Principal Component Analysis

IS382 - Predictive Modeling

Jason Bryer

Week 10

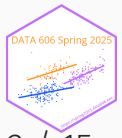
Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsuppervised (i.e. no outcome or dependent variable used) statistical and linear algebra technique used to reduce the dimensionality of a dataset while keeping as much important information and variance as possible.

How it works:

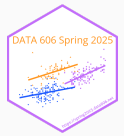
- First Principal Component (PC1): The new axis or direction that captures the largest amount of total variance (spread) in the data.
- Second Principal Component (PC2): The direction that captures the next highest amount of variance while remaining orthogonal (at a 90-degree angle) to PC1.
- Subsequent PCs: Additional orthogonal axes that capture progressively smaller amounts of remaining variance.

The math behind PCA relies on standardizing the data, computing a covariance matrix, and extracting eigenvectors (which define the new directions) and eigenvalues (which measure how much variance each direction holds).



Applications of PCA

- Data Visualization: Compresses high-dimensional datasets into 2D or 3D scatter plots so you can visually spot clusters, trends, or outliers.
- Machine Learning Preprocessing: Reduces the number of features to speed up training time and prevent overfitting.
- Noise Filtering & Compression: Drops minor components that carry little variance, stripping away random noise while preserving the core signal.



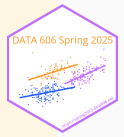
Example: Protein Data

The protein dataset is a real-valued multivariate dataset describing the average protein consumption by citizens of 25 European countries.

For each country, there are ten columns. The first eight correspond to the different types of proteins. The last one corresponds to the total value of the average values of proteins.

```
protein_data <- read.csv("../data/protein.csv")
str(protein_data)
```

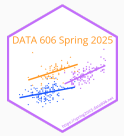
```
## 'data.frame':    25 obs. of  10 variables:
##  $ Country  : chr  "Albania" "Austria" "Belgium" "Bulgaria" ...
##  $ RedMeat  : num  10.1 8.9 13.5 7.8 9.7 10.6 8.4 9.5 18 10.2 ...
##  $ WhiteMeat: num  1.4 14 9.3 6 11.4 10.8 11.6 4.9 9.9 3 ...
##  $ Eggs     : num  0.5 4.3 4.1 1.6 2.8 3.7 3.7 2.7 3.3 2.8 ...
##  $ Milk     : num  8.9 19.9 17.5 8.3 12.5 25 11.1 33.7 19.5 17.6 ...
##  $ Fish     : num  0.2 2.1 4.5 1.2 2 9.9 5.4 5.8 5.7 5.9 ...
##  $ Cereals  : num  42.3 28 26.6 56.7 34.3 21.9 24.6 26.3 28.1 41.7 ...
##  $ Starch   : num  0.6 3.6 5.7 1.1 5 4.8 6.5 5.1 4.8 2.2 ...
##  $ Nuts     : num  5.5 1.3 2.1 3.7 1.1 0.7 0.8 1 2.4 7.8 ...
##  $ Fr.Veg   : num  1.7 4.3 4 4.2 4 2.4 3.6 1.4 6.5 6.5 ...
```



Normalize the data

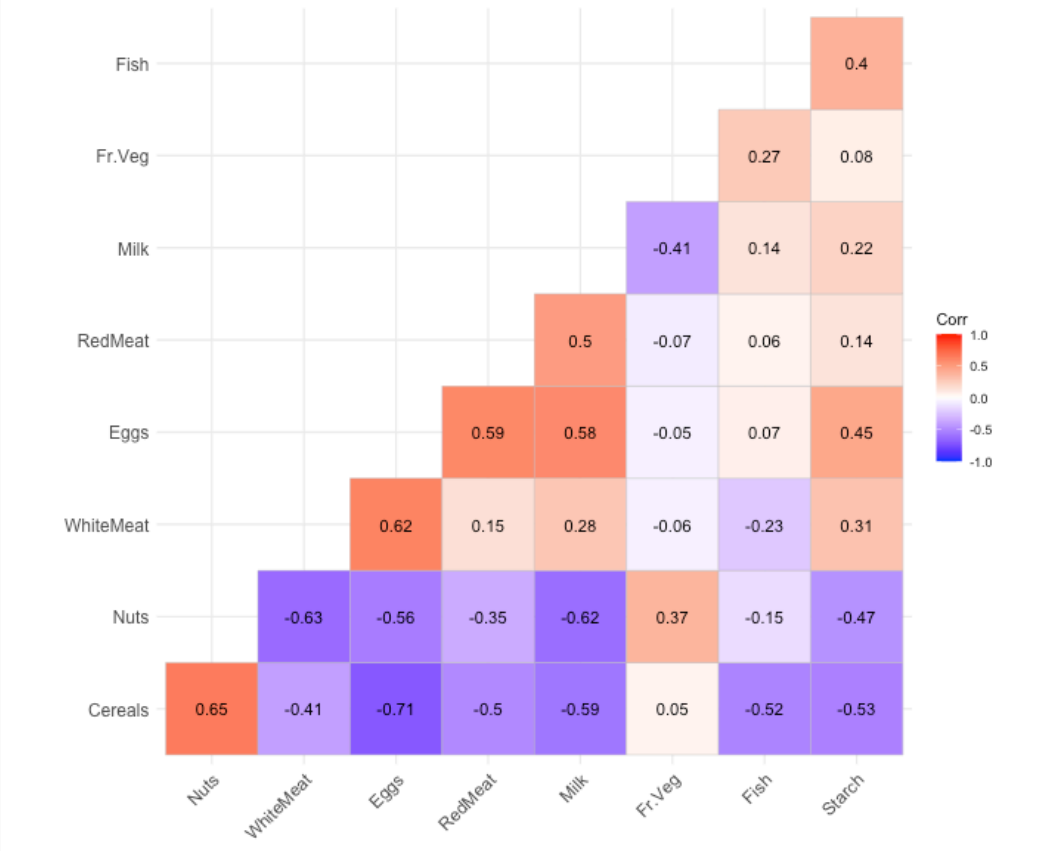
Because we are we are working with variances across all the variables, it is very important that they all have the same variance. Let's convert all the numeric columns to z-scores (i.e. mean = 0, standard deviation = 1).

```
data_normalized <- protein_data |>
    dplyr::select(!Country) |>
    dplyr::mutate(dplyr::across(where(is.numeric), scale))
```



Visualizing the correlation matrix

```
corr_matrix <- cor(data_normalized)
ggcorrplot::ggcorrplot(corr_matrix, hc.order = TRUE,
                       type = "lower", lab = TRUE)
```



The heatmap reveals strong positive correlations between animal protein sources (red meat, white meat, eggs, and milk), which explains why the first principal component captures nearly 77% of total variance. This correlation structure is exactly what PCA is designed to exploit.

Applying PCA

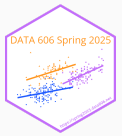
```
data.pca <- princomp(data_normalized)
summary(data.pca)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6     Comp.7     Comp.8     Comp
## Standard deviation     1.9611680 1.2528366 1.0405781 0.9573283 0.6672967 0.55868214 0.51062910 0.33412607 0.308459
## Proportion of Variance 0.4451597 0.1816666 0.1253244 0.1060738 0.0515376 0.03612566 0.03017848 0.01292132 0.011012
## Cumulative Proportion  0.4451597 0.6268263 0.7521507 0.8582245 0.9097621 0.94588776 0.97606624 0.98898757 1.000000
```

Loadings

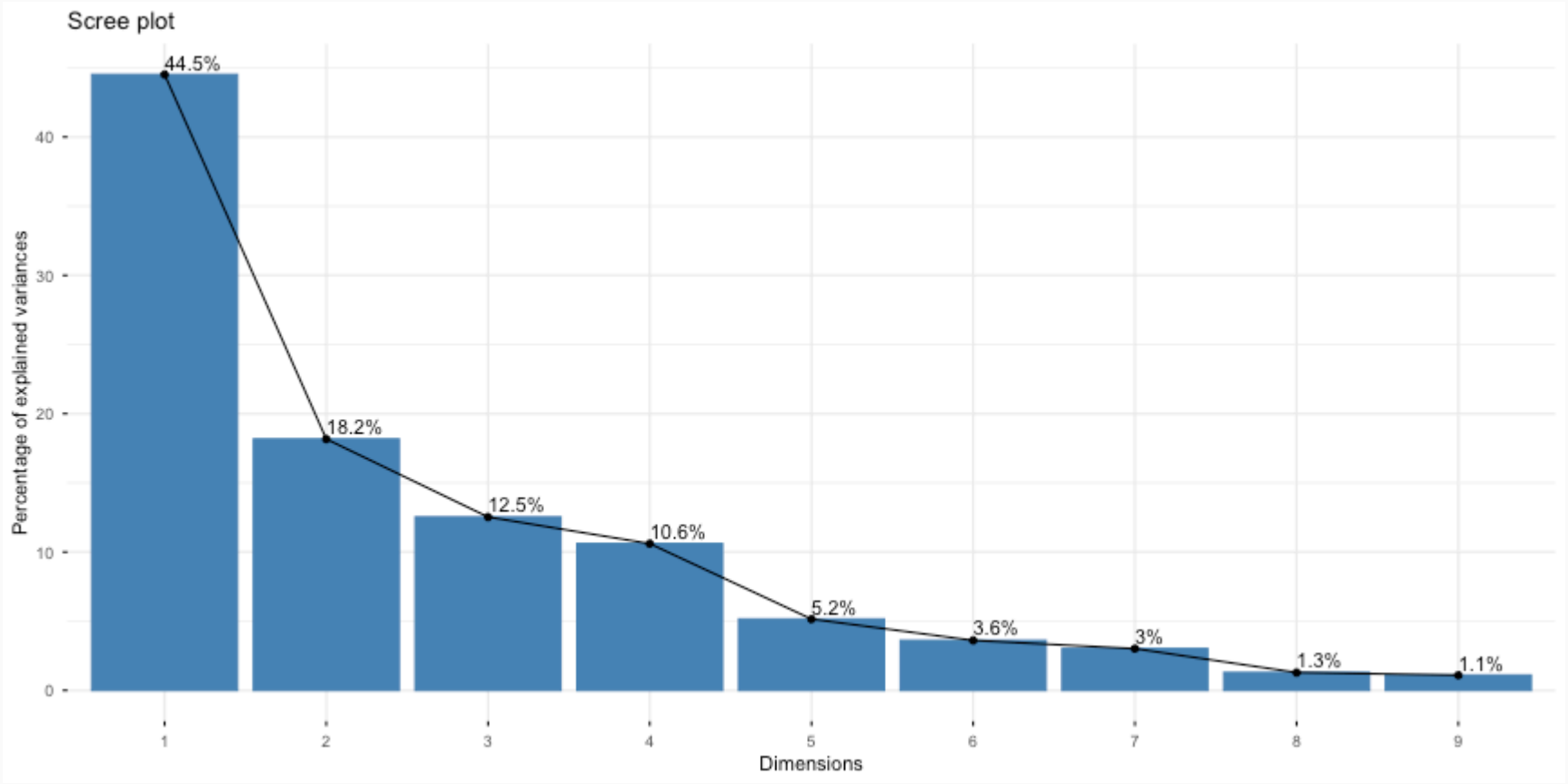
```
data.pca$loadings[, 1:2]
```

```
##                Comp.1      Comp.2
## RedMeat     0.3026094  0.05625165
## WhiteMeat   0.3105562  0.23685334
## Eggs        0.4266785  0.03533576
## Milk        0.3777273  0.18458877
## Fish        0.1356499 -0.64681970
## Cereals    -0.4377434  0.23348508
## Starch      0.2972477 -0.35282564
## Nuts       -0.4203344 -0.14331056
## Fr.Veg     -0.1104199 -0.53619004
```



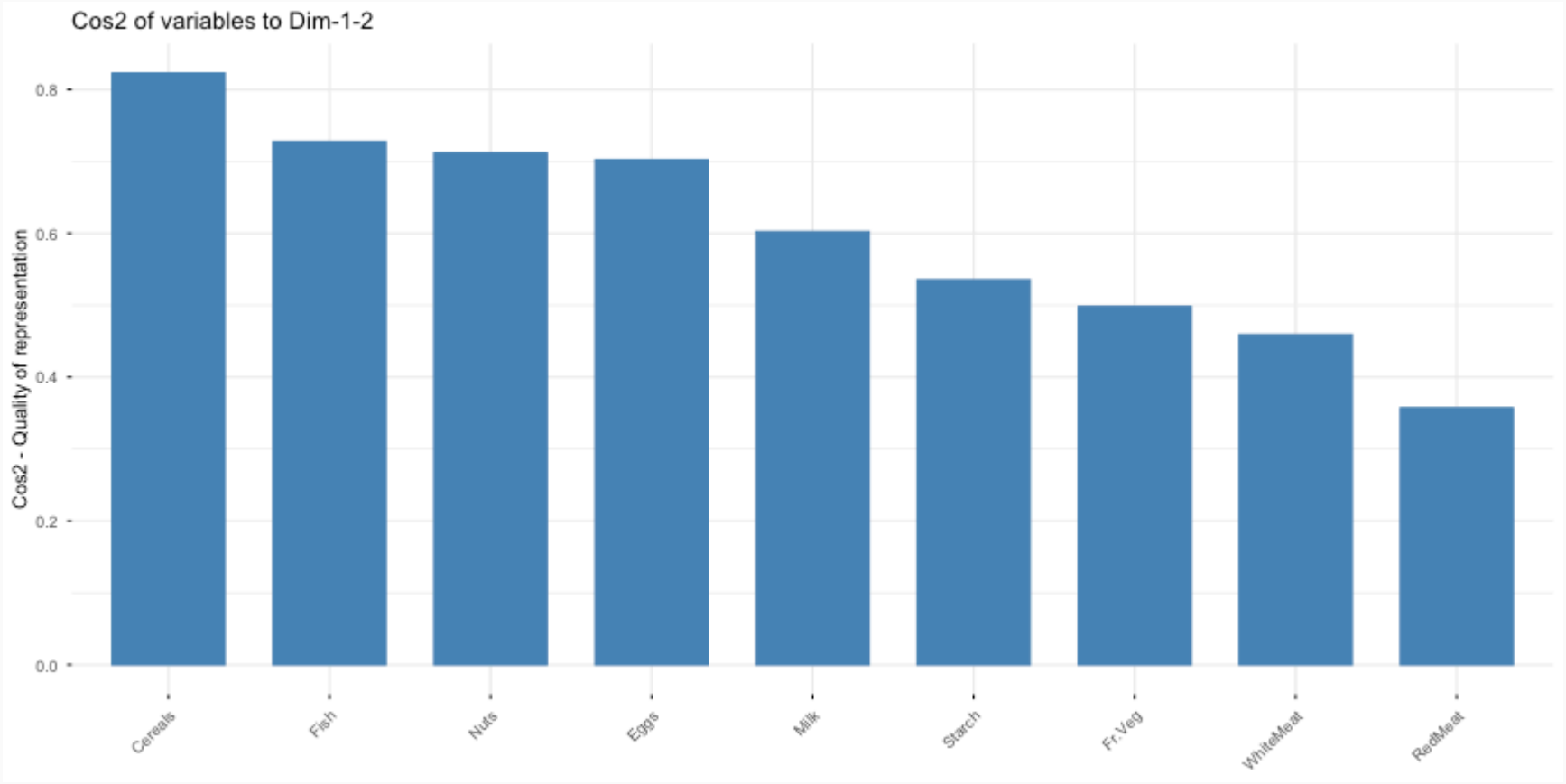
Scree Plot

```
factoextra::fviz_eig(data.pca, addlabels = TRUE)
```



Contribution of each variable

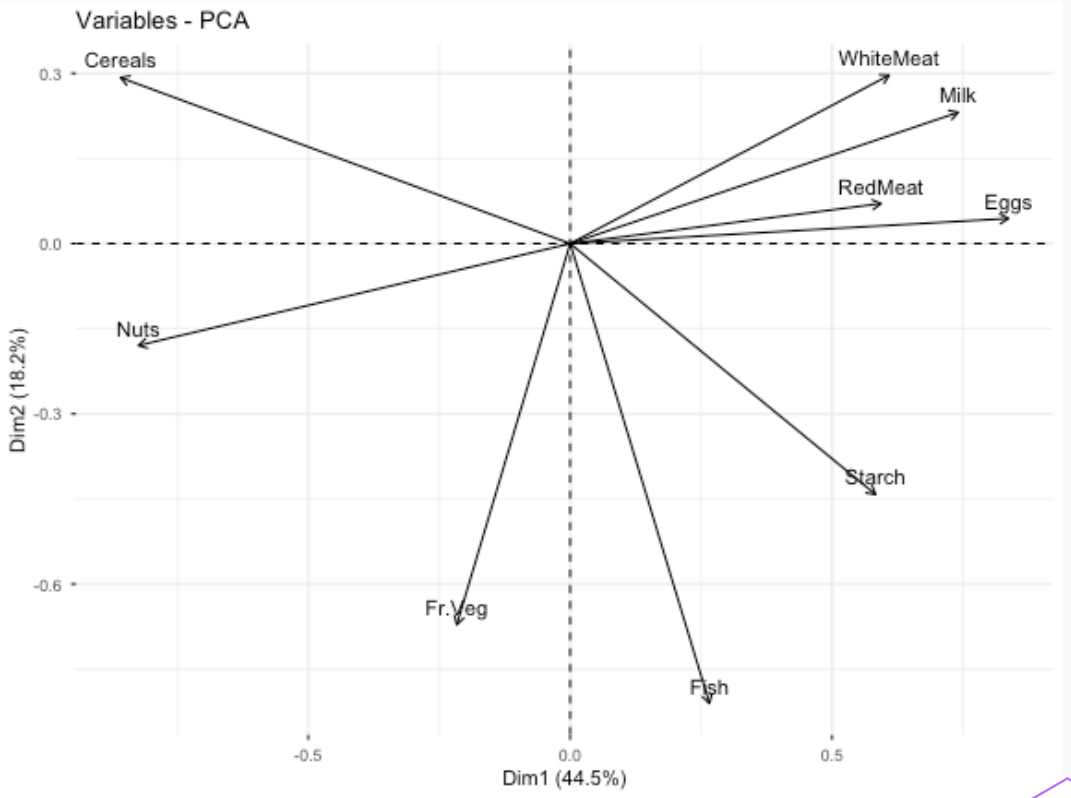
```
factoextra::fviz_cos2(data.pca, choice = "var", axes = 1:2)
```



Biplot of the attributes

With the biplot, it is possible to visualize the similarities and dissimilarities between the samples, and further shows the impact of each attribute on each of the principal components.

```
factoextra::fviz_pca_var(data.pca, col.var = "black")
```

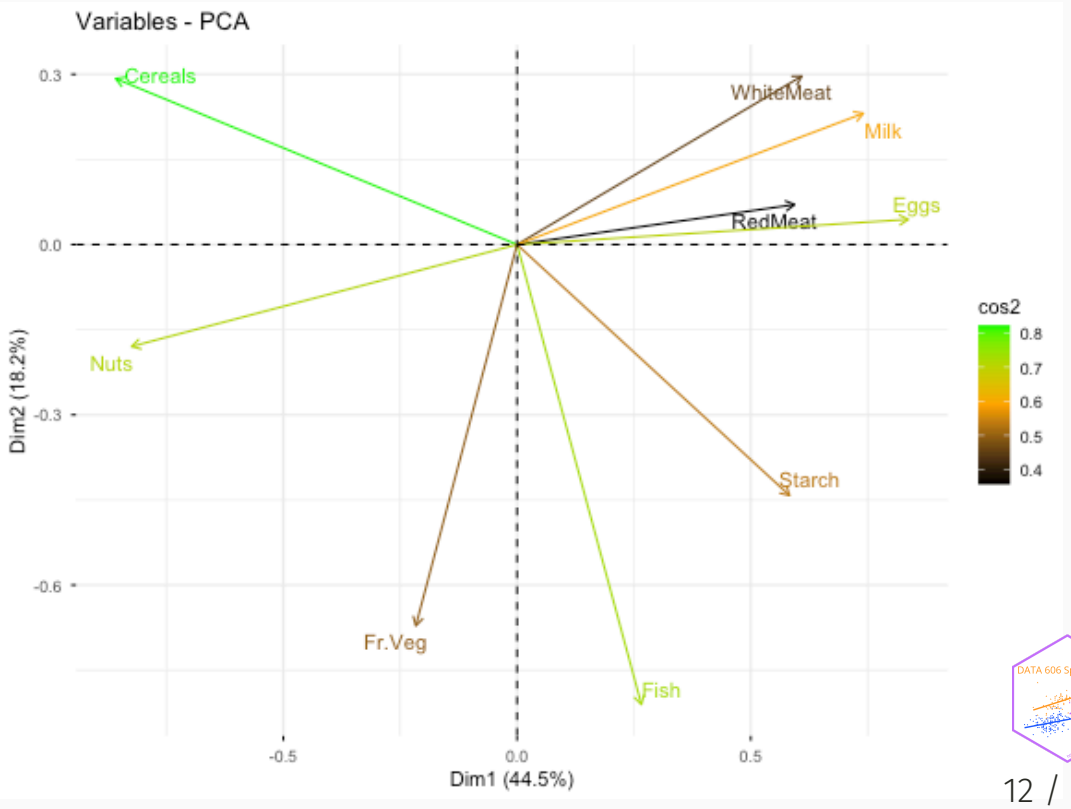


Biplot combined with cos2

The biplot and attributes importance can be combined to create a single biplot, where attributes with similar cos2 scores will have similar colors.

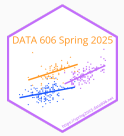
- High cos2 attributes are colored in green: Cereals, pulses, oilseeds, eggs, and milk.
- Mid cos2 attributes have an orange color: white meat, starchy food, fish, and red meat.
- Finally, low cos2 attributes have a black color: fruits and vegetables,

```
factoextra::fviz_pca_var(
    data.pca, col.var = "cos2",
    gradient.cols = c("black", "orange", "green"),
    repel = TRUE)
```



How to choose the number of components

- **Elbow rule**: Look at the scree plot and find where the curve bends sharply. Components to the right of the elbow contribute little additional variance.
- **Variance threshold**: Retain enough components to explain 80% to 90% of the total variance. In this dataset, the first two components already explain about 89%.



Limitations of PCA

1. Assumption of Linearity

2. Loss of Interpretability (try Exploratory Factor Analysis if interpretability is important)

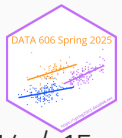
   - Combined features: The new principal components are linear combinations of all original features.
   - Obscured meaning: It becomes difficult or impossible to map a specific principal component back to an intuitive, real-world meaning from the original variables.

3. Sensitivity to Scale and Standardization

   - Scale-dependent: PCA relies heavily on variance. If features are on vastly different scales (e.g., measuring one variable in millimeters and another in kilometers), the feature with the larger numerical variance will dominate the principal components.
   - Requirement: Data must be mean-centered and standardized (scaled to unit variance) before applying PCA.

4. Variance Does Not Equal Importance

   - Low-variance traps: PCA assumes that directions with the highest variance contain the most important signal.
   - Ignored structure: Sometimes, critical patterns, anomaly classes, or target-relevant information reside in low-variance dimensions. Additionally, PCA is an unsupervised technique—it does not look at target/label variables to see if high-variance components actually help classification or prediction.



One Minute Paper

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?

Scan the QR code or click here to complete the One Minute Paper