DSO530 Statistical Learning Methods

Lecture 9 : Unsupervised Learning part I

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
xint@marshall.usc.edu

- very subjective
- no gold standard of output present.

Introduction

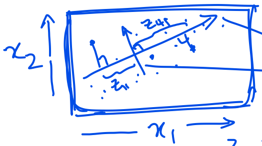
In classification, you care about test data.
In clustering, you care about how well you can group.

- We will talk about unsupervised learning topics:
  - *principal components analysis* (I): for data visualization or data pre-processing before supervised (or other unsupervised) techniques are applied
  - *clustering* (II): discovering unknown subgroups in data
- Recall the difference between supervised learning and unsupervised learning
- Name a few unsupervised learning examples?
- It is not so obvious to assess unsupervised learning
- Unsupervised learning is developing fast these days.

Introduction

- We will talk about unsupervised learning topics:
  - *principal components analysis* (I): for data visualization or data pre-processing before supervised (or other unsupervised) techniques are applied
  - *clustering* (II): discovering unknown subgroups in data
- Recall the difference between supervised learning and unsupervised learning
- Name a few unsupervised learning examples?
- It is not so obvious to assess unsupervised learning
- Unsupervised learning is developing fast these days.

Principal Components Analysis



data varies along x_1 more.
direction of data variation ①
other direction of variance ②
∴ Data varies more along ①
$z_{11}, z_{21} \ldots z_{n1}$: PCs where variability is maintained

- The principal component directions are directions in feature space along which the original data are highly variable
- Principal component analysis (PCA) refers to the process by which principal components are computed, and the subsequent use of these components in understanding the data
- If we were to do scatterplot for every pair of variables when $p = 10$, how many scatterplot do we need? $\binom{10}{2}$
- Clearly, a better method is required to visualize the n observations when p is large

y	x_1	x_2	z_1	x_3
	0	0	z_{11}	
	8	0	z_1	
	9	0	⋮	
	14	⋮	⋮	
	100	⋮	⋮	
	⋮	0	z_{n1}	

variability is necessary to be group predictor but not a necessarily one.

- First principal component of a set of features $X_1, X_2, \cdots, X_p$ is the *normalized* linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p\,, \text{ where } \sum_{j=1}^{p} \phi_{j1}^2 = 1\,.$$

- We refer to $\phi_{11}, \cdots, \phi_{p1}$ as the loadings of the first principal component.
- Given a $n \times p$ data set X, how do we compute the first principal component?
- Assume that each of the variables in X has been centered to have mean zero
- We then look for the linear combination of the sample feature values of the form

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

normalization.

that has largest sample variance, subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

- In other words, the first principal component loading vector solves the optimization problem direction that maximizes variability.

$$\max_{\phi_{11},\cdots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} z_{i1}^2 \right\}, \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1 .$$

- Need some matrix knowledge to communicate the solution. (Dr. Gilbert Strang's linear algebra course: https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/)

z_{11} → number of component
→ data pt.

- We refer to $z_{11}, \cdots, z_{n1}$ as the scores of the first principal component
- The loading vector $\phi_1 = (\phi_{11}, \cdots, \phi_{p1})^T$ defines a direction in feature space along which the data vary the most
- If we project n data points $x_1, \cdots, x_n$ on to the first PC, the projected values are $z_{11}, \cdots, z_{n1}$
- The second principal component is the linear combination of $X_1, \cdots, X_p$ that has *maximal* variance out of all linear combinations that are *uncorrelated* with the first component Z_1.

2^{nd} is orthogonal to 1^{st} PC.

- The second principal components scores $z_{12}, z_{22}, \ldots, z_{n2}$ take the form

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip}$$

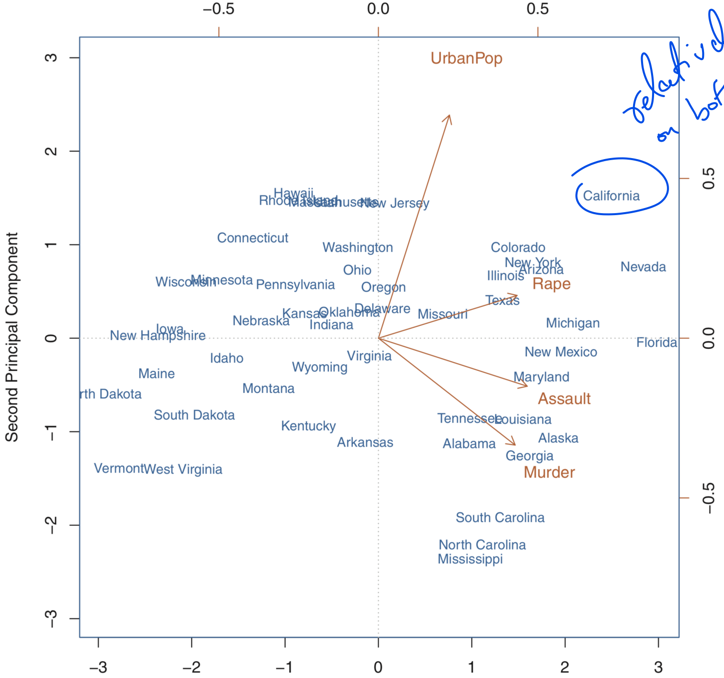
- "Z_2 and Z_1 are uncorrelated" is equivalent to "ϕ_1 is perpendicular to ϕ_2"
- The maximum number of PCs is $\boxed{\min(n-1, p)}$ (think about 2 points on a plane)

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

-ve sign of a PC is still the same direction.

TABLE 10.1. *The principal component loading vectors, ϕ_1 and ϕ_2, for the* USArrests *data. These are also displayed in Figure 10.1.*

- A *biplot* (optional): represents both pc scores and loading vectors



relatively high on both PCAs.

Another Interpretation of Principal Components (Optional)

- Principal components provide low-dimensional linear surfaces that are *closest* to the observations
- The first principal component loading vector represents the line in p-dimensional space that is closest to the n observations
- The first two principal components of a data set span the plane that is closest to the n observations
- $x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}$
- When $M = \min(n-1, p)$, $x_{ij} = \sum_{m=1}^{M} z_{im}\phi_{jm}$

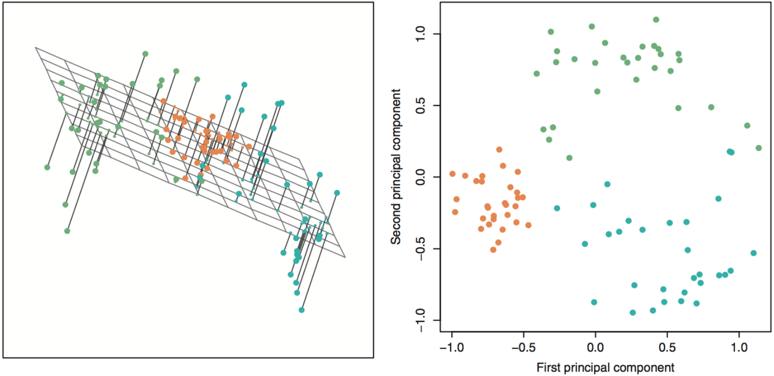


FIGURE 10.2. *Ninety observations simulated in three dimensions.* Left: *the first two principal component directions span the plane that best fits the data. It minimizes the sum of squared distances from each point to the plane.* Right: *the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane. The variance in the plane is maximized.*

- Usually we standardize variables (mean 0 and SD 1) before doing PCA

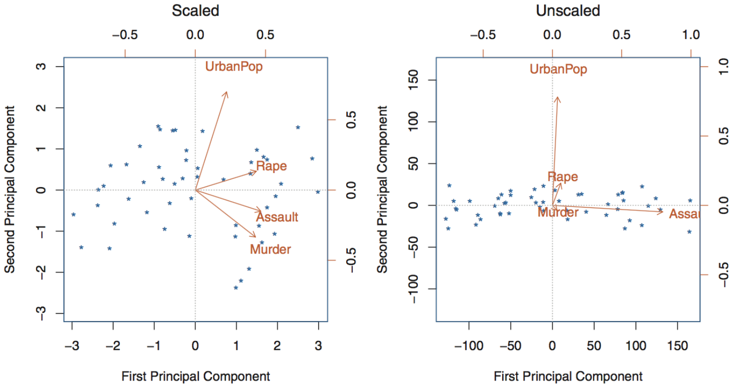


FIGURE 10.3. *Two principal component biplots for the* USArrests *data.* Left: *the same as Figure 10.1, with the variables scaled to have unit standard deviations.* Right: *principal components using unscaled data.* Assault *has by far the largest loading on the first principal component because it has the highest variance among the four variables. In general, scaling the variables to have standard deviation one is recommended.*

- In certain settings, variables may be measured in the same units. Then, we might choose not to scale the variables before PCA

exception to scaling

Proportion of Variance Explained

- We are interested in knowing the *proportion of variance explained* (PVE) by each principal component
- The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as

$$\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

- The variance explained by the mth principal component

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

- The PVE of the mth principal component is given by

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

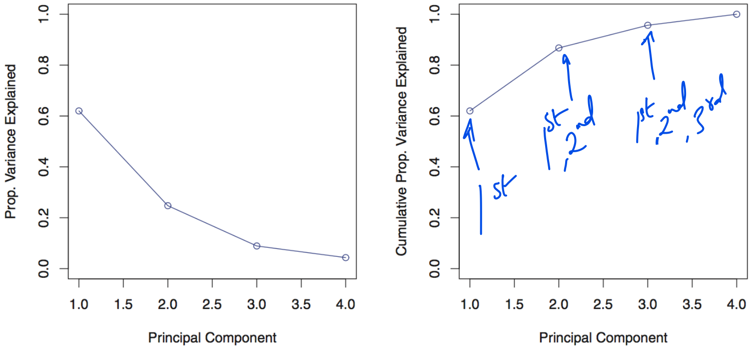


FIGURE 10.4. Left: *a scree plot depicting the proportion of variance explained by each of the four principal components in the* `USArrests` *data.* Right: *the cumulative proportion of variance explained by the four principal components in the* `USArrests` *data.*

- The question of how many principal components are enough is inherently ill-defined, and will depend on the specific area of application and the specific data set

options to decide PCs : r^2 threshold, elbow γ e, etc.