Understanding PCA Components and Explained Variance Ratio

What Are Components/Loadings in PCA?

Components in Principal Component Analysis (PCA) are new variables created by combining original data, such as internet access, device ownership, or digital literacy in villages. Think of them as "recipes" that mix these indicators to capture key patterns in the data. Loadings are the "ingredients" in these recipes, showing how much each original indicator contributes to a component. A higher loading means that indicator has a greater influence.

Example

Consider data on three indicators for villages:

- Internet speed (Mbps)
- Device ownership (% of households with smartphones)
- Digital literacy (% of people who can use online tools)

PCA might create a **first component** that's a mix like:

- 50% internet speed + 40% device ownership + 10% digital literacy

This component might represent "overall digital access." The numbers (50%, 40%, 10%) are the **loadings**, showing each indicator's contribution. A village scoring high on this component likely has good internet, many devices, and decent literacy.

A **second component** might focus on a different pattern:

- 10% internet speed + 20% device ownership + 70% digital literacy

This could represent "digital skills dominance," with loadings showing digital literacy as the main driver.

What Is Explained Variance Ratio?

Explained variance ratio indicates how much of the data's total variation each component captures. It's like saying, "How much of the story about digital access does this component explain?" Each component gets a percentage, and they add up to 100% across all components.

Example

Using the village data:

- The **first component** (overall digital access) might have an explained variance ratio of 0.6 (60%), capturing 60% of the differences in digital access across villages.
- The **second component** (digital skills dominance) might have an explained variance ratio of 0.25 (25%), capturing another 25%.
- A **third component** might capture 0.15 (15%), and so on.

Using only the first component for the Village Digital Accessibility Index (VDAI) summarizes 60% of the data's variation. Adding the second component covers 85% (60% + 25%), providing a fuller picture.

Why Does This Matter?

- **Components/loadings** help understand what drives digital access (e.g., is it mostly internet speed or literacy?).
- **Explained variance ratio** helps decide how many components to use for the VDAI to capture enough of the data's story without overcomplicating things.

This approach simplifies the data while keeping the most important patterns, making it easier to compare villages' digital accessibility.