STA 35C: Statistical Data Science III

Lecture 23: Principal Component Analysis (cont'd)

Dogyoon Song

Fall 2026, UC Davis

Announcements

Homework 7 is posted

- Due Thu, Dec 3, 11:59 PM PT
- Please submit on time and follow the submission instructions
- Review the problems early so you have time to ask questions

Agenda

Quick recap: Principal component analysis (PCA) Overview & intuition

- Objective: dimension reduction while retaining as much variance as possible
- Intuition: projection that retains maximum variance

Today: Formalizing PCA and interpreting its output

- PCA formulation and properties
  - General formulation
  - PCA as a change of basis
    - Principal component scores
  - Proportion of variance explained
    - Choosing number of PCs via scree plot
  - (`Optional`) Additional details (scaling, uniqueness, etc.)
- Example applications

Recap: Unsupervised learning

Two branches of statistical learning:

- *Supervised learning*
  - Setup/goal: We observe (X, Y) and want to learn a function $f : X \to Y$
  - Examples: regression, classification, ...
- *Unsupervised learning*
  - Setup/goal: We observe only X (no Y) and aim to discover patterns or structures in X
  - Examples:
    - PCA: find a few directions that capture most variation (=information) in the data
    - Clustering: identify subgroups (clusters) among observations

Why unsupervised learning?

- We may have data only on features X; or we want to do exploratory analysis
- Often a preliminary step before supervised tasks

Recap: PCA overview & intuition

Problem Setup:

- We have data of $X \in \mathbb{R}^p$, where p is possibly large
- We want to reduce dimension to $r \ll p$ while retaining most "information" in data

PCA approach:

- Project data (X) onto an r-dimensional subspace (spanned by r vectors)
- These r vectors (=PCs) are chosen to capture maximum variance in X
- Unsupervised learning: no Y is used

Outcome:

- A few linear combinations of $X_1, \ldots, X_p$ that explain most variation
- Useful for dimension reduction, model interpretation, and data visualization

Recap: PCA illustration 1: $p = 2$ to $r = 1$

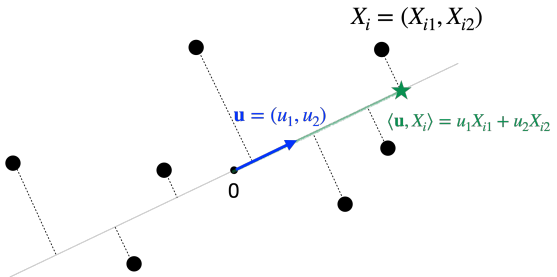


Figure: PCA finds the direction **u** such that the projected scores have the largest variance.

PCA as a 2D $\rightarrow$ 1D projection:

- Each centered data point $X_i = (x_{i1}, x_{i2})$ is mapped to $\langle \mathbf{u}, X_i \rangle = u_1 x_{i1} + u_2 x_{i2}$
- PCA picks **u** (with $\|\mathbf{u}\| = 1$) that *maximizes* the variance of $\langle \mathbf{u}, X_i \rangle$, $\frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{u}, X_i \rangle^2$
- Geometrically, the "major axis" of the data cloud is identified

PCA formulation: First principal component

Assumption: The data are centered:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = 0.$$

For any unit direction $\mathbf{u} \in \mathbb{R}^p$, the projected **score** of observation i is

$$z_i(\mathbf{u}) = \langle \mathbf{u}, \mathbf{x}_i \rangle.$$

The empirical variance of the projected scores is

$$\frac{1}{n}\sum_{i=1}^{n} \left(z_i(\mathbf{u}) - \bar{z}(\mathbf{u})\right)^2 = \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{u}, \mathbf{x}_i \rangle^2.$$

First principal component direction:

$$\mathbf{u}_1 = \underset{\|\mathbf{u}\|=1}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{u}, \mathbf{x}_i \rangle^2.$$

PCA formulation: Later principal components

Second PC direction:

$$\mathbf{u}_2 = \underset{\substack{\|\mathbf{v}\|=1 \\ \langle \mathbf{v}, \mathbf{u}_1 \rangle = 0}}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{x}_i \rangle^2.$$

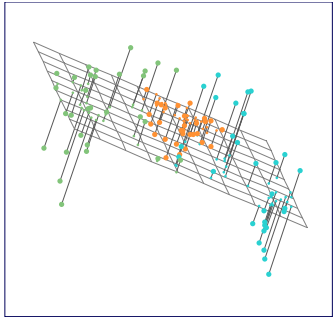
- The second PC captures the largest remaining variance subject to being orthogonal to $\mathbf{u}_1$
- Subsequent PCs are defined similarly, each orthogonal to all previous PC directions
- The resulting PC scores are empirically uncorrelated: e.g., $\sum_{i=1}^{n} z_i(\mathbf{u}_1)\, z_i(\mathbf{u}_2) = 0$.

Terminology:

- **Loadings**: entries of the PC direction vector $\mathbf{u}_k$, telling how much original variables contribute to PC k
- **Scores**: projected coordinates $z_{ik} = \langle \mathbf{u}_k, \mathbf{x}_i \rangle$, telling where observation i lies along PC k

(`Optional`) The PC directions are the eigenvectors of the sample covariance matrix

PCA illustration: $p = 3$ reduced to $r = 2$



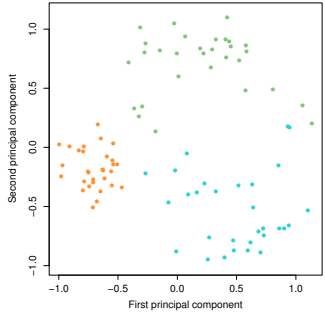


Figure: Ninety observations in $\mathbb{R}^3$. **Left:** The first two PC directions span a plane that best fits the data, minimizing total squared distance. **Right:** Data are "flattened" onto that 2D plane, forming their PC scores [JWHT21, Figure 12.2].

Key idea in higher dimension:

- Find a subspace of dimension r that capture maximal variance (=minimizing residuals)
- $\mathbf{u}_1$ is the top PC direction, $\mathbf{u}_2$ is second, etc., each orthogonal

PCA as a change of basis

If $\mathbf{u}_1, \ldots, \mathbf{u}_p$ are the PC directions, then the k-th PC score for observation i is

$$z_{ik} = \langle \mathbf{u}_k, \mathbf{x}_i \rangle = \sum_{j=1}^{p} u_{jk} x_{ij}.$$

Interpretation:

- PCA rotates the coordinate system

$$(X_1, \ldots, X_p) \quad \longrightarrow \quad (Z_1, \ldots, Z_p)$$

- The new variables $Z_1, Z_2, \ldots$ are ordered by how much variance they capture
- For dimension reduction, keep only the first r scores:

$$(Z_1, \ldots, Z_r), \qquad r \ll p.$$

(`Optional`) **Matrix form:** If $U = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$, then $Z = XU$.

Illustration: PCA as a rotated coordinate system

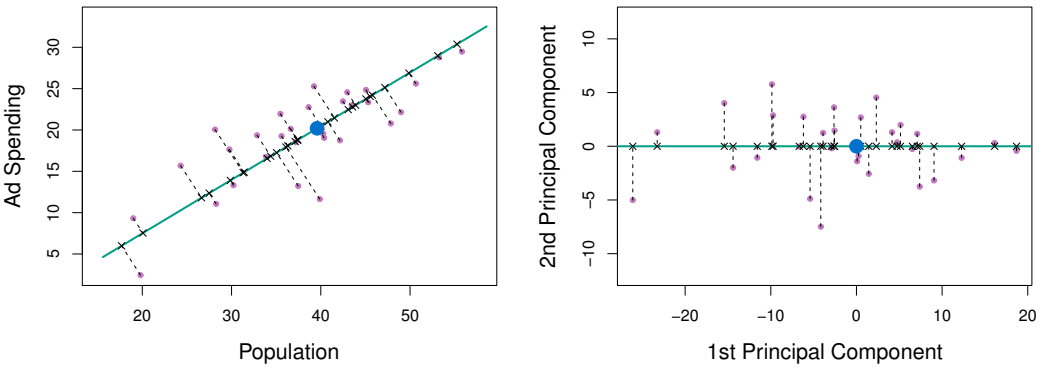


Figure: A subset of the `advertising` data, with mean population (`pop`) and advertising spending (`ad`) shown as a blue circle. **Left:** The first principal component direction (green) captures the greatest data variation and defines the line that best fits all observations (distances shown by dashed lines). **Right:** The plot is rotated so that this principal component aligns with the horizontal x-axis. [JWHT21, Figure 6.15].

Pop-up quiz: PCA basics

Question: Which statement about PCA is **false**?

A) PCA is unsupervised and uses only X, not Y.

B) The first principal component direction maximizes the variance of the projected data.

C) The second principal component direction is found with no constraints.

D) PCA can reduce dimension by keeping only the first few PC scores.

Answer: C. The second PC direction must be orthogonal to the first PC direction; subsequent PC directions are constrained to be orthogonal to all earlier ones.

Follow-up: If software reports $\mathbf{u}_1 = (0.6, 0.8)$ but another package reports $(-0.6, -0.8)$, which output is wrong?

Answer: Neither. PC directions are unique only up to sign; the projected scores also flip sign.

Proportion of variance explained (PVE)

Question: If we only keep r PCs, how much total variance remains?

- For centered data, **total variance** is

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

- The variance explained by the k-th PC is

$$\mathrm{Var}(\langle \mathbf{u}_k, X\rangle) = \frac{1}{n}\sum_{i=1}^{n} z_{ik}^2 \quad \text{where} \quad z_{ik} = \langle \mathbf{u}_k, \mathbf{x}_i\rangle$$

- The **proportion of variance explained** (PVE) by the k-th PC is

$$\mathrm{PVE}_k = \frac{\sum_{i=1}^{n} z_{ik}^2}{\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2} = \frac{\sum_{i=1}^{n}\|z_{ik}\mathbf{u}_k\|^2}{\sum_{i=1}^{n}\|\mathbf{x}_i\|^2}$$

- The **cumulative PVE** for the first r PCs is

$$\mathrm{PVE}_{1:r} = \sum_{k=1}^{r}\mathrm{PVE}_k = 1 - \frac{\sum_{i=1}^{n}\left\|\mathbf{x}_i - \sum_{k=1}^{r} z_{ik}\mathbf{u}_k\right\|^2}{\sum_{i=1}^{n}\|\mathbf{x}_i\|^2} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

Example: Proportion of variance explained (1/2)

Example

Dataset: Let $X \in \mathbb{R}^3$. Suppose we have five centered points:

$$\mathcal{X} = \big\{ (0,0,0),\ (0,-1,0),\ (0,1,0),\ (0,0,-3),\ (0,0,3) \big\}.$$

One can verify $\sum_i \mathbf{x}_i = (0,0,0)$, so these are already mean-centered.

Step 1: Compute total variance.

$$\begin{aligned} \mathrm{Var}(X) &= \frac{1}{5} \sum_{i=1}^{5} \|\mathbf{x}_i\|^2 = \frac{1}{5} \sum_{i=1}^{5} \left(x_{i1}^2 + x_{i2}^2 + x_{i3}^2 \right) \\ &= \frac{1}{5} \left(0^2 + (-1)^2 + 1^2 + (-3)^2 + 3^2 \right) = \frac{1+1+9+9}{5} = \frac{20}{5} = 4. \end{aligned}$$

Step 2: Identify the first principal component.
A simple inspection shows the direction of greatest variance is along the z-axis:

$$\mathbf{u}_1 \ = \ (0,\ 0,\ 1).$$

Indeed, points $(0,0,\pm 3)$ have the largest spread among the three coordinates.

Example: Proportion of variance explained (2/2)

Example

Step 3: Variance along $\mathbf{u}_1$ and PVE. Since $\mathbf{u}_1 = (0, 0, 1)$, the first PC score of $\mathbf{x}_i$ is equal to x_{i3}.

$$\mathrm{Var}\big(\mathbf{u}_1 \cdot X\big) \;=\; \frac{1}{5}\sum_{i=1}^{5}\big(\langle \mathbf{u}_1, \mathbf{x}_i\rangle\big)^2 \;=\; \frac{1}{5}\sum_{i=1}^{5}(x_{i3})^2 \;=\; \frac{1}{5}\big(0^2 + 0^2 + 0^2 + (-3)^2 + 3^2\big) \;=\; \frac{18}{5} \;=\; 3.6.$$

Hence the proportion of variance explained by the first PC is

$$\mathrm{PVE}_1 \;=\; \frac{3.6}{4.0} \;=\; 0.9 \quad \text{(i.e., 90\% of total variance).}$$

Additional remarks.

- Similarly, we can verify that the second PC direction is $\mathbf{u}_2 = (0, 1, 0)$.
- Hence,

  $$\mathrm{PVE}_2 \;=\; \frac{0.4}{4.0} \;=\; 0.1 \quad\Longrightarrow\quad \mathrm{PVE}_{1:2} = \mathrm{PVE}_1 + \mathrm{PVE}_2 = 1.$$

  That is, all information about the dataset $\mathcal{X}$ is explained by the first two PC scores.

PVE: A small numerical example

Suppose PCA gives the following proportions of variance explained:

PC	1	2	3	4
PVE_k	0.55	0.25	0.12	0.08

Cumulative PVE:

PCs kept	1	1:2	1:3	1:4
Cumulative PVE	0.55	0.80	0.92	1.00

Interpretation:

- Keeping 2 PCs preserves 80% of the total variance
- Keeping 3 PCs preserves 92% of the total variance
- The choice of r depends on the goal: visualization, compression, or downstream analysis

Scree plot: PVE vs. number of PCs

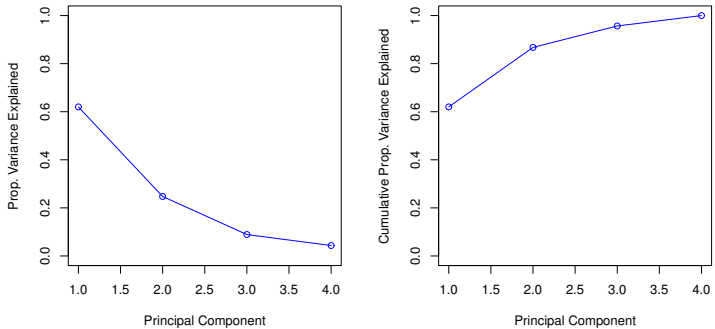


Figure: A scree plot for the `USArrests` data. **Left:** proportion of variance explained by each PC. **Right:** cumulative PVE [JWHT21, Figure 12.3].

How to read a scree plot:

- Look for an "elbow" where additional PCs add little variance
- Alternatively, choose r so cumulative PVE exceeds a target, such as 80% or 90%
- There is no universal best cutoff

Scree plot: How many PCs to retain?

Trade-off:

- Smaller r: easier visualization and interpretation, more compression
- Larger r: retains more variation, less information loss

Typical criteria:

- Keep enough PCs to explain a large fraction of variance
- Choose r so the **cumulative PVE** is "high enough," or identify an "elbow" in the scree plot
- No universal formula for the "best" r

Important caveat:

- PCA preserves variation in X, not necessarily information most relevant for a response Y

Choosing the number of PCs using scree plot example

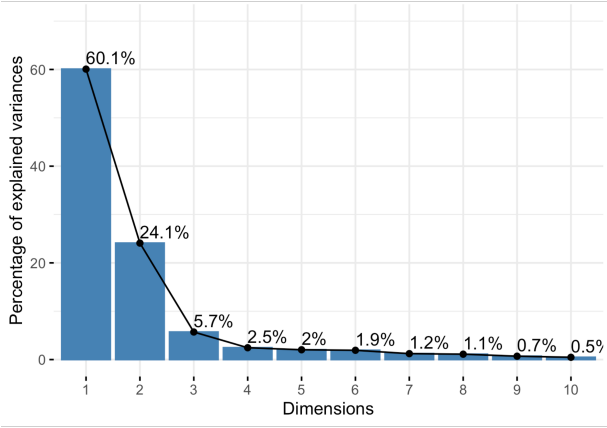


Figure: A scree plot from `mtcars` dataset in R. The elbow appears to occur at the third principal component, which suggests keeping the first three components (source: Statistics Globe).

Trade-off:

- Smaller dimension r is easier to interpret and visualize
- Larger r retains more variance in the data

(`Optional`) Additional PCA details

Scaling variables?

- If predictors have very different scales (e.g. `height` in cm vs. `income` in $), standardizing them to unit variance can drastically alter PCA directions
- Whether to scale depends on context: retain raw scales when meaningful, or standardize to give each variable unit variance before PCA

Uniqueness:

- Principal component directions are unique up to a sign ($\mathbf{u}$ vs. $-\mathbf{u}$)
- This sign usually does not affect interpretation, so software packages pick a sign convention automatically

Computation:

- Solve for eigenvectors/eigenvalues of the sample covariance (or correlation) matrix
- **In R:** `prcomp(..., scale=TRUE)` or `princomp(...)`

Pop-up quiz: PVE and number of PCs

Suppose the first five PCs have PVE:

PC	1	2	3	4	5
PVE_k	0.45	0.25	0.15	0.10	0.05

Question: What is the smallest number of PCs needed to explain at least 80% of the total variance?

A) 1
B) 2
C) 3
D) 4

Answer: C. The cumulative PVE is 0.45, 0.70, 0.85, so the first 3 PCs are enough to exceed 80%.

PCA application in high-dimensional genomics

Example: Genomics data [NJB+08]

- 1,387 individuals from Europe
- Each individual has genotype measurements at 197,146 loci
- PCA reduces the data from nearly 200,000 variables to a few principal components
- The first two PCs reveal meaningful geographic structure

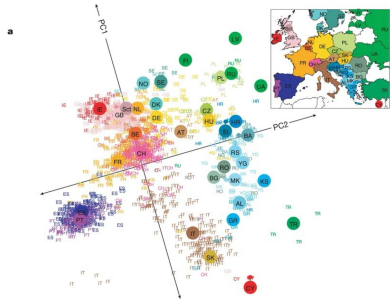


Figure: First two PCs of genetic variation among 1,387 Europeans. Small colored points are individuals; large dots mark country medians in PC1–PC2 space [NJB+08, Figure 1a].

PCA application in image compression

Example: Compressing a grayscale image via PCA

- Original image has 372×492 pixels, each a grayscale intensity in $[0, 255]$
- Partition the image into 12×12 blocks, so each block is a $12 \times 12 = 144$-dimensional "vector"
- There are $N = \frac{372}{12} \times \frac{492}{12} = 1271$ such vectors (observations)
- PCA approximates each block using only the first r PC scores for the dimension reduction

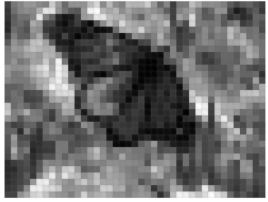


Figure: Compressing an image by PCA. **Left:** original image. **Right:** PCA rank-1 approximation. With $r = 1$, almost all details are lost, but the main global contrast is still visible.

PCA application in image compression (cont'd)



Original



Rank-1 PCA ($r = 1$)



Rank-3 PCA ($r = 3$)



Rank-6 PCA ($r = 6$)



Rank-16 PCA ($r = 16$)



Rank-60 PCA ($r = 60$)

Figure: PCA-based image compression. Larger r yields better reconstruction quality.

PCA: Limitations and cautions

- **PCA is unsupervised:** It preserves variation in X, not necessarily information relevant to a response Y.
- **Scaling matters:** Variables with larger scale or variance can dominate the first PCs.
- **Interpretability can be difficult:** PCs are linear combinations of many original variables.
- **PCA captures linear structure:** Nonlinear structure may require other methods.
- **Outliers can matter:** Extreme observations can strongly influence directions of maximum variance.

Wrap-up: Takeaways

Principal Component Analysis (PCA):

- Finds a few PC directions that capture maximum variance in the data
- The first few PCs often capture most of the total variation, enabling dimension reduction
- PCA is *unsupervised*, commonly used for exploratory analysis or as a pre-processing step

Proportion of Variance Explained (PVE):

- Cumulative PVE quantifies how much of the total variance is retained by the first r PCs
- A scree plot of PVE vs. PC index can guide how many PCs to keep

Additional remarks:

- In R, use `prcomp(...)` or `princomp(...)`
- Predictor scaling can affect PCA
- Once you learn linear algebra & eigendecomposition, the definitions and details of PCA will become much clearer

Suggested reading: [JWHT21, Ch. 12.2]

Next lecture: Clustering: K-means

References



Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.

An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.

Springer, New York, NY, 2nd edition, 2021.



John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al.

Genes mirror geography within europe.

Nature, 456(7218):98–101, 2008.