# STA 35C: Statistical Data Science III

## Lecture 22: Principal Component Analysis

Dogyoon Song

Spring 2026, UC Davis

# Agenda

**Quick recap:** smoothing splines

- Why the solution is a natural cubic spline
- What effective degrees of freedom mean

**Today:**

- Unsupervised learning
  - Supervised vs. unsupervised learning
  - What unsupervised learning is for
- Principal component analysis (PCA): Intuition
  - PCA as dimension reduction
  - PCA in $p = 2$: projection and maximum variance
  - Why centering and standardization matter

# Recap: Smoothing splines

**Smoothing spline:** estimate a smooth function $g$ by solving

$$\min_{g \in \mathcal{G}} \left\{ \underbrace{\sum_{i=1}^{n} \left(y_i - g(x_i)\right)^2}_{\text{RSS; data fidelity}} + \lambda \underbrace{\int \left(g''(t)\right)^2 dt}_{\text{smoothness penalty}} \right\}$$

- The first term rewards fitting the observed data.
- The second term penalizes curvature, discouraging excessive wiggles.
- $\lambda \geq 0$ controls the tradeoff between fit and smoothness.
  - $\lambda = 0$: interpolation
  - $\lambda \to \infty$: least-squares line

Although the optimization ranges over many smooth functions $g$, the minimizer has a special form: it is a **natural cubic spline** with knots at the observed $x_i$'s.

- This does *not* mean we manually choose $n$ knots and then fit an ordinary regression spline.
- The penalty term shrinks the fit toward smoother curves, so the *effective* flexibility can be far smaller than the number of knots suggests for (unpenalized) regression splines.

# Recap: Effective degrees of freedom

**Key point:** A smoothing spline solution lives in a natural cubic spline space with knots at the observed $x_i$'s, but the penalty ($\lambda$) controls how flexible the fitted curve actually is.

**Effective degrees of freedom for smoothing splines:**

- A natural cubic spline with $K$ (interior) knots has $K + 2$ degrees of freedom.
- If the knots are placed at all observed $x_i$'s, then there are $K = n - 2$ interior knots.
- Thus the full natural-spline space has $(n - 2) + 2 = n$ degrees of freedom (free parameters).
- Despite $n$ nominal degrees of freedom, these are constrained/shrunk down

$$\begin{aligned} \lambda = 0 \quad &\Rightarrow \quad \mathrm{df}_\lambda = n \qquad \text{(interpolation)}, \\ \lambda \to \infty \quad &\Rightarrow \quad \mathrm{df}_\lambda = 2 \qquad \text{(least-squares line)}. \end{aligned}$$

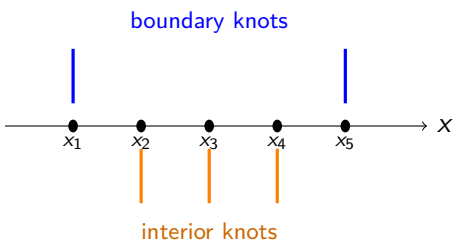
- For $0 < \lambda < \infty$, $\mathrm{df}_\lambda$ lies between 2 and $n$ (larger $\lambda \leftrightarrow$ lower $\mathrm{df}_\lambda$).

For more technical details, see [JWHT21, Chapter 7.4.4 & 7.5.2].

# Recap: Further elaboration on the df and the effective df

Assume $x_1, \ldots, x_n$ are distinct and ordered.

**Smoothing-spline fact:** The solution is a natural cubic spline with boundary knots at $x_1, x_n$ and interior knots at $x_2, \ldots, x_{n-1}$.




For $n = 5$:

- Interior knots: $K = n - 2 = 3$.
- Natural cubic spline df:

$$K + 2 = (n - 2) + 2 = n.$$

- Thus the full natural-spline space has df $n$, not $n + 2$.

**Effective df:** The penalty controls how much of this $n$-nominal df is actually used.

# Illustration of smoothing splines

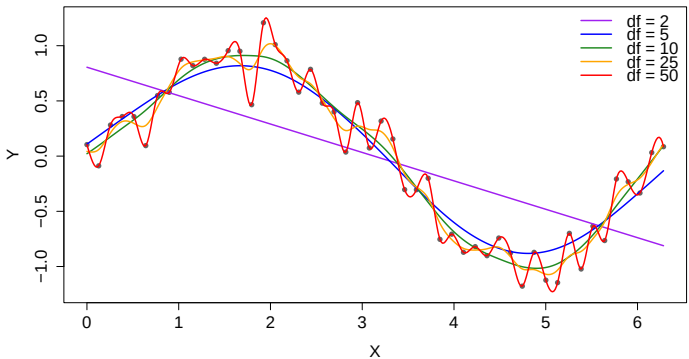



Figure: Smoothing splines fit to $n = 50$ data points from $Y = \sin(X) + \epsilon$ on $[0, 2\pi]$ at various $\lambda$ values. For smaller $\lambda$, the fit nearly interpolates all points (wiggly) and has high effective df. As $\lambda$ grows larger, the fit becomes smoother with lower effective df; for very large $\lambda$, it approaches a straight line.

# Example code: Smoothing splines in R

```
set.seed(2026)

n <- 50
x <- seq(0, 2*pi, length.out = n)
y <- sin(x) + rnorm(n, sd = 0.2)

df_values <- c(2, 5, 10, 25, 50)
fits <- lapply(df_values, function(df0) {
  smooth.spline(x, y, df = df0)
})

xx <- seq(min(x), max(x), length.out = 400)

preds <- sapply(fits, function(fit) {
  predict(fit, x = xx)$y
})

colors <- c("purple", "blue", "forestgreen", "
     orange", "red")
```

For additional examples, see [JWHT21, Sec. 7.8.1 & 7.8.2]

```
plot(x, y,
     pch = 16,
     col = "gray40",
     main = "",
     xlab = "X",
     ylab = "Y",
     cex.axis = 1.4,
     cex.lab = 1.4,
     ylim = range(y, preds))

for (i in seq_along(df_values)) {
  lines(xx, preds[, i],
        col = colors[i],
        lwd = 2)
}

legend("topright",
       legend = paste0("df = ", df_values),
       col = colors,
       lwd = 2,
       bty = "n",
       cex = 0.9)
```

# Pop-up quiz: Smoothing splines

Suppose a smoothing spline is fit to $n$ observations with distinct $x_i$'s.

**Question:** Which statement is most accurate?

A) The solution has $n+2$ effective degrees of freedom because it is a natural cubic spline with $n$ observed $x_i$'s.

B) The solution is a natural cubic spline with knots at the observed $x_i$'s, but $\lambda$ controls the effective flexibility.

C) Increasing $\lambda$ adds more knots and makes the fitted curve more wiggly.

D) When $\lambda \to \infty$, the smoothing spline becomes a step function.

**Answer: B.** The smoothing-spline solution is a natural cubic spline with boundary knots at the smallest/largest $x_i$'s and interior knots at the remaining $x_i$'s. Its full spline space has nominal df $n$, and the penalty controls the effective df: from $n$ when $\lambda = 0$ toward 2 as $\lambda \to \infty$.

# Transition: From supervised to unsupervised learning

**So far: Supervised learning**

$$(X, Y) \quad \Longrightarrow \quad \text{learn how } X \text{ predicts or explains } Y.$$

**Now: Unsupervised learning**

$$X = (X_1, \ldots, X_p) \quad \Longrightarrow \quad \text{discover structure in } X \text{ without a response } Y.$$

**Key difference:** There is no response variable telling us what the "right answer" is.

**New questions:**

- Can we summarize high-dimensional data using fewer variables?
- Can we find natural groupings among observations?

# Unsupervised learning

**Examples:**

- Group patients using gene-expression profiles.
- Summarize thousands of measurements with a few new variables.
- Visualize high-dimensional observations in two dimensions.
- Segment customers based on behavior or purchases.

**Important caveat:** Unsupervised learning is often exploratory; there may not be a single objectively "correct" answer.

**Two major tools we will study:**

- **PCA for dimension reduction:** find low-dimensional directions capturing most variation.
- **Clustering for grouping observations:** find groups of similar observations.

# Principal component analysis (PCA): Overview

**Goal:** Reduce $X = (X_1, \ldots, X_p)$ from $p$ variables to $r \ll p$ new variables while preserving as much variation in $X$ as possible

**Reasons for dimensionality reduction:**

- *Preprocessing for modeling:* Reduce the number of variables before downstream analysis
- *Computational benefits:* Easier/faster to store and process fewer variables
- *Visualization:* Plotting or interpreting data in 2D or 3D
- *Noise reduction:* Focusing on major signals in the data

**How does PCA work?**

- Creates new variables, called *principal component scores*
- Each score is a projection of the data onto a direction of high variance
- These directions are called the *principal component directions*, or *principal components* (PCs)

# PCA: Visual Illustration for $p = 2$

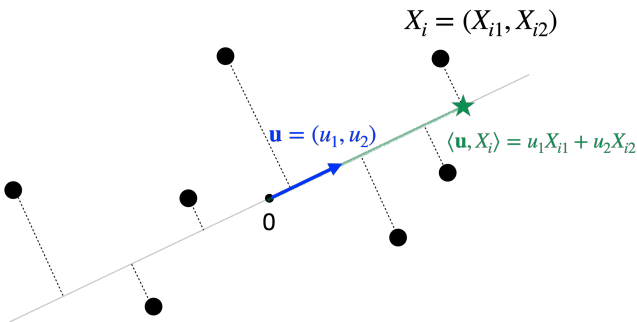



Figure: PCA finds the direction **u** such that the projections of the data onto that direction have the largest variance.

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

- For any unit vector $\mathbf{u} = (u_1, u_2)$, the projection of an observed point $\mathbf{x}_i$ onto the line spanned by **u** is

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

- The first PC is the direction (line) where projected points are most spread out; the variance of $z_i = \langle \mathbf{u}, x_i \rangle$ is maximized
- Geometrically, it is the "major axis" of the data cloud

# Toy example: Finding the direction of largest spread (1/2)

## Example

Consider a small 2D dataset:

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

These three points lie on the line spanned by $(2,1)$ and are already centered: their sample mean is $(0,0)$.

Compare the variances along three directions:

$$\mathbf{u}_x = (1,0), \qquad \mathbf{u}_y = (0,1), \qquad \mathbf{u}_* = \frac{1}{\sqrt{5}}(2,1).$$

**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

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

- 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 $\frac{1}{3}\left((-\sqrt{5})^2 + 0^2 + \sqrt{5}^2\right) = \frac{10}{3}$.

# Toy example: Finding the direction of largest spread (2/2)

## Example (Continued)

**Projected variances:**

| Direction | Variance of projected data |
|---|---|
| $(1, 0)$ | $8/3$ |
| $(0, 1)$ | $2/3$ |
| $\frac{1}{\sqrt{5}}(2, 1)$ | $10/3$ |

Thus, among the three directions, the variance is maximized along

$$\mathbf{u}_* = \frac{1}{\sqrt{5}}(2, 1).$$

**Question:** How can we formulate and find the maximum-variance direction (among infinitely many directions)?

# PCA formulation in $p = 2$

**Assumption:** The data are centered, i.e. $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = 0$.

**Key idea of PCA:**

- For a *direction* $\mathbf{u} = (u_1, u_2)$ with $\|\mathbf{u}\|^2 = u_1^2 + u_2^2 = 1$, the projected value of observation $i$ is
$$z_i(\mathbf{u}) = \langle \mathbf{u}, \mathbf{x}_i \rangle = u_1 x_{i1} + u_2 x_{i2}.$$
- The empirical variance[1] of the projected scores along $\mathbf{u}$ is
$$\mathrm{Var}(\langle \mathbf{u}, X \rangle) = \frac{1}{n}\sum_{i=1}^{n} \big(z_i(\mathbf{u}) - \bar{z}(\mathbf{u})\big)^2 = \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{u}, \mathbf{x}_i \rangle^2 = \frac{1}{n}\sum_{i=1}^{n} \big(u_1 x_{i1} + u_2 x_{i2}\big)^2.$$
- The **1st principal component** is the direction $\mathbf{u}$ that *maximizes* this variance:
$$\text{maximize } \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{u}, \mathbf{x}_i \rangle^2 \quad \text{subject to} \quad \|\mathbf{u}\| = \sqrt{u_1^2 + u_2^2} = 1.$$
- **In words:** choose the unit direction whose projected scores have the largest variance.

[1] We use the PCA convention of computing variance with denominator $n$.

# General formulation of PCA beyond 2 dimensions

**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 $\mathbf{u}_1$,

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

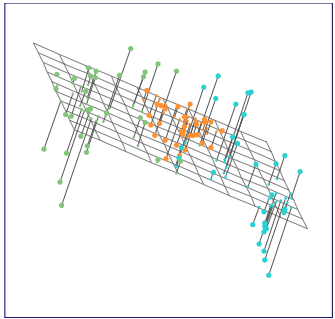
- $\mathbf{u}_1 \perp \mathbf{u}_2 \implies$ the random variables $Z_1 = \langle \mathbf{u}_1, X \rangle$ and $Z_2 = \langle \mathbf{u}_2, X \rangle$ are *uncorrelated*

**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

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



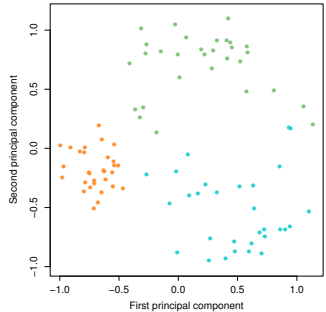



Figure: Ninety observations in $\mathbb{R}^3$. **Left:** The first two PC directions span a plane that best fits the data, minimizing total squared distance. **Right:** Data are "flattened" onto that 2D plane, forming their PC scores [JWHT21, Figure 12.2].

**Key idea in higher dimension:**

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

# PCA and standardization

**PCA is based on variance.** Variables with larger numerical scale can dominate the first principal components.

**Example:** Without standardization, $X_1$ may dominate simply because it has larger units.

- $X_1 =$ income measured in dollars.
- $X_2 =$ age measured in years.

**Common practice of standardization:**

$$\widetilde{X}_j = \frac{X_j - \bar{X}_j}{s_j}.$$

- Centering is essential: PCA studies variation around the mean.
- Standardization is often used when variables are measured on different scales or have very different variances.

# Pop-up quiz: PCA intuition

Suppose a centered two-dimensional dataset has the following projected variances along three unit directions:

| Direction | $(1,0)$ | $(0,1)$ | $\frac{1}{\sqrt{2}}(1,1)$ |
|---|---|---|---|
| Projected variance | 2 | 1 | 5 |

**Question:** Which statement is most accurate?

A) The first PC among these directions is $(0,1)$, because it has the smallest variance.

B) The first PC among these directions is $\frac{1}{\sqrt{2}}(1,1)$, because it has the largest projected variance.

C) PCA chooses the direction that best predicts a response variable $Y$.

D) PCA does not require centering or scaling under any circumstances.

**Answer: B.** PCA is unsupervised: it chooses directions of large variance in $X$, not directions that predict $Y$. Centering is essential, and standardization may be important when variables have different units or scales.

# Wrap-up

- **Smoothing splines:** The solution is a natural cubic spline with knots at the observed $x_i$'s, but $\lambda$ controls the effective flexibility.
  - **Effective degrees of freedom:** For distinct $x_i$'s, $\lambda = 0$ gives effective df $n$, while $\lambda \to \infty$ gives effective df 2.
- **Unsupervised learning:** We study structure in $X$ without a response $Y$.
- **PCA:** Finds directions of maximum variance and uses projections onto those directions for dimension reduction.
  - **PCA scores:** Each observation is represented by its projections onto PC directions; keeping the first few scores gives a lower-dimensional representation.
  - **Standardization matters:** Since PCA is based on variance, variables with larger scales can dominate unless variables are standardized.

*Suggested reading:* [JWHT21, Ch. 12.1–12.2]

# 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.