STA 35C: Statistical Data Science III

Lecture 23: Principal Component Analysis (cont'd)

Dogyoon Song

Spring 2025, UC Davis

Announcement

Final exam on Fri, June 6 (1:00 pm–3:00 pm) in classroom

- **Be on time**: The exam starts at 1:00 pm and ends at 3:00 pm sharp
- **Three hand-written cheat sheets allowed:** Letter-size (8.5"$\times$11"), double-sided, brief formulas/notes
- **Calculator**: A simple (non-graphing) scientific calculator is allowed
- **No other materials** beyond the cheat sheets (no textbooks, etc.)
- **SDC accommodations:** Confirm scheduling to take on Thu, June 5 with AES *ASAP*

Preparation:

- The exam is *cumulative* (Lectures 1–25)
- A practice final exam and brief answer key will be provided on the course webpage
- Office hours this week:
  - Instructor: Wed, 4:30–5:30pm ; Thu, 2:30–3:00pm
  - TA: Mon/Thu 1–2pm

Today's topics

Principal component analysis (PCA)

- **Overview & intuition**
  - Objective: dimension reduction with minimal information loss
  - Intuition: projection that retains maximum variance
- **Formalism & properties**
  - Principal components (PCs)
  - PCA as a change of basis
  - Proportion of variance explained
  - Choosing number of PCs via scree plot
  - (`Optional`) Additional details (scaling, uniqueness, etc.)
- **Applications of PCA**

Quick review: 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 within 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

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"

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

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

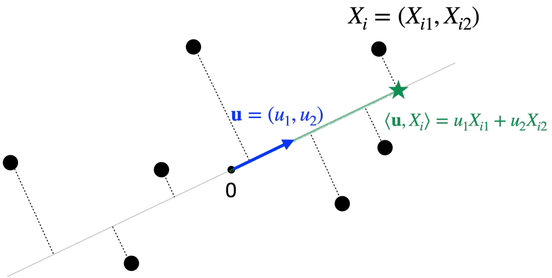


Figure: For a unit vector $\mathbf{u} = (u_1, u_2)$, consider the orthogonal projection of X_i onto the line spanned by $\mathbf{u}$, which is $\langle \mathbf{u}, X_i \rangle = u_1 X_{i1} + u_2 X_{i2}$. PCA finds the direction $\mathbf{u}$ maximizing the variance of projection $\langle \mathbf{u}, X_i \rangle$.

2D $\rightarrow$ 1D projection:

- Each 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 $\mathbf{u}$ (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 example: $p = 2$ data to $r = 1$

Example

Consider a small 2D dataset:

$$\mathcal{X} = \{ (-2, -1), (0, 0), (2, 1) \}.$$

These three points lie on the line spanned by $(2, 1)$.

Projection: For a unit vector $\mathbf{u} = (u_1, u_2)$, the projection of $X_i = (x_{i1}, x_{i2})$ onto (the line spanned by) $\mathbf{u}$ is

$$\langle \mathbf{u}, X_i \rangle = u_1 x_{i1} + u_2 x_{i2}.$$

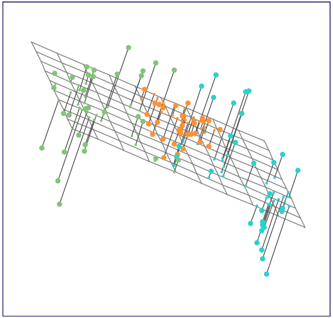
Key idea: PCA finds $\mathbf{u}$ that maximizes the variance of these projected values $\langle \mathbf{u}, X_i \rangle$.

Observe that

- If $\mathbf{u} = (1, 0)$, the variance in this direction is $\frac{1}{3}\left((-2)^2 + 0^2 + 2^2\right) = \frac{8}{3}$.
- If $\mathbf{u} = (0, 1)$, the variance in this direction is $\frac{1}{3}\left((-1)^2 + 0^2 + 1^2\right) = \frac{2}{3}$.
- If $\mathbf{u} = \frac{1}{\sqrt{5}}(2, 1)$, the variance in this direction is $\frac{1}{3}\left((-\sqrt{5})^2 + 0^2 + \sqrt{5}^2\right) = \frac{10}{3}$ (the maximum).

Hence $\mathbf{u}^* = \frac{1}{\sqrt{5}}(2, 1)$ is the PCA direction.

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



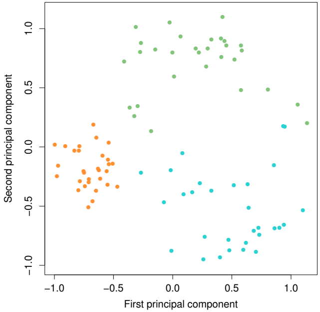


Figure: Ninety observations in $\mathbb{R}^3$. **Left:** The first two PC directions span a plane that best fits the data, minimizing total squared distance. **Right:** Data are "flattened" onto that 2D plane, forming their PC scores [JWHT21, Figure 12.2].

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: Formulation ($p = 2,\ r = 1$)

Assumption: Data is *centered*, i.e.

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i = 0$$

First principal component direction = the direction $\mathbf{u} = (u_1, u_2)$ that solves

$$\text{maximize } \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}\cdot\mathbf{x}_i)^2 \quad \text{subject to} \quad \|\mathbf{u}\| = \sqrt{u_1^2 + u_2^2} = 1$$

- The *variance* of X along $\mathbf{u}$ is

$$\operatorname{Var}(\langle \mathbf{u}, X\rangle) = \mathbb{E}\Big[(\langle \mathbf{u}, \mathbf{X}\rangle - \underbrace{\mathbb{E}\langle \mathbf{u}, X\rangle}_{=0})^2\Big] = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}\cdot\mathbf{x}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(u_1 x_{i1} + u_2 x_{i2})^2.$$

- Geometrically, the solution is the "major axis" in $\mathbb{R}^2$ that explains the largest spread

PCA: General formulation ($p \geq 2$, $r \geq 1$)

First PC: a unit vector $\mathbf{u}_1 \in \mathbb{R}^p$ that maximizes variance, i.e.,

$$\mathbf{u}_1 = \underset{\|\mathbf{u}\|=1}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{u} \cdot \mathbf{x}_i\right)^2$$

Second PC: a unit vector $\mathbf{u}_2 \in \mathbb{R}^p$ maximizing variance, subject to being orthogonal to $\mathbf{u}_1$,

$$\mathbf{u}_2 = \underset{\substack{\|\mathbf{v}\|=1 \\ \langle \mathbf{v}, \mathbf{u}_1 \rangle = 0}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{v} \cdot \mathbf{x}_i\right)^2$$

- $\mathbf{u}_1 \perp \mathbf{u}_2 \implies$ the random variables $Z_1 = \langle \mathbf{u}_1, X \rangle$ and $Z_2 = \langle \mathbf{u}_2, X \rangle$ are *uncorrelated*
- Subsequent PCs $\mathbf{u}_3, \ldots, \mathbf{u}_p$ are defined analogously, each orthogonal to all preceding PCs

Interpretation:

- The k-th PC is orthogonal to all prior ones, ensuring uncorrelatedness among the PC scores
- (`Optional`) The PC directions correspond to the eigenvectors of the sample covariance matrix

PCs as linear combinations (change of basis)

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

$$Z_{ik} = \langle \mathbf{u}_k, X_i \rangle = \sum_{j=1}^{p} u_{k,j} X_{i,j}.$$

- We can write X_i as a combination of the $\mathbf{u}_k$ basis:

$$X_i = \sum_{j=1}^{p} X_{ij} \mathbf{e}_j = \sum_{k=1}^{p} Z_{ik} \, \mathbf{u}_k$$

- For dimension reduction, we might keep only $Z_{i1}, \ldots, Z_{ir}$ for $r \ll p$, compressing the data

(`Optional`):

- $(Z_1, \ldots, Z_p)$ is just a linear transformation of $(X_1, \ldots, X_p)$
- In matrix form, $Z = X\,U$, where U is the orthonormal matrix whose columns are $\mathbf{u}_1, \ldots, \mathbf{u}_p$

PCs as linear combinations (change of basis)

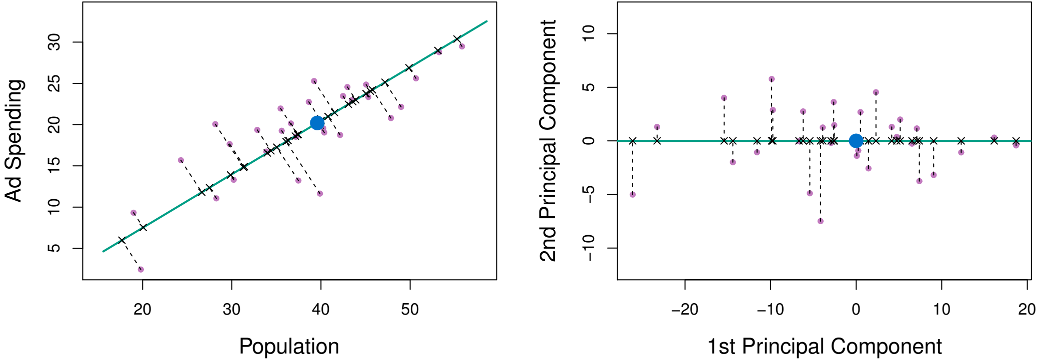


Figure: A subset of the `advertising` data, with the mean `pop` and `ad` budgets shown as a blue circle. **Left:** The first principal component direction (green) captures the greatest data variation and defines the line that best fits all observations (distances shown by dashed segments). **Right:**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

Pop-up quiz #1: Basic PCA understanding

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

A) PCA is unsupervised, using only $\{X_j\}$, not Y.

B) The first principal component is the direction in predictor space along which the projected data has the largest variance.

C) The second principal component must be found by maximizing projected variance with no extra constraint.

D) PCA can serve as dimension reduction by keeping only a few top PCs capturing most variance.

Answer: (**C**) is false.

- For the second principal component, there is an extra requirement that it is *uncorrelated* (and hence orthogonal, in geometric terms) to the first principal component.

Pop-up quiz #2: Interpreting the first principal component

Question: Suppose you have p predictors and you compute the first principal component. Which choice **best describes** how to interpret that first component?

A) It is always the average of all the predictors, so it has little to do with variance.

B) It is the unit-length direction that maximizes how spread out (variable) the data is after projecting onto that direction.

C) It represents a decision boundary for separating classes in your dataset.

D) It is guaranteed to pass exactly through every data point if we use all observations.

Answer: (**B**) is correct.

- The first principal component is the direction along which the data points show the greatest variance; it does not necessarily pass through every data point, nor does it reflect a classification decision boundary.

Proportion of variance explained (PVE)

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

- For centered data, *total variance* is

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

- The variance explained by the k-th PC is

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

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

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

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

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

Scree plot: PVE vs. number of PCs

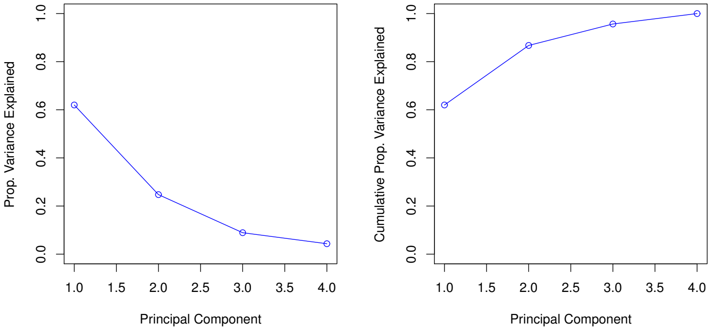


Figure: A scree plot for the `USArrests` data. **Left:** proportion of variance explained by each PC. **Right:** cumulative PVE [JWHT21, Figure 12.3].

Scree plot:

- Plot PVE or cumulative PVE vs. PC index k
- Often look for an "elbow" beyond which additional PCs yield minimal gains

Scree plot: How many PCs to retain?

Trade-off:

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

Question: How many principal components do we need?

- No universal formula for the “best” r
- Typically choose r so the **cumulative PVE** is “high enough,” or identify an “elbow” in the scree plot
- Larger r retains more variance (less information loss) but can be less interpretable

In practice, use scree plot to find an "elbow" and retain the PCs on the left

Choosing the number of PCs using scree plot example

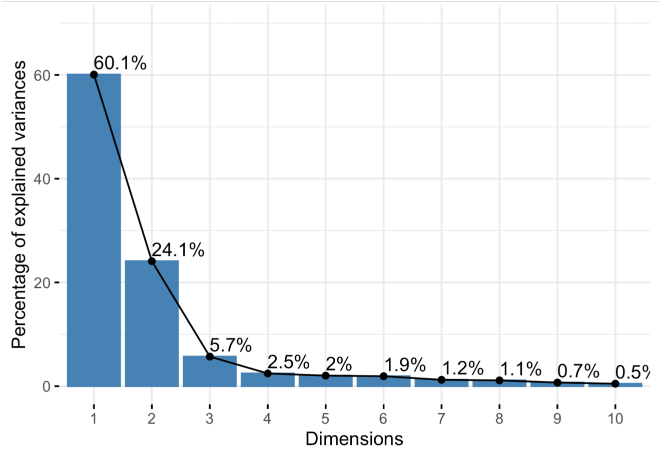


Figure: A scree plot from `mtcars` dataset in R. The elbow appears to occur at the third principal component, which suggests keeping the first three components (source: Statistics Globe).

(`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: if raw scales matter, do not standardize; if you want each feature to contribute equally, do scale

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(...)`

PCA application in high-dimensional genomics

Example: Genomics data [NJB+08]

- 1,387 individuals from Europe, each with genotype data at 197,146 loci
- Apply PCA $\rightarrow$ reduce dimension from p =197k to 2 principal components
- Two PCs remarkably recapitulate Europe's geography in "genetic space," demonstrating how PCA can drastically compress data while still capturing meaningful structure

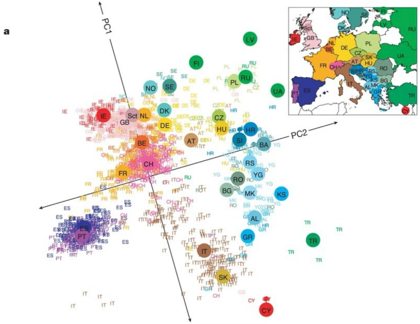


Figure: First two principal components of genetic variation among 1,387 Europeans. Small colored points are individuals; large dots mark country medians in PC1–PC2 space [NJB+08, 0Figure 1-a].

PCA application in image compression

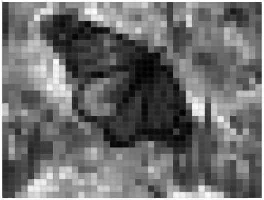
Example: Compressing a grayscale image via PCA

- Original image has 372×492 pixels, each a grayscale intensity in $[0, 255]$
- The image is partitioned 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)
- Apply PCA with rank $r \in \{1, 3, 6, 16, 60\}$ 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.

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):

- Quantifies how much of the total variance is retained by a chosen number r of 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

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.