Yale University

Principal Component Analysis and Low-Rank Approximation

S&DS 265 · Introductory Machine Learning · Lecture 14

Zhuoran Yang

Fall 2026

AI usage: figures were generated, and these slides polished, with the help of AI tools.

Outline

Learning Objectives

In this lecture you will learn:

1. Why a low-dimensional linear subspace can summarize correlated measurements;
2. How variance maximization and reconstruction minimization define the same principal directions;
3. Why eigenvectors and singular vectors compute the fitted PCA subspace;
4. How rank controls variance retained, storage cost, and recognition accuracy;
5. Why centering, scaling, and outliers change the representation PCA learns.

Outline

Motivating Example: Many College Measurements

For each of 777 colleges, the table records applications, enrollment, tuition, spending, graduation, and a dozen more quantities, all on different scales.

Seventeen numbers per college is more than anyone can hold at once. Can a handful of numbers describe how the colleges differ?

Example and data: ISLP Chapter 12 (James et al., 2023).

The College Data

college	Apps	Accept	Enroll	Top10perc	Outstate	Expend	⋯	Grad.Rate
Abilene Christian	1660	1232	721	23	7440	7041	⋯	60
Adelphi	2186	1924	512	16	12280	10527	⋯	56
Adrian	1428	1097	336	22	11250	8735	⋯	54
Agnes Scott	417	349	137	60	12960	19016	⋯	59
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮

The table has $n = 777$ colleges and $d = 17$ numeric measurements each, on units that range from dollars to percentages to counts. Seven columns are shown; ten are hidden behind the dots.

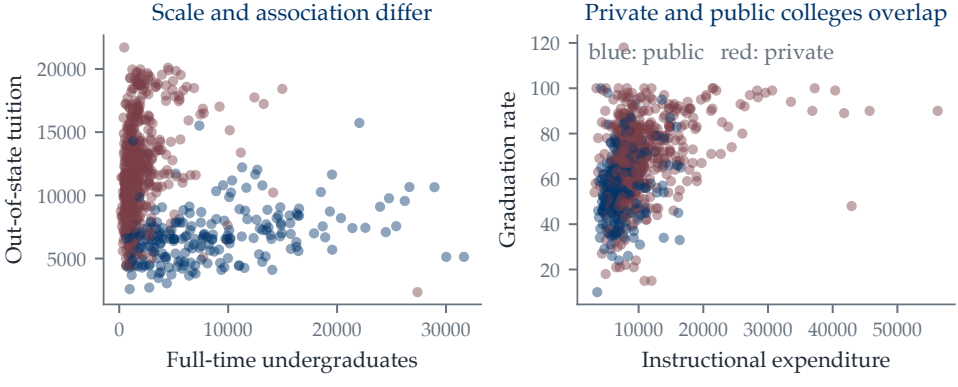
Source: First four rows of College.csv from the ISLP resources.

Reading One Row, in Full

Abilene Christian University					
applications	1660	part-time undergrad	537	PhD faculty (%)	70
accepted	1232	out-of-state tuition	7440	terminal degree (%)	78
enrolled	721	room and board	3300	student/faculty	18.1
top 10% of class (%)	23	books	450	alumni donating (%)	12
top 25% of class (%)	52	personal spending	2200	spending per student	7041
full-time undergrad	2885			graduation rate (%)	60

Seventeen numbers, and together they are the coordinates of one point $x_i \in \mathbb{R}^{17}$. Nothing here is a response: we are describing the colleges, not predicting anything about them.

No Pair of Columns Summarizes a College



Two of the $\binom{17}{2} = 136$ pairs one could plot. Tuition against enrollment half-separates the two kinds of college; spending against graduation rate barely does. Either way a pair shows two columns out of seventeen.

College data from ISLP, plotted in the recorded units.

Why Seek a Low-Dimensional Representation?

Visualization

expose the dominant directions, and the unusual observations, in a picture we can actually look at

Compression

replace d coordinates by $r \ll d$ scores, cheaper to store, send, and search

Both ask for the same object: a few coordinates that keep what matters about an observation. The work of this lecture is to say what "matters" means precisely enough to compute.

Center the Data First

We are looking for a direction, not a location, so move the origin to the middle of the cloud:

$$x_i \leftarrow x_i - \bar{x}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Center the Data First

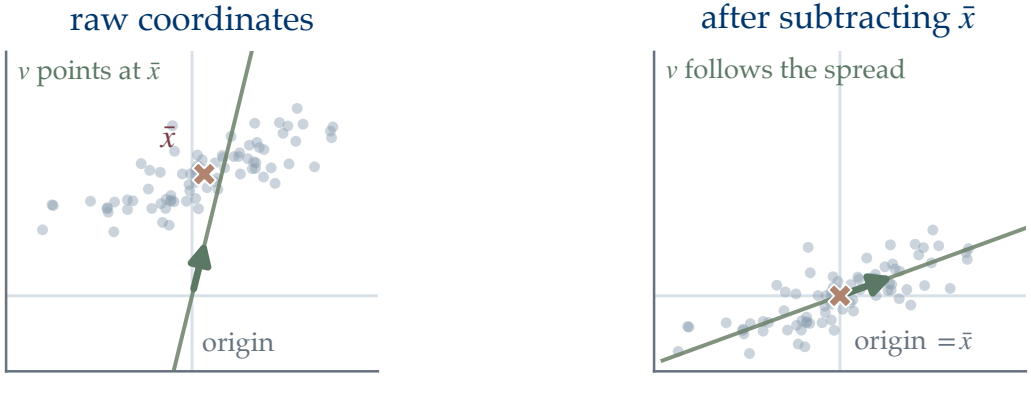
We are looking for a direction, not a location, so move the origin to the middle of the cloud:

$$x_i \leftarrow x_i - \bar{x}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The rows now sum to zero, so the scores along any direction have mean zero and their average square is their variance.

From now on x_i denotes the centered observation and X the centered matrix.

Skip It and the Best Direction Just Points at the Mean



The same cloud twice, on the same axes. With the origin left where the units put it, the direction of largest average squared score runs from the origin to $\bar{x}$ and cuts across the cloud. Move the origin to $\bar{x}$ and the same criterion swings 56° to follow the spread.

Original course simulation, seed 26517; in each panel the arrow maximizes $\frac{1}{n}\sum_i (v^\top x_i)^2$ over unit v.

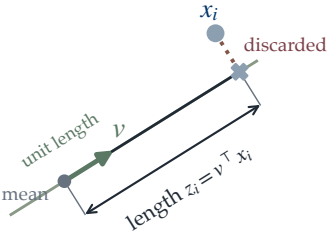
Summarizing a College by One Number

Describe each college by a single number: pick a unit direction v and record how far along it the college sits,

$$z_i = v^\top x_i.$$

Everything perpendicular to v is discarded.

One score is a signed distance along v



Summarizing a College by One Number

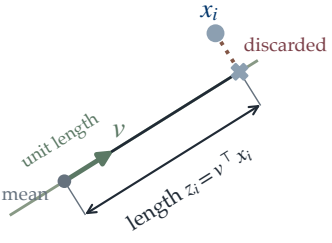
Describe each college by a single number: pick a unit direction v and record how far along it the college sits,

$$z_i = v^\top x_i.$$

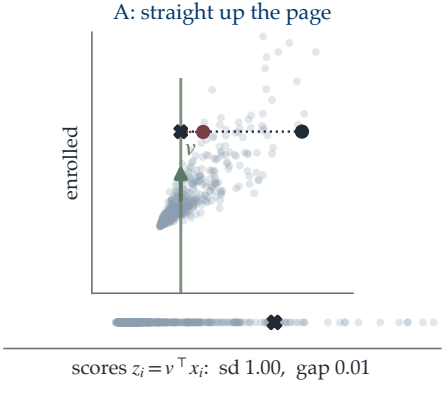
Everything perpendicular to v is discarded.

So the question is concrete: which v?

One score is a signed distance along v



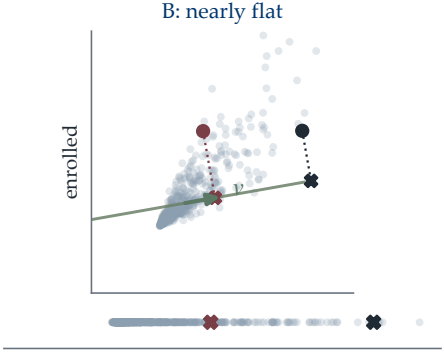
Candidate A: Straight Up



The cloud lies across this direction, so the scores pile up and colleges as different as Berkeley (black) and Southwest Missouri (red) collapse onto the same number.

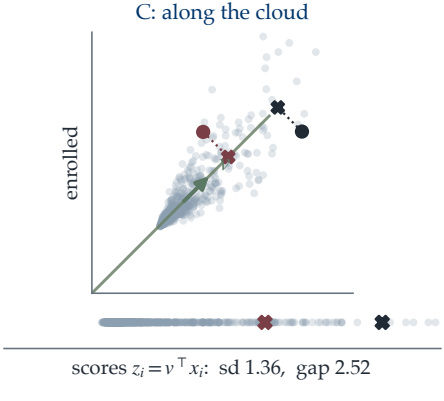
Source: Two standardized ISLP `College` measurements; scores $z_i = v^\top x_i$ shown beneath each cloud.

Candidate B: Nearly Flat



Turning toward applications spreads the scores wider, but the cloud still leans away from the line.

Candidate C: Along the Cloud



Pointing the way the cloud is stretched spreads the scores the most of the three, and leaves the shortest dotted segments.

The One-Dimensional Problem

A unit direction turns an observation into one number, $v^\top x$. Ask for the direction along which that number varies the most.

- **Population.** With x being a centered random vector, $\mathbb{E}[x] = 0$:

$$v_1 \in \arg\max_{\|v\|_2=1} \operatorname{Var}(v^\top x).$$

The One-Dimensional Problem

A unit direction turns an observation into one number, $v^\top x$. Ask for the direction along which that number varies the most.

- **Population.** With x being a centered random vector, $\mathbb{E}[x] = 0$:

$$v_1 \in \arg\max_{\|v\|_2=1} \operatorname{Var}(v^\top x).$$

- **Sample.** We see only 777 colleges, so use the average square:

$$\hat{v}_1 \in \arg\max_{\|v\|_2=1} \frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2.$$

Without the constraint $\|v\|_2 = 1$, scaling v inflates either objective without changing the direction.

From One Direction to r

One number is rarely enough. Asking for r directions and adding up what they capture,

$$\max_{\|v_1\|_2=\cdots=\|v_r\|_2=1} \sum_{j=1}^{r} \mathrm{Var}(v_j^\top x),$$

is useless: every one comes back as v_1.

From One Direction to r

One number is rarely enough. Asking for r directions and adding up what they capture,

$$\max_{\|v_1\|_2=\cdots=\|v_r\|_2=1} \sum_{j=1}^{r} \mathrm{Var}(v_j^\top x),$$

is useless: every one comes back as v_1.

So stop asking for r vectors. Ask for an r-dimensional subspace S, and keep the variance of the projection onto it:

$$\max_{\dim S=r} \mathbb{E}\|P_S x\|_2^2 \qquad \text{(the population problem of } \textbf{PCA}\text{)}$$

From One Direction to r

One number is rarely enough. Asking for r directions and adding up what they capture,

$$\max_{\|v_1\|_2=\cdots=\|v_r\|_2=1} \sum_{j=1}^{r} \mathrm{Var}(v_j^\top x),$$

is useless: every one comes back as v_1.

So stop asking for r vectors. Ask for an r-dimensional subspace S, and keep the variance of the projection onto it:

$$\max_{\dim S=r} \mathbb{E}\|P_S x\|_2^2 \qquad \text{(the population problem of } \textbf{PCA}\text{)}$$

A subspace has no duplicates to hand back.

Problem Setup of PCA

- Take any orthonormal basis $V = [v_1, \dots, v_r]$ of S; then $\|P_S x\|_2^2 = \sum_j (v_j^\top x)^2$.

Problem Setup of PCA

- Take any orthonormal basis $V = [v_1, \dots, v_r]$ of S; then $\|P_S x\|_2^2 = \sum_j (v_j^\top x)^2$.
- The sample problem is then

$$\widehat{V} \in \underset{V^\top V = I_r}{\arg\max} \; \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (v_j^\top x_i)^2$$

Problem Setup of PCA

- Take any orthonormal basis $V = [v_1, \dots, v_r]$ of S; then $\|P_S x\|_2^2 = \sum_j (v_j^\top x)^2$.
- The sample problem is then

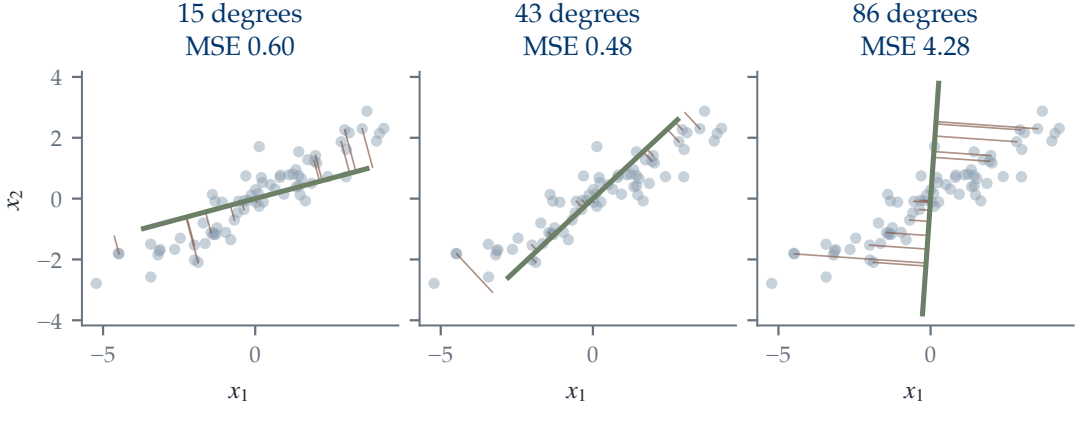
$$\widehat{V} \in \operatorname*{arg\,max}_{V^\top V = I_r} \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (v_j^\top x_i)^2$$

- The constraint only picks a basis: rotating V inside S changes nothing, so what we fit is the subspace itself.

Principal component analysis is the name for this pair of problems. Next we single out one basis inside S: the **principal directions**, whose scores are the **principal components**.

Outline

Projection onto a Candidate Direction



Original course simulation, seed 26517; red segments are what each direction discards.

Encode, Decode, and Residual

The scores keep part of each observation and throw the rest away. Name the three pieces. Fixing a unit direction v,

$$\underbrace{z_i = v^\top x_i}_{\text{encode: one score}}, \qquad \underbrace{\hat{x}_i = z_i v = v v^\top x_i}_{\text{decode: back to } \mathbb{R}^d}, \qquad \underbrace{r_i = x_i - \hat{x}_i}_{\text{residual}}.$$

Encode, Decode, and Residual

The scores keep part of each observation and throw the rest away. Name the three pieces. Fixing a unit direction v,

$$\underbrace{z_i = v^\top x_i}_{\text{encode: one score}}, \qquad \underbrace{\hat{x}_i = z_i v = v v^\top x_i}_{\text{decode: back to } \mathbb{R}^d}, \qquad \underbrace{r_i = x_i - \hat{x}_i}_{\text{residual}}.$$

The residual carries no component along v, since $v^\top v = 1$ gives

$$v^\top r_i = v^\top x_i - (v^\top v)\, v^\top x_i = 0.$$

Encode, Decode, and Residual

The scores keep part of each observation and throw the rest away. Name the three pieces. Fixing a unit direction v,

$$\underbrace{z_i = v^\top x_i}_{\text{encode: one score}}, \qquad \underbrace{\hat{x}_i = z_i v = v v^\top x_i}_{\text{decode: back to } \mathbb{R}^d}, \qquad \underbrace{r_i = x_i - \hat{x}_i}_{\text{residual}}.$$

The residual carries no component along v, since $v^\top v = 1$ gives

$$v^\top r_i = v^\top x_i - (v^\top v)\, v^\top x_i = 0.$$

So each observation splits into what the line keeps and what it discards, at a right angle. That right angle is what the next page exploits.

Pythagorean Energy Decomposition

derive

- The kept part $vv^\top x$ and the discarded part $(I - vv^\top)x$ meet at a right angle, so Pythagoras applies to every observation:

$$\|x\|_2^2 = \underbrace{(v^\top x)^2}_{\text{captured}} + \underbrace{\|x - (v^\top x)v\|_2^2}_{\text{lost}}.$$

Pythagorean Energy Decomposition

- The kept part $vv^\top x$ and the discarded part $(I - vv^\top)x$ meet at a right angle, so Pythagoras applies to every observation:

$$\|x\|_2^2 = \underbrace{(v^\top x)^2}_{\text{captured}} + \underbrace{\|x - (v^\top x)v\|_2^2}_{\text{lost}}.$$

- Averaging over the observations makes it a statement about the whole matrix:

$$\tfrac{1}{n}\|X\|_F^2 = \tfrac{1}{n}\|Xv\|_2^2 + \tfrac{1}{n}\|X - Xvv^\top\|_F^2.$$

Pythagorean Energy Decomposition

derive

- The kept part $vv^\top x$ and the discarded part $(I - vv^\top)x$ meet at a right angle, so Pythagoras applies to every observation:

$$\|x\|_2^2 = \underbrace{(v^\top x)^2}_{\text{captured}} + \underbrace{\|x - (v^\top x)v\|_2^2}_{\text{lost}}.$$

- Averaging over the observations makes it a statement about the whole matrix:

$$\tfrac{1}{n}\|X\|_F^2 = \tfrac{1}{n}\|Xv\|_2^2 + \tfrac{1}{n}\|X - Xvv^\top\|_F^2.$$

The left side does not depend on v. So keeping the most variance and losing the least are one problem, and solving either solves both.

One Direction: a Quadratic Form

- Take $r = 1$. The objective is a quadratic form in the sample covariance,

$$\frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 = v^\top \widehat{\Sigma} v, \qquad \widehat{\Sigma} = \frac{1}{n}X^\top X.$$

One Direction: a Quadratic Form

- Take $r = 1$. The objective is a quadratic form in the sample covariance,

$$\frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 = v^\top \widehat{\Sigma} v, \qquad \widehat{\Sigma} = \frac{1}{n}X^\top X.$$

- Recall: v is an **eigenvector** with **eigenvalue** λ when $\widehat{\Sigma}v = \lambda v$, so the matrix only rescales it.

One Direction: a Quadratic Form

- Take $r = 1$. The objective is a quadratic form in the sample covariance,

$$\frac{1}{n}\sum_{i=1}^{n}(v^\top x_i)^2 = v^\top \widehat{\Sigma} v, \qquad \widehat{\Sigma} = \frac{1}{n}X^\top X.$$

- Recall: v is an **eigenvector** with **eigenvalue** λ when $\widehat{\Sigma}v = \lambda v$, so the matrix only rescales it.
- The maximizer is the eigenvector of the largest eigenvalue, and the variance it captures is that eigenvalue: $\hat{v}_1 = v_1$ and $v_1^\top \widehat{\Sigma} v_1 = \lambda_1$.

Many Directions: One Eigendecomposition

- $\widehat{\Sigma}$ is symmetric, so its eigenvectors can be taken orthonormal and it factors as

$$\widehat{\Sigma} = V \Lambda V^{\top}, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

Many Directions: One Eigendecomposition

- $\widehat{\Sigma}$ is symmetric, so its eigenvectors can be taken orthonormal and it factors as

$$\widehat{\Sigma} = V \Lambda V^{\top}, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

- The best r-dimensional subspace is spanned by the leading r eigenvectors, and it captures $\lambda_1 + \cdots + \lambda_r$.

Many Directions: One Eigendecomposition

- $\widehat{\Sigma}$ is symmetric, so its eigenvectors can be taken orthonormal and it factors as

$$\widehat{\Sigma} = V \Lambda V^{\top}, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

- The best r-dimensional subspace is spanned by the leading r eigenvectors, and it captures $\lambda_1 + \cdots + \lambda_r$.

No search at any r: one eigendecomposition answers every version of the problem at once. Appendix A.1 proves it.

Successive Components

- **A greedy search finds the same basis:** the best direction, then the best one orthogonal to it, and so on,

$$v_j \in \arg\max_{v} \; v^\top \widehat{\Sigma} v \quad \text{subject to} \quad \|v\|_2 = 1, \quad v^\top v_\ell = 0 \;\; (\ell < j).$$

Successive Components

- **A greedy search finds the same basis:** the best direction, then the best one orthogonal to it, and so on,

$$v_j \in \arg\max_{v} \; v^\top \widehat{\Sigma} v \quad \text{subject to} \quad \|v\|_2 = 1, \quad v^\top v_\ell = 0 \;\; (\ell < j).$$

- **Each eigenvalue is a variance:** component j captures λ_j, and together they account for all of it, $\frac{1}{n}\|X\|_F^2 = \lambda_1 + \cdots + \lambda_d$.

Successive Components

- **A greedy search finds the same basis:** the best direction, then the best one orthogonal to it, and so on,

$$v_j \in \arg\max_{v} \; v^\top \widehat{\Sigma} v \quad \text{subject to} \quad \|v\|_2 = 1, \quad v^\top v_\ell = 0 \;\; (\ell < j).$$

- **Each eigenvalue is a variance:** component j captures λ_j, and together they account for all of it, $\frac{1}{n}\|X\|_F^2 = \lambda_1 + \cdots + \lambda_d$.
- **Variance explained by the first r:** $\dfrac{\lambda_1 + \cdots + \lambda_r}{\lambda_1 + \cdots + \lambda_d}$, the number a scree plot displays.

The New Coordinates Are Uncorrelated

- That basis buys more than tidiness. For $j \neq q$,

$$\mathrm{Cov}(v_j^\top x,\ v_q^\top x) = v_j^\top \widehat{\Sigma} v_q = \lambda_q\, v_j^\top v_q = 0,$$

using $\widehat{\Sigma} v_q = \lambda_q v_q$ and then $v_j^\top v_q = 0$.

The New Coordinates Are Uncorrelated

- That basis buys more than tidiness. For $j \neq q$,

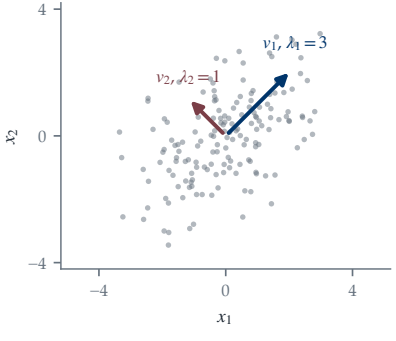
$$\mathrm{Cov}(v_j^\top x,\ v_q^\top x) = v_j^\top \widehat{\Sigma} v_q = \lambda_q\, v_j^\top v_q = 0,$$

  using $\widehat{\Sigma} v_q = \lambda_q v_q$ and then $v_j^\top v_q = 0$.
- So d correlated measurements become r scores that share no linear information: each component reports something the others do not.

Uncorrelated is not independent, and it is not a claim that the components are separate mechanisms.

A Two-Dimensional Example

- Two variables of variance 2, covariance 1: $\widehat{\Sigma} = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right]$.
- $\det(\widehat{\Sigma} - \lambda I) = (\lambda - 3)(\lambda - 1)$: $\lambda_1 = 3$, $\lambda_2 = 1$.
- $v_1 = \frac{1}{\sqrt{2}}(1, 1)^\top$ is the sum, v_2 the difference.
- Total variance $\lambda_1 + \lambda_2 = 4$, so the first component explains $3/4 = 75\%$.



Source: Course simulation, seed 26517, reshaped so its sample covariance is exactly $\widehat{\Sigma}$.

Where Would a Gap in the Spectrum Come From?

Suppose the measurements really are built from r hidden factors,

$$x = Vy + \varepsilon, \qquad y \sim \mathcal{N}(0, \Gamma), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

with $V^\top V = I_r$, $\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_r)$, and ε independent of y. Then

$$\Sigma = V \Gamma V^\top + \sigma^2 I.$$

Where Would a Gap in the Spectrum Come From?

Suppose the measurements really are built from r hidden factors,

$$x = Vy + \varepsilon, \qquad y \sim \mathcal{N}(0, \Gamma), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

with $V^\top V = I_r$, $\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_r)$, and ε independent of y. Then

$$\Sigma = V \Gamma V^\top + \sigma^2 I.$$

Its eigenvalues are $\gamma_j + \sigma^2$ along the columns of V and σ^2 in every other direction.

Where Would a Gap in the Spectrum Come From?

Suppose the measurements really are built from r hidden factors,

$$x = Vy + \varepsilon, \qquad y \sim \mathcal{N}(0, \Gamma), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

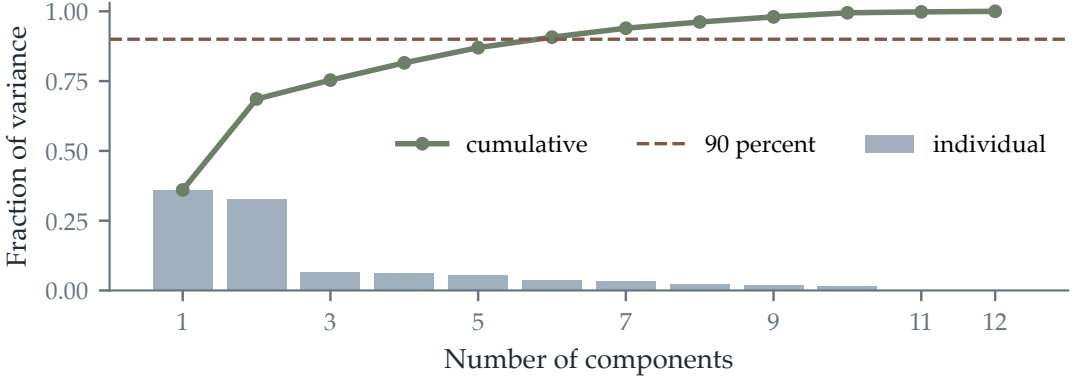
with $V^\top V = I_r$, $\Gamma = \mathrm{diag}(\gamma_1, \dots, \gamma_r)$, and ε independent of y. Then

$$\Sigma = V \Gamma V^\top + \sigma^2 I.$$

Its eigenvalues are $\gamma_j + \sigma^2$ along the columns of V and σ^2 in every other direction.

Under such a model the spectrum is r raised values sitting on a flat floor, and PCA recovers V. Lecture 17 makes this model explicit and then lets a network replace V.

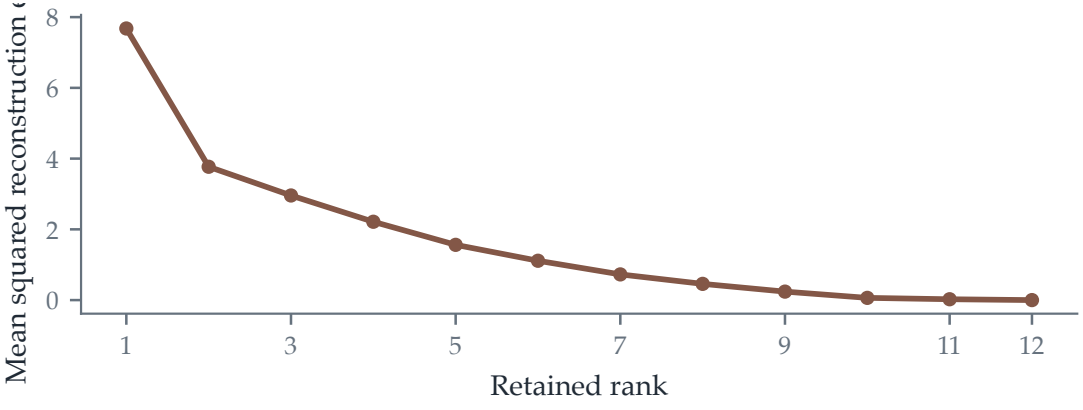
Scree Plot for Standardized College Measurements



A scree plot asks where the raised values stop and the floor begins. Two components stand clear of the rest (36% and 33%, then a drop to 7%), but the tail slopes rather than sitting flat, so there is no single σ^2 here and six components are still needed for 90%.

Course calculation from 12 standardized College variables; no outcome is used.

Reconstruction Error by Rank



Course calculation on the same standardized College matrix.

Rank Is a Modeling Choice

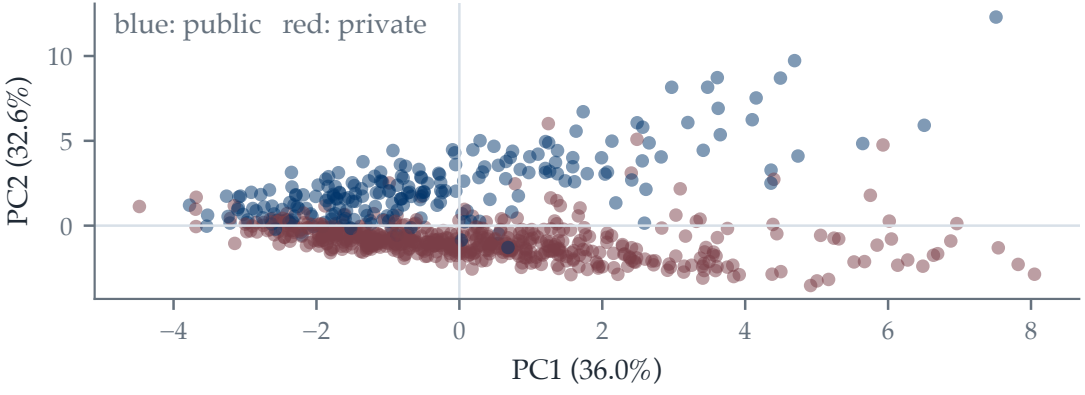
Candidate evidence includes

- Retained variance $\sum_{j \le r} \lambda_j / \sum_j \lambda_j$;
- Discarded variance $\lambda_{r+1} + \cdots + \lambda_d$;
- Stability of the estimated subspace across resamples;
- Downstream validation when the representation serves another task;
- Storage, latency, interpretability, and failure costs.

In the College example, six components retain about 91% of standardized variance; that number does not settle every downstream use.

Outline

College Scores on the First Two Components



Course PCA of 12 standardized College variables; sector colors are shown only after fitting.

Scores Locate Observations

- The score of observation i on component j is $z_{ij} = v_j^\top x_i$ for $j = 1, \dots, r$: its position along that axis.

Scores Locate Observations

- The score of observation i on component j is $z_{ij} = v_j^\top x_i$ for $j = 1, \dots, r$: its position along that axis.
- Distance splits into what the plot shows and what it hides,

$$\|x_i - x_k\|_2^2 = \|V_r^\top (x_i - x_k)\|_2^2 + \|(I - V_r V_r^\top)(x_i - x_k)\|_2^2.$$

Scores Locate Observations

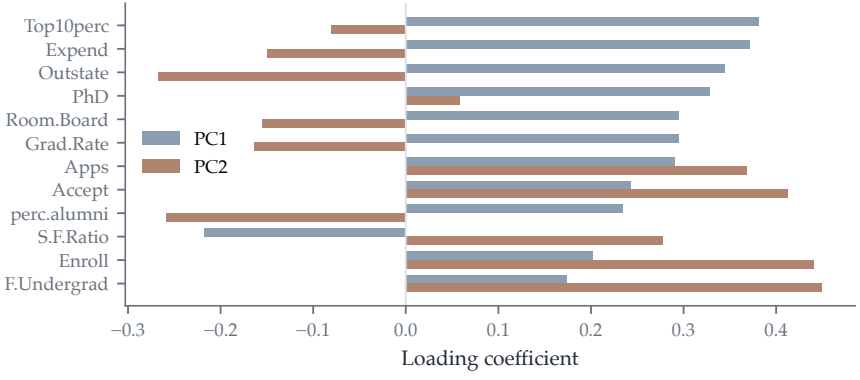
- The score of observation i on component j is $z_{ij} = v_j^\top x_i$ for $j = 1, \dots, r$: its position along that axis.
- Distance splits into what the plot shows and what it hides,

$$\|x_i - x_k\|_2^2 = \|V_r^\top (x_i - x_k)\|_2^2 + \|(I - V_r V_r^\top)(x_i - x_k)\|_2^2.$$

- So two colleges that look adjacent on a two-component plot may be far apart in the $d - r$ directions the plot dropped.

A visual cluster in two components need not be a cluster in the original space.

College Loadings



Source: Course PCA coefficients; component signs are arbitrary.

Loadings Describe Co-Variation

- A score is a weighted sum of the measured columns,

$$z_{ij} = \sum_{k=1}^{d} x_{ik} v_{kj}, \qquad j = 1, \ldots, r,$$

and the weight v_{kj} is the **loading** of feature k on component j.

Loadings Describe Co-Variation

- A score is a weighted sum of the measured columns,

$$z_{ij} = \sum_{k=1}^{d} x_{ik} v_{kj}, \qquad j = 1, \ldots, r,$$

  and the weight v_{kj} is the **loading** of feature k on component j.

- Large loadings mark the features that move that score, in the units chosen.

Loadings Describe Co-Variation

- A score is a weighted sum of the measured columns,

$$z_{ij} = \sum_{k=1}^{d} x_{ik} v_{kj}, \qquad j = 1, \dots, r,$$

  and the weight v_{kj} is the **loading** of feature k on component j.
- Large loadings mark the features that move that score, in the units chosen.
- The sign is arbitrary: v_j and $-v_j$ are the same axis, and with tied eigenvalues any rotation inside that eigenspace fits equally well.

Loadings Describe Co-Variation

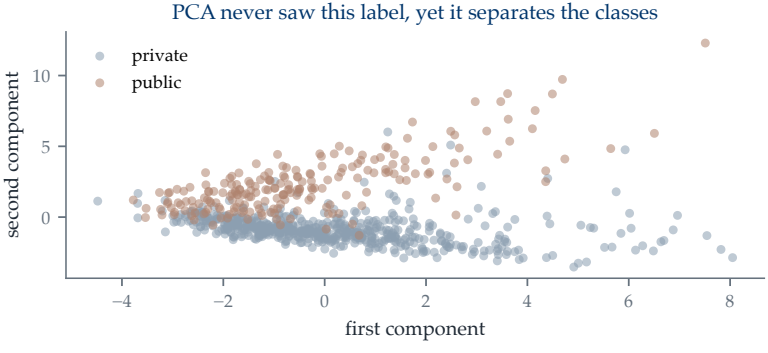
- A score is a weighted sum of the measured columns,

$$z_{ij} = \sum_{k=1}^{d} x_{ik} v_{kj}, \qquad j = 1, \dots, r,$$

  and the weight v_{kj} is the **loading** of feature k on component j.
- Large loadings mark the features that move that score, in the units chosen.
- The sign is arbitrary: v_j and $-v_j$ are the same axis, and with tied eigenvalues any rotation inside that eigenspace fits equally well.

Loadings summarize association, not causes or interventions.

PCA Never Saw the Label



Color the same two scores by private against public, a column PCA was never given, and the picture splits almost cleanly: two components carry 92.3% five-fold accuracy for that label, against 93.8% from all twelve variables. Read it as luck. The components were chosen to maximize variance and never saw this column, so nothing ties them to any particular label.

Logistic regression, five-fold cross-validation on the College data.

Outline

The Same Method on Photographs

- 400 photographs of 40 people at 64×64 pixels. Read row by row, one photograph is a point

$$x_i \in \mathbb{R}^{4096}, \qquad n = 400, \quad d = 4096.$$

The Same Method on Photographs

- 400 photographs of 40 people at 64×64 pixels. Read row by row, one photograph is a point

$$x_i \in \mathbb{R}^{4096}, \qquad n = 400, \quad d = 4096.$$

- Nothing in the method changes: center the columns, keep the leading r directions.

The Same Method on Photographs

- 400 photographs of 40 people at 64×64 pixels. Read row by row, one photograph is a point

$$x_i \in \mathbb{R}^{4096}, \qquad n = 400, \quad d = 4096.$$

- Nothing in the method changes: center the columns, keep the leading r directions.
- Two things do change: a loading vector is now itself an image, and with more pixels than photographs there are at most 399 directions to find.

Source: AT&T Laboratories Cambridge "ORL" database, 400 images of 40 subjects.

Center the Pixels, Do Not Standardize Them

- The college table mixed dollars with percentages, so each column was divided by its own standard deviation.

Center the Pixels, Do Not Standardize Them

- The college table mixed dollars with percentages, so each column was divided by its own standard deviation.
- Pixels are already in one unit, so no column needs rescuing.

Center the Pixels, Do Not Standardize Them

- The college table mixed dollars with percentages, so each column was divided by its own standard deviation.
- Pixels are already in one unit, so no column needs rescuing.
- And a pixel that barely varies is saying something true: nothing happens in that corner of the frame.

Center the Pixels, Do Not Standardize Them

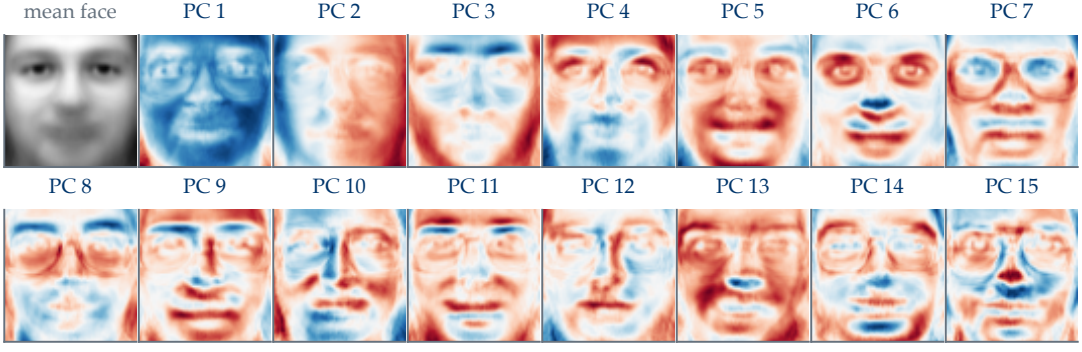
- The college table mixed dollars with percentages, so each column was divided by its own standard deviation.
- Pixels are already in one unit, so no column needs rescuing.
- And a pixel that barely varies is saying something true: nothing happens in that corner of the frame.
- Dividing by its standard deviation would inflate exactly those dead corners into full-sized coordinates.

Center the Pixels, Do Not Standardize Them

- The college table mixed dollars with percentages, so each column was divided by its own standard deviation.
- Pixels are already in one unit, so no column needs rescuing.
- And a pixel that barely varies is saying something true: nothing happens in that corner of the frame.
- Dividing by its standard deviation would inflate exactly those dead corners into full-sized coordinates.

Standardize when the columns carry different units; center only when they share one.

The Mean Face and the Leading Eigenfaces



Course calculation on 280 training images; each loading vector is drawn as an image, red positive and blue negative. The Yale Face Database grew out of this line of work (Belhumeur, Hespanha, and Kriegman, 1997).

Reading an Eigenface

- Every face in the collection is the mean face plus a weighted sum of those images,

$$x \approx \bar{x} + z_1 v_1 + \cdots + z_r v_r,$$

and the weights z_j are the r numbers that now stand for the face.

Reading an Eigenface

- Every face in the collection is the mean face plus a weighted sum of those images,

$$x \approx \bar{x} + z_1 v_1 + \cdots + z_r v_r,$$

  and the weights z_j are the r numbers that now stand for the face.
- Red and blue are opposite signs, so each eigenface is a contrast. PC 2 darkens one side of the face against the other, which is lighting rather than identity.

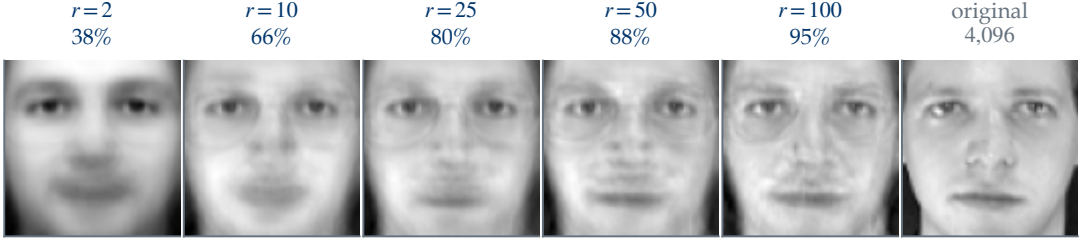
Reading an Eigenface

- Every face in the collection is the mean face plus a weighted sum of those images,

$$x \approx \bar{x} + z_1 v_1 + \cdots + z_r v_r,$$

  and the weights z_j are the r numbers that now stand for the face.
- Red and blue are opposite signs, so each eigenface is a contrast. PC 2 darkens one side of the face against the other, which is lighting rather than identity.
- Later ones are local: eyebrows, the mouth, and in PC 7 a pair of glasses shared by several subjects.

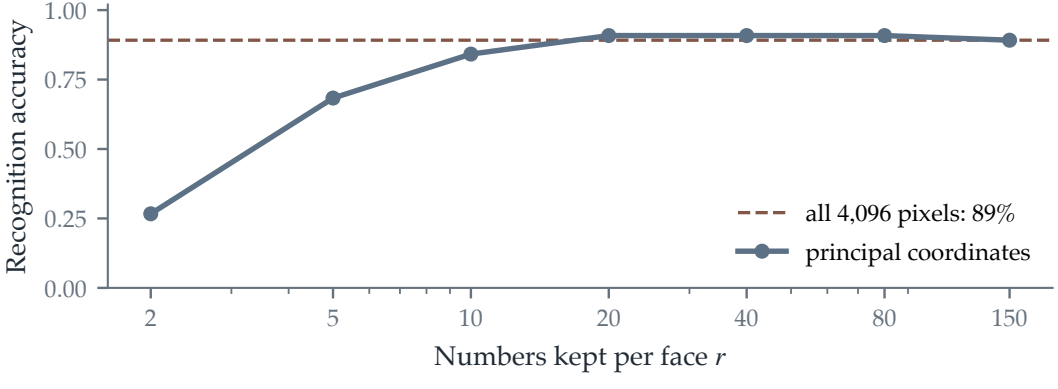
Rank Buys Identity



A held-out face rebuilt from r numbers. Two components give a generic face, ten give this person's build, and the fine detail that identifies them arrives late. Note the glasses borrowed from the basis at $r = 10$ and 25, which the original does not wear.

Components fitted on the training faces only, percentages are cumulative variance.

Fewer Numbers, Better Recognition



Each held-out photograph is matched to its nearest of the 280 training photographs, and counts as correct when that neighbor is the same person, one of 40, so chance would give 2.5%. All 4,096 pixels get it right 89% of the time. The first 20 principal coordinates do slightly better, 91%.

One nearest neighbor, 7 training and 3 test photographs of each of 40 people, seed 26517.

Why the Scores Are Cheaper to Search

- One distance per stored face, per query:

$$\underbrace{\Theta(nd) = 400 \times 4096 \approx 1.6 \text{ million}}_{\text{pixels}} \qquad \underbrace{\Theta(nr) = 400 \times 20 = 8{,}000}_{\text{scores}}$$

Why the Scores Are Cheaper to Search

- One distance per stored face, per query:

$$\underbrace{\Theta(nd) = 400 \times 4096 \approx 1.6 \text{ million}}_{\text{pixels}} \qquad \underbrace{\Theta(nr) = 400 \times 20 = 8{,}000}_{\text{scores}}$$

- Storage shrinks with it: 20 numbers per face instead of 4,096, or 0.5%.

Why the Scores Are Cheaper to Search

- One distance per stored face, per query:

$$\underbrace{\Theta(nd) = 400 \times 4096 \approx 1.6 \text{ million}}_{\text{pixels}} \qquad \underbrace{\Theta(nr) = 400 \times 20 = 8{,}000}_{\text{scores}}$$

- Storage shrinks with it: 20 numbers per face instead of 4,096, or 0.5%.
- With the shared directions counted too, the database falls to 90,000 numbers.

Why the Scores Are Cheaper to Search

- One distance per stored face, per query:

$$\underbrace{\Theta(nd) = 400 \times 4096 \approx 1.6 \text{ million}}_{\text{pixels}} \qquad \underbrace{\Theta(nr) = 400 \times 20 = 8{,}000}_{\text{scores}}$$

- Storage shrinks with it: 20 numbers per face instead of 4,096, or 0.5%.
- With the shared directions counted too, the database falls to 90,000 numbers.

Accuracy improved because the discarded directions held mostly noise, which is an empirical finding here, not a law.

Source: Sirovich and Kirby (1987); Turk and Pentland (1991).

Outline

A Reminder: the Singular Value Decomposition

Every matrix factors, square or not, symmetric or not. For a centered $X \in \mathbb{R}^{n \times d}$ of rank q,

$$\underset{n\times d}{X} = \underset{n\times q}{U}\,\underset{q\times q}{D}\,\underset{q\times d}{V^\top}.$$

- V holds orthonormal directions in feature space, $V^\top V = I_q$. These are the objects PCA has been looking for.
- $D = \mathrm{diag}(\sigma_1, \dots, \sigma_q)$ with $\sigma_1 \geq \cdots \geq \sigma_q \geq 0$ records the strength of each.
- U is orthonormal too, and UD holds the coordinates of the rows in that basis.

The middle factor is written D here; $\widehat{\Sigma}$ stays the covariance matrix.

The SVD Already Contains the Answer

derive

- Substitute $X = UDV^\top$ into the covariance. Since $U^\top U = I$ the middle collapses:

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} V D \underbrace{U^\top U}_{=I} D V^\top = V\Big(\frac{D^2}{n}\Big) V^\top.$$

The SVD Already Contains the Answer

derive

- Substitute $X = UDV^\top$ into the covariance. Since $U^\top U = I$ the middle collapses:

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} V D \underbrace{U^\top U}_{=I} D V^\top = V\Big(\frac{D^2}{n}\Big) V^\top.$$

- That is an eigendecomposition, so the columns of $V = [v_1, \dots, v_q]$ are the principal directions, already ordered, with $\lambda_j = \sigma_j^2 / n$. Keeping r means $V_r = [v_1, \dots, v_r]$.

The SVD Already Contains the Answer

derive

- Substitute $X = UDV^\top$ into the covariance. Since $U^\top U = I$ the middle collapses:

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} V D \underbrace{U^\top U}_{=I} D V^\top = V\Big(\frac{D^2}{n}\Big) V^\top.$$

- That is an eigendecomposition, so the columns of $V = [v_1, \dots, v_q]$ are the principal directions, already ordered, with $\lambda_j = \sigma_j^2 / n$. Keeping r means $V_r = [v_1, \dots, v_r]$.
- The scores come free: $Z_r = X V_r = U_r D_r$, where U_r holds the first r columns of U and $D_r = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$. No projection is computed at all.

Why Compute It This Way

route to the leading directions	cost	the faces
form $\widehat{\Sigma}$, then eigendecompose it	$O(nd^2 + d^3)$	7×10^{10}
singular value decomposition of X	$O(nd \min(n, d))$	7×10^{8}
iterate for the leading $r = 20$ only	$O(ndr)$	3×10^{7}

With $n = 400$ photographs of $d = 4096$ pixels, the covariance alone has 16.8 million entries; the SVD never forms it, and works on the smaller side instead.

Appendix A.3 introduces power iteration, the iterative method in the last row.

The SVD Yields the Best Low-Rank Approximation

- Truncating the factorization keeps the r strongest rank-one pieces, $\widehat{X}_r = U_r D_r V_r^\top = \sum_{j \le r} \sigma_j u_j v_j^\top$.

The SVD Yields the Best Low-Rank Approximation

- Truncating the factorization keeps the r strongest rank-one pieces, $\widehat{X}_r = U_r D_r V_r^\top = \sum_{j \le r} \sigma_j u_j v_j^\top$.
- Among all rank-r matrices, none is closer to X in squared error, and the error left behind is exactly the discarded tail:

$$\widehat{X}_r \in \operatorname*{arg\,min}_{\operatorname{rank}(M) \le r} \|X - M\|_F^2, \qquad \|X - \widehat{X}_r\|_F^2 = \sum_{j > r} \sigma_j^2$$

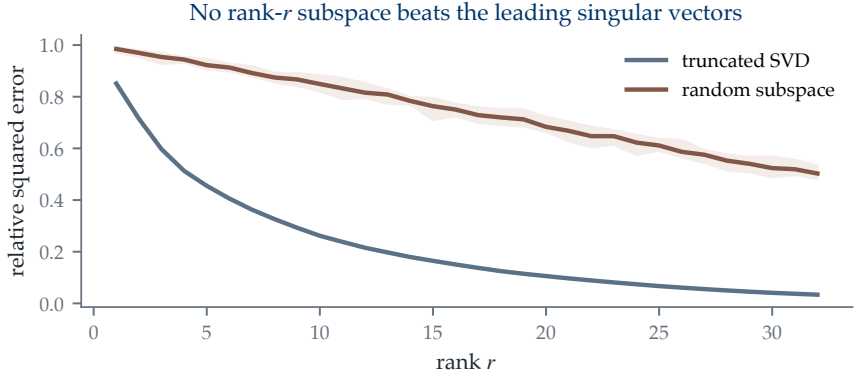
The SVD Yields the Best Low-Rank Approximation

- Truncating the factorization keeps the r strongest rank-one pieces, $\widehat{X}_r = U_r D_r V_r^\top = \sum_{j \le r} \sigma_j u_j v_j^\top$.
- Among all rank-r matrices, none is closer to X in squared error, and the error left behind is exactly the discarded tail:

$$\widehat{X}_r \in \operatorname*{arg\,min}_{\operatorname{rank}(M) \le r} \|X - M\|_F^2, \qquad \|X - \widehat{X}_r\|_F^2 = \sum_{j > r} \sigma_j^2$$

This is the Eckart–Young theorem, and it holds for squared error; another loss can prefer a different rank-r matrix.

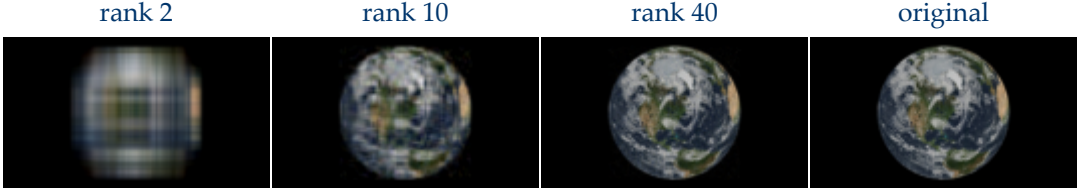
Checking Optimality against Random Subspaces



Project the digit images onto 12 random rank-r subspaces and onto the leading principal directions. At $r = 10$ the random subspaces leave 85% of the squared error and the truncated SVD leaves 26%, and no subspace can do better.

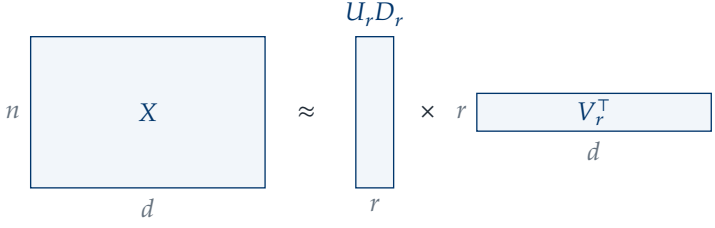
Course computation on the 1,797 × 64 digits matrix; 12 random draws, seed 26517.

One Image, Compressed by Its Own Singular Values



Derived from NASA SVS Blue Marble 2002; here the matrix is a single image and each RGB channel is truncated separately.

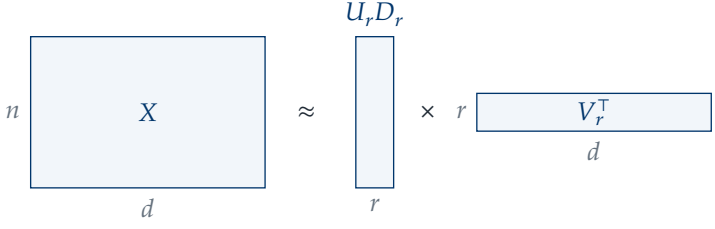
Storage Cost of a Low-Rank Matrix



nd numbers

$r(n + d)$ numbers

Storage Cost of a Low-Rank Matrix



The rank-40 Earth keeps $40(216+384) = 24{,}000$ numbers per channel against 82,944, a factor of 3.5. Real savings also depend on precision, metadata, and the cost of encoding and decoding.

The Same Idea at Modern Scale

- A rank-r pair describes an $n \times d$ matrix with $r(n + d)$ numbers instead of nd. At $n = d = 4096$ and $r = 8$ that is 65,536 numbers rather than 16.8 million, a factor of 256.

The Same Idea at Modern Scale

- A rank-r pair describes an $n \times d$ matrix with $r(n + d)$ numbers instead of nd. At $n = d = 4096$ and $r = 8$ that is 65,536 numbers rather than 16.8 million, a factor of 256.
- The same trick is now used while fitting large models, not just while storing them. Fine-tuning learns an update ΔW to a weight matrix, and **LoRA** restricts that update to be low rank.

The Same Idea at Modern Scale

- A rank-r pair describes an $n \times d$ matrix with $r(n + d)$ numbers instead of nd. At $n = d = 4096$ and $r = 8$ that is 65,536 numbers rather than 16.8 million, a factor of 256.
- The same trick is now used while fitting large models, not just while storing them. Fine-tuning learns an update ΔW to a weight matrix, and **LoRA** restricts that update to be low rank.

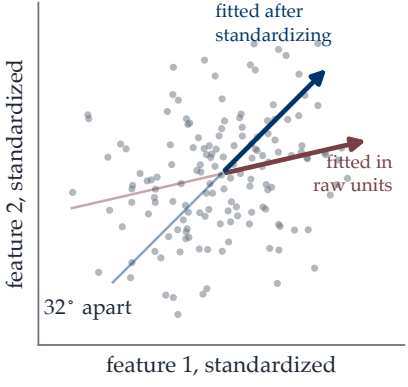
$$W \leftarrow W + BA, \qquad B \in \mathbb{R}^{n \times r}, \quad A \in \mathbb{R}^{r \times d}, \quad r \ll \min(n, d).$$

Source: Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," 2021.

Outline

1. Principal Directions
2. Solving the PCA Optimization Problem
3. Reading the Components
4. Eigenfaces
5. The SVD and Low-Rank Approximation
6. Preprocessing and Evidence

Units Decide the Leading Direction



One cloud, drawn in standardized coordinates, with both answers on the same axes. Feature 1 carries the larger unit, so fitting the raw columns gives a direction 32° away from the one fitted after standardizing. Neither is wrong. They answer different questions.

Original course simulation, seed 26517; the raw-unit direction is carried into the standardized picture.

To Scale or Not to Scale

Centering is not optional, but scaling is a decision.

Center only

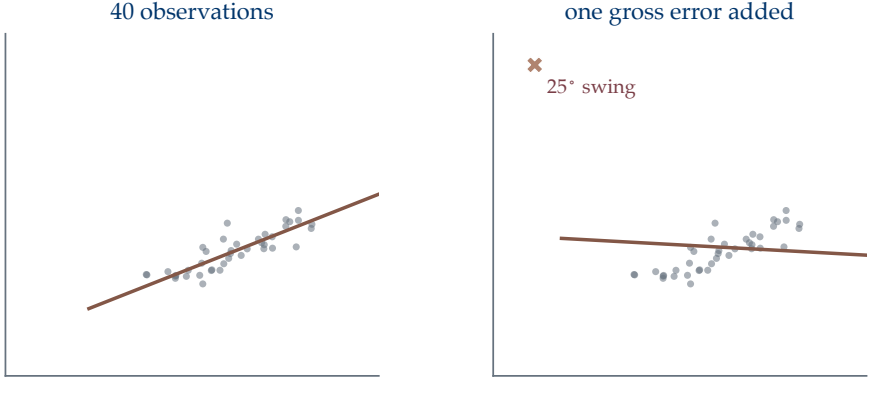
keep the original units when they are comparable and a large movement really is a large movement

Standardize

divide by each standard deviation when units are arbitrary, which fits PCA to the correlation matrix instead

Scaling is part of the scientific question: it fixes what one unit of squared distance means, and the College components above are the standardized ones.

One Bad Point Rotates the Fit



The same 40 observations, plus a single gross error. The fitted first direction swings 25° and no longer follows the cloud at all. Squared distance is what did it: the criterion pays $\|x_i\|_2^2$ for every point, so one remote row can outbid forty ordinary ones.

Original course simulation, seed 26517; both panels on the same axes and equal aspect.

Outliers, and What to Do about Them

- Each row enters through an outer product, $\widehat{\Sigma} = \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$, so leverage grows with the square of its distance from the mean.

Outliers, and What to Do about Them

- Each row enters through an outer product, $\widehat{\Sigma} = \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$, so leverage grows with the square of its distance from the mean.
- To find one, look at the score plot: a gross error usually sits alone. To measure it, refit without that row and see how far the direction moves.

Outliers, and What to Do about Them

- Each row enters through an outer product, $\widehat{\Sigma} = \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$, so leverage grows with the square of its distance from the mean.
- To find one, look at the score plot: a gross error usually sits alone. To measure it, refit without that row and see how far the direction moves.
- Repair it with a model rather than with deletion. **Robust PCA** splits the matrix as $X = L + S$, a low-rank L plus a sparse S holding the gross errors, and fits both.

Outliers, and What to Do about Them

- Each row enters through an outer product, $\widehat{\Sigma} = \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$, so leverage grows with the square of its distance from the mean.
- To find one, look at the score plot: a gross error usually sits alone. To measure it, refit without that row and see how far the direction moves.
- Repair it with a model rather than with deletion. **Robust PCA** splits the matrix as $X = L + S$, a low-rank L plus a sparse S holding the gross errors, and fits both.

Dropping a point because it changed the answer is not a rationale. A homework problem returns to this.

The Two Reasons, Answered

Visualization

777 colleges on two axes that hold 69% of their variance, and every face as a point in 20

Compression

a face in 20 numbers rather than 4,096, recognized better than at full resolution; the Earth at rank 40

The Two Reasons, Answered

Visualization

777 colleges on two axes that hold 69% of their variance, and every face as a point in 20

Compression

a face in 20 numbers rather than 4,096, recognized better than at full resolution; the Earth at rank 40

Both came from one criterion, chosen without looking at any task. That is what makes the representation reusable, and also what makes it worth checking against the task you actually have.

Summary: Formulation and Solution

Setup. Centered $x_1, \dots, x_n \in \mathbb{R}^d$ and no response: the task is to describe variation, not to predict a label.

Objective. Find the r-dimensional subspace keeping the most variance, $\max_{\dim S = r} \mathbb{E}\|P_S x\|_2^2$; in a basis, $\max_{V^\top V = I_r} \frac{1}{n} \sum_j \sum_i (v_j^\top x_i)^2$.

Solution. The leading eigenvectors of $\widehat{\Sigma}$, and λ_j is the variance component j captures.

Duality. Keeping the most variance and losing the least in squared reconstruction are the same problem.

Summary: Computation and Limits

Computation. $X = UDV^\top$ supplies the directions V, the scores UD, and $\lambda_j = \sigma_j^2/n$, without ever forming $\widehat{\Sigma}$.

Rank. Truncation is the best rank-r approximation (Eckart–Young); rank costs $r(n + d)$ numbers and is chosen from variance kept, stability, and the intended use.

Preprocessing. Centering is required; scaling is a decision that changes which direction is leading.

Claim boundary. Variance is not relevance. Components are fitted without a task, and one outlier can rotate them.

Next Lecture

PCA summarizes an observation by its coordinates along a few directions. It never says which observations belong together.

Lecture 15 — Clustering replaces directions with group assignments. The definition of similarity, meaning representation, distance, and scale, becomes the central modeling choice.

Reading for this lecture: ISLP Chapter 12.

Appendix

Four notes skipped in the lecture hour.

A.1 Why the top eigenvector wins. The bound behind the claim that one eigendecomposition solves the problem.

A.2 Why truncation is optimal. The Eckart–Young bound for squared Frobenius error.

A.3 How the eigenvectors are actually computed. Power iteration and deflation, for matrices too large to decompose.

A.4 When is the answer reproducible? The eigengap, and comparing fitted subspaces rather than loadings.

A.1 Why the Largest Eigenvalue Wins

derive

Because $\widehat{\Sigma}$ is symmetric its eigenvectors form an orthonormal basis of $\mathbb{R}^d$, so any unit v can be written $v = \sum_j c_j v_j$ with $\sum_j c_j^2 = 1$. The objective is then a weighted average of eigenvalues:

$$v^\top \widehat{\Sigma} v = \sum_j \lambda_j c_j^2 \;\le\; \lambda_1 \sum_j c_j^2 = \lambda_1,$$

because every $\lambda_j \le \lambda_1$. The bound is attained at $c_1 = 1$, that is at $v = v_1$.

The same argument inside the orthogonal complement of $v_1, \dots, v_{j-1}$ gives the jth component, with value λ_j.

A.2 Why Truncation Is Optimal

derive

For any matrix M with rank at most r, its column space can capture at most r orthogonal left-singular directions. The variational characterization gives

$$\|X - M\|_F^2 \geq \sum_{j>r} \sigma_j^2.$$

The truncated SVD attains that bound:

$$\boxed{X_r \in \underset{\operatorname{rank}(M)\leq r}{\arg\min} \|X - M\|_F^2}.$$

This is the Eckart–Young theorem for squared Frobenius error; a different loss can imply a different optimum.

A.3 Power Iteration: the Leading Eigenvector

derive

Decomposing a $d \times d$ matrix costs $O(d^3)$, impossible at scale. When only the top directions are wanted, they can be reached iteratively: start from a random $v^{(0)}$ and repeat

$$v^{(t+1)} = \frac{\widehat{\Sigma} v^{(t)}}{\|\widehat{\Sigma} v^{(t)}\|_2}.$$

Write $v^{(0)} = \sum_j c_j v_j$ in the eigenbasis. Each multiplication scales coordinate j by λ_j, so after t steps

$$\widehat{\Sigma}^t v^{(0)} = \sum_j \lambda_j^t c_j v_j = \lambda_1^t \Big(c_1 v_1 + \sum_{j>1} (\tfrac{\lambda_j}{\lambda_1})^t c_j v_j \Big).$$

Every ratio λ_j / λ_1 is below one, so the tail dies and $v^{(t)} \to v_1$. Only products $\widehat{\Sigma} v$ are needed, each costing $O(nd)$ as $X^\top (Xv)$.

A.3 Deflation: the Next Eigenvector

derive

With v_1 and $\lambda_1 = v_1^\top \widehat{\Sigma} v_1$ in hand, subtract that component from the matrix:

$$\widehat{\Sigma}^{(2)} = \widehat{\Sigma} - \lambda_1 v_1 v_1^\top.$$

The eigenvectors are unchanged and v_1 now has eigenvalue 0, while every other v_j keeps its own λ_j. The same iteration on $\widehat{\Sigma}^{(2)}$ therefore returns v_2, and r repetitions return the retained basis in $O(ndr)$.

Convergence is governed by λ_2/λ_1: a near-tie makes the iteration slow and the direction unstable It is the same gap that decides whether a scree plot has a visible elbow.

A.4 The Eigengap and Subspace Stability

Individual loading vectors can be unstable when adjacent eigenvalues are close. Compare two fitted rank-r projectors through

$$P = V_r V_r^\top, \qquad P' = V_r' V_r'^\top, \qquad \|P - P'\|_F.$$

Bootstrap or split-sample variation in this distance reveals whether the retained subspace is reproducible even when signs or bases change.