Unsupervised Learning
STAT 391 – Quantitative Introductory Statistics for Data Science

Néhémy Lim

University of Washington

Spring 2017
Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
Outline

Principle Components Analysis
In supervised learning, we have a set of \( p \) features \( X_1, \ldots, X_p \) measured on \( n \) observations with associated responses \( y_1, \ldots, y_n \).

In **unsupervised learning**, measurements do not have associated responses.

Goal is to discover interesting things about the measurements:

- Is there an informative way to visualize the data?
- Can we discover subgroups among the variables or the observations?

Examples:
- better understanding of cancers from gene expression levels in patients
- identify groups of shoppers based on their purchase histories
Outline

Principle Components Analysis
What is PCA?

- Principle components allow us to summarize a large set of correlated variables with a smaller number of representative variables.
- Principle component directions are directions along which the original data are highly variable.
- Principle components analysis (PCA) refers to the process by which principle components are computed.
- PCA is a tool for data visualization or data pre-processing before supervised techniques are applied.
Suppose we wish to visualize \( n \) observations with measurements on a set of \( p \) features, \( X_1, \ldots, X_p \).

One option is to look at two-dimensional scatterplots (impossible if \( p \) is large).

Most likely, none of them is informative.

We want to find a low-dimensional representation of data that captures as much of the information as possible.

In PCA, the information is measured by the amount that the observations vary along each dimension.

Each dimension (or principle component) found by PCA is a linear combination of the \( p \) features:

- The first principle component is the normalized linear combination:

\[
Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p
\]

that has the largest variance where \( \sum_{j=1}^{p} \phi_{j1}^2 = 1 \).

\( \phi_{11}, \ldots, \phi_{p1} \) are the loadings of the first principle component.

\( \phi_1 = (\phi_{11}, \ldots, \phi_{p1})^T \) is the loading vector of the first principle component.
Given a \( n \times p \) data set \( \mathbf{X} \), assume that each variable \( X_j \) has been centered to have mean 0.

The first principle component loading vector solves the optimization problem

\[
\begin{align*}
\text{maximize} & \quad \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \\
\text{subject to} & \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1
\end{align*}
\]

Note that the objective in (2) is the sample variance of the values of \( z_{i1} = \phi_{11} x_{i1} + \ldots + \phi_{p1} x_{ip} \).

We refer to \( z_{11}, \ldots, z_{n1} \) as the scores of the first principle component.

The loading vector \( \phi_1 \) defines a direction in feature space along which the data vary most: if we project the \( n \) data points onto this direction, the projected values are the principle component scores \( z_{11}, \ldots, z_{n1} \).
The second principle component is the linear combination of $X_1, \ldots, X_p$ with largest variance, out of all linear combinations that are uncorrelated with $Z_1$.

The second principle component scores are

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

$\phi_2 = (\phi_{12}, \ldots, \phi_{p2})^T$ is the loading vector of the second principle component.

Constraining $Z_2$ to be uncorrelated with $Z_1$ is equivalent to constraining the direction $\phi_2$ to be orthogonal to $\phi_1$.

Once we have computed the principle components, we can plot them against each other.
Figure 1: The population size (pop) and ad spending (ad) for 100 different cities (purple circles). The green solid line indicates the first principal component ($\phi_{11} = 0.839$ and $\phi_{21} = 0.544$), and the blue dashed line indicates the second principal component ($\phi_{21} = 0.544$ and $\phi_{22} = -0.839$).
Figure 2: Biplot for the USArests data. The blue state names represent the scores for the first two principal components. The orange arrows indicate the first two principal component loading vectors.
Principle components provide low-dimensional surfaces that are closest to the observations.

The first principle component is the line in $p$-dimensional space that is closest to the $n$ observations for the Euclidean distance.

Figure 3: 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. Right: The left-hand panel has been rotated so that the first principal component direction coincides with the $x$-axis.
The first two principle components span the plane that is closest to the $n$ observations.

Figure 2: Ninety observations simulated in three dimensions. Left: The plane spanned by the first two principal component direction. Right: The 90 observations projected onto that plane.
The first $M$ principal component score vectors and the first $M$ principal component loading vectors provide the best $M$-dimensional approximation (in terms of Euclidean distance) to the $i$th observation:

$$x_{ij} \approx \sum_{m=1}^{M} z_{im} \phi_{jm}$$  \hspace{1cm} (3)

When $M = \min(n - 1, p)$, the above representation is exact.
The results obtained by PCA depend on how the variables have been individually scaled.

Figure 4: Two principal component biplots for the **USArrests** data: Murder, Rape, Assault in per 100,000 people, and UrbanPop in % of the population that lives in urban area. Left: the variables were scaled to have unit standard deviations. Right: principal components using unscaled data.
Figure 5: 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.