consider increasing many
ggrepel: 47 unlabeled data points (too many overlaps). Consider increasing max.overlap
ggrepel: 47 unlabeled data points (too many overlaps). Consider increasing max.overlap
ggrepel: 47 unlabeled data points (too many overlaps). Consider increasing max.overlaps
ggrepel: 47 unlabeled data points (too many overlaps). Consider increasing max.overlap
ggrepel: 47 unlabeled data points (too many overlaps). Consider increasing max.overlap
            Infant.Mortality
                                         Education
      Swiss data on first two PCswiss data on first two PCs
      centered, scaled
                                  centered, spaled
                20
                                             30
    <del>-52020</del>5
                                 -5202505
                                             20
  .fittedPC1
                                .fittedPC1
               15
camination
                                            tholic
      Swiss data on first two PCswiss data@n first two PCs
fittedPC2
      centered, scalled
                                  centered, scaled
                               = 5
                 20
                                              50
    <del>-52026</del>05
                                 -5202505
  .fittedPC1
                               .fittedPC1
                                             rti<del>lit</del>y
                rit@lture
      Swiss data on first two PCswiss data on first two PCs
      centered, scaled
                                  centered, scalled
                  60
                                              70
     5202505
                  40
                                 -5202505
                                              60
   .fittedPC1
                               .fittedPC1
                                              50
                  20
```

Investigate eigenvalues of covariance matrix

Question

How many axes should we keep?

p_screeplot %+% (pco_cs%>% tidy(matrix="pcs")) # plot(pco_cs) Elbow rule: keep the first three PCs

Provide an interpretation of the first two principal axes

i Question

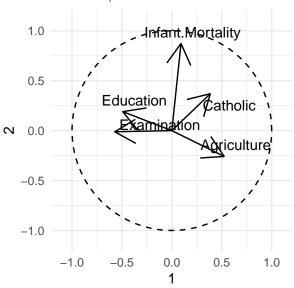
Which variables contribute to the two first principal axes?

solution

This comes from the correlation circle. We rely on function prep_co_circle and on the graphical pipeline co_circle_ppl.

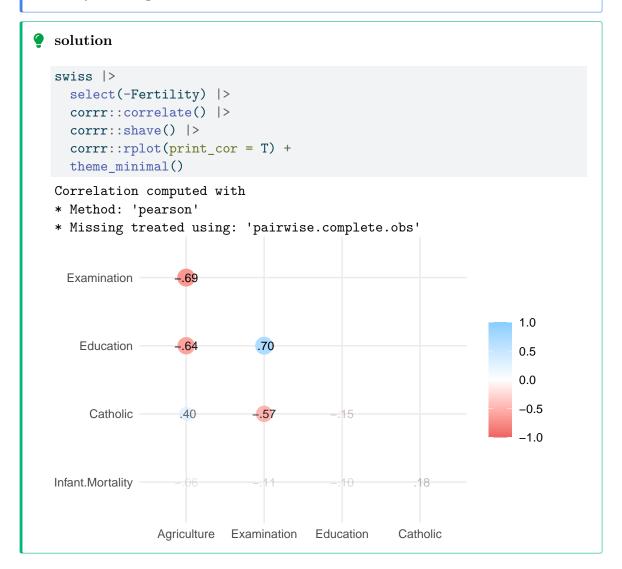
Swiss, correlation circle

centered, scaled



Question

Analyze the signs of correlations between variables and axes?



Add the Fertility variable

i Question

Plot again the correlation circle using the same principal axes as before, but add the Fertility variable. How does Fertility relate with covariates? with principal axes?

```
solution
U <- pco_cs %$% # exposition pipe
  (1/sqrt(nrow(x)-1) *x %*% diag((sdev)^(-1)))
Uprime <- with(pco_cs,</pre>
  1/sqrt(nrow(x)-1) *x %*% diag((sdev)^(-1)))
t(U) %*% U
              [,1]
                            [,2]
                                         [,3]
                                                       [,4]
                                                                     [,5]
[1,] 1.000000e+00 -1.023487e-16 3.469447e-17 -1.162915e-15 2.706169e-16
[2,] -1.023487e-16 1.000000e+00 9.367507e-17 4.781874e-16 4.805184e-16
[3,] 3.469447e-17 9.367507e-17 1.000000e+00 4.454933e-15 3.122502e-16
[4,] -1.162915e-15 4.781874e-16 4.454933e-15 1.000000e+00 -6.543431e-16
[5,] 2.706169e-16 4.805184e-16 3.122502e-16 -6.543431e-16 1.000000e+00
t(Uprime) %*% Uprime
                            [,2]
                                         [,3]
                                                       [,4]
              [,1]
                                                                     [,5]
[1,] 1.000000e+00 -1.023487e-16 3.469447e-17 -1.162915e-15 2.706169e-16
[2,] -1.023487e-16 1.000000e+00 9.367507e-17 4.781874e-16 4.805184e-16
[3,] 3.469447e-17 9.367507e-17 1.000000e+00 4.454933e-15 3.122502e-16
[4,] -1.162915e-15 4.781874e-16 4.454933e-15 1.000000e+00 -6.543431e-16
[5,] 2.706169e-16 4.805184e-16 3.122502e-16 -6.543431e-16 1.000000e+00
(norm(U,type = "F"))^2
[1] 5
```

Display individuals (districts)

```
i Question
```

Comment

Biplot

i Question

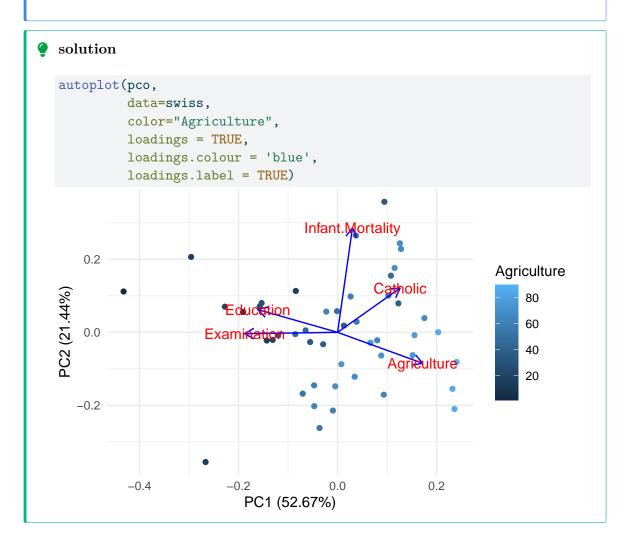
```
pco <- swiss %>%
  select(-Fertility) %>%
  prcomp(scale.=T)
df_cocirc <- pco %>%
  tidy(matrix="v") %>%
  pivot_wider(id_cols =column,
              names_from = PC,
              values_from = value)
augment(pco, data=swiss) %>%
  ggplot() +
  geom_point(aes(x=.fittedPC1,
                 y=.fittedPC2,
                 color=Fertility, label=.rownames)) +
  coord_fixed() +
  ggrepel::geom_text_repel(aes(x=.fittedPC1,
                               y=.fittedPC2,
                               color=Infant.Mortality,
                               label=.rownames)) +
  geom_segment(data=df_cocirc,
               mapping=aes(x= 4* 1,
                           y = 4 * ^2,
                           linetype=factor(column),
                           label=column,
                           xend=0,
                           yend=0),
               arrow = grid::arrow(ends = "first",
                                    unit(.1, "inches")
                                  )) +
  scale_color_viridis_c() +
  xlim(c(-5,5)) +
  ylim(c(-5,5)) #+
```

Warning in geom_point(aes(x = .fittedPC1, y = .fittedPC2, color = Fertility, : Ignoring unknown aesthetics: label
Warning in geom_segment(data = df_cocirc, mapping = aes(x = 4 * `1`, y = 4 * : Ignoring unknown aesthetics: label
Warning: ggrepel: 37 unlabeled data points (too many overlaps). Consider increasing max.overlaps

5.0 Pertility

-- 75

i Question



References

https://scholar.google.com/citations?user=xbCKOYMAAAAJ&hl=fr&oi=ao