

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Data Reduction at Scale: Sampling & PCA

Lecture 8 · Thursday 1 October 2026

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–5 min) Recap and roadmap
2. (5–10 min) Why reduce data? Rows vs columns; population vs sample
3. (10–22 min) Simple random, stratified and cluster sampling — worked on our 500 farms
4. (22–28 min) How big a sample? Standard error and the √n rule
5. (28–32 min) Reservoir sampling for data streams
6. (32–55 min) PCA from zero: covariance → projection → eigenvectors → worked 2 × 2 example → real farms
7. (55–62 min) Scaling: PCA's cost and randomized PCA
8. (62–75 min) Exercise, summary, questions

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Distinguish	simple random, stratified and cluster sampling, and compute a stratified allocation by hand
Explain	standard error, SE = σ / √n, and why 4× more data only halves the error
Describe	reservoir sampling and why it gives every stream item the same chance
Compute	a 2-attribute PCA by hand: eigenvalues, eigenvectors, % variance explained
Interpret	a scree plot and PCA loadings on real data, and name a scalable PCA method

Recap: Yesterday Changed Scale, Today Changes Size

IIT Madras Zanzibar

Lecture 7 (yesterday)

Same size, new form

Normalization: new ruler per attribute

Discretization: numbers → categories

Rows and columns unchanged

Lecture 8 (today)

Smaller data, same insight

Sampling: fewer rows

PCA: fewer columns

Keep what matters for mining

Remember: PCA needs normalized data. Today you will see what happens on our farms if you skip it.

Data Reduction at Scale: Sampling & PCA

IIT Madras Zanzibar

→ **Foundations: Why Reduce, and What a Sample Is**

Sampling Methods, Step by Step

How Big Should a Sample Be?

Sampling a Stream: Reservoir Sampling

PCA from Zero

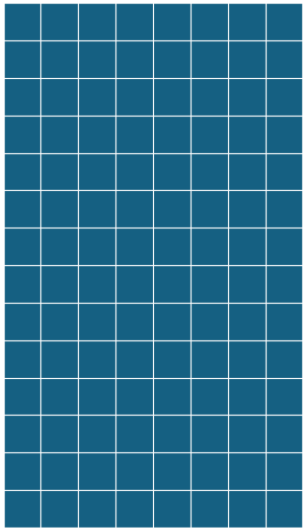
PCA on Our Farms, and at Scale

In-Class Exercise

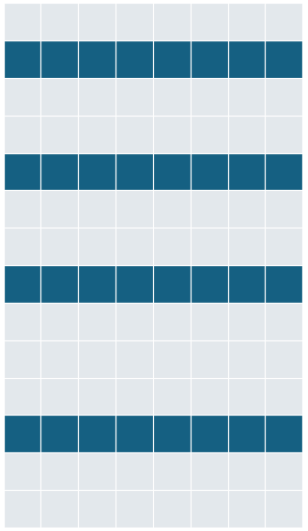
What Is Data Reduction?

Definition. Data reduction produces a much smaller version of the data that gives (almost) the same analytical results.

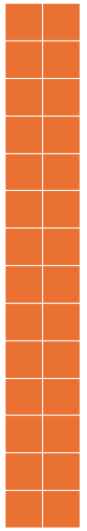
Original: 500 rows × 5+ columns



Sampling → fewer ROWS



PCA → fewer COLUMNS



1 **Why**
faster algorithms, less memory, lower cost

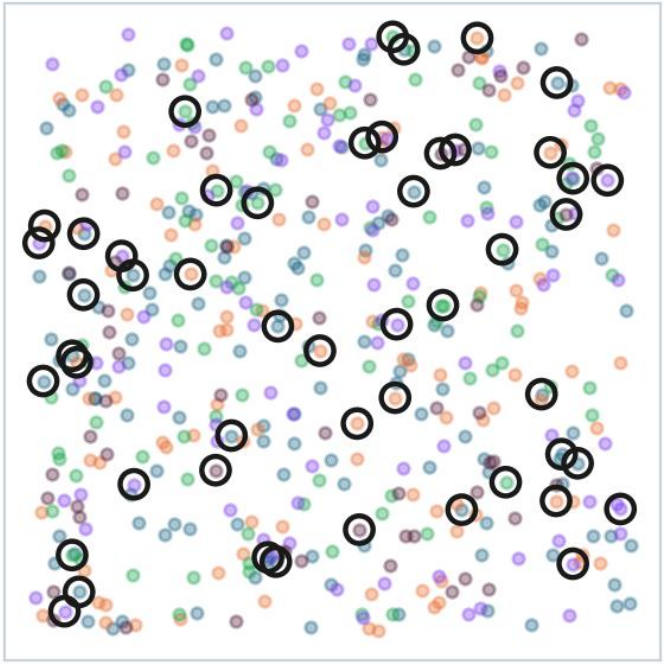
2 **Fewer rows**
sampling: 500 farms → 50

3 **Fewer columns**
PCA: 5 attributes → 2 components

Population vs Sample: Tasting the Soup

IIT Madras Zanzibar

Population: all 500 farms (faint) · Sample: 50 circled



Symbol	What it means	In our example
N	population size: everyone	500 farms
n	sample size: the ones we examine	50 farms
n / N	sampling fraction	50 / 500 = 10%

Analogy: tasting soup

One spoonful can tell you about the whole pot — but only if the pot is well stirred.

Sampling methods are ways of 'stirring'.

Goal of every sampling method: a sample that REPRESENTS the population, so that answers from n records match answers from N records.

Data Reduction at Scale: Sampling & PCA

IIT Madras Zanzibar

Foundations: Why Reduce, and What a Sample Is

→ **Sampling Methods, Step by Step**

How Big Should a Sample Be?

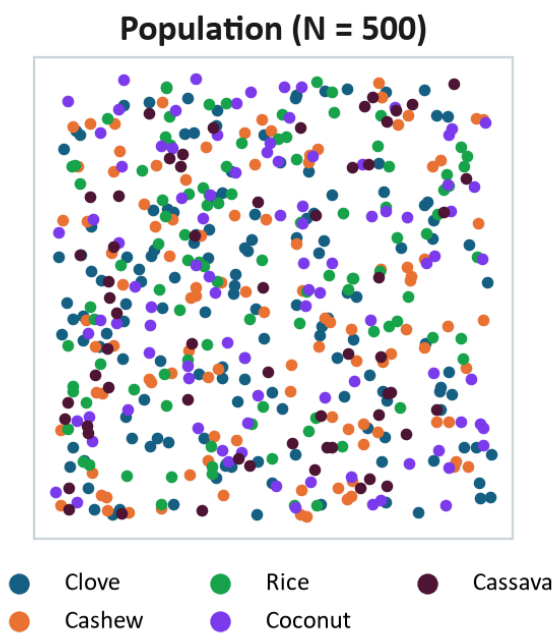
Sampling a Stream: Reservoir Sampling

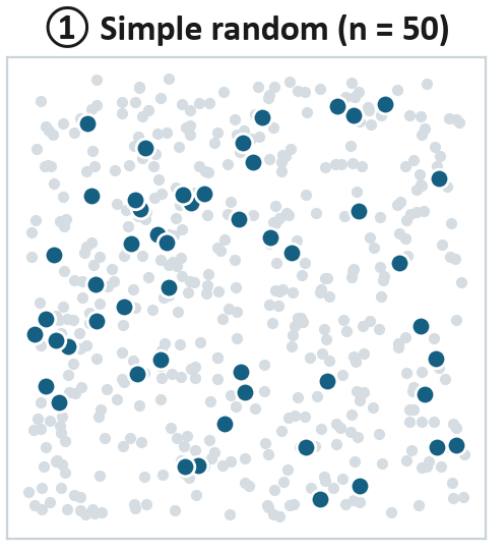
PCA from Zero

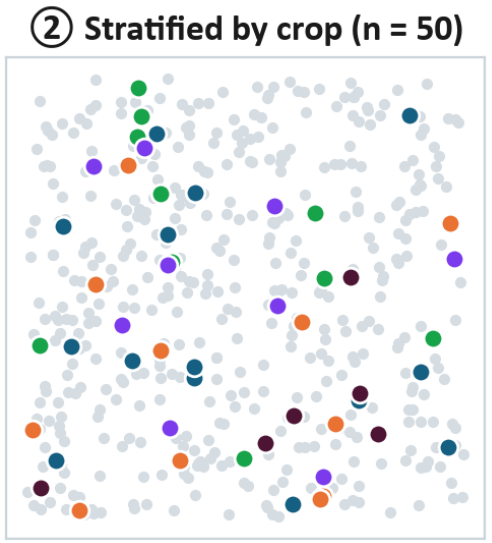
PCA on Our Farms, and at Scale

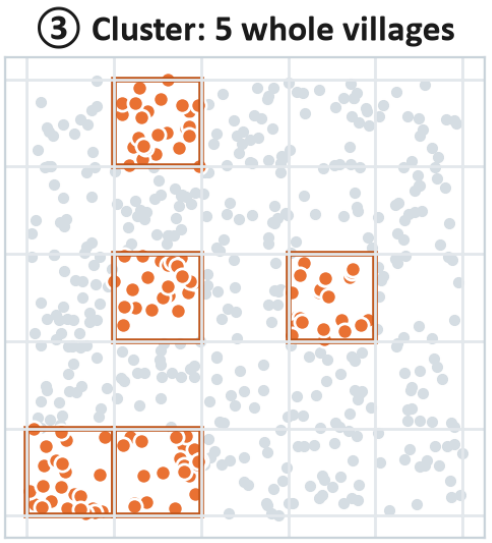
In-Class Exercise

Four Ways to Draw a Sample









1 Simple random
every record equally likely — names from a hat

2 Stratified
split into groups, sample each group fairly

3 Cluster
pick whole groups (villages), take everyone in them

Simple Random Sampling: With or Without Replacement

IIT Madras Zanzibar

Without replacement (SRSWOR)

like dealing cards

- A chosen farm is set aside
- No farm can appear twice
- Draw 1: 1/500, draw 2: 1/499 …
- The usual default

With replacement (SRSWR)

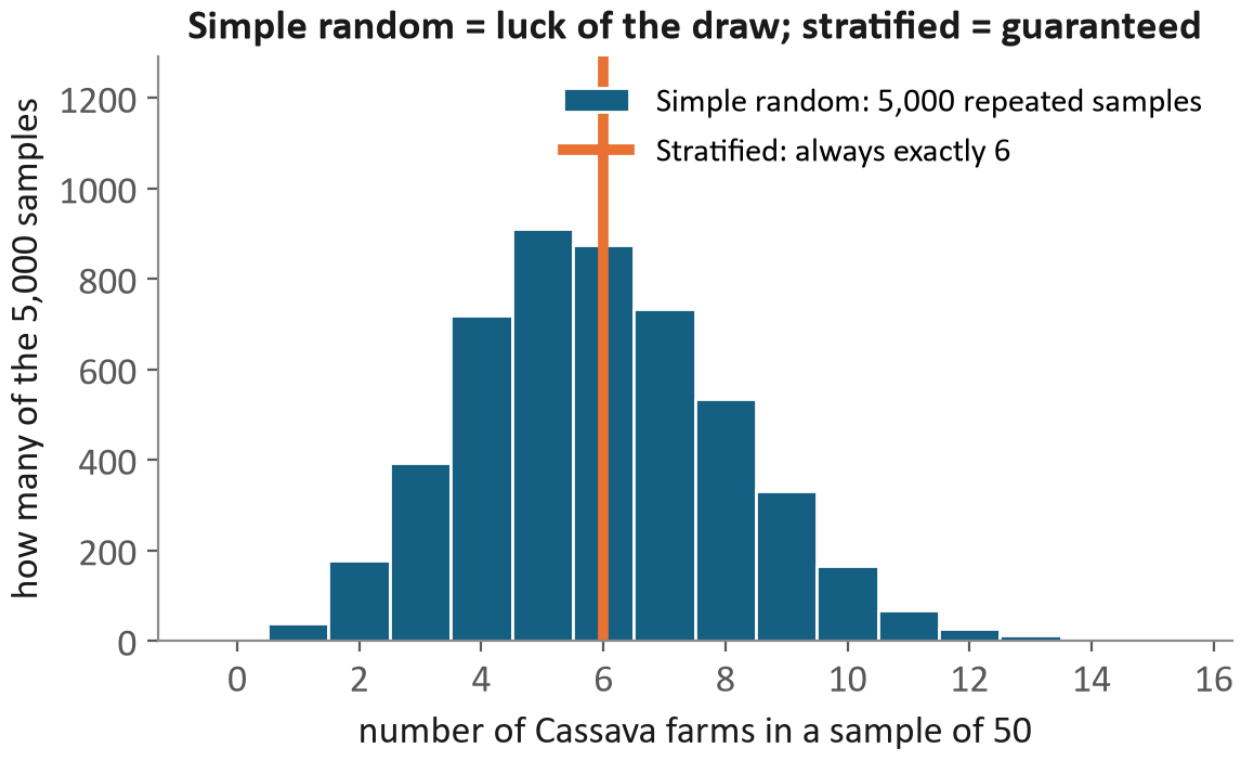
like rolling a die

- A chosen farm is put back
- The same farm can appear again
- Every draw: 1/500
- Used by the bootstrap (Week 8)

```
without:  df.sample(n=50)          with:  df.sample(n=50, replace=True)
```

Weakness of both: pure luck decides how many farms of each crop you get — next slide.

The Problem With Pure Luck



Cassava: 59 of 500 farms (11.8%)

1. **Fair share in n = 50**
   50 × 59 / 500 = 5.9 ≈ 6 farms

2. **Simple random, 5,000 tries**
   anywhere from 0 to 15 farms

3. **12% of samples**
   have 3 or fewer — half the fair share

✓ **Stratified**
exactly 6, every time

Lesson: when subgroups matter (crops, regions, genders), don't leave their representation to luck.

Stratified Sampling: The Formula

IIT Madras Zanzibar

$$n_h = n \times N_h / N$$

each group's share of the sample = its share of the population

1 **Share of population**
142 / 500 = 0.284 (28.4%)

2 **Same share of sample**
50 × 0.284 = 14.2

3 **Round, then draw randomly**
14 Clove farms, chosen at random

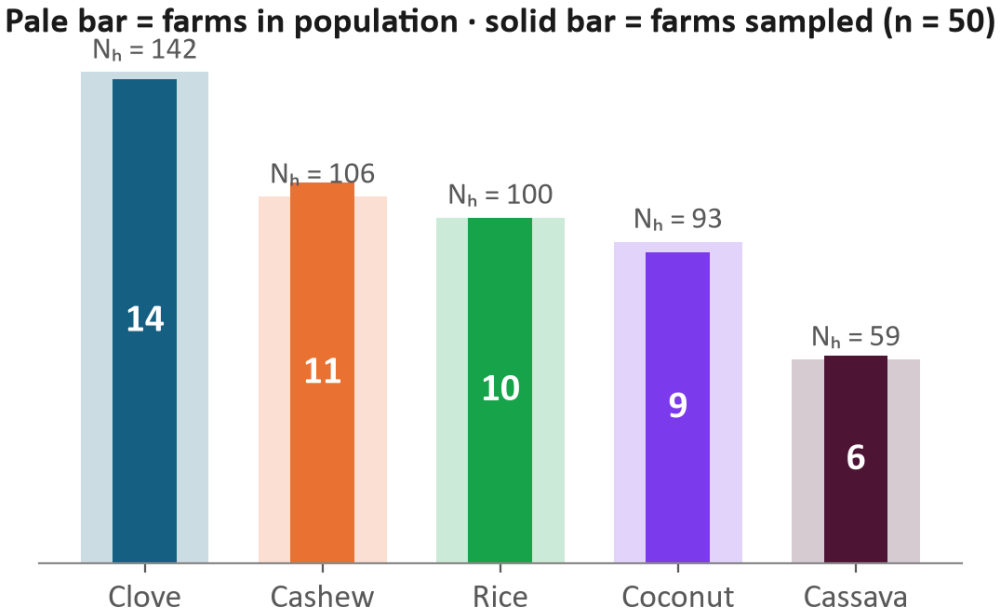
Symbol	What it means	In our example
h	one stratum (group)	Clove
N_h	records in that stratum	142
N	population size	500
n	total sample size wanted	50
n_h	how many to draw from h	?

In words: every group gets its fair share of the sample — no group can be left out by bad luck.

Worked Example: Stratify All 500 Farms by Crop

IIT Madras Zanzibar

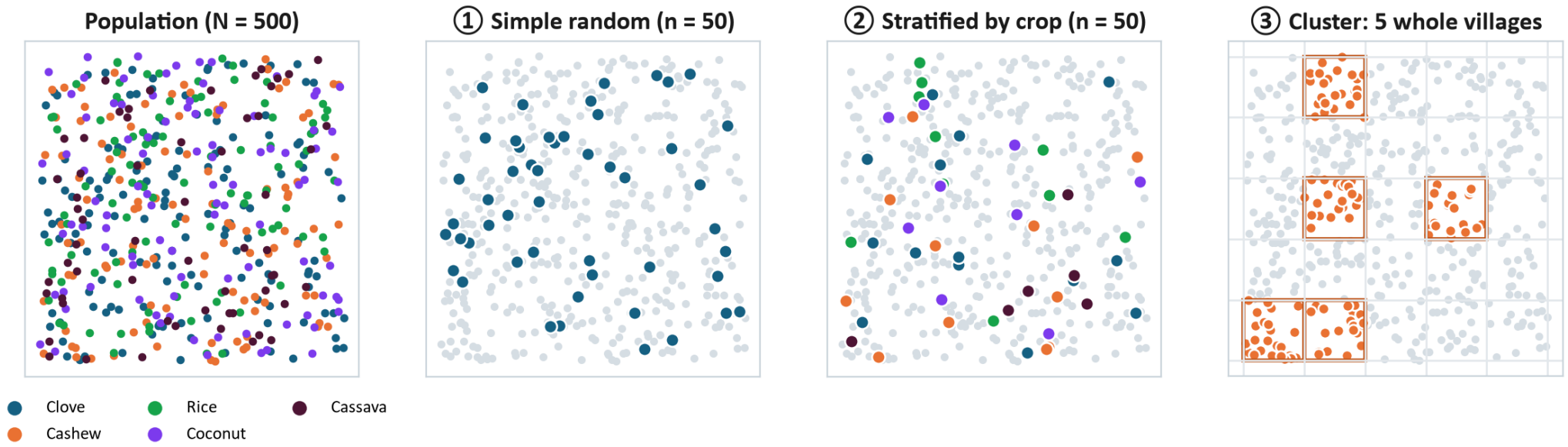
Crop	N_h	N_h / N	$n \times N_h / N$	n_h
Clove	142	0.284	14.2	**14**
Cashew	106	0.212	10.6	**11**
Rice	100	0.200	10.0	**10**
Coconut	93	0.186	9.3	**9**
Cassava	59	0.118	5.9	**6**
Total	**500**	**1.000**	**50.0**	**50 ✓**



If the total is off by one: round all down, then give the leftover places to the largest decimal parts (largest-remainder rule).

```
pandas:   df.groupby('crop').sample(frac=0.10, random_state=1)
```

Cluster Sampling: Cheap to Collect, Less Precise



How

1. Divide into natural clusters (villages)
2. Pick a few clusters at random
3. Survey every record inside them

Stratified vs cluster

Stratified: a FEW from EVERY group

Cluster: EVERY one from a FEW groups

Cluster is cheaper but less precise

Sampling in the Real World

IIT Madras Zanzibar

1 National surveys (e.g. DHS)

stratify by region → pick clusters → survey households

stratified + cluster

2 Opinion polls

stratify by region and age so every group is heard

stratified

3 Factory quality control

test a random sample of items, not all of them

simple random

4 Big-data prototyping

develop on 1% of the table, then run on 100%

simple random

5 Machine learning

stratified train/test split keeps rare classes in both

stratified

6 Financial audits

check a sample of transactions

random / stratified

Common thread: a well-drawn sample answers the question at a fraction of the cost.

Data Reduction at Scale: Sampling & PCA

IIT Madras Zanzibar

Foundations: Why Reduce, and What a Sample Is

Sampling Methods, Step by Step

→ **How Big Should a Sample Be?**

Sampling a Stream: Reservoir Sampling

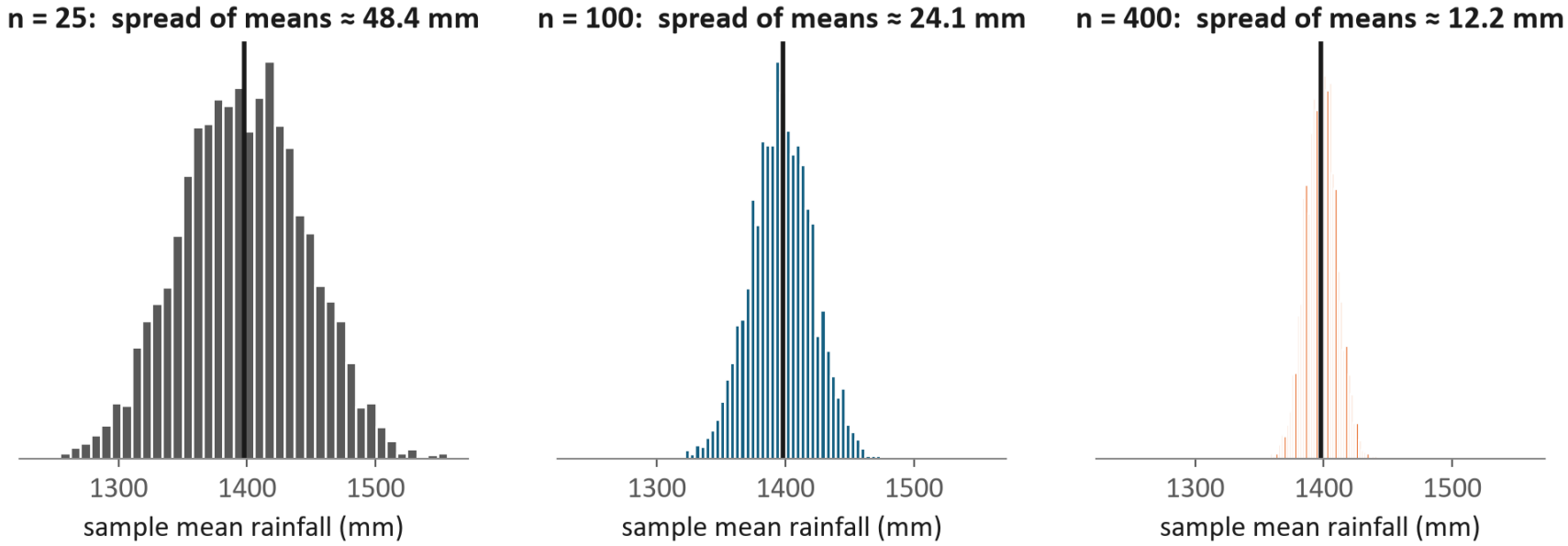
PCA from Zero

PCA on Our Farms, and at Scale

In-Class Exercise

Every Sample Wobbles: The Sampling Distribution

Draw a sample, compute its mean, repeat 3,000 times — bigger samples wobble less



Standard error (SE)

how far a sample's answer typically lands from the true answer

n	25	100	400
spread of means (mm)	48.4	24.1	12.2

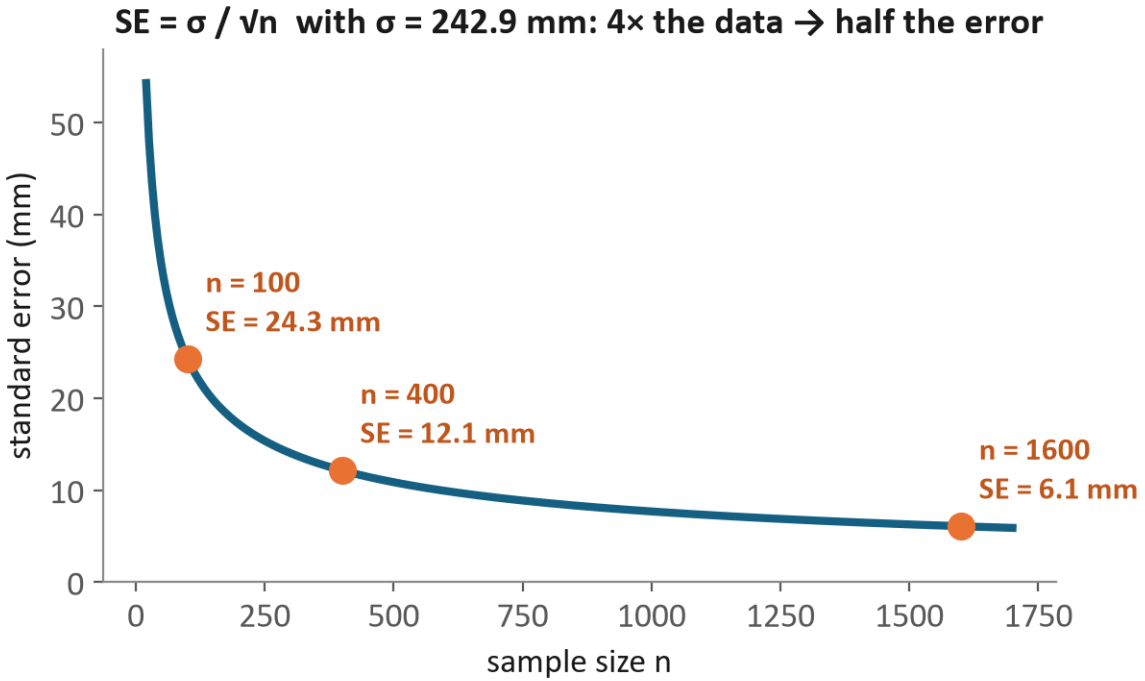
The √n Rule: SE = σ / √n

$$SE = \sigma / \sqrt{n}$$

to plan a study: $n = (\sigma / SE)^2$

Symbol	What it means	In our example
σ	spread of the attribute	242.9 mm
n	sample size	25, 100, 400
SE	typical error of the sample mean	?

n	√n	SE = σ/√n
100	10	24.3 mm
400	20	12.1 mm
1,600	40	6.1 mm



Diminishing returns: 4× the data → half the error; 100× the data → one-tenth the error. Planning: SE = 20 mm needs $n = (243/20)^2 \approx 148$ farms.

Data Reduction at Scale: Sampling & PCA

IIT Madras Zanzibar

Foundations: Why Reduce, and What a Sample Is

Sampling Methods, Step by Step

How Big Should a Sample Be?

→ **Sampling a Stream: Reservoir Sampling**

PCA from Zero

PCA on Our Farms, and at Scale

In-Class Exercise

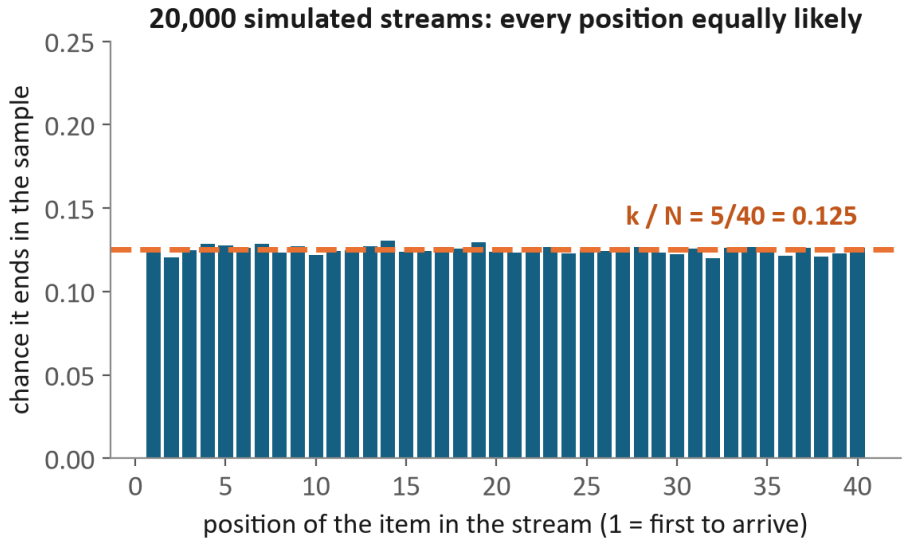
Reservoir Sampling: A Fair Sample From a Stream

IIT Madras Zanzibar

item n	rule (k = 2)	chance it is kept	each item's chance now
1, 2	fill the reservoir	1	2/2 = 1
3	keep with prob 2/3	2/3	2/3
4	keep with prob 2/4	1/2	2/4 = 1/2
n	keep with prob k/n	k/n	k/n

1 **Item 1 still there after item 3?**
$1 - (2/3 \times 1/2) = 2/3 = k/n$ ✓

2 **…and after item 4?**
$2/3 \times (1 - 2/4 \times 1/2) = 2/3 \times 3/4 = 1/2 = k/n$ ✓



Memory: only k slots, whether the stream has a thousand items or a billion.

Real use: sampling live sensor readings, network packets or transaction logs for quick analysis.

Data Reduction at Scale: Sampling & PCA

IIT Madras Zanzibar

Foundations: Why Reduce, and What a Sample Is

Sampling Methods, Step by Step

How Big Should a Sample Be?

Sampling a Stream: Reservoir Sampling

→ **PCA from Zero**

PCA on Our Farms, and at Scale

In-Class Exercise

PCA in Plain Words

IIT Madras Zanzibar

Definition

PCA replaces the original attributes by a few new ones — principal components — that are weighted combinations of the originals and capture as much of the spread (variance) as possible.

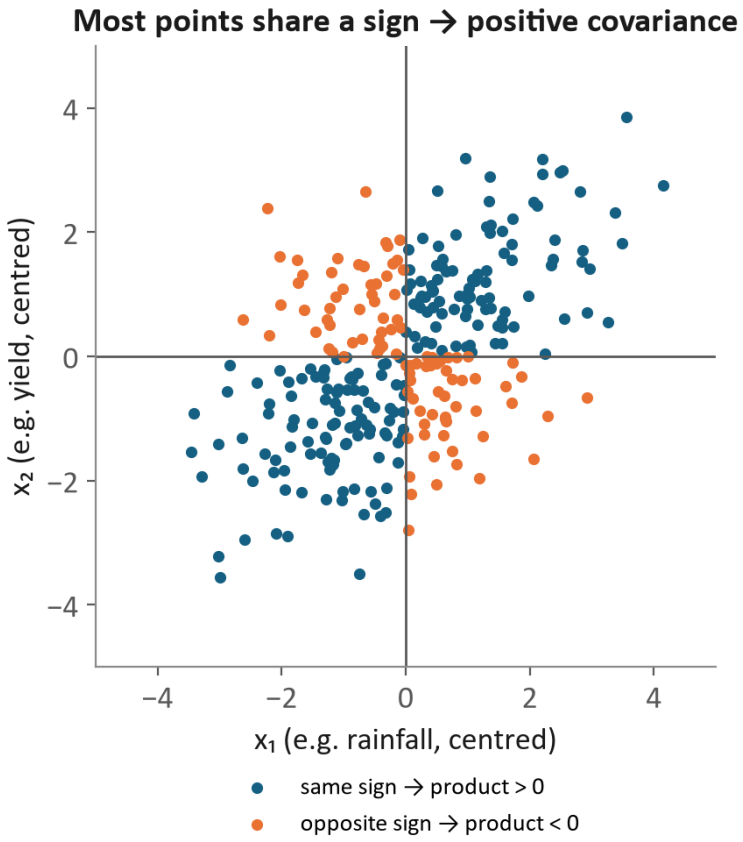
Analogy: photographing a fish

From the side you see almost everything; head-on you see a blob. PCA finds the side view of your data.

1. **PC1**
   direction of largest spread
2. **PC2**
   largest remaining spread, at 90° to PC1
3. **Uncorrelated**
   components never repeat each other
4. **Keep a few**
   fewer columns, most of the information

Why data mining cares: on our farms, price and revenue have r = 0.92 — two columns telling almost the same story. PCA merges them automatically.

Refresher: Covariance Tells How Two Attributes Move Together



$$\mathrm{Cov}(X_1, X_2) = \Sigma\,(x_1 - \mu_1)(x_2 - \mu_2) \,/\, n$$

the average product of the two centred values

1. **Positive**
   they rise and fall together (teal dominates)
2. **Negative**
   one rises as the other falls (orange dominates)
3. **Cov(X, X) = Var(X)**
   covariance of an attribute with itself is its variance
4. **Standardized data**
   covariance = correlation r

Hold on to this: PCA is built entirely from variances and covariances.

The Covariance Matrix: Every Pair in One Table

IIT Madras Zanzibar

Worked example (2 attributes)

Σ	x_1	x_2
x_1	2 (variance)	1 (covariance)
x_2	1 (covariance)	2 (variance)

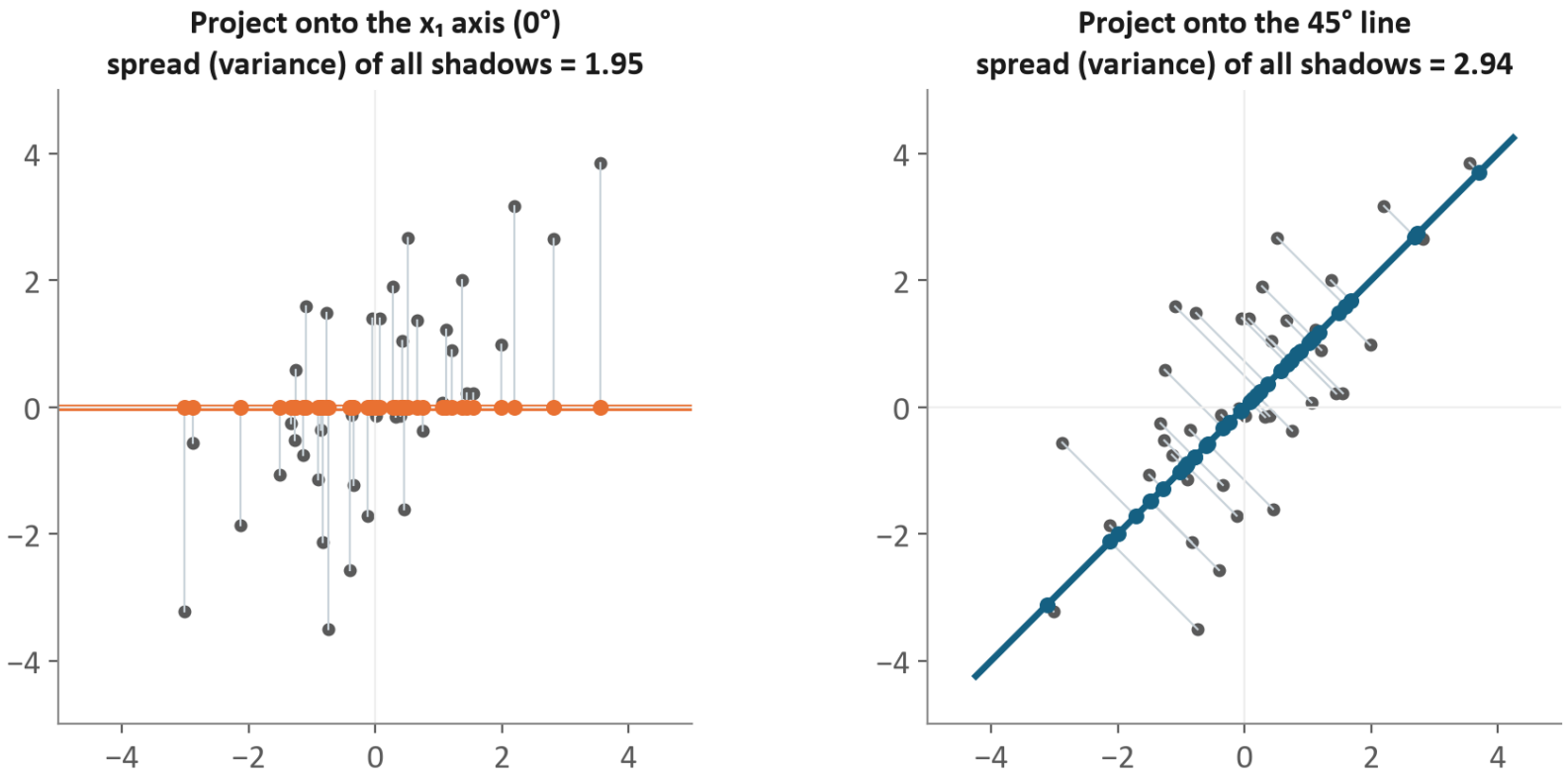
- Diagonal = each attribute's variance
- Off-diagonal = covariances
- Symmetric: Cov(A, B) = Cov(B, A)

Our 500 farms, standardized (= correlation matrix)

500 farms (z)	size	yield	rain	price	revenue
size	1.00	-0.02	0.02	-0.01	-0.03
yield	-0.02	1.00	0.01	-0.42	-0.23
rain	0.02	0.01	1.00	-0.01	-0.01
price	-0.01	-0.42	-0.01	1.00	0.92
revenue	-0.03	-0.23	-0.01	0.92	1.00

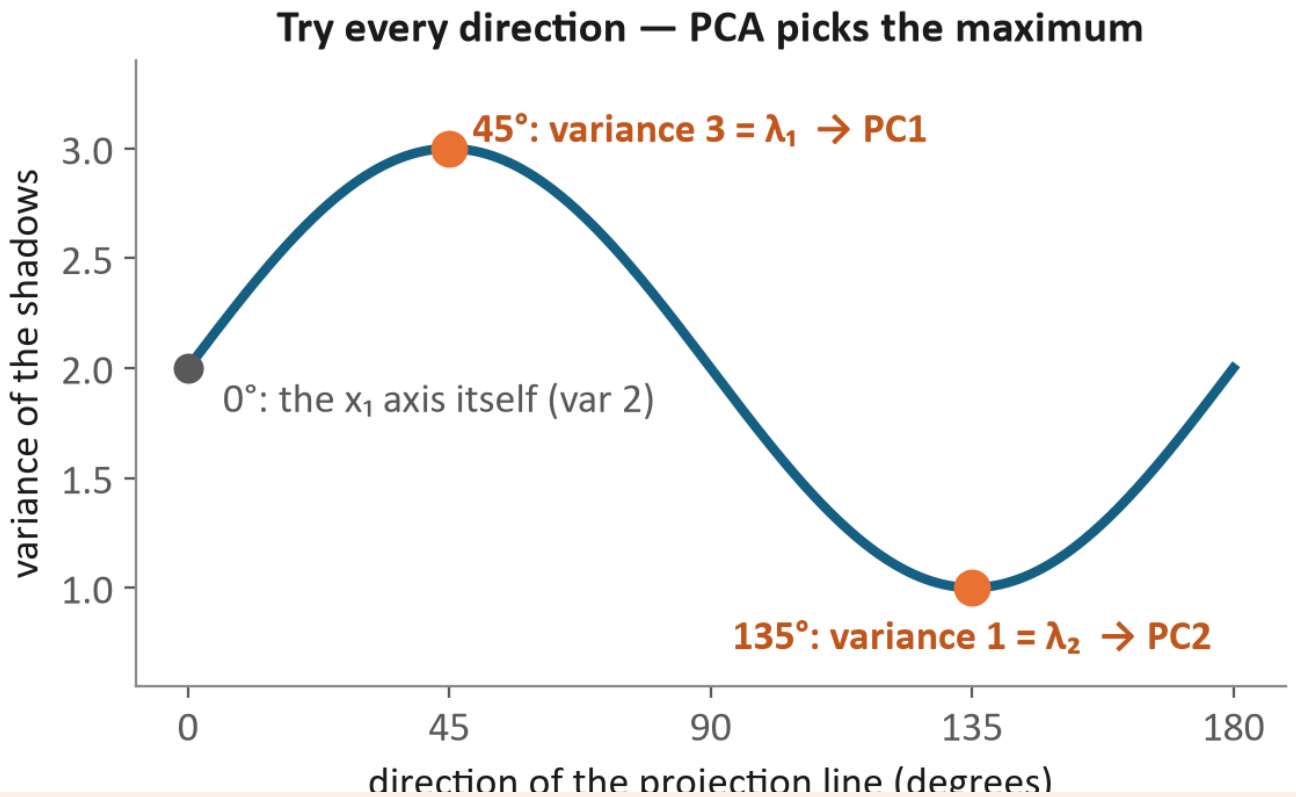
Spot the pattern: price & revenue r = 0.92; yield & price r = -0.42. Rainfall and farm size are nearly independent of the rest.

The Key Idea: Project the Data Onto a Line



The PCA question: which line makes the shadows most spread out? Spread = information; piled-up shadows lose it.

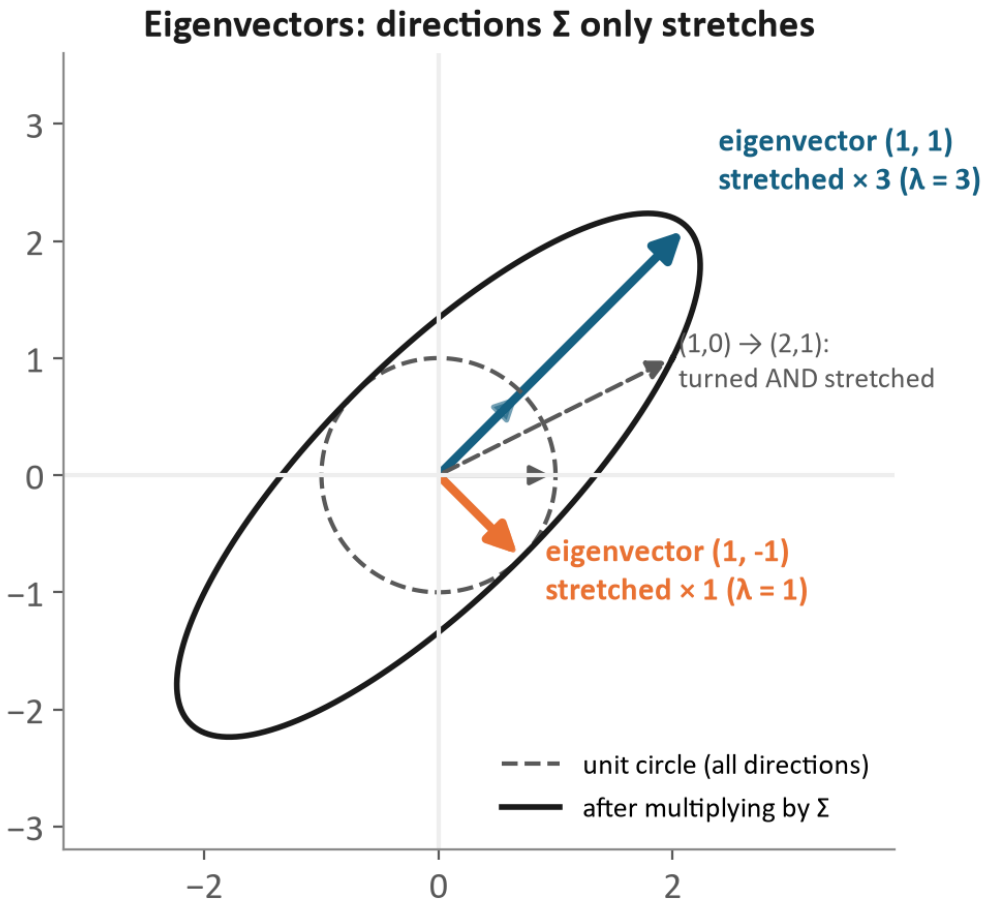
Try Every Direction: PCA Picks the Maximum



1. **Best direction: 45°**
   shadow variance 3 → this is PC1
2. **Worst direction: 135°**
   shadow variance 1 → this is PC2
3. **At right angles**
   PC1 and PC2 are always perpendicular
4. **Shortcut**
   eigenvalues give 3 and 1 without searching

Names: the best directions are the EIGENVECTORS of Σ; the variance along each is its EIGENVALUE (λ, 'lambda').

Eigenvectors in Plain Words



$$\Sigma v = \lambda v$$

multiplying by Σ only stretches v, by the factor λ

Symbol	What it means	In our example
v	eigenvector: a special direction	(1, 1)
λ	eigenvalue: the stretch factor	3
Σ	covariance matrix	[[2, 1], [1, 2]]

✓ **Check (1, 1)**

Σ(1, 1) = (2+1, 1+2) = (3, 3) = 3 × (1, 1)

✓ **Check (1, −1)**

Σ(1, −1) = (2−1, 1−2) = (1, −1) = 1 × (1, −1)

Worked Example (1 of 3): Find the Eigenvalues

IIT Madras Zanzibar

$$\det [[a, b], [c, d]] = ad - bc$$

tool: the determinant of a 2 × 2 matrix

Symbol	What it means	In our example
I	identity matrix	[[1, 0], [0, 1]]
λ	unknown eigenvalue	?
det	determinant: ad − bc	—

1 **Subtract λ from the diagonal**
$\Sigma - \lambda I = [[2 - \lambda, 1], [1, 2 - \lambda]]$

2 **Determinant**
$(2 - \lambda)(2 - \lambda) - (1)(1)$

3 **Set to zero**
$(2 - \lambda)^2 = 1 \rightarrow 2 - \lambda = \pm 1$

4 **Solve**
$\lambda_1 = 3,\ \lambda_2 = 1$ (largest first)

Check: eigenvalues add up to the total variance on the diagonal: 3 + 1 = 2 + 2 = 4. Variance explained by PC1 = 3 / 4 = 75%; PC2 = 1 / 4 = 25%.

Worked Example (2 of 3): Find the Eigenvectors

IIT Madras Zanzibar

$\lambda_1 = 3 \rightarrow$ PC1

1. **$\Sigma - 3I$**
   $[[-1, 1], [1, -1]]$
2. **Solve $-v_1 + v_2 = 0$**
   $v_2 = v_1 \rightarrow (1, 1)$
3. **Make length 1: ÷ √2**
   PC1 = (0.707, 0.707)

Read PC1

equal weight on both attributes: an overall 'productivity' score (the sum)

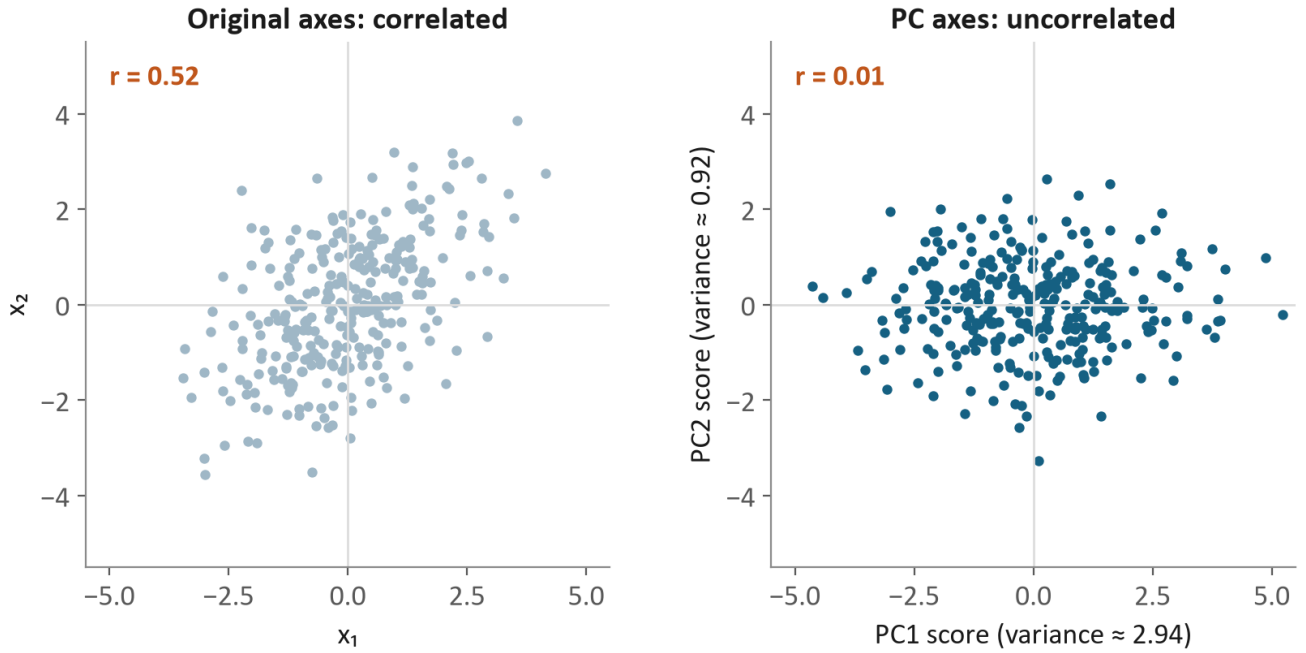
$\lambda_2 = 1 \rightarrow$ PC2

1. **$\Sigma - 1I$**
   $[[1, 1], [1, 1]]$
2. **Solve $v_1 + v_2 = 0$**
   $v_2 = -v_1 \rightarrow (1, -1)$
3. **Make length 1: ÷ √2**
   PC2 = (0.707, −0.707)

Read PC2

the difference between them: high rainfall but low yield, or the reverse

Worked Example (3 of 3): Scores, and What We Lose



One farm: x = (1.0, 0.5)

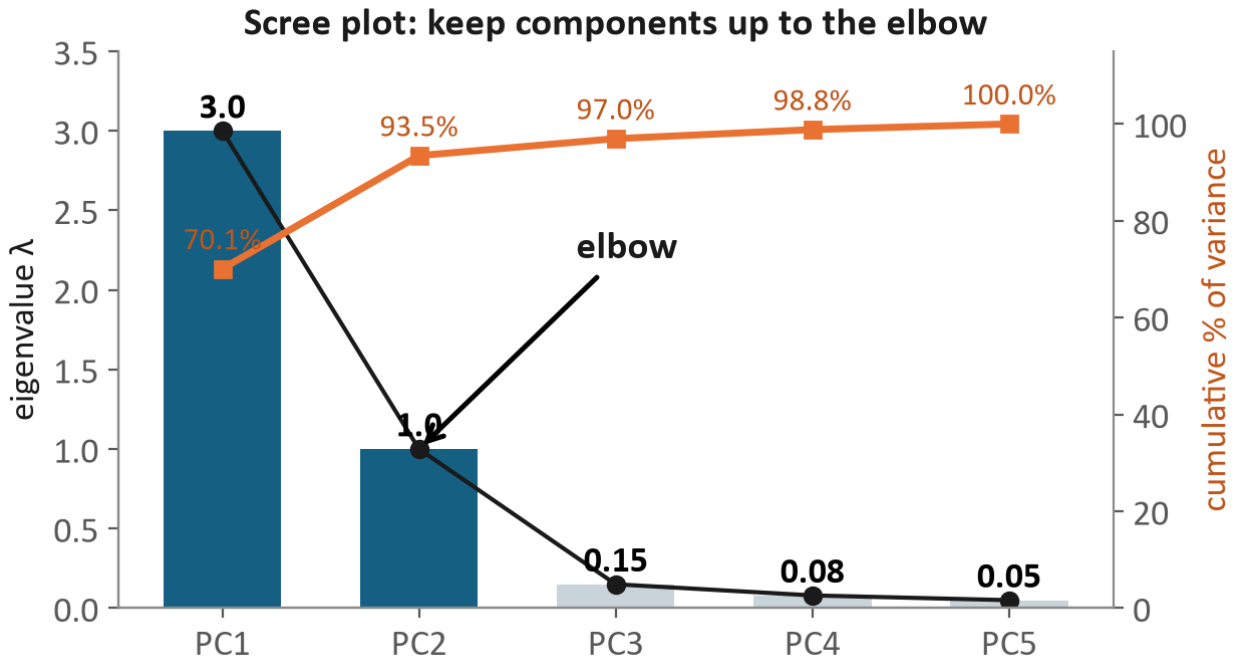
1. **PC1 score**
   0.707·1.0 + 0.707·0.5 = 1.06

2. **PC2 score**
   0.707·1.0 − 0.707·0.5 = 0.35

3. **Keep PC1 only → rebuild**
   1.06 × (0.707, 0.707) = (0.75, 0.75)

What we lose: true point (1.0, 0.5) vs rebuilt (0.75, 0.75) — a small error, equal to the dropped PC2 part. We kept 75% of the variance with half the columns. PCA is a rotation of the axes; dropping components is the reduction.

How Many Components to Keep? The Scree Plot



1 **Elbow rule**
keep components before the curve flattens

2 **Cumulative target**
keep enough for 80–90% of variance

3 **Kaiser rule (standardized)**
keep components with $\lambda > 1$

✓ **Here: keep 2**
(3.0 + 1.0) / 4.28 = 93.5% of variance

Formula: % variance explained by PCk = $\lambda k / (\lambda 1 + \lambda 2 + \ldots + \lambda d)$.

Data Reduction at Scale: Sampling & PCA

IIT Madras Zanzibar

Foundations: Why Reduce, and What a Sample Is

Sampling Methods, Step by Step

How Big Should a Sample Be?

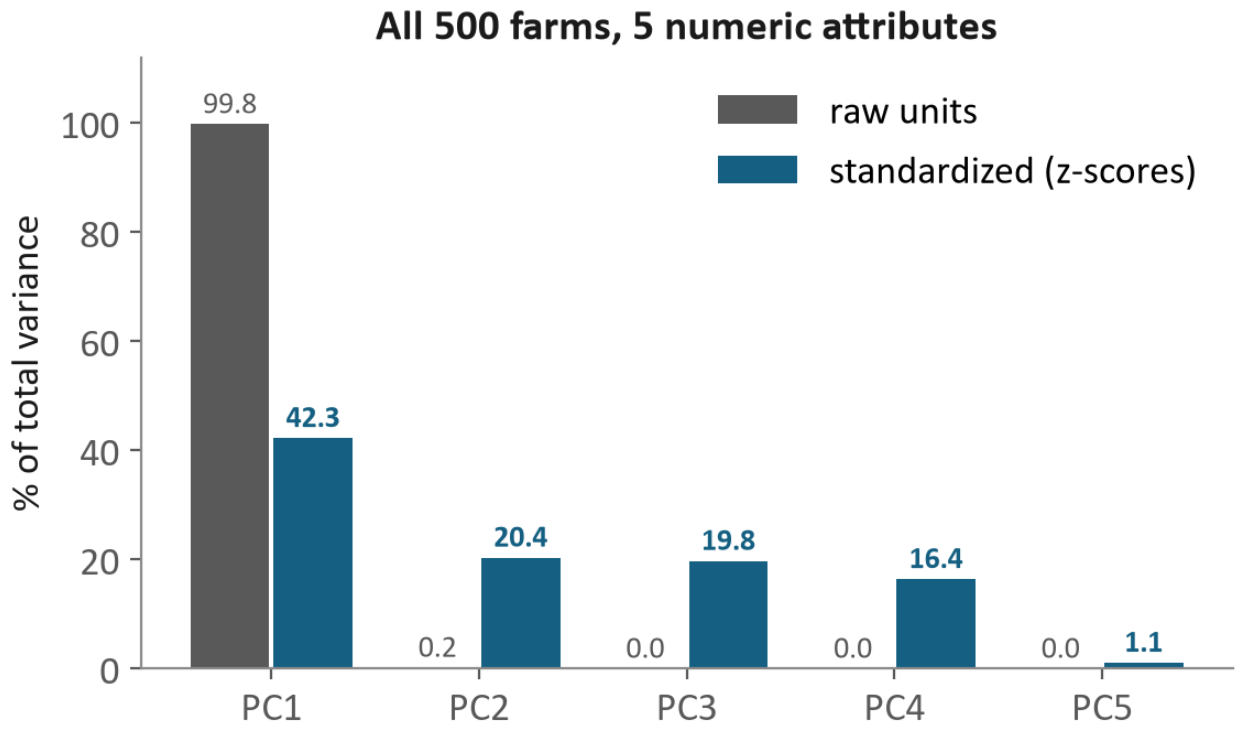
Sampling a Stream: Reservoir Sampling

PCA from Zero

→ **PCA on Our Farms, and at Scale**

In-Class Exercise

PCA on Our 500 Farms: Why You Must Standardize First



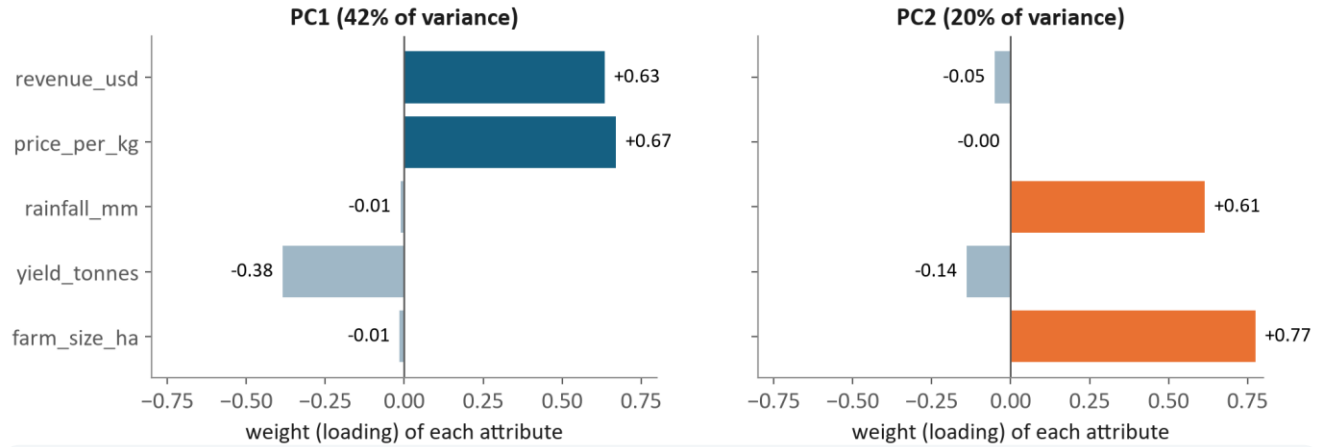
1. **Raw units: PC1 = 99.8%**
   because revenue's variance ≈ 33,000,000 vs farm size ≈ 2

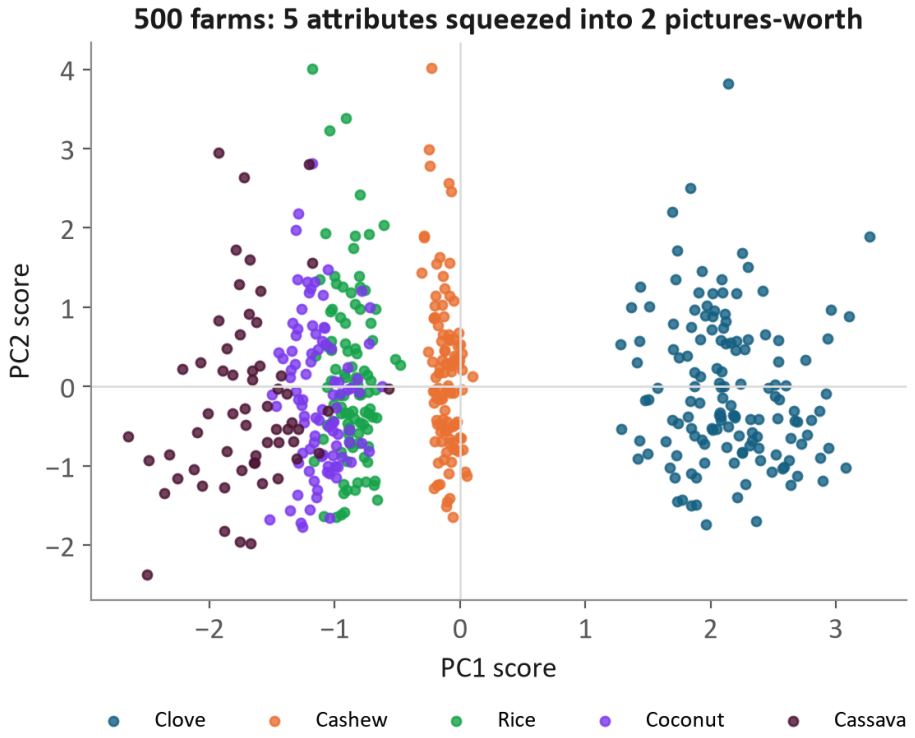
2. **That is a scale artefact**
   PC1 is just 'revenue' — yesterday's lesson again

3. **Standardized: 42 / 20 / 20 / 16 / 1 %**
   an honest split of the information

Rule: standardize (z-score) before PCA unless every attribute is in the same unit. In scikit-learn: StandardScaler() then PCA().

Reading the Components: Loadings





1 **PC1 (42%): 'high-value crop'**
price +0.67, revenue +0.63, yield -0.38

2 **PC2 (20%): 'big, wet farm'**
farm size +0.77, rainfall +0.61

Notice: crops separate along PC1 in the scatter — even though PCA never saw the crop labels.

How Many to Keep for Our Farms?

IIT Madras Zanzibar

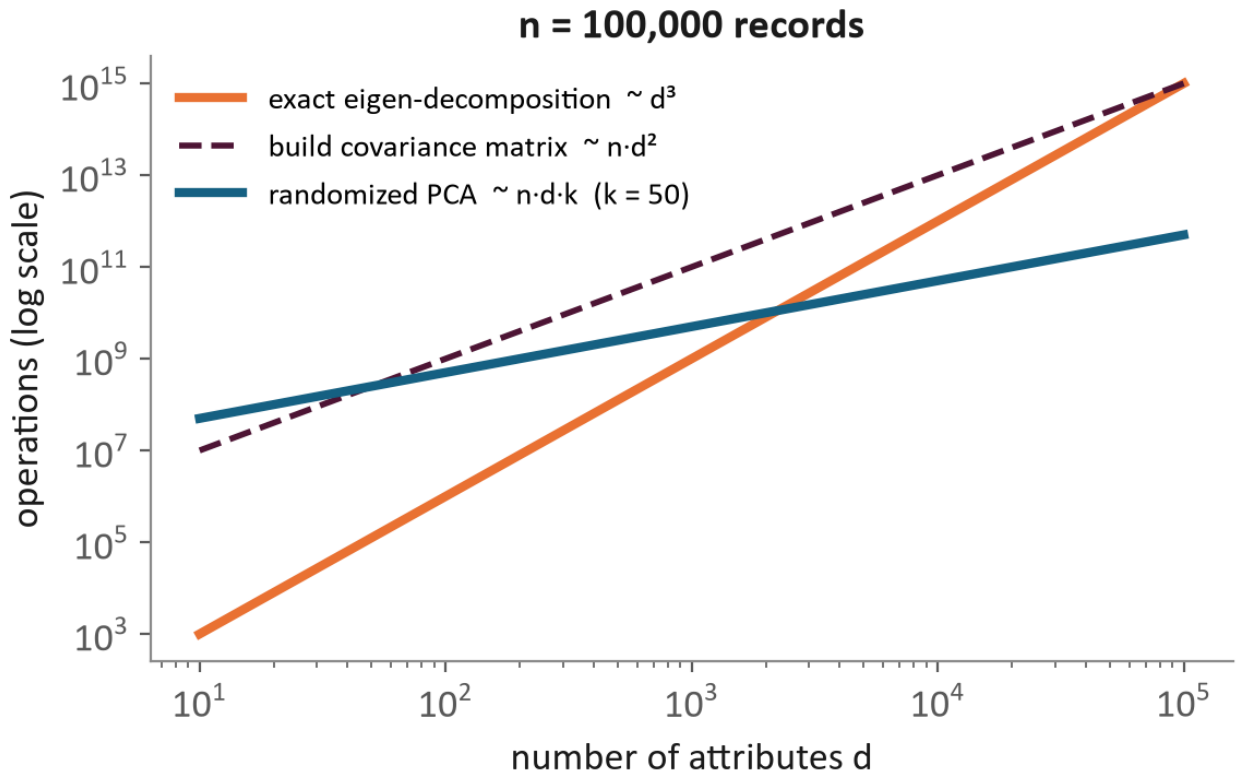
	PC1	PC2	PC3	PC4	PC5
eigenvalue λ	2.12	1.02	0.99	0.82	0.06
% of variance	42.3%	20.4%	19.8%	16.4%	1.1%
cumulative %	42.3%	62.7%	82.4%	98.9%	100.0%
Kaiser (λ > 1)?	keep	keep	borderline	drop	drop

1 **Keep 2 → 63%**
Kaiser rule; good for 2-D plots

2 **Keep 4 → 99%**
drops only the redundant price/revenue part

Honest lesson: real data compresses less than textbook examples. Choosing k is a judgement you must justify — part of the mini-project rubric.

Why Full PCA Struggles at Scale



step	cost	d = 10,000
covariance matrix	$n \cdot d^2$	10^{13} ($n = 10^5$)
exact eigen-decomposition	d^3	10^{12}
randomized PCA (k = 50)	$n \cdot d \cdot k$	5×10^{10}

Fix: randomized PCA finds only the top k components directly. sklearn: PCA(n_components=50, svd_solver='randomized').

Course theme: exact-but-expensive vs approximate-but-scalable — the same trade-off as reservoir sampling and streaming statistics.

PCA in the Real World

IIT Madras Zanzibar

1 Genomics

thousands of gene measurements → a few components

2 Face recognition

'eigenfaces': faces as a few components of pixels

3 Finance

hundreds of stock returns → a few market factors

4 Satellite imagery

many spectral bands (crop monitoring) compressed

5 Surveys

many related questions → a few underlying factors

6 Visualization

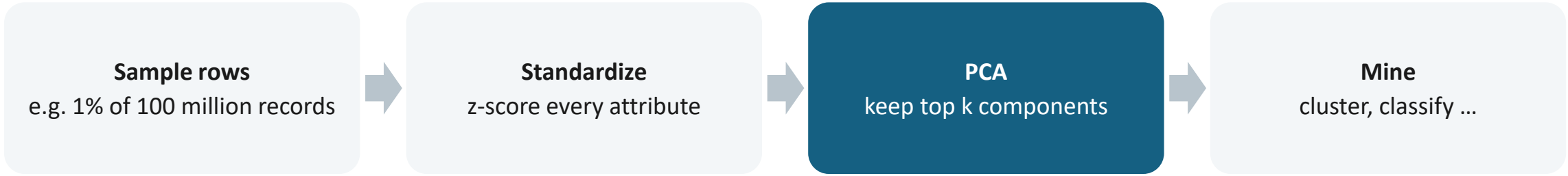
any high-dimensional data drawn in 2-D (PC1, PC2)

Common thread: many correlated measurements → a few components that carry most of the information.

Sampling vs PCA: Two Tools That Combine

IIT Madras Zanzibar

	Reduces	Preserves	Use when
Sampling	rows (records)	each attribute's distribution	too many records to process
PCA	columns (attributes)	most of the total variance	many, correlated attributes



Standard pipeline: sample → standardize → PCA → mine. Each step makes the next one cheaper.

Data Reduction at Scale: Sampling & PCA

IIT Madras Zanzibar

Foundations: Why Reduce, and What a Sample Is

Sampling Methods, Step by Step

How Big Should a Sample Be?

Sampling a Stream: Reservoir Sampling

PCA from Zero

PCA on Our Farms, and at Scale

→ **In-Class Exercise**

In-Class Exercise (12 minutes, pairs)

IIT Madras Zanzibar

IN-CLASS TASK

Our 500 farms by region: Pemba North 109 · Pemba South 107 · North Unguja 98 · Zanzibar Urban 97 · South Unguja 89

1. Compute a proportional stratified sample of n = 40 by region. Check that it adds to 40.

2. Rainfall has σ = 242.9 mm. What is the SE at n = 100? What n halves it? What n cuts it to a quarter?

3. $\Sigma = [[4, 0], [0, 1]]$. Eigenvalues? % of variance explained by PC1? (Hint: diagonal matrix.)

4. $\Sigma = [[5, 2], [2, 2]]$. Find both eigenvalues with the determinant method, and PC1's % of variance.

Solution

Region	N_h	$40 \times N_h / 500$	round down	+ leftover	n_h
Pemba North	109	8.72	8	+1	**9**
Pemba South	107	8.56	8		**8**
North Unguja	98	7.84	7	+1	**8**
Zanzibar Urban	97	7.76	7	+1	**8**
South Unguja	89	7.12	7		**7**
Total	500	40.00	37	+3	**40 ✓**

2 **SE**
242.9/√100 = 24.3 mm; halve: n = 400; quarter: n = 1,600

3 **Diagonal Σ**
λ = 4, 1 → PC1 = 4/5 = 80%

4 **$(5-\lambda)(2-\lambda) - 4 = 0$**
$\lambda^2 - 7\lambda + 6 = 0 \rightarrow \lambda = 6, 1 \rightarrow$ PC1 = 6/7 = 85.7%

Q1 lesson: ordinary rounding gives 41 — one too many. Round down (37 farms), then give the 3 leftover places to the largest decimal parts.

Q3 lesson: uncorrelated attributes → PCA simply returns the original axes. PCA helps only when attributes are correlated.

Summary & Exam Checkpoints

IIT Madras Zanzibar

Sampling: fewer rows

Stratified guarantees every group its fair share: $n_h = n \cdot N_h / N$. Cluster is cheaper but less precise.

SE = σ / √n

4× the data halves the error. Plan with $n = (\sigma / SE)^2$. Reservoir sampling handles streams with k slots.

PCA: fewer columns

Eigenvectors of Σ = directions of greatest spread; eigenvalues = variance along them; % = $\lambda / \Sigma\lambda$.

Standardize first

On our farms raw PCA gave a fake 99.8%; standardized gave an honest 42/20/20/16/1%.

Common Questions

IIT Madras Zanzibar

Q **Is a bigger sample always better?**

A More accurate, yes — but with diminishing returns (√n). A biased sample stays biased however big it is.

Q **Does PCA delete attributes?**

A No. Every component is a mix of ALL attributes. It replaces columns; it does not choose among them (that is feature selection).

Q **Can I interpret a principal component?**

A Often, from its loadings — e.g. PC1 on our farms is a 'high-value crop' axis. Sometimes no clear meaning exists.

Q **Should I use PCA before a decision tree?**

A Usually not needed — trees handle many attributes and lose readability on PCs. It helps k-NN, k-means and plotting.

How This Feeds Your Assessment — and What's Next

IIT Madras Zanzibar

Assessment

Quiz 1: stratified allocation, SE calculations, 2 × 2 eigenvalues and % variance

Mini-project: if you use PCA — standardize, show the scree plot, justify k

Tutorial 3 (today)

Normalization, discretization and PCA in Python on the Zanzibar data — Mini-Project announced

Friday: Lecture 9

Data warehousing: storing cleaned, reduced data for fast analysis

Bring to the tutorial: a laptop with Python, pandas and scikit-learn, and today's worked examples.

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed. Morgan Kaufmann — Sections 2.4 (data reduction), PCA and sampling.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2(11), 559–572.
- Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A, 374.
- Halko, N., Martinsson, P.-G. & Tropp, J. A. (2011). Finding structure with randomness. SIAM Review, 53(2), 217–288.
- Vitter, J. S. (1985). Random sampling with a reservoir. ACM Transactions on Mathematical Software, 11(1), 37–57.
- Cochran, W. G. (1977). Sampling Techniques, 3rd ed. Wiley.