Principal Components Analysis[1]

[1] Based on materials in ISLR 10.2

Overview

- Principal component analysis (PCA) produces a **low-dimensional** representation of a dataset. It finds a sequence of linear combinations of the variables that have **maximal variance**, and are mutually **uncorrelated**.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data **visualization**.

Overview: continued

See PCA_handwritten_notes.pdf

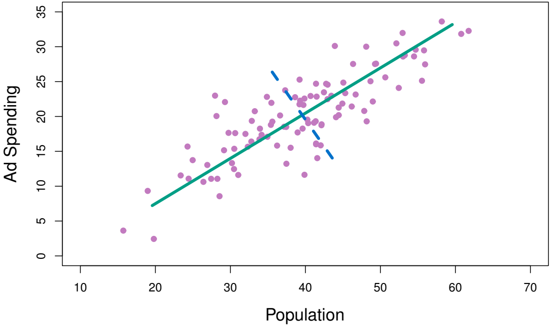
PCA: details

- The **first principal component** (PC$_1$) of a set of features $X_1, \ldots, X_p$ is the normalized linear combination of the features

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

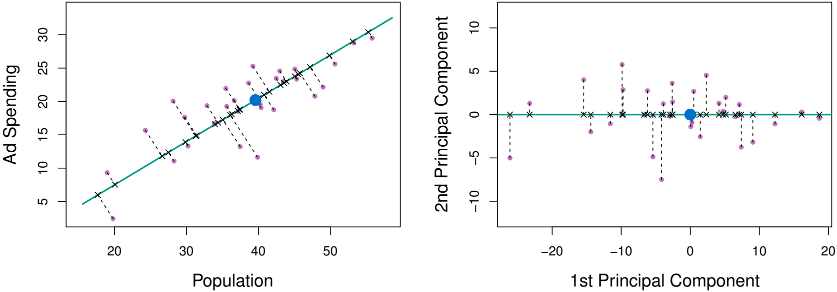
  that has the largest variance. By normalized, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.
- We refer to the elements $\phi_{11}, \ldots, \phi_{p1}$ as the **loadings** of the first principal component; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})^T$.
- We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

Example



The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

Example: continued



A subset of the advertising data. The mean pop and ad budgets are indicated with a blue circle.

Left: The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all *n* of the observations. The distances from each observation to the principal component are represented using the black dashed line segments.

Right: The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.

Computation of Principal Components

- Suppose we have a $n \times p$ data set $\mathbf{X}$. Since we are only interested in variance, we assume that each of the variables in $\mathbf{X}$ has been centered to have mean zero (that is, the column means of $\mathbf{X}$ are zero).
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{p1} x_{ip} \qquad (1)$$

  for $i = 1, \ldots, n$ that has largest sample variance, subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.
- Since each of the x_{ij} has mean zero, then so does z_{i1} (for any values of ϕ_{j1}). Hence the sample variance of z_{i1} can be written as $\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$.

Computation: continued

- Plugging in (1) the first principal component loading vector solves the optimization problem
$$\max_{\phi_{11},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^{2} \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^{2}=1$$
- This problem can be solved via a singular-value decomposition of the matrix $\mathbf{X}$, a standard technique in linear algebra.
- We refer to Z_1 as the first principal component, with realized values $z_{11},\ldots,z_{n1}$.

Geometry of PCA

- The **loading vector** ϕ_1 with elements $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ defines a direction in feature space along which the data vary the most.
- If we project the n data points $x_1, \ldots, x_n$ onto this direction, the projected values are the **principal component scores** $z_{11}, ..., z_{n1}$ themselves.

Further principal components

- The second principal component is the linear combination of $X_1, \ldots, X_p$ that has maximal variance among all linear combinations that are **uncorrelated** with Z_1.
- The second principal component scores $z_{12}, z_{22}, \ldots, z_{n2}$ take the form

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

  where ϕ_2 is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$.

Further principal components: continued

- It turns out that constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction ϕ_1. And so on.
- The principal component directions $\phi_1, \phi_2, \phi_3, \ldots$ are the ordered sequence of eigenvectors of the matrix $\mathbf{X}^T\mathbf{X}$, and the variances of the components are the eigenvalues.

Illustration

- USAarrests data: For each of the fifty states in the United States, the data set contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. We also record UrbanPop (the percent of the population in each state living in urban areas).
- The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

USAarrests data: PCA plot

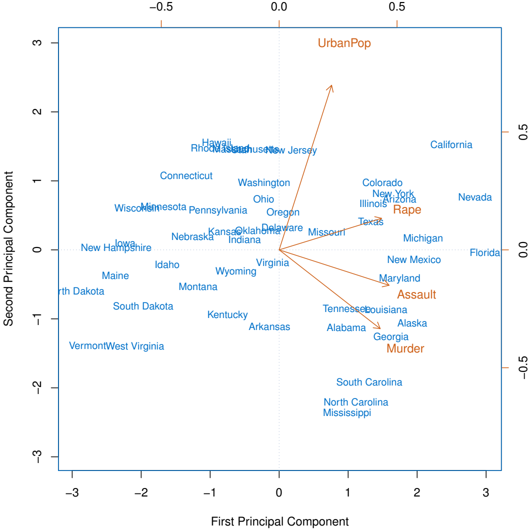


Figure details

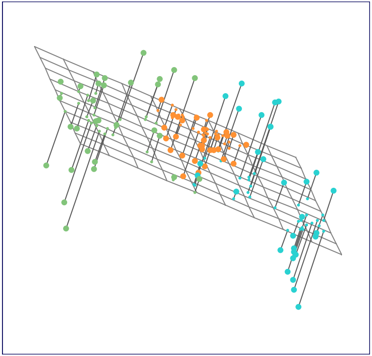
The first two principal components for the USArrests data.

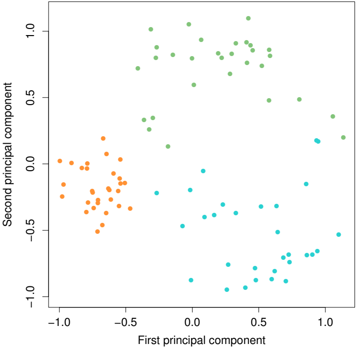
- The blue state names represent the score (z_{i1}, z_{i2}) of each observation i for the first two principal components.
- The orange arrows indicate loading vectors (ϕ_{j1}, ϕ_{j2}) of each variable j for the first two principal component with axes on the top and right. For example, the loading for Rape on the first component is 0.54, and its loading on the second principal component 0.17 [the word Rape is centered at the point (0.54, 0.17)].
- This figure is known as a **biplot**, because it displays both the principal component scores and the principal component loadings.

PCA loadings

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

Another Interpretation of Principal Components



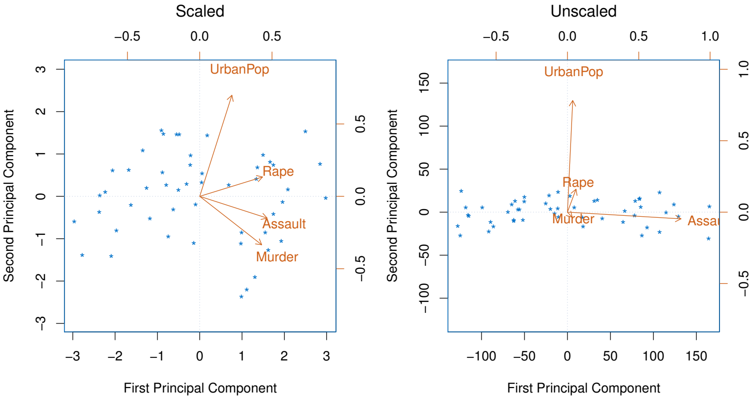


PCA find the hyperplane closest to the observations

- The first principal component loading vector has a very special property: it defines the line in p-dimensional space that is **closest** to the n observations (using average squared Euclidean distance as a measure of closeness)
- The notion of principal components as the dimensions that are closest to the n observations extends beyond just the first principal component.
- For instance, the first two principal components of a data set span the plane that is closest to the n observations, in terms of average squared Euclidean distance.

Scaling of the variables matters

- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.
- If they are in the same units, you might or might not scale the variables.



Proportion Variance Explained

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.
- The **total variance** present in a data set (assuming that the variables have been centered to have mean zero) is defined as

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

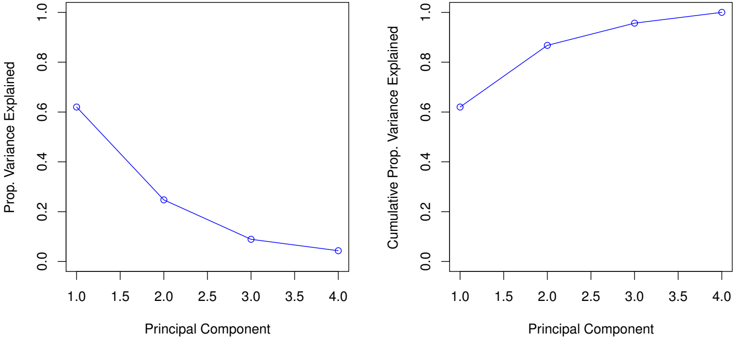
  and the variance explained by the *m*th principal component is

$$\mathrm{Var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2$$

- It can be shown that $\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{m=1}^{M} \mathrm{Var}(Z_m)$, with $M = \min(n-1, p)$.

Proportion Variance Explained: continued

- Therefore, the PVE of the *m*th principal component is given by the positive quantity between 0 and 1
$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} z_{ij}^2}$$
- The PVEs sum to one. We sometimes display the cumulative PVEs.



How many principal components should we use?

If we use principal components as a summary of our data, how many components are sufficient?

- No simple answer to this question, as cross-validation is not available for this purpose.
  - Why not?
  - When could we use cross-validation to select the number of components?
- The "scree plot" on the previous slide can be used as a guide: we look for an "elbow".

What's next?

How to use the principle components?

- Short answer: replacing the original data by the the principal component scores. We can, for example,
  - regress response variable y on the scores (also known as principle component regression);
  - classify observations based on scores;
  - cluster the scores;
  - etc...

Other dimension reduction tools

We've only discussed one linear dimension reduction technique. There are many other dimension reduction tools:

- CUR matrix approximation
- Non-negative matrix factorization (NMF)
- Kernel PCA
- T-distributed stochastic neighbor embedding (t-SNE)
- Isomap
- Autoencoder
- Self-Organizing Maps (SOM)
- Sammon mapping
- Locally Linear Embedding (LLE)