Unsupervised Learning - Principal Compnent Analysis (PCA)

Mark Asch - IMU/VLP/CSU

2023

Program

1. Data Analysis
   (a) Introduction: the 4 identifiers of “big data” and “data science”
   (b) Supervised learning methods: regression—advanced, k-NN, linear classification methods, SVM, NN, decision trees.
   (c) Unsupervised learning methods: principal component analysis, k-means, clustering.

Introduction

- Supervised learning:

  ⇒ we have p characteristics $X_1, X_2, \ldots, X_p$ mesured on n observations

  ⇒ one response Y mesured on the same observations

  ⇒ objective: predict Y using $X_1, X_2, \ldots, X_p$

  ⇒ methods: regression and classification

- Unsupervised learning:

  ⇒ we only have p characteristics $X_1, X_2, \ldots, X_p$ mesured on n observations

  ⇒ we want to make predictions, but we do not have an associated response variable...

  ⇒ objective: discover interesting effects with respect to the observations of $X_1, X_2, \ldots, X_p$

  → Can we visualize or represent the data in a more informative way?

  → Can we discover sub-groups or clusters among the variables or the observations?

Warnings!

- Unsupervised learning is more difficult, since it is subjective.

✘ Unsupervised learning is often part of an initial phase of exploratory data analysis (EDA), where we compute elementary statistics and plot basic histograms and boxplots---see Introductory lectures.

✘ There are no universal cross-validation methods for unsupervised learning.

✘ There is no response variable that can be used to test our models.

✔ However, there are a large number of application domains where unsupervised learning is widely used.

Introduction to PCA

- PCA is a common approach for finding a low dimensional ensemble of characteristics from a large set of variables.

- It can be used for dimension reduction of an $(n \times p)$ dimensional data matrix X. Once this is done,

  ⇒ the direction of the first principal component vector is that along which the observations vary the most, or it is the direction that captures or explains the most variance in the observations.

  ⇒ the principal components can form the basis for a regression—this is principal component regression—see Regression lectures.

- Principal component analysis is, by definition, the computation of these principal components (directions) and their use for understanding the data.

- PCA can also serve as a tool for visualizing the data, as an additional technique of EDA.

- A simple case is shown in the Figure below.

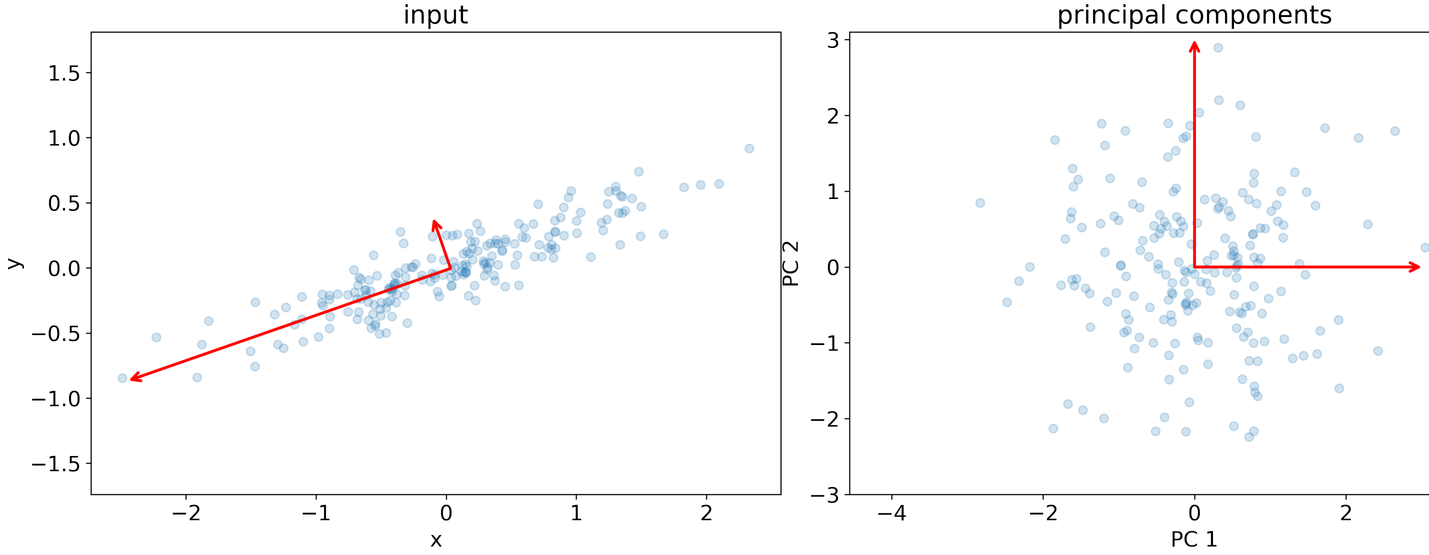


Figure 1: PCA for a dataset with two features. In the left plot, we depict the original data. The vectors represent the directions of the first and the second principal components, and their lengths represent their importance in terms of the amount of variance explained. The projection of the data onto the two PCs is depicted in the right plot.

Theory

- We want to visualize n observations over a set of p characteristics, $X_1, X_2, \ldots, X_p$.

  ⇒ We could do a pairwise scatterplot of the characteristics, but the number of combinations,

  $$\binom{p}{2} = p(p-1)/2,$$

  soon becomes very large.

- The PCA helps us to identify the smallest possible number of dimensions (in the case where p is large) that represent the largest possible proportion of the variation.

- The first principal component, Z_1, is the linear combination of the characteristics,

  $$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p,$$

  with the largest variance, where

⇒ $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ (normalization), and
⇒ $\phi_{11}, \ldots, \phi_{p1}$ are the loads (coefficients).

Computation of PCs

- To compute the principal components, we suppose that

  ⇒ X is a dataset of dimension $n \times p$ and that
  ⇒ each variable in X is centered (average equals zero), column-by-column.

- We seek a linear combination of the characteristics of the form

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$$

  that has a maximal variance, and whose coefficients are normalized.

- Mathematically, we want to find

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\},$$

such that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

- We remark the following:
  - ⇒ The quantity that we maximize is precisely the empirical variance of z.
  - ⇒ The optimization problem is solved by a decomposition into eigenvalues and eigenvectors of the matrix $X^{\mathrm{T}}X$,
  - ⇒ The second principal component is the linear combination of maximal variance that is non-correlated with the first principal component, and thus orthogonal to it, and so on for the third, fourth, etc.
  - ⇒ Once the principal components have been computed, we can plot them 2-by-2 to obtain low dimensional views of the data. This is the EDA aspect of PCA.

PVE

- The principal components represent the proportion of variance explained (PVE).

⇒ Total variance for centered variables is, by definition,

$$V_T = \sum_{j=1}^{p} \mathrm{Var}\left(X_j\right) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2,$$

⇒ Variance explained by the m-th principal component is

$$V_m = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2 .$$

⇒ Then the proportion

$$\mathrm{PVE} = V_m / V_T$$

gives us the variance explained.

How many PCs ?

Please see the examples below.

- A matrix $\boldsymbol{X}$ of dimension $n \times p$ has $\min(n-1,p)$ distinct principal components

- The major question is: how many components one needs to adequately describe the data. T

- To decide this, we plot the PVE as a function of the principal component number, in what is known as a "scree plot."

  ⇒ The curve obtained will usually exhibit an "elbow," where the curvature changes sign.

  ⇒ This point gives us the optimal number of principal components to retain.

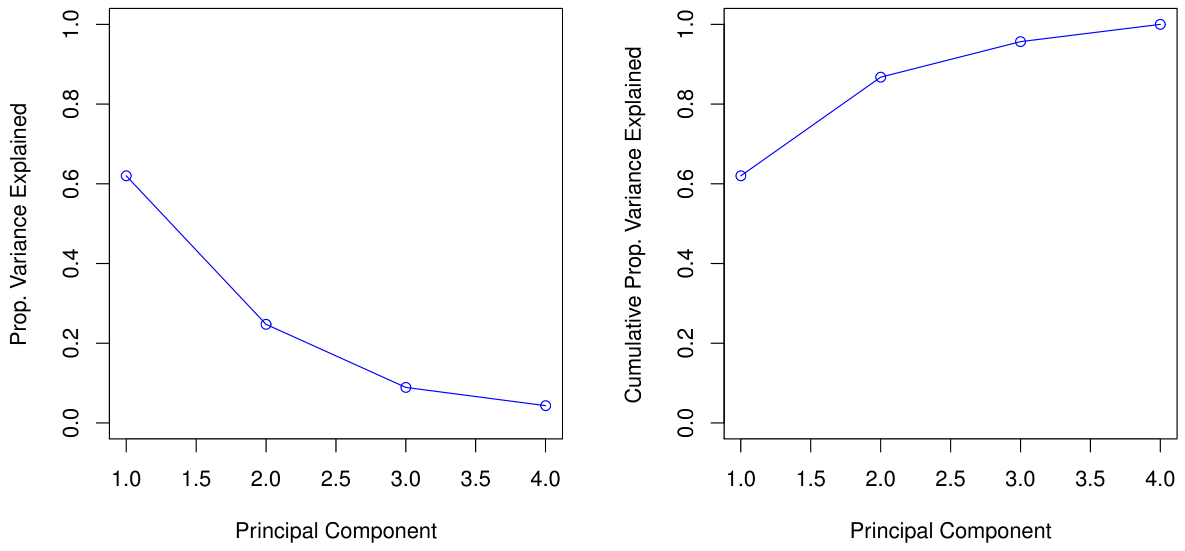


Figure 2: Scree Plot: left, the proportion of variance explained by each of the 4 principal components; right, the cumulative proportion.

Example: arrests in USA

- the database USArrests contains 50 states and 4 variables: Murder, Assault, UrbanPop, Rape

- the rows contain the 50 states in alphabetical order

```
> states =row.names(USArrests )
> states
[1] "Alabama"          "Alaska"          "Arizona"
[4] "Arkansas"         "California"      "Colorado"
[7] "Connecticut"      "Delaware"        "Florida"
[10] "Georgia"          "Hawaii"          "Idaho"
[13] "Illinois"         "Indiana"         "Iowa"
[16] "Kansas"           "Kentucky"        "Louisiana"
[19] "Maine"            "Maryland"        "Massachuse
[22] "Michigan"         "Minnesota"       "Mississipp
[25] "Missouri"         "Montana"         "Nebraska"
...
[46] "Virginia"         "Washington"      "West Virgi
[49] "Wisconsin"        "Wyoming"
```

- the columns contain the 4 variables

```
> names(USArrests)
[1] "Murder " "Assault " "UrbanPop " "Rape"
```

- EDA: Let us examine the data. The averages and variances differ a lot.

```
> apply(USArrests, 2, mean)
 Murder Assault UrbanPop  Rape
  7.79   170.76   65.54   21.23
> apply(USArrests, 2, var)
Murder  Assault  UrbanPop   Rape
  19.0   6945.2    209.5     87.7
```

- Hence, we need to apply a scaling when using the function `prcomp()` that computes the principal components.

```
> pr.out=prcomp(USArrests, scale=TRUE)
```

- The output contains a number of useful quantities

```
> names(pr.out)
[1] "sdev" "rotation " "center " "scale" "x"
```

- The variables `centre` and `scale` correspond to the averages and standard deviations used in the scaling

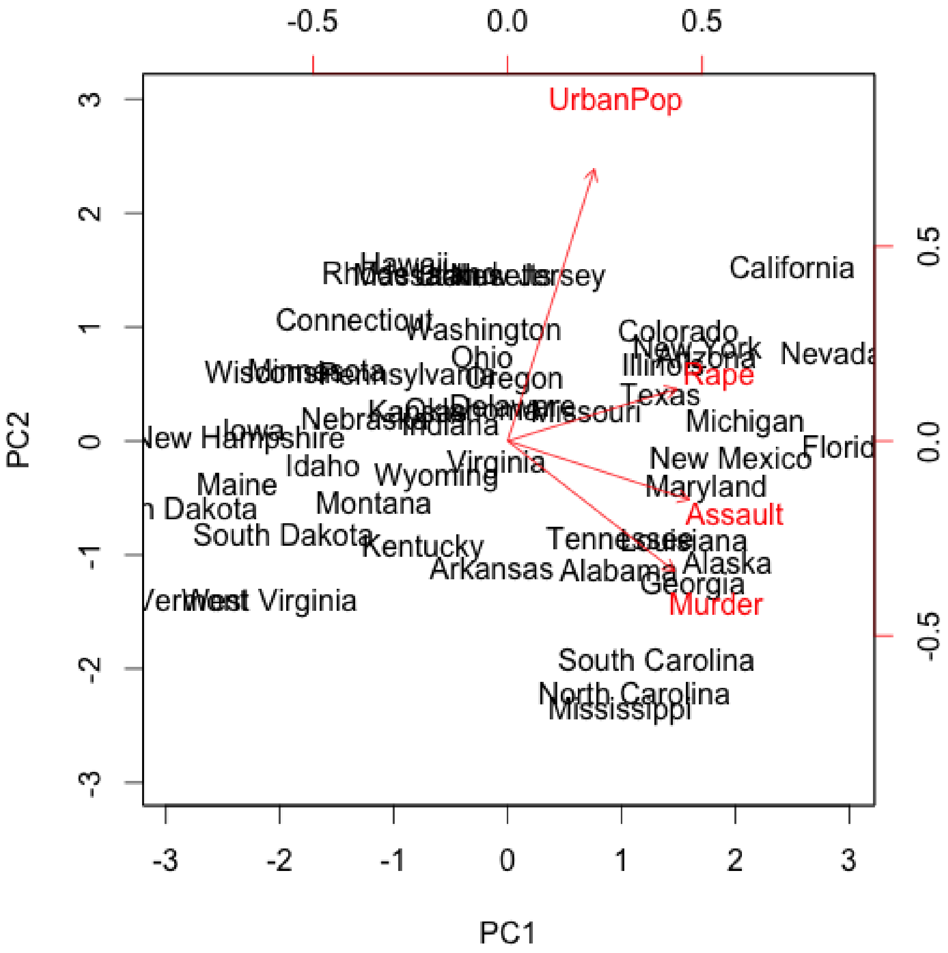
```
> pr.out$center
Murder  Assault  UrbanPop  Rape
7.79    170.76    65.54    21.23
> pr.out$scale
Murder  Assault  UrbanPop  Rape
4.36    83.34     14.47     9.37
```

- The matrx `rotation` provides the coefficients of the 4 principal components

```
> pr.out$rotation
            PC1    PC2    PC3    PC4
Murder   -0.536  0.418 -0.341  0.649
Assault  -0.583  0.188 -0.268 -0.743
UrbanPop -0.278 -0.873 -0.378  0.134
Rape     -0.543 -0.167  0.818  0.089
```

- To change the orientation and plot the projections of the original variables d'origine on the first 2 principal directions/axes

```
> pr.out$rotation=-pr.out$rotation
> pr.out$x=-pr.out$x
> biplot(pr.out, scale =0)
```



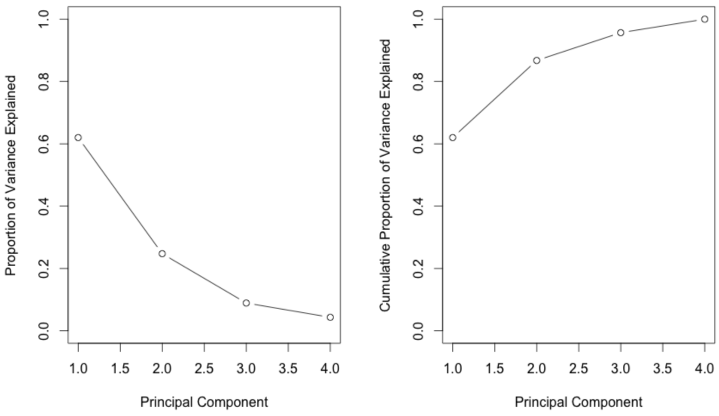
- Compute the variance and the PVE

```
> pr.out$sdev
```

```
[1] 1.575 0.995 0.597 0.416
> pr.var=pr.out$sdev^2
> pr.var
[1] 2.480 0.990 0.357 0.173
> pve=pr.var/sum(pr.var)
> pve
[1] 0.6201 0.2474 0.0891 0.0434
```

- and plot them

```
> par(mfrow=c(1,2))
> plot(pve,xlab="Principal Component",ylab="Proportion
of Variance Explained",ylim=c(0,1),type='b')
> plot(cumsum(pve),xlab="Principal Component",ylab ="Cumulat
Proportion of Variance Explained",ylim=c(0,1),type='b')
```



Interpretation of PCA

- After a PCA, the original explnanatory variables are "lost"

- But, the reduction of dimension can elucidate and facilitate the interpretaion and comprehension of all the data, especially when the number of explanatory variables is (too) big

- The interpretation always depends on the context, and should be done with the assistance of a domain expert:

  ⇒ the first component is associated to all the violent crimes;

  ⇒ the second component opposes assaults and murders with the population (more violent crimes in the less populated states) ;

  ⇒ there is also a correlation between southern states and violent crimes, with the exception of rape.

Examples

1. The dataset state.x77 contains 8 explanatory variables (4 more than USArrests) and allows to relate crimes with some socio-economical factors.

2. PCA_cars04 is a very complete dataset of cars from a consumer ascociation in the USA.

3. PCA_cancer contains breast cancer data. We analyze it with scikit-learn.

References

1. M. DeGroot, M. Schervish, *Probability and Statistics*, Addison Wesley, 2002.

2. Spiegel, Murray and Larry Stephens, *Schaum's Outline of Statistics*, 6th edition, McGraw Hill. 2017.

3. G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer. 2013.

4. T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*. Springer. 2009.

5. Rachel Schutt and Cathy O'Neil. *Doing Data Science.* O'Reilly. 2014.