Data Science: Unsupervised Learning

Adriano Z. Zambom

Unsupervised Learning

Unsupervised Learning: Principal Components Analysis

PCA

- Objective:

PCA

- Objective: Perform dimension reduction when we have many variables

PCA

- Objective: Perform dimension reduction when we have many variables
- That is, explain the variance-covariance structure of the data through a few linear combinations

PCA

- Objective: Perform dimension reduction when we have many variables
- That is, explain the variance-covariance structure of the data through a few linear combinations
- With 12 variables, for example, there will be more than 200 three-dimensional scatterplots to be studied

PCA

- Objective: Perform dimension reduction when we have many variables
- That is, explain the variance-covariance structure of the data through a few linear combinations
- With 12 variables, for example, there will be more than 200 three-dimensional scatterplots to be studied
- Summary: Dimension reduction + interpretation

PCA

- Helpful when?

PCA

- Helpful when? we have multicolinearity: correlations between the x-variables

Principal Components Analysis

PCA

- Helpful when? we have multicolinearity: correlations between the x-variables
- Ex: What would you see in a plot of two x-variables that have a nearly perfect correlation?

PCA

- Helpful when? we have multicolinearity: correlations between the x-variables
- Ex: What would you see in a plot of two x-variables that have a nearly perfect correlation? around a line

PCA

- Helpful when? we have multicolinearity: correlations between the x-variables
- Ex: What would you see in a plot of two x-variables that have a nearly perfect correlation? around a line
- That line could be used as a new (one-dimensional) axis to represent the variation among data points.

PCA

- Helpful when? we have multicolinearity: correlations between the x-variables
- Ex: What would you see in a plot of two x-variables that have a nearly perfect correlation? around a line
- That line could be used as a new (one-dimensional) axis to represent the variation among data points.
- Ex: Suppose that we have verbal, math, and total SAT scores. Note that (at most) two dimensions of the data are necessary because total= verbal +math

PCA

- Helpful when? we have multicolinearity: correlations between the x-variables
- Ex: What would you see in a plot of two x-variables that have a nearly perfect correlation? around a line
- That line could be used as a new (one-dimensional) axis to represent the variation among data points.
- Ex: Suppose that we have verbal, math, and total SAT scores. Note that (at most) two dimensions of the data are necessary because total= verbal +math
  'at most' two dimensions and not exactly two: if there is a strong correlation between verbal and math, then it may be possible that there is only one true dimension to the data.

PCA

- The total variability is given by all variables, say we have p variables.
- However, sometimes we can find just a few (less than p) that explain almost all variability contained in all these p variables.

PCA

- The total variability is given by all variables, say we have p variables.
- However, sometimes we can find just a few (less than p) that explain almost all variability contained in all these p variables.
- For example: if you have a large number of variables, say p, for your regression analysis, it may be difficult to fit the model, or you may not want to use all these variables. Instead, you can use only a few components $k < p$, and the prediction of your response variable Y may be satisfactory.

PCA Suppose we have random variables $X_1, \ldots, X_p$.

PCA Suppose we have random variables $X_1, \ldots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ

PCA Suppose we have random variables $X_1, \ldots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

PCA Suppose we have random variables $X_1, \ldots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.Consider the linear combinations

$$Y_1 = \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

Principal Components Analysis

PCA Suppose we have random variables $X_1, \ldots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.Consider the linear combinations

$$Y_1 = \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$Y_2 = \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

PCA Suppose we have random variables $X_1, \ldots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.Consider the linear combinations

$$\begin{aligned} Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \\ Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned}$$

Then
$Var(Y_i) =$

Principal Components Analysis

PCA Suppose we have random variables $X_1, \ldots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.Consider the linear combinations

$$\begin{aligned}
Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\
&\vdots \\
Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p
\end{aligned}$$

Then
$Var(Y_i) = \mathbf{a}_i'\Sigma\mathbf{a}_i, i = 1, 2, \ldots, p$
$Cov(Y_i, Y_k) =$

Principal Components Analysis

PCA Suppose we have random variables $X_1, \ldots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.Consider the linear combinations

$$
\begin{aligned}
Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\
&\vdots \\
Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p
\end{aligned}
$$

Then
$Var(Y_i) = \mathbf{a}_i'\Sigma\mathbf{a}_i, i = 1, 2, \ldots, p$
$Cov(Y_i, Y_k) = \mathbf{a}_i'\Sigma\mathbf{a}_k, i, k = 1, 2, \ldots, p$

Principal Components Analysis

PCA Suppose we have random variables $X_1, \dots, X_p$. Assume these random variables together $\mathbf{X}' = [X_1, X_2, \dots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.Consider the linear combinations

$$\begin{aligned}
Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p \\
&\vdots \\
Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p
\end{aligned}$$

Then
$Var(Y_i) = \mathbf{a}_i'\Sigma\mathbf{a}_i, i = 1, 2, \dots, p$
$Cov(Y_i, Y_k) = \mathbf{a}_i'\Sigma\mathbf{a}_k, i, k = 1, 2, \dots, p$
The principal components are those *uncorrelated* linear combinations $Y_1, \dots, Y_p$ whose variances are as large as possible.

PCA

- **First principle component** = linear combination $\mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$ that maximizes $Var(\mathbf{a}_1'\mathbf{X})$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$.

PCA

- **First principle component** = linear combination $\mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$ that maximizes $Var(\mathbf{a}_1'\mathbf{X})$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$.
- **Second principle component** = linear combination $\mathbf{a}_2'\mathbf{X}$ that maximizes $Var(\mathbf{a}_2'\mathbf{X})$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $Cov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = 0$.

PCA

- **First principle component** = linear combination $\mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$ that maximizes $Var(\mathbf{a}_1'\mathbf{X})$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$.
- **Second principle component** = linear combination $\mathbf{a}_2'\mathbf{X}$ that maximizes $Var(\mathbf{a}_2'\mathbf{X})$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $Cov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = 0$.

  ...
- **i-th principle component** = linear combination $\mathbf{a}_i'\mathbf{X}$ that maximizes $Var(\mathbf{a}_i'\mathbf{X})$ subject to $\mathbf{a}_i'\mathbf{a}_i = 1$ and $Cov(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_k'\mathbf{X}) = 0 \forall k < i$.

PCA

Result The variances of the Principal components have the following property

$$\begin{aligned}\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} &= \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p \\ &= \sum_{i=1}^{p} Var(Y_i)\end{aligned}$$

PCA

Result The variances of the Principal components have the following property

$$\begin{aligned}\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} &= \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p \\ &= \sum_{i=1}^{p} Var(Y_i)\end{aligned}$$

That is, total population variance is equal to the sum of the eigen-values. Hence, The proportion of total variance explained by the k-th principal component is

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

PCA

Result The vectors $\mathbf{a}_i$ that define the principal components are actually the eigen-vectors associated with the matrix Σ

PCA

Example 8.1 (Calculating the population principal components) Suppose t
random variables X_1, X_2 and X_3 have the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

It may be verified that the eigenvalue–eigenvector pairs are

$$\begin{aligned} \lambda_1 &= 5.83, & \mathbf{e}_1' &= [.383, -.924, 0] \\ \lambda_2 &= 2.00, & \mathbf{e}_2' &= [0, 0, 1] \\ \lambda_3 &= 0.17, & \mathbf{e}_3' &= [.924, .383, 0] \end{aligned}$$

Therefore, the principal components become

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = .383X_1 - .924X_2 \\ Y_2 &= \mathbf{e}_2'\mathbf{X} = X_3 \\ Y_3 &= \mathbf{e}_3'\mathbf{X} = .924X_1 + .383X_2 \end{aligned}$$

PCA

In this example we have

$$\sigma_1 + \sigma_2 + \sigma_3 = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2 + 0.17$$

So the proportion of variability explained by the first principal component is

$$\frac{5.83}{5.83 + 2 + 0.17} = 0.728$$

PCA

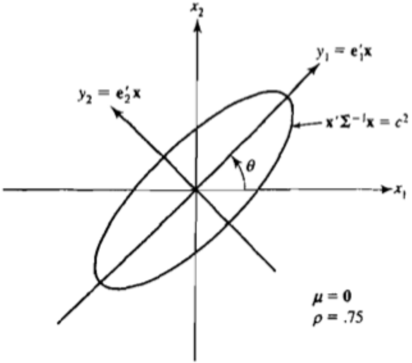

Figure 8.1 The constant density ellipse $\mathbf{x}'\Sigma^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector $\mathbf{X}$ having mean $\mathbf{0}$.

PCA

Example: (Summarizing sample variability with two sample principal components) A census provided information, by tract, on five socioeconomic variables for Madison, Wisconsin, area.

PCA

Example: (Summarizing sample variability with two sample principal components) A census provided information, by tract, on five socioeconomic variables for Madison, Wisconsin, area.

Coefficients for the Principal Components
(Correlation Coefficients in Parentheses)

Variable	$\hat{e}_1\ (r_{\hat{y}_1, x_k})$	$\hat{e}_2\ (r_{\hat{y}_2, x_k})$	$\hat{e}_3$	$\hat{e}_4$	$\hat{e}_5$
Total population	−0.039(−.22)	0.071(.24)	0.188	0.977	−0.058
Profession	0.105(.35)	0.130(.26)	−0.961	0.171	−0.139
Employment (%)	−0.492(−.68)	0.864(.73)	0.046	−0.091	0.005
Government employment (%)	0.863(.95)	0.480(.32)	0.153	−0.030	0.007
Medium home value	0.009(.16)	0.015(.17)	−0.125	0.082	0.989
Variance ($\hat{\lambda}_i$):	107.02	39.67	8.37	2.87	0.15
Cumulative percentage of total variance	67.7	92.8	98.1	99.9	1.000

PCA in R:

PCA in R:

Let's use the dataset:
?USArrests

PCA in R:

Let's use the dataset:
?USArrests

There are a few functions to run PCA in R: ?
?prcomp (stats)
?princomp (stats)
?PCA (FactoMineR)
?dudi.pca (ade4)
?acp (amap).

Principal Components Analysis

PCA in R:

Let's use the dataset:
?USArrests

There are a few functions to run PCA in R: ?
?prcomp (stats)
?princomp (stats)
?PCA (FactoMineR)
?dudi.pca (ade4)
?acp (amap).

Let's focus on ?prcomp and ?princomp

PCA in R:

```
fit = prcomp(USArrests, scale. = TRUE)
```

PCA in R:

fit = prcomp(USArrests, scale. = TRUE)

```
Standard deviations (1, .., p=4):
[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation (n x k) = (4 x 4):
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

Principal Components Analysis

```
fit2 = princomp(USArrests, cor = TRUE)
pca2$sdev
unclass(pca2$loadings)
```

PCA
The number of Principal Components

PCA

The number of Principal Components

There is no definitive answer to this question. Consider:

- the amount of total sample variance explained

PCA

The number of Principal Components

There is no definitive answer to this question. Consider:

- the amount of total sample variance explained
- the relative sizes of the eigenvalues (the variances of the sample components)

PCA

The number of Principal Components

There is no definitive answer to this question. Consider:

- the amount of total sample variance explained
- the relative sizes of the eigenvalues (the variances of the sample components)
- the subject-matter interpretation of the components

PCA

The number of Principal Components

There is no definitive answer to this question. Consider:

- the amount of total sample variance explained
- the relative sizes of the eigenvalues (the variances of the sample components)
- the subject-matter interpretation of the components
- A component associated with an eigenvalue near zero and, hence deemed unimportant, may indicate an unsuspected linear dependency in the data.

PCA

The number of Principal Components

There is no definitive answer to this question. Consider:

- the amount of total sample variance explained
- the relative sizes of the eigenvalues (the variances of the sample components)
- the subject-matter interpretation of the components
- A component associated with an eigenvalue near zero and, hence deemed unimportant, may indicate an unsuspected linear dependency in the data.
- A useful visual aid to determining an appropriate number of principal components is a scree plot. With the eigenvalues ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus i-the magnitude of an eigenvalue versus its number. **look for the elbow!**

PCA

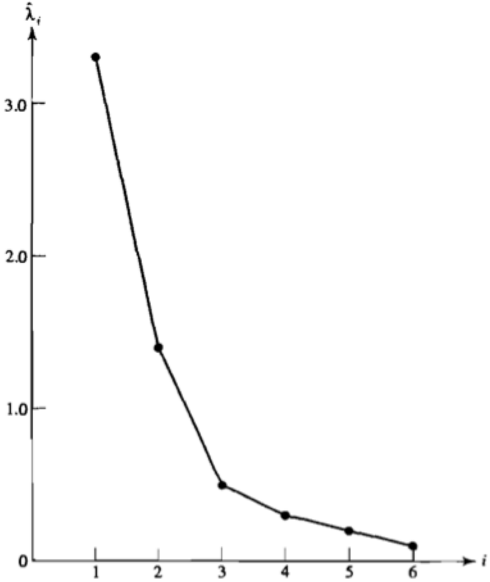

Figure 8.2 A scree plot.

PCA
Use *screeplot* function in R
screeplot(fit)
screeplot(fit2)

PCA

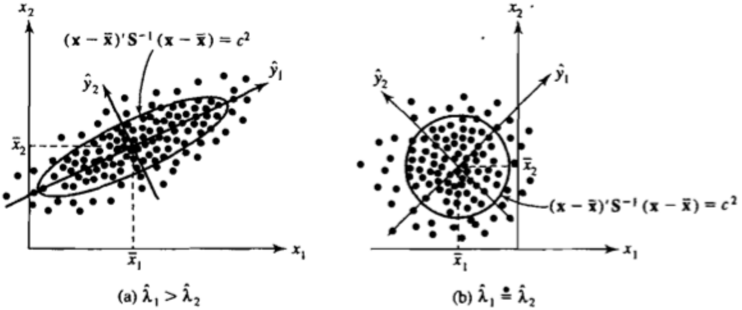

Figure 8.4 Sample principal components and ellipses of constant distance.

Principal Components Analysis

Videos:
https://www.youtube.com/watch?v=_UVHneBUBW0
https://www.youtube.com/watch?v=HMOI_lkzW08
https://www.youtube.com/watch?v=kw9R0nD69OU
https://www.youtube.com/watch?v=NLrb41ls4qo
Exercises:
1. Apply PCA on the wine data:
wine <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data",
sep=",")
2. https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/pca-exercises
Remember to take a look at the loadings, plots, variability explained, choosing the number of components, etc