Recap: Dimensionality Reduction — PCA and t-SNE

DS701 Session 15 — Mon Oct 26, 2026

Today's plan

Knowledge-Check Review

KC 1: The first step, and why

What is the very first step of PCA, and in one sentence, why is it required?

KC 2: PCA and the SVD

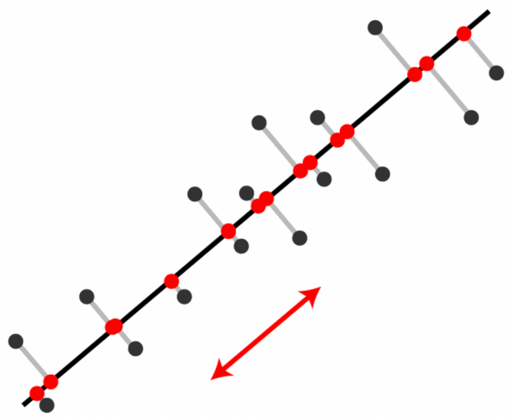
If the centered data matrix has SVD $X = U\Sigma V^T$*, where do the principal components live, and how do the singular values relate to the eigenvalues of the covariance matrix* S*?*

KC 3: Two things a t-SNE plot does *not* tell you

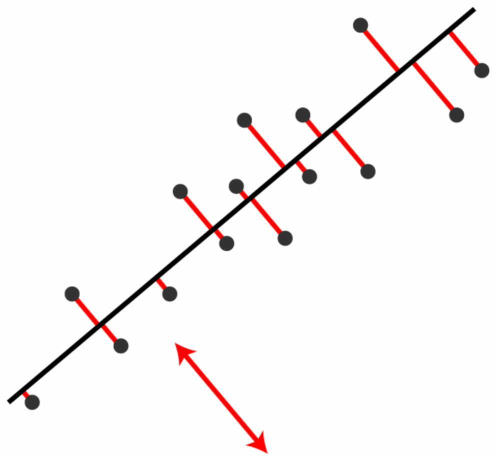
Name two things you should not *conclude from a t-SNE plot, and one thing t-SNE genuinely does preserve.*

Highlights

1. PCA in one picture



The sum of **variances** of the projected points is a **maximum**.



The sum of **residuals** (squared distances from points to the line) is a **minimum**.

2. The PCA recipe

3. What $Y = XV'$ actually does

Rotation, then projection.

- V' defines d new orthogonal axes pointing along the directions of maximum variance.
- Each point becomes its coordinates along those axes:

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{x}_i \cdot \mathbf{v}_1 & \mathbf{x}_i \cdot \mathbf{v}_2 & \cdots & \mathbf{x}_i \cdot \mathbf{v}_d \end{bmatrix}$$

- The discarded directions are the low-variance ones.

4. In practice: PCA *is* the SVD of the centered data

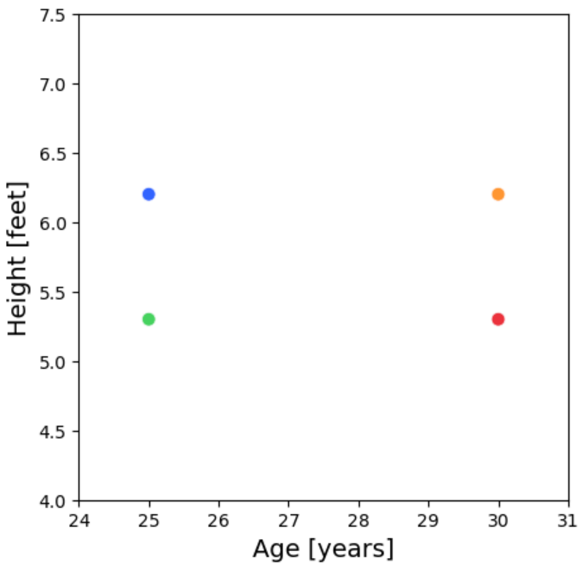
$$X = U\Sigma V^T \quad \implies \quad X^T X = V\Sigma^2 V^T$$

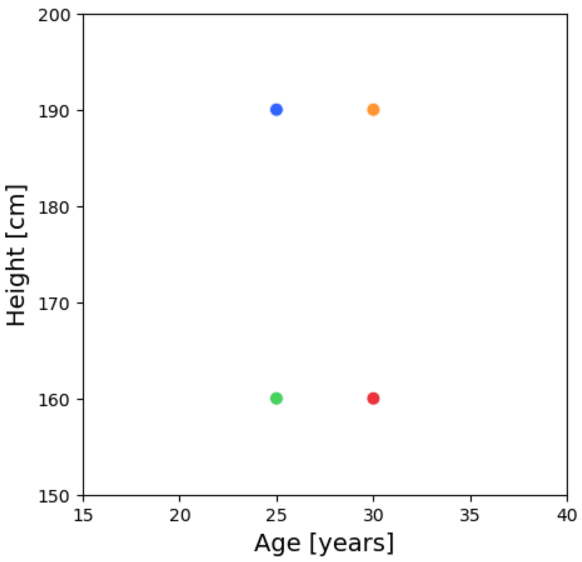
5. Choosing d: explained variance and the scree plot

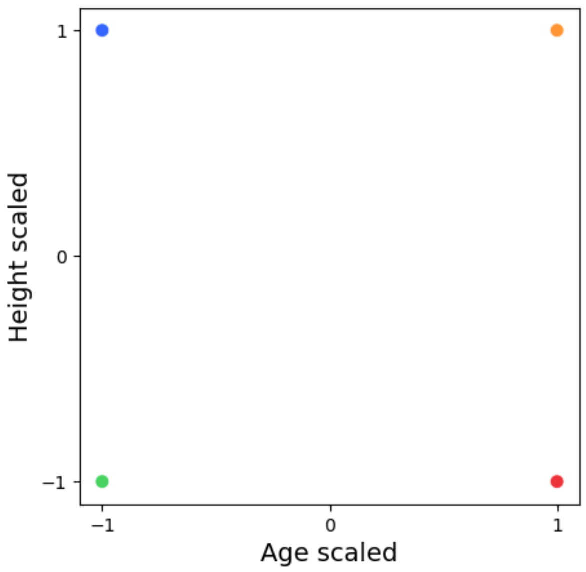
Total variance $T = \lambda_1 + \lambda_2 + \cdots + \lambda_n$.

6. To scale or not to scale

Same four people, height in feet vs. centimeters:







Height in **feet** — variance runs horizontally (age).

Height in **cm** — variance now runs vertically.

Standardized — variance shared evenly.

7. Why this matters: genes mirror geography

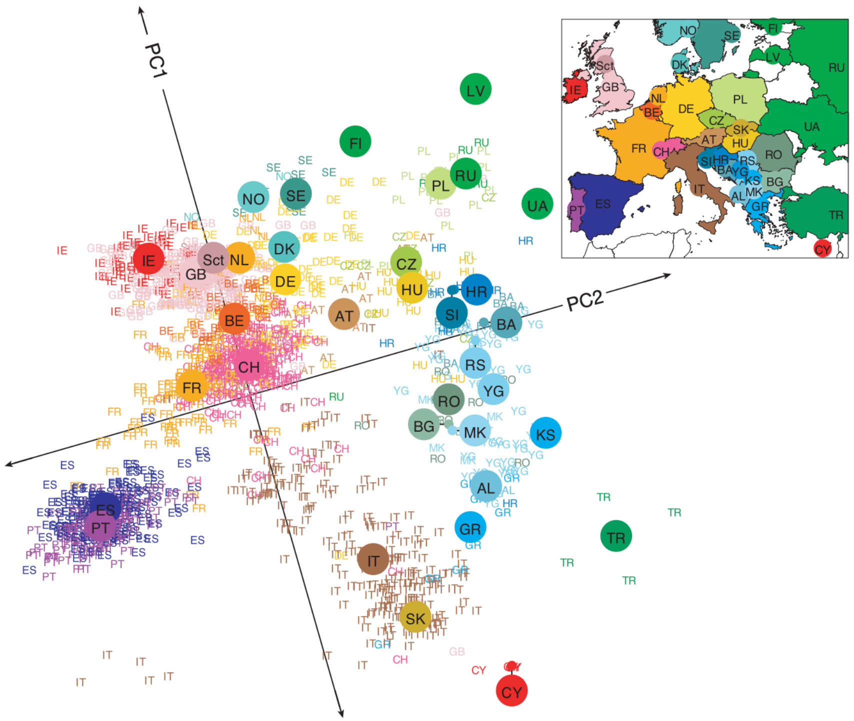


Image credit: Novembre et al., 2008

8. t-SNE: a different objective entirely

9. PCA vs. t-SNE

Feature	PCA	t-SNE
Speed	Fast	Slow (>10k points hurts)
Purpose	Dimensionality reduction	Visualization
Preserves	Global variance	Local neighborhoods
Linear?	Yes	No
Deterministic?	Yes	No (random init)
Distances	Meaningful	Not meaningful
Reusable transform?	Yes	No

10. t-SNE cautions — read this before you interpret a plot

Don’t trust

- cluster **sizes** (dense regions get inflated)
- **distances between** clusters
- any structure that isn’t stable across seeds
- the coordinates as model features

Do

- standardize / PCA-preprocess first
- try 3–5 perplexities (5, 15, 30, 50)
- rerun with different seeds
- read only **local** neighborhoods

In-Class Activity

Activity: PCA and t-SNE on real data

Goal: run the full PCA pipeline yourself (standardize, choose d from explained variance, reconstruct, verify against the SVD), then **stretch**: run t-SNE at four perplexities and two seeds, and put a *number* on "local neighborhoods are preserved" while cluster sizes and gaps are not.

- Work in groups of 2–3.
- Parts marked **(autograded)** are submitted to Gradescope; the rest is participation.
- Part 4 asks for your group's **rule for reading a t-SNE plot** — staff will circulate and ask, so agree on your answer.

Section A1 Open in Colab **Section B1** Open in Colab Open on GitHub

> ⓘ The activity notebook goes live on the day of the lecture. Colab is optional — you can also open it on GitHub and run it locally.

Going deeper

Full lecture notes: Dimensionality Reduction — PCA + t-SNE

- Covariance matrix, derived step by step
- Spectral decomposition
- Geometric interpretation of $Y = XV'$
- Relationship to the SVD
- To scale or not to scale
- Choosing σ / perplexity
- KL divergence objective
- t-SNE limitations

Prerequisite refresher: Eigendecomposition