Section 10: Dimension Reduction

(UC Davis, Spring 2026) — STA 141A: Fundamentals of Statistical Data Science

Instructor: Akira Horiguchi

Overview

Based on Chapter 12 of ISL book James et al. (2021). For more R code examples, see R Markdown files in book website.

Dimension reduction

Principal component analysis (PCA)

Dimension reduction

Principal component analysis (PCA)

Dimension reduction

Suppose we have n observations on a set of p features $X_1, X_2, \ldots, X_p$.

- That is, suppose we have n data points

$$\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1p})^\top, \quad \mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2p})^\top, \quad \ldots, \quad \mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{np})^\top.$$

- *Dimension reduction:* reduce dimensionality of data while retaining as much information about the data as possible.
- Idea: not all p dimensions are equally interesting. If the jth feature has almost the same value for all n observations, do we really need to keep track of the jth feature?

There are many approaches to dimension reduction: this course will consider only PCA.

Dimension reduction: some applications

1. Visualize high-dimensional data ($p \gg 2$) as 2-dimensional (scatter)plots of data points.
2. Image compression (reduce computer memory storage requirements).

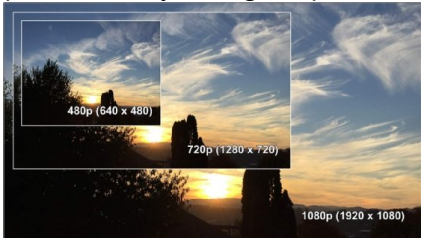

Figure: Image from techtarget.com

3. Learning lower-dimensional ($\ll p$) latent structures of images:
   - Classifier: handwritten digits (MNIST dataset).
   - Generative model: a variational autoencoder (VAE) uses *Encoder*-Decoder architecture.
     - The *Encoder* portion compresses high-dimensional data (e.g., thousands of features) into a low-dimensional "latent space."
     - The Decoder portion takes a sampled latent representation and attempts to reconstruct the original high-dimensional data from it.

Scaling the variables

We are usually interested in analyzing the impact of certain features in relation to their variation.

- The variance of features can be large solely because their values are large.
- Not an issue if features are measured in the same units.
- So, by scaling (by their standard deviation), the variation among all predictors is comparable, independently of their magnitude.
- In general, scaling the variables to have standard deviation one is recommended.

General linear algebra idea: orthogonal

Consider two p-dimensional vectors $\mathbf{a} = (a_1, \ldots, a_p)$ and $\mathbf{b} = (b_1, \ldots, b_p)$.

- The *dot product*

$$\mathbf{a} \cdot \mathbf{b} \coloneqq \sum_{j=1}^{p} a_j b_j \quad \text{which also equivalent to the matrix multiplication} \quad \mathbf{a}^\top \mathbf{b}$$

  measures how much one vector "points" in the same direction as another.
- The vectors $\mathbf{a}$ and $\mathbf{b}$ are called *orthogonal* or *perpendicular* to each other if and only if their dot product $\mathbf{a} \cdot \mathbf{b}$ is zero.

General linear algebra idea: linear combination

A *linear combination* of the original p features $X_1, X_2, \ldots, X_p$ is defined as

$$\phi_1 X_1 + \phi_2 X_2 + \cdots + \phi_p X_p$$

for some *coefficient* values $\phi_1, \phi_2, \ldots, \phi_p$.

- Such a linear combination can be represented by the vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)$.
- We can think of each ϕ_j as the contribution of X_j to the linear contribution.
- A linear combination is called *normalized* if $\|\boldsymbol{\phi}\|_2^2 = 1$. (also called a *unit vector*)

Principal component analysis (PCA)

PCA – Idea

PCA seeks a small number of dimensions that are as "interesting" as possible.

- "Interesting" is measured by how much the observations vary along the dimension.
- Each dimension (i.e., *principal component (PC)*) found by PCA is a *linear combination* of the original p features:

$$\phi_1 X_1 + \phi_2 X_2 + \cdots + \phi_p X_p$$

- Coefficients $\phi_1, \phi_2, \ldots, \phi_p$ of a PC are called its *loadings*;
  the normalized vector $\boldsymbol{\phi} \coloneqq (\phi_1, \phi_2, \ldots, \phi_p)$ is called the PC's *loading vector*.
- PC loading vectors, by construction, are always orthogonal to each other.
- PCA computes p orthogonal PC loading vectors in decreasing order of "interestingness."

Let's start by constructing the first PC.

First principal component: how to compute?

Given an $n \times p$ data matrix $\boldsymbol{X}$, how to compute the *first PC*?

- Because we are only interested in variance, assume that each column in $\boldsymbol{X}$ has mean zero.
- For any (normalized) vector $\boldsymbol{\phi}_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1}) \in \mathbb{R}^p$, we can project each data point $\mathbf{x}_i \in \mathbb{R}^p$ onto the direction given by the unit vector $\boldsymbol{\phi}_1$:

$$z_{i1} = \boldsymbol{\phi}_1^\top \mathbf{x}_i = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots \phi_{p1} x_{ip}, \qquad \text{for } i = 1, \ldots, n. \tag{1}$$

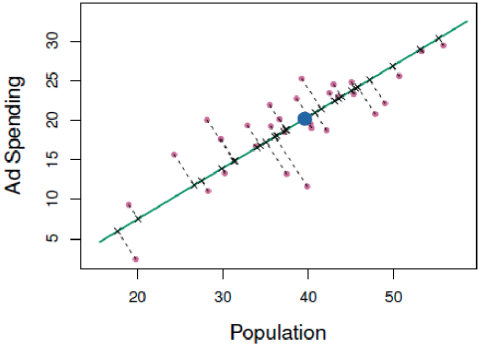

Figure: Figure by James et al. (2021). A subset of the advertising data ($p = 2$). Each data point $\mathbf{x}_i$ is displayed as a red point. The mean pop and ad budgets are indicated with a blue circle.

First principal component: how to compute?

- The first PC loading vector is defined to be the *normalized* vector $\boldsymbol{\phi}_1$ that maximizes

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \,. \tag{2}$$

- *Interpretation*: Each column of $\boldsymbol{X}$ is centered, so the mean of $z_{11}, z_{21}, \ldots, z_{n1}$ is also zero. Hence (2) is (almost) just the sample variance of $z_{11}, z_{21}, \ldots, z_{n1}$.
  Hence we are trying to find the direction $\boldsymbol{\phi}_1 \in \mathbb{R}^p$ along which the data vary the most.
- (Normalization ensures the variance is not arbitrarily large.)
- Calculation: (2) can be maximized via *eigen decomposition* or *singular value decomposition*. (See lab, but you don't need to know this for exam).

First principal component: geometric interpretation

The 1st PC loading vector defines a direction in feature space along which the data vary the most.

- For all $i = 1, \ldots, n$, the data point $\mathbf{x}_i$ projected onto this direction gives the PC score z_{i1}.

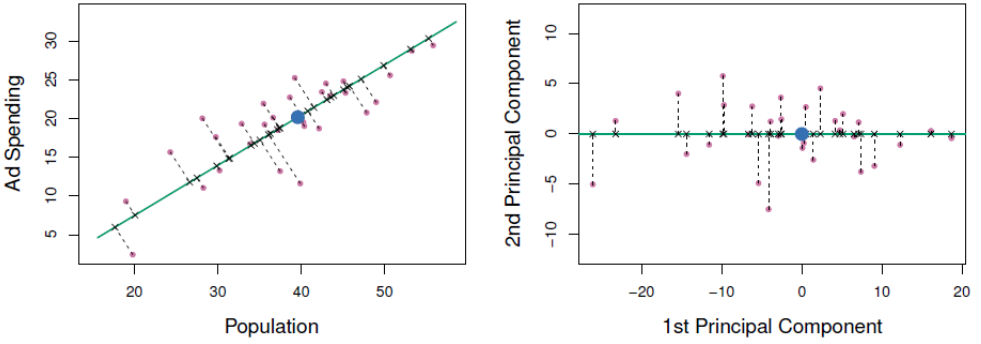

Figure: Figure by James et al. (2021). A subset of the advertising data. The mean pop and ad budgets are indicated with a blue circle. Left: The 1st PC direction (in green). It is the dimension along which the data vary the most, and it defines the line that is closest to all n observations. Right: The left-hand panel has been rotated so that the 1st PC direction coincides with the x-axis.

Remaining principal components

Suppose we have computed the first PC loading vector $\boldsymbol{\phi}_1$. *How to compute the 2nd PC vector?*

- PCA computes the second PC by finding the normalized vector $\boldsymbol{\phi}_2$ that maximizes

$$\frac{1}{n}\sum_{i=1}^{n} z_{i2}^2 \;\coloneqq\; \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\phi}_2^\top \mathbf{x}_i\right)_{i2}^2, \tag{3}$$

  but now this vector $\boldsymbol{\phi}_2$ must be orthogonal to the vector $\boldsymbol{\phi}_1$.
- If $p = 2$, then there is only one direction orthogonal to $\boldsymbol{\phi}_1$, and we immediately obtain $\boldsymbol{\phi}_2$.
- If $p > 2$, there are infinitely many directions orthogonal to $\boldsymbol{\phi}_1$.

How to compute the remaining PC vectors? Third, fourth, and so on.

- Once $\boldsymbol{\phi}_2$ is computed, PCA finds $\boldsymbol{\phi}_3$ by maximizing (2) with the additional constraint that $\boldsymbol{\phi}_3$ has to be orthogonal to both $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_1$.
- And so on to find $\boldsymbol{\phi}_4$, $\boldsymbol{\phi}_5, \ldots, \boldsymbol{\phi}_p$.

Geometric interpretation: alternative interpretation

PCs provide low-dimensional linear surfaces that are closest to the n observations.

- The first PC loading vector is the line in p-dimensional space closest to the n observations.
- The first two PC loading vectors span the plane closest to the n observations.
- The first three PC loading vectors span the 3-dim hyperplane closest to the n observations.

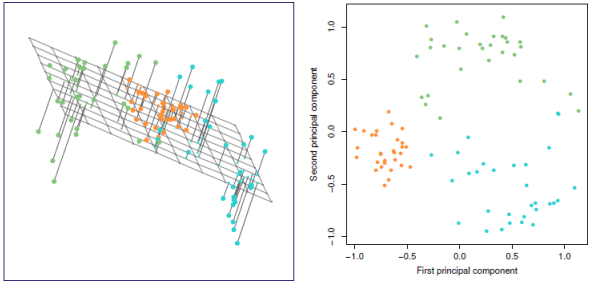

Figure: Figure by James et al. (2021). Observations simulated in three dimensions, displayed in color for ease of visualization. Left: The first two PC directions span the plane that best fits the data (in sense of minimizing RSS). Right: The first two PC score vectors give the coordinates of the projection of the observations onto the plane.

Geometric interpretation: alternative interpretation

For any positive integer $M \leq \min\{n-1, p\}$, together the first M PC score vectors and the first M PC loading vectors provide an M-dimensional approximation to the ith observation:

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}. \tag{4}$$

How good is this approximation? It is the best, in the sense of the following optimization problem.

- Suppose we have a data matrix $\boldsymbol{X}$ that is column-centered, and have chosen a value of M.
- For any M, consider which values $\{a_{im}\}$ and $\{b_{jm}\}$ minimize the sum of squares

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \hat{x}_{ij}^{(M)}\right)^2, \qquad \text{where} \quad \hat{x}_{ij}^{(M)} = \sum_{m=1}^{M} a_{im}b_{jm}. \tag{5}$$

- It can be shown that for any value of M, the minimizers of (5) are $a_{im} = z_{im}$ and $b_{jm} = \phi_{jm}$.
- When $M = \min\{n-1, p\}$, the representation (4) becomes exact.

The proportion of variance explained (PVE)

How much information is retained by projecting onto a few PCs? That is, how much variance in the data is contained in PCs? Here, assume centered data as usual.

- The *total variance* in the data can be decomposed into the sum of two terms:

$$\underbrace{\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2}_{\text{Variance of data}} = \underbrace{\sum_{m=1}^{M} \frac{1}{n} \sum_{i=1}^{n} z_{im}^2}_{\text{Var. of first } M \text{ PCs}} + \underbrace{\frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_{ij} - \sum_{m=1}^{M} z_{im} \phi_{jm} \right)^2}_{\text{MSE of } M\text{-dimensional approximation}} \tag{6}$$

- The *proportion of variance explained (PVE)* of the mth PC is the quotient

$$PVE_m = \frac{\text{variance explained by the } m\text{th PC}}{\text{variance of data}} = \frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}. \tag{7}$$

- The proportion of variance explained by the first M PCs is (rearrange terms in (6))

$$\sum_{m=1}^{M} PVE_m = \frac{\text{variance explained by the first } M \text{ PCs}}{\text{variance of data}} = 1 - \frac{RSS}{TSS}, \tag{8}$$

where TSS is total sum of squared elements of $\boldsymbol{X}$, and RSS is residual sum of squares of the M-dimensional PC approximation.

PVE is usually visualized with the *screeplot*

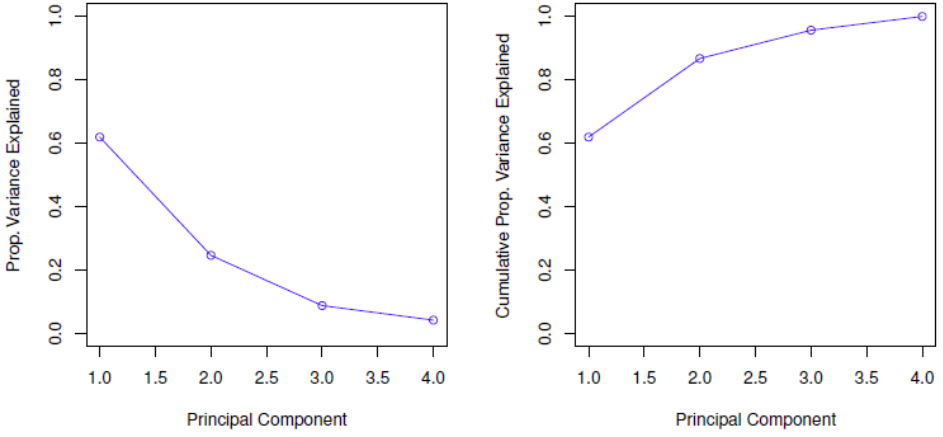

Figure: Figure by James et al. (2021). Left (7): A scree plot depicting the proportion of variance explained (PVE) by each of four principal components in `USArrests` data. Right (8): The cumulative PVE by four principal components.

Number of predictors

Usually, we only want a few dimensions in order to better visualize or understand the data.

- How many principal components do we need? How can we justify using only e.g., three PCs vs four PCs vs five PCs?
- There is no simple answer to this! There is no formula that can be applied universally that gives us the optimum value.
- We can intuitively decide to choose the number of predictors by eyeballing the *screeplot* which depicts the proportion of variance explained (PVE).
- In the screeplot, we look for a point at which the PVE by each subsequent principal component drops significantly off (this is subjective). Such a drop is often referred to as an *elbow* in the screeplot, and the rule by choosing the point is thus called *elbow rule*.

Other uses for PCA

Many statistical techniques, such as regression, classification, and clustering, use the full $n \times p$ data matrix.

- Can instead use the $n \times M$ matrix whose columns are the first $M \ll p$ PC score vectors.
- This can lead to less noisy results, since often the signal in a data set is concentrated in its first few principal components.