SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 8 – Methods of dealing with multicollienarity (MPV 9.5, HTF 3.5, HTW 8.2)

February 14, 2022

Todays lecture

- Finding a linear relationships between pairs of variables
  - Variance influence factors (VIF)
  - Eigensystem analysis
- Dealing with multicollinearity
  - Collecting more data
  - Removing regressors
  - Principal component analysis (PCA)
  - Sparse principal component analysis (sparse PCA)

Assumptions

In this lecture, we assume that the vectors of observations of each regressor is centered and normalized, and that the vector of responses is also centered and normalized.
In other words, we assume that the data has been transformed so that

$$\bar{\mathbf{y}} = \overline{X_{\cdot 1}} = \ldots = \overline{X_{\cdot (k+1)}} = 0$$

and

$$\|\mathbf{y}\|_2^2 = \|X_{\cdot 1}\|_2^2 = \ldots = \|X_{\cdot (k+1)}\|_2^2 = 1.$$

$\rightarrow$ In regression models for X and $\mathbf{y}$ in *standard form*, we will have no intercept term, i.e., $\beta_0 = 0$. For this reason, this regression coefficient will be omitted, and when going back to the original scaling after a regression model has been fitted, one usually adds a term $\beta_0 = \bar{\mathbf{y}} - \sum \hat{\beta}_k \bar{\mathbf{x}}_j$.

Finding linear relationships between two variables

Motivation

Assume that $X_{\cdot i}$ and $X_{\cdot j}$ are nearly linearly dependent. Since X is on standard form, this implies that either $X_{\cdot i} \approx X_{\cdot j}$ or $X_{\cdot i} \approx -X_{\cdot j}$.
Assume that $X_{\cdot i} \approx \tau X_{\cdot j}$, where $\tau \in \{-1, 1\}$. Then

$$X^T X(i,j) = X_{\cdot i}^T X_{\cdot j} \approx \tau X_{\cdot i}^T X_{\cdot i} = \tau \underbrace{X^T X(i,i)}_{=1}.$$

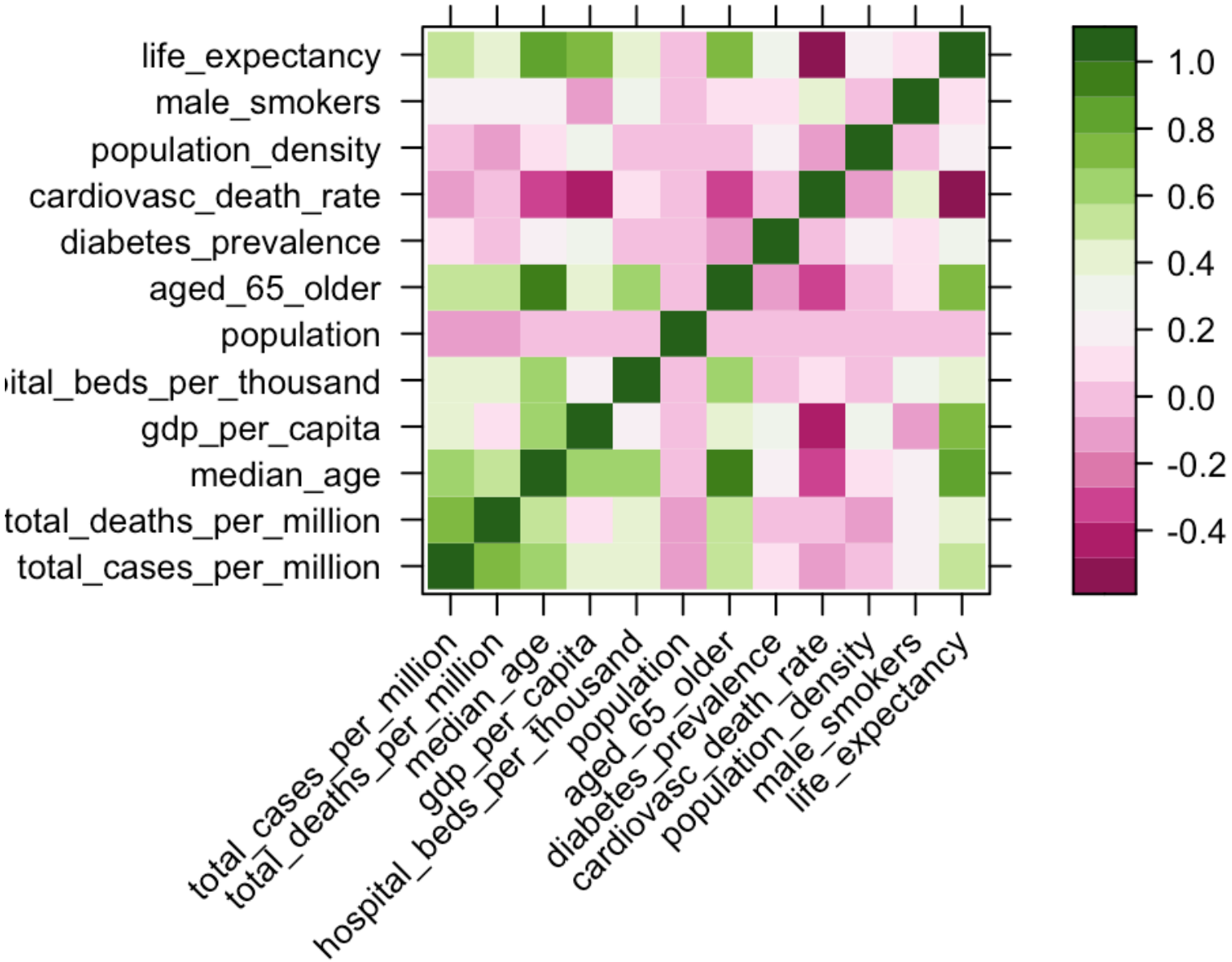
Idea

If $X_{\cdot i}$ and $X_{\cdot j}$ are almost linearly dependent, then $|X^T X(i,j)| \approx 1$.

Downside

This method cannot detect linear dependencies involving more than two regressors.

Finding linear relationships between two variables



Variance inflation factors

Variance inflation factors

$$VIF_j := \frac{\mathrm{Var}(\hat{\beta}_j)}{\sigma^2} = (X^T X)^{-1}(j,j) = (1 - R^2_{(j)})^{-1},.$$

where $R^2_{(j)} = 1 - SS^{(j)}_{Res}/SS^{(j)}_T$ is the coefficient of determination for the model obtained by removing the jth regressor from the model. Hence VIF_j is a measure of to which extent the jth regressor is linearly dependent on the other regressors.

Rule

There is multicollinearity if $VIF_j \geq 5$ for some j.

Downside

This method can detect multicollinearity, but will not tell you what the linear relationship is. However, you could proceed with studying the model $x_j = b_0 + \sum_{i \neq j} b_i x_i$ more closely.

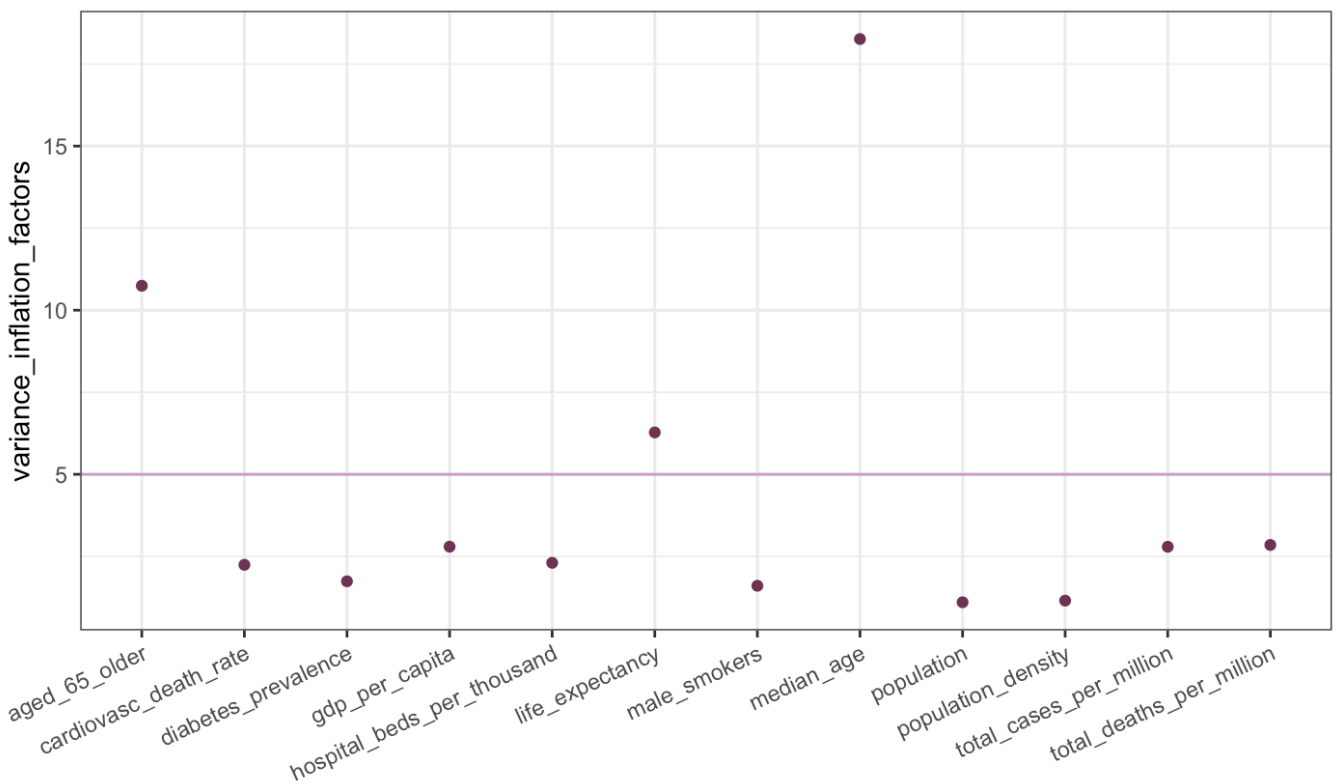
Example

```r
library("car")

df00.model1 <- lm(people_fully_vaccinated_per_hundred~ I(gdp
    _per_capita^.16) + I(diabetes_prevalence^3.62), data =
    df00)
vif(df00.model1)
```

I(gdp_per_capita^0.16)	I(diabetes_prevalence^3.62)
1.000066	1.000066

Example

```r
library("car")

df00.model2 <- lm(people_fully_vaccinated_per_hundred~ total
    _cases_per_million+total_deaths_per_million+median_age+
    gdp_per_capita+hospital_beds_per_thousand+population+
    aged_65_older+diabetes_prevalence+cardiovasc_death_rate+
    population_density+male_smokers+life_expectancy , data =
     df00)
vif(df00.model2)
```



Eigenvalue analysis of $X^T X$

Motivation

If there is multicollinearity, then $X^T X$ will be ill-conditioned and have small determinant. Since $\det X^T X = \prod \lambda_i$, it follows that at least one of the eigenvalues will be small. This motivates the following two measures of multicollinearity.

Condition number

$$\kappa := \lambda_{max}/\lambda_{min}$$

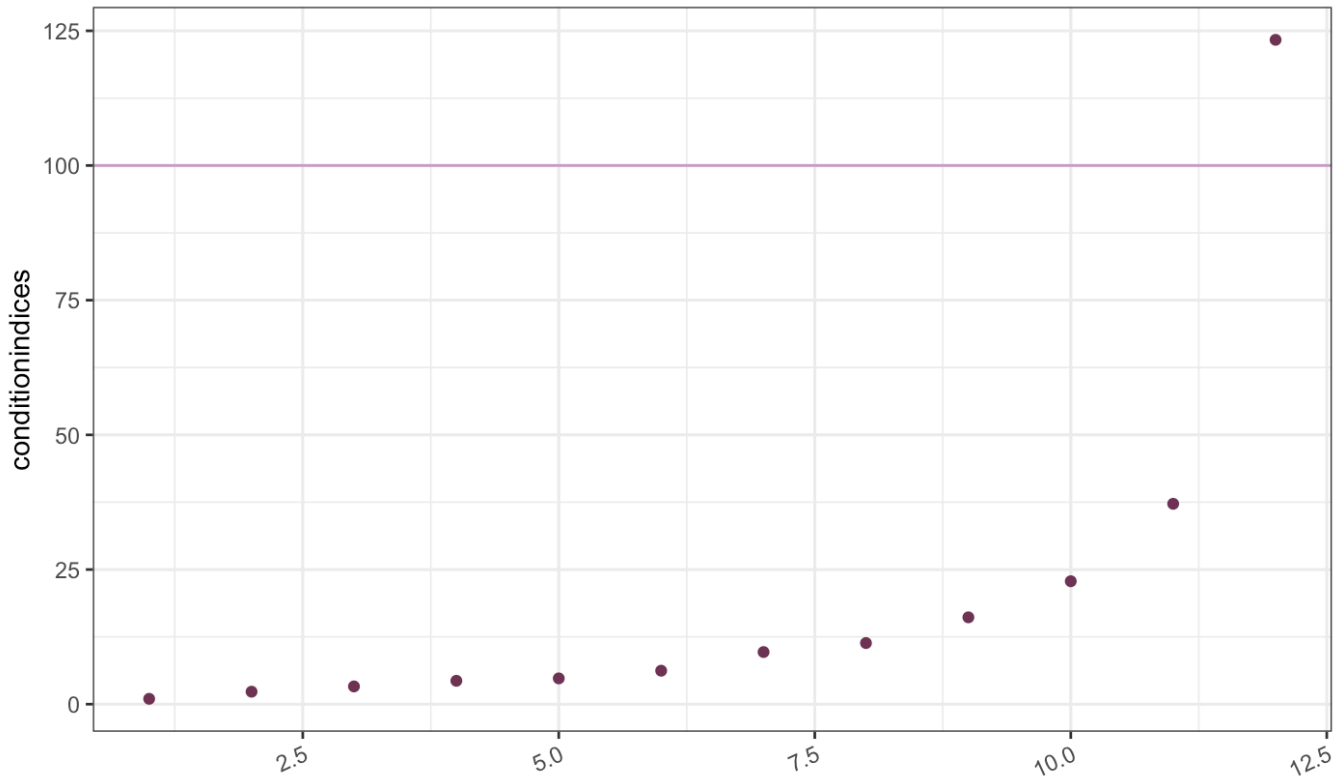
Rule: Moderate to strong multicollinearity if $100 \leq \kappa < 1000$, and severe multicollinearity if $\kappa \geq 1000$.

Condition indices

$$\kappa_j := \lambda_{max}/\lambda_j$$

Rule: Multicollinearity if $\kappa_j \geq 100$ for some j

Example



Eigensystem analysis of X^TX

Motivation

Assume that $X^TX\mathbf{v}_j = \lambda_j \mathbf{v}_j$. Then

$$\|X\mathbf{v}_j\|_2^2 = (X\mathbf{v}_j)^T X\mathbf{v}_j = \mathbf{v}_j^T \underbrace{X^TX\mathbf{v}_j}_{=\lambda_j\mathbf{v}_j} = \mathbf{v}_j^T \lambda_j \mathbf{v}_j = \lambda_j \|\mathbf{v}_j\|_2^2 = \lambda_j,$$

and hence $X\mathbf{v}_j \approx 0$ if $\lambda_j \approx 0$.

Idea

If an eigenvector λ_j of X^TX is very small, and $\mathbf{v}_j$ is the corresponding eigenvector, then the linear relationship $X\mathbf{v}_j = 0$ approximately holds between the regressors. In other words, we can use the eigensystem of X^TX to identify linear relationships between the regressors.

Downside

The linear relationships we find in this way might not be the most natural ones, and we sometimes have to consider linear combinations of these linear relationships to find simpler ones.

Example

```r
X <- data.matrix(df01[,c("total_cases_per_million", "total_
    deaths_per_million","median_age", "gdp_per_capita","
    hospital_beds_per_thousand","population","aged_65_older"
    ,"diabetes_prevalence","cardiovasc_death_rate","
    population_density", "male_smokers","life_expectancy" )
    ])

X <- scale(X)/sqrt(nrow(X)-1)
XtX <- t(X)%*%X

eigen(XtX)$values[12]
eigen(XtX)$vectors[,12]
```

```
[1] 0.0365042

[1] -0.005747818 -0.054576192  0.801328033 -0.109266477
    -0.042011810 -0.035261603 -0.537725723 -0.102005698
    -0.075914278 -0.017250500 -0.028309156 -0.183025047
```

$$-0.01x_1 + 0.05x_2 - 0.80x_3 + 0.11x_4 + 0.04x_5 + 0.04x_6 + 0.54x_7$$
$$+ 0.10x_8 + 0.08x_9 + 0.02x_{10} + 0.03x_{11} + 0.18x_{12} \approx 0$$

Collecting additional data

Assume that we have a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 and x_2, in the data collected, seem to have a near linear relationship, so that in almost all samples, $x_1 \approx x_2$.

Sources of collinearity

- Constraints on the model or population
- Model specification
- An overdefined model
- **The data collection method**

Idea

Collect more samples (x_{1j}, x_{2j}, y_j) where x_{1j} and x_{2j} are chosen to be different.

Comments

Only works if we can collect more data, and data as above exists and are not too unusual. Also, we should ensure that the new points will automatically be very influential.

Model respecifications

Assume that we have a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 and x_2, in the data collected, seem to have a near linear relationship, so that in almost all samples, $x_1 \approx x_2$.

Sources of collinearity

- Constraints on the model or population
- **Model specification**
- **An overdefined model**
- The data collection method

Redefine the variables

Define new variables, e.g., $x_1 + x_2$, $x_1 - x_2$, or $x_1 x_2$, which are chosen such that they preserve the important information in the previous variables but reduces the ill-conditioning of $X^T X$.

Eliminate variables

Try to eliminate variables which causes linear dependence, and at the same time does not have significant explanatory power.

Principal components regression (PCR)

Idea

Recall the decomposition $X^TX = PDP^T$. Since the columns of P are orthonormal eigenvectors of X^TX, we have $P^TP = I$, and hence also $PP^T = I$.

Write

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = XPP^T\boldsymbol{\beta} + \varepsilon = \underbrace{XP}_{=:Z}\underbrace{P^T\boldsymbol{\beta}}_{=:\boldsymbol{\alpha}} + \boldsymbol{\varepsilon} = Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

Then

$$Z^TZ = (XP)^T(XP) = P^T\underbrace{X^TX}_{=PDP^T}P = P^T(PDP^T)P = D,$$

and hence the columns of Z are orthogonal. Consequently, $\mathbf{y} = Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ is a regression model with orthogonal regressors.

Principal components

The columns in P, i.e., the orthogonal eigenvectors of X^TX, are referred to as the *principal components*.

Principal components analysis (PCA)

PCA refers to analyzing the principal components to detect patterns in the data.

Principal components regression (PCR)

Idea

$$X^T X = PDP^T, \qquad Z = XP, \qquad \boldsymbol{\alpha} = P^T \boldsymbol{\beta}, \qquad Z^T Z = D$$

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

Observations

- $\hat{\alpha} = (Z^T Z)^{-1} Z^T \mathbf{y} = D^{-1} Z^T \mathbf{y}$
- $\mathrm{Var}(\hat{\alpha}) = \sigma^2 (Z^T Z)^{-1} = \sigma^2 D^{-1}$
- $\|Z_{\cdot j}\|_2^2 = (Z\mathbf{e}_j)^T (Z\mathbf{e}_j) = \mathbf{e}_j^T Z^T Z \mathbf{e}_j = \mathbf{e}_j^T D \mathbf{e}_j = D(j,j) = \lambda_j$

$\rightarrow$ If λ_j is small, then the variance of $\hat{\alpha}_j$ is large, and the jth principal component corresponds to a linear combination of the original regressors that is almost equal to zero, i.e. to a multicollinearity in the original model.

Idea

Remove the principal components with "small" eigenvalues from the model.

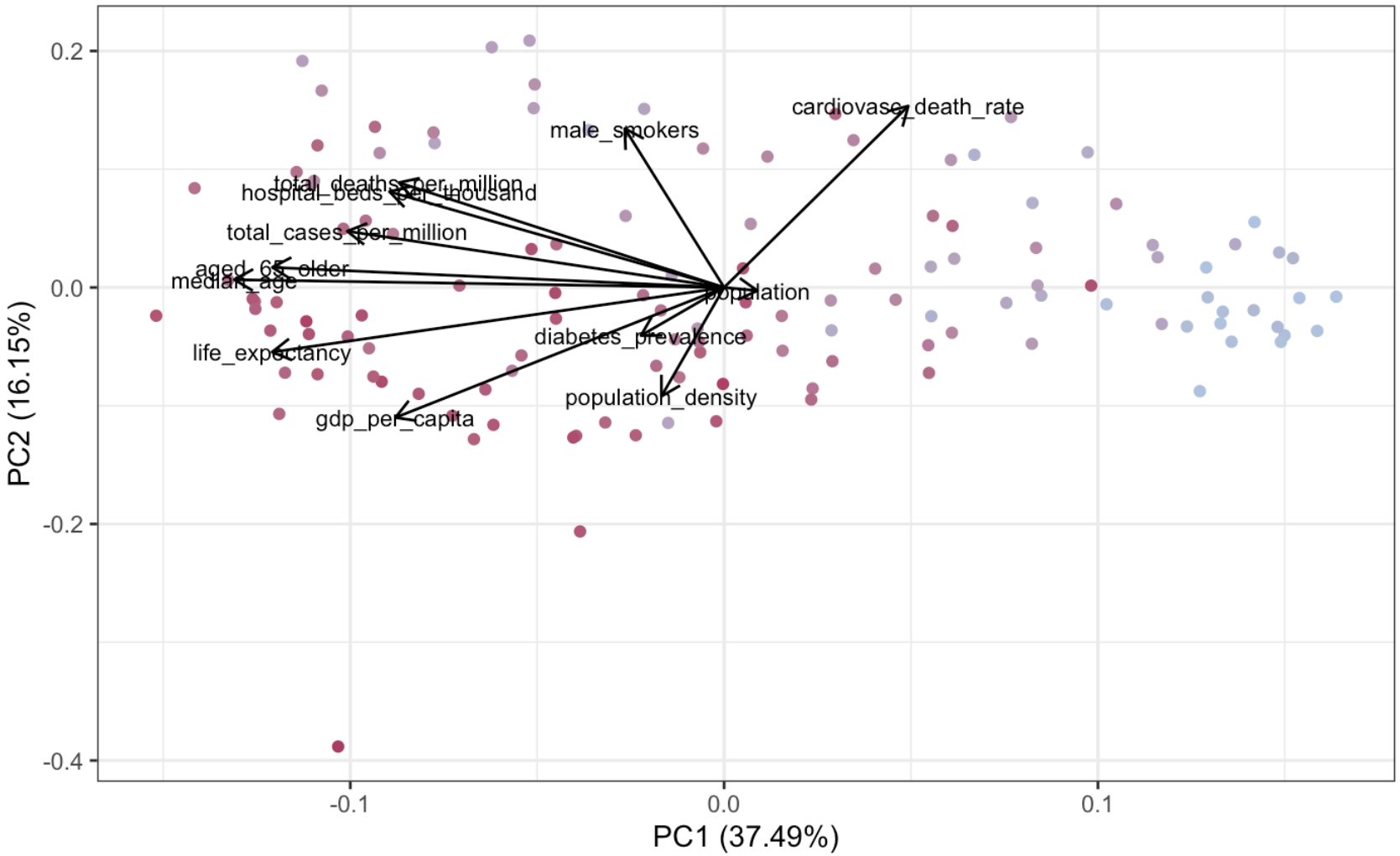
Example

```
df01.pca <- prcomp(~ total_cases_per_million+total_deaths_
    per_million+median_age+gdp_per_capita+hospital_beds_per_
    thousand+population+aged_65_older+diabetes_prevalence+
    cardiovasc_death_rate+population_density+male_smokers+
    life_expectancy , data = df01, center = TRUE,scale. =
    TRUE)
summary(df01.pca)
```

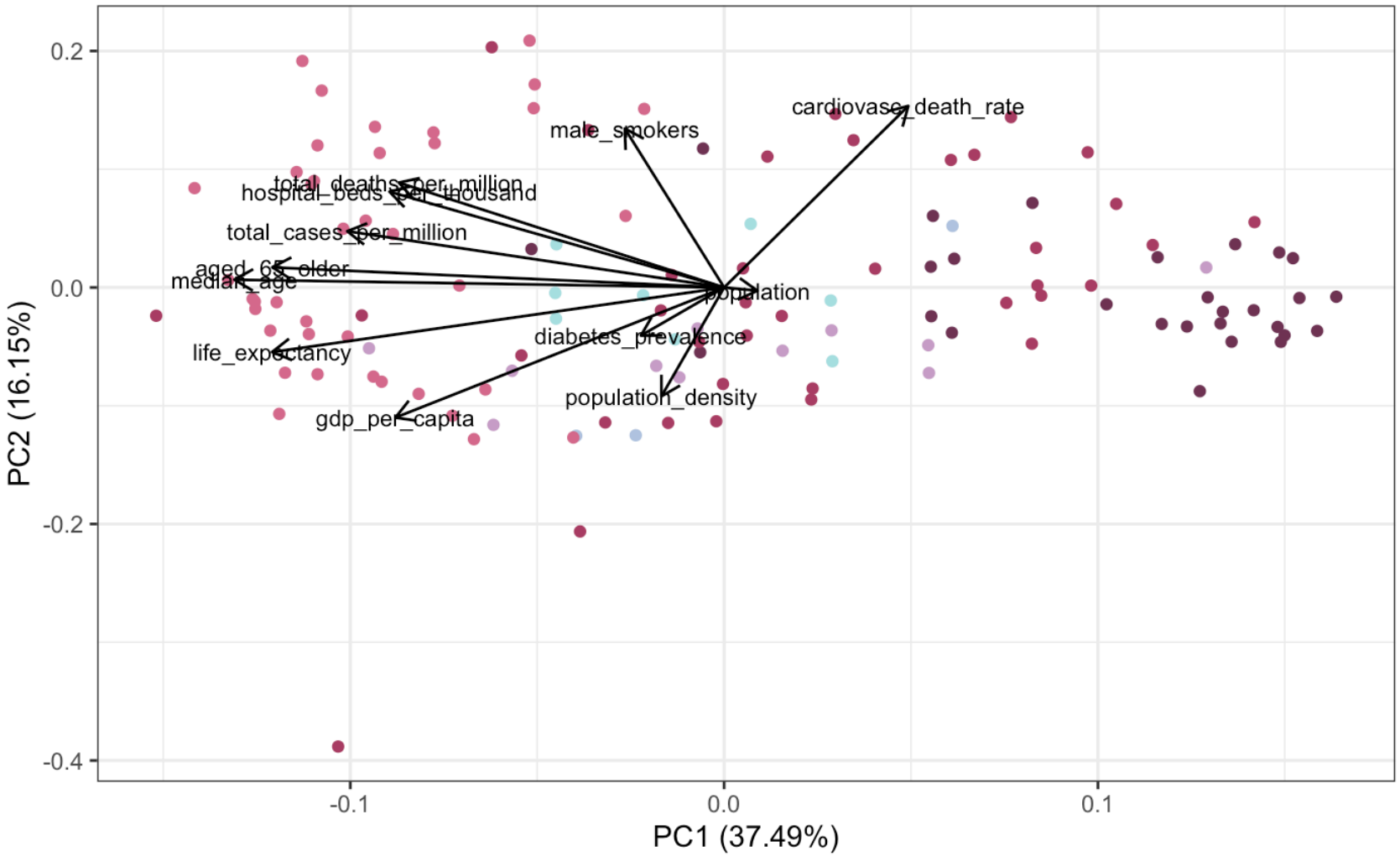
	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.1218	1.3927	1.1679	1.0178	0.9690	0.8509
Proportion of Variance	0.3752	0.1616	0.1137	0.0863	0.0783	0.0603
Cumulative Proportion	0.3752	0.5368	0.6505	0.7368	0.8150	0.8753

	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.6819	0.6297	0.5284	0.4441	0.3479	0.1931
Proportion of Variance	0.0388	0.0331	0.0233	0.0164	0.0101	0.0031
Cumulative Proportion	0.9141	0.9472	0.9704	0.9869	0.9970	1.0000

Example


0.2
0.0
-0.2
-0.4
PC2 (16.15%)
-0.1
0.0
0.1
PC1 (37.49%)
male_smokers
cardiovasc_death_rate
total_deaths_per_million
hospital_beds_per_thousand
total_cases_per_million
aged_65_older
median_age
population
life_expectancy
diabetes_prevalence
population_density
gdp_per_capita

Example


0.2
0.0
-0.2
-0.4
PC2 (16.15%)
-0.1
0.0
0.1
PC1 (37.49%)
male_smokers
cardiovase_death_rate
total_deaths_per_million
hospital_beds_per_thousand
total_cases_per_million
aged_65_older
median_age
population
life_expectancy
diabetes_prevalence
population_density
gdp_per_capita

Example

Below, we calculate the coefficients for the original regressors from the coefficients for the principal components.

```
df01.pc$rotation[,1:4] %*% data.matrix(df01.pcmodel$
    coefficients)[2:5,]
```

```
total_cases_per_million      1.3113413
total_deaths_per_million    -0.1636393
median_age                   4.2009415
gdp_per_capita               5.4896563
hospital_beds_per_thousand   1.3495715
population                   1.5913492
aged_65_older                3.1278789
diabetes_prevalence          2.9784789
cardiovasc_death_rate       -4.3522086
population_density           3.5986983
male_smokers                -0.9235646
life_expectancy              5.2658193
```

Example

```
df01.pcrmodel <- lm(df01$people_fully_vaccinated_per_hundred
    ~df01.pc$x[,1]+df01.pc$x[,2]+df01.pc$x[,3]+df01.pc$x
    [,4]) # can have no missing data in y for this to work
summary(df01.pcrmodel)
```

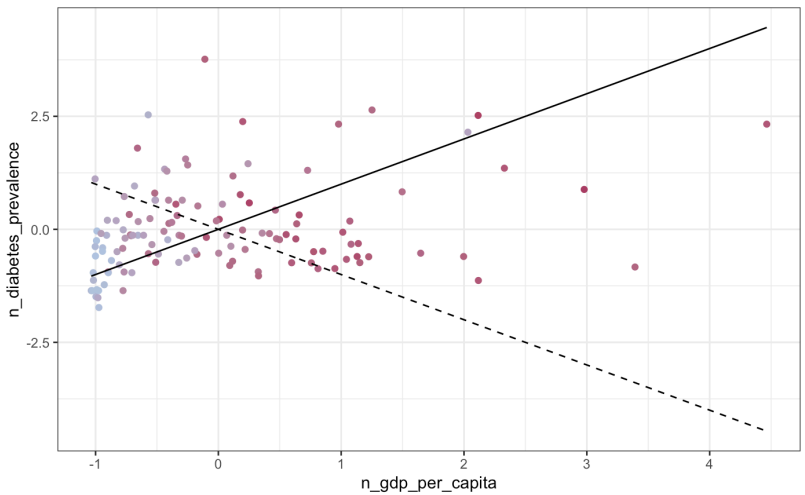
```
Residuals:
    Min      1Q  Median      3Q     Max
-41.990 -10.864   0.306   8.726  55.291

                Estimate Std. Error t value Pr(>|t|)
(Intercept)      45.1360     1.3864  32.557  < 2e-16 ***
df01.pc$x[, 1]   -8.9052     0.6560 -13.575  < 2e-16 ***
df01.pc$x[, 2]   -6.5900     0.9994  -6.594 1.18e-09 ***
df01.pc$x[, 3]    2.9026     1.1918   2.435   0.0163 *
df01.pc$x[, 4]   -1.3628     1.3675  -0.997   0.3210
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.56 on 121 degrees of freedom
Multiple R-squared: 0.6598, Adjusted R-squared:  0.6486
F-statistic: 58.67 on 4 and 121 DF,  p-value: < 2.2e-16
```

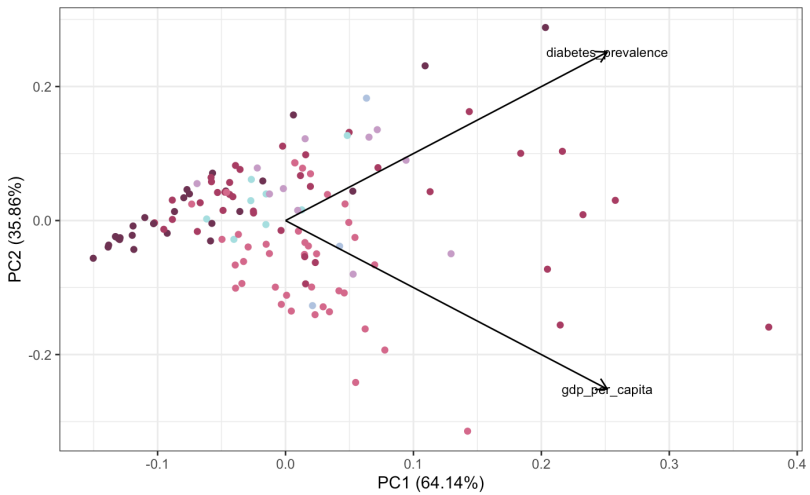
Example

```
df01.pc <- prcomp(~ gdp_per_capita+ diabetes_prevalence,
    data = df01, center = TRUE,scale. = TRUE)
summary(df01.pc)
```

	PC1	PC2
Standard deviation	1.1326	0.8469
Proportion of Variance	0.6414	0.3586
Cumulative Proportion	0.6414	1.0000

Example


n_diabetes_prevalence
2.5
0.0
-2.5
-1
0
1
2
3
4
n_gdp_per_capita


diabetes_prevalence
gdp_per_capita
PC2 (35.86%)
0.2
0.0
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
PC1 (64.14%)

The geometry of PCA

Observation 1

If A is a quadratic matrix, then the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots$ of A and be found by

$$\mathbf{v}_j = \underset{\substack{\mathbf{v}\colon \|\mathbf{v}\|_2=1,\\ \mathbf{v}\perp\mathbf{v}_i\ \forall i<j}}{\arg\max}\ \mathbf{v}^T A\mathbf{v} \qquad \text{and} \qquad \lambda_j = \underset{\substack{\mathbf{v}\colon \|\mathbf{v}\|_2=1,\\ \mathbf{v}\perp\mathbf{v}_i\ \forall i<j}}{\max}\ \mathbf{v}^T A\mathbf{v}$$

If $A = X^T X$, then $\mathbf{v}^T A\mathbf{v} = \mathbf{v}^T X^T X\mathbf{v} = \|X\mathbf{v}\|_2^2$.

Observation 2

The squared distance between $\mathbf{x}_j$ and the line $0 + t \cdot \mathbf{v}$, where $\mathbf{v}^T\mathbf{v} = 1$, is given by

$$\begin{aligned}
\|\mathbf{x}_j - \mathrm{proj}_{\mathbf{v}}\mathbf{x}_j\|_2^2 &= \left\|\mathbf{x}_j - \left((\mathbf{v}^T\mathbf{v})^{-1}\mathbf{v}^T\mathbf{x}_j\right)\mathbf{v}\right\|_2^2 = \|\mathbf{x}_j - (\mathbf{v}^T\mathbf{x}_j)\mathbf{v}\|_2^2 \\
&= \left(\mathbf{x}_j - (\mathbf{v}^T\mathbf{x}_j)\mathbf{v}\right)^T\left(\mathbf{x}_j - (\mathbf{v}^T\mathbf{x}_j)\mathbf{v}\right) \\
&= \mathbf{x}_j^T\mathbf{x}_j - (\mathbf{v}^T\mathbf{x}_j)\mathbf{v}^T\mathbf{x}_j - \mathbf{x}_j^T(\mathbf{v}^T\mathbf{x}_j)\mathbf{v} + (\mathbf{v}^T\mathbf{x}_j)\mathbf{v}^T(\mathbf{v}^T\mathbf{x}_j)\mathbf{v} \\
&= 1 - (\mathbf{x}_j^T\mathbf{v})^2,
\end{aligned}$$

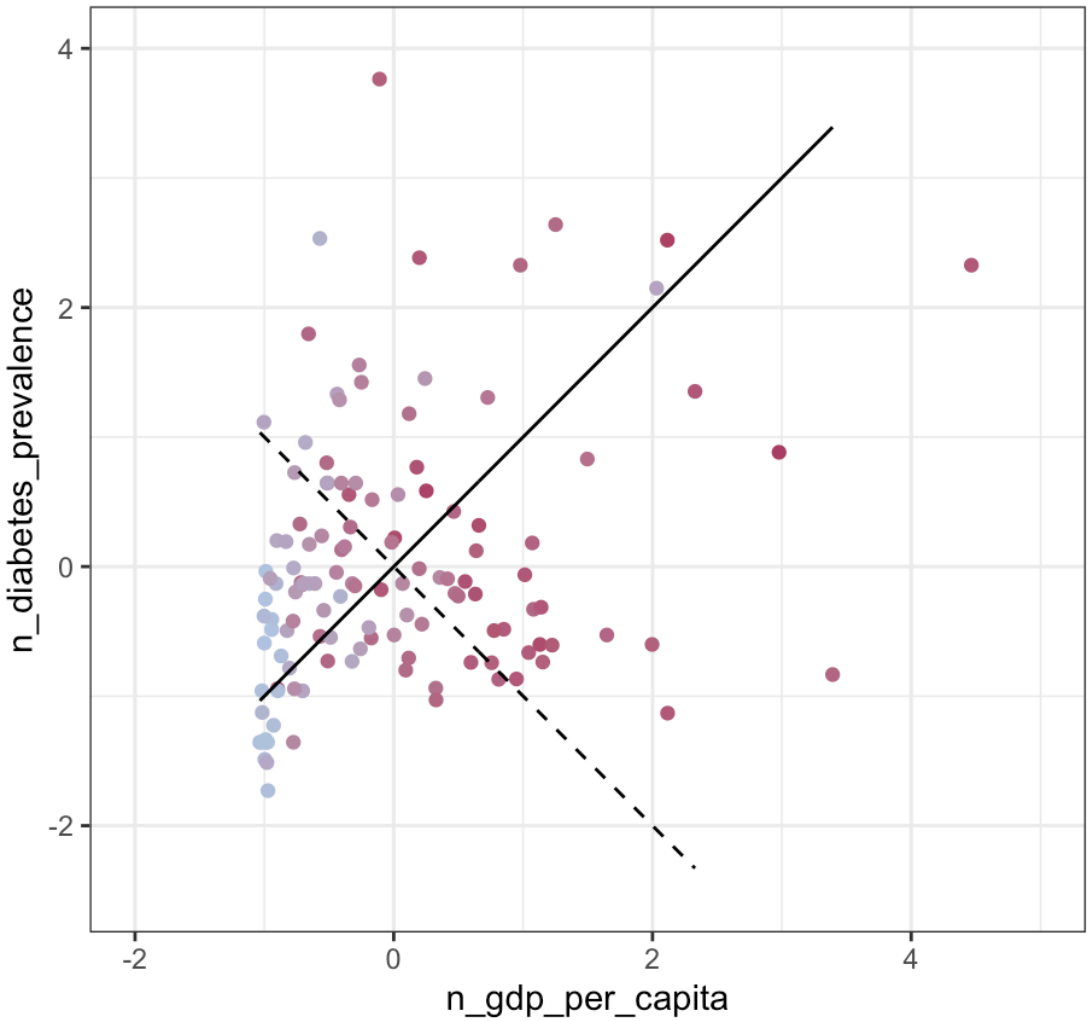
and hence

$$\sum_j \left\|\mathbf{x}_j - \mathrm{proj}_{\mathbf{v}}\mathbf{x}_j\right\|_2^2 = \sum\left(1 - (\mathbf{x}_j^T\mathbf{v})^2\right) = k - \mathbf{v}^T X^T X\mathbf{v} = k - \|X\mathbf{v}\|_2^2.$$

Observation 3

When we are talking about PCA, $A = X^T X$, $P = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$, $Z = XP$, and hence $\mathbf{z}_j(k) = \mathbf{x}_j^T\mathbf{v}_k = (\mathbf{v}_k^T\mathbf{v})^{-1}\mathbf{v}_k^T\mathbf{x}_j$.

Example



Sparse principal component analysis (sparse PCA)

Motivation

When $k+1 \gg n$, the eigenvectors of $X^T X$ can be sensitive to the particular sample.

Idea

Note that principal components of X are given by $Z\mathbf{e}_j = XP\mathbf{e}_j$, where

$$P\mathbf{e}_j = \underset{\substack{\mathbf{z}:\ \|\mathbf{z}\|_2=1,\\ \mathbf{z}\perp Z\mathbf{e}_i\ \forall i<j}}{\arg\min}\ \mathbf{z}^T X^T X\mathbf{z} = \underset{\substack{\mathbf{z}:\ \|\mathbf{z}\|_2=1,\\ \mathbf{z}\perp Z\mathbf{e}_i\ \forall i<j}}{\arg\min}\ \|X\mathbf{z}\|_2^2$$

To this formula we can add a penalty, which will make the principal components more stable.

$$P^{(t)}\mathbf{e}_j := \underset{\substack{\mathbf{z}:\ \|\mathbf{z}\|_2=1,\\ \mathbf{z}\perp Z_{\cdot i}\forall i<j,\\ \|\mathbf{z}\|_1\le t}}{\arg\min}\ \|X\mathbf{z}\|_2^2 = \underset{\substack{\mathbf{z}:\ \|\mathbf{z}\|_2=1,\\ \mathbf{z}\perp Z_{\cdot i}\forall i<j}}{\arg\min}\ \|X\mathbf{z}\|_2^2 + \lambda(t)\|\mathbf{z}\|_1$$

Sparse principal component analysis (sparse PCA)

Idea

Define

$$P^{(t)}\mathbf{e}_j := \underset{\substack{\mathbf{z}:\ \|\mathbf{z}\|_2=1,\\ \mathbf{z}\perp Z_{\cdot i}\forall i<j,\\ \|\mathbf{z}\|_1\leq t}}{\arg\min}\ \|X\mathbf{z}\|_2^2$$

and

$$\lambda_j^{(t)} := \min_{\substack{\mathbf{z}:\ \|\mathbf{z}\|_2=1,\\ \mathbf{z}\perp Z_{\cdot i}\forall i<j,\\ \|\mathbf{z}\|_1\leq t}} \|X\mathbf{z}\|_2^2.$$

Note that $\lambda_j^{(t)}$ will only be an eigenvalue of X^TX if $t=\infty$.

Sparse PCA

1. Pick $t \geq 0$.
2. Calculate the pseudo-eigenvalues $\lambda_j^{(t)}$.
3. Perform regression analysis using the weighed principal components $Z^{(t)}\mathbf{e}_j := XP^{(t)}$ with the largest pseudo-eigenvalues $\lambda_j^{(t)}$.

Example

```r
library("elasticnet")
df01.spc <- spca(X, K = 5, type = "predictor", sparse = "
    penalty", para = c(.2, .2, .2, .5, .5))
df01.spca
```

```
5 sparse PCs
Pct. of exp. var. : 23.2 12.7  8.2  8.2  7.9
Num. of non-zero loadings :  6 3 1 1 1
Sparse loadings
                                PC1    PC2 PC3 PC4 PC5
total_cases_per_million      -0.324  0.000   0   0   0
total_deaths_per_million     -0.111  0.000   0   0   0
median_age                   -0.867  0.000   0   0   0
gdp_per_capita                0.000 -0.206   0   0   0
hospital_beds_per_thousand   -0.187  0.000   0   0   0
population                    0.000  0.000   0  -1   0
aged_65_older                -0.184  0.000   0   0   0
diabetes_prevalence           0.000  0.000   1   0   0
cardiovasc_death_rate         0.000  0.813   0   0   0
population_density            0.000  0.000   0   0   1
male_smokers                  0.000  0.545   0   0   0
life_expectancy              -0.247  0.000   0   0   0
```

Example

```
df01.spc$x <- X%*%df01.spc$loadings

df01.pcrmodel <- lm(df01$people_fully_vaccinated_per_hundred
    ~df01.spc$x[,1]+df01.spc$x[,2]+df01.spc$x[,3]+df01.spc$x
    [,4])
summary(df01.pcrmodel)
```

```
Residuals:
   Min      1Q  Median      3Q     Max
-40.30  -12.06    1.26   10.08   56.98

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        45.136      1.489  30.315  < 2e-16 ***
df01.spc$x[, 1] -105.991     10.272 -10.319  < 2e-16 ***
df01.spc$x[, 2]  -77.830     13.568  -5.736 7.28e-08 ***
df01.spc$x[, 3]   59.546     16.939   3.515 0.000619 ***
df01.spc$x[, 4]  -15.092     16.829  -0.897 0.371613
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.71 on 121 degrees of freedom
Multiple R-squared: 0.6076, Adjusted R-squared:  0.5946
F-statistic: 46.84 on 4 and 121 DF,  p-value: < 2.2e-16
```

Example

Below, we calculate the coefficients for the original regressors from the coefficients for the sparse principal components.

```
df01.spc$loadings %*% df01.spcmodel$coefficients
```

```
                                [,1]
total_cases_per_million    -14.635267
total_deaths_per_million    -5.029128
median_age                 -39.153592
gdp_per_capita              21.784591
hospital_beds_per_thousand  -8.455239
population                 -59.545870
aged_65_older               -8.302296
diabetes_prevalence        -77.830356
cardiovasc_death_rate      -86.163607
population_density         -15.092154
male_smokers               -57.753132
life_expectancy            -11.151146
```