Principal Components Analysis (PCA)

From last time:

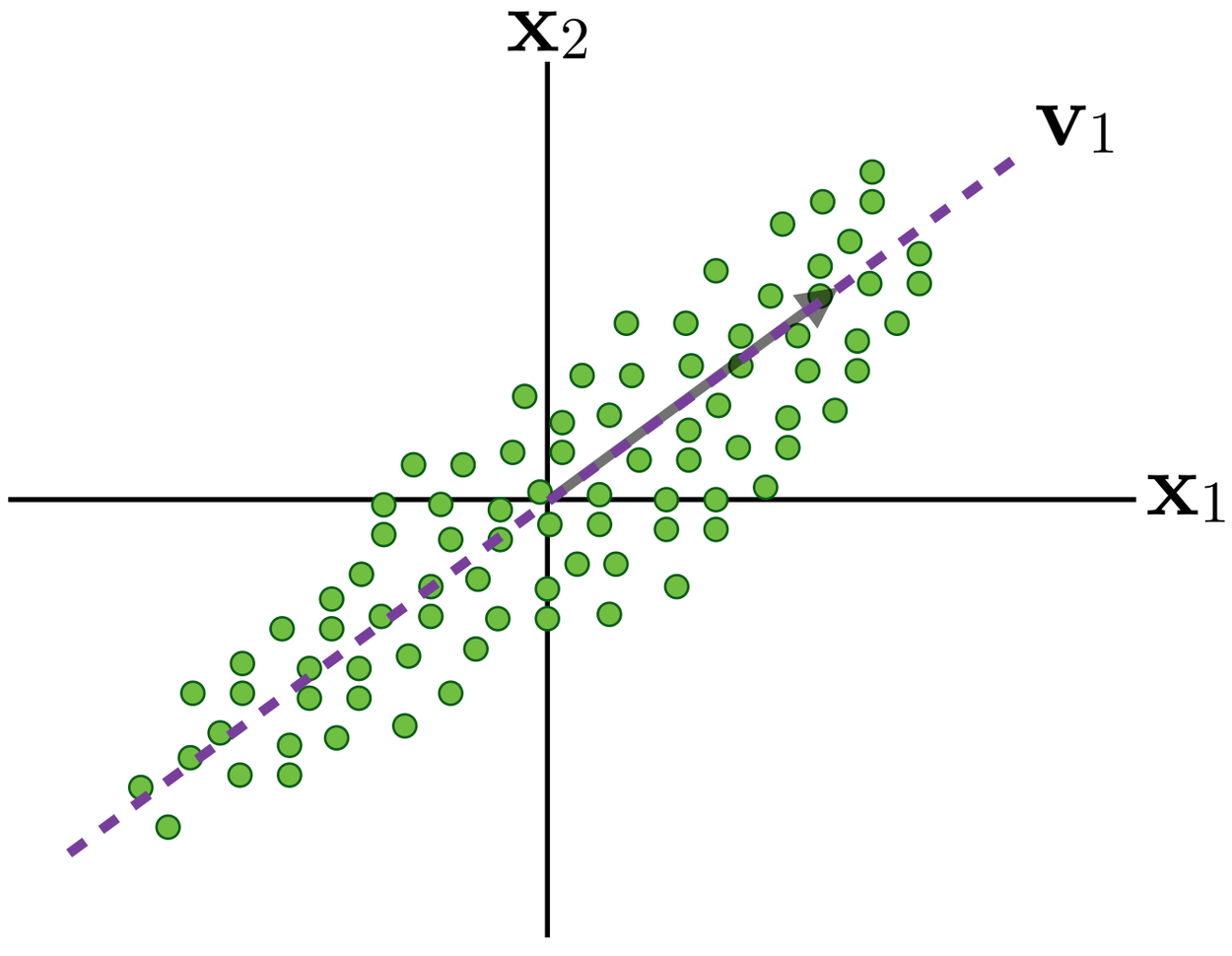
- Covariance/Correlation matrices are symmetric

- Symmetry ⇒ eigenvectors are orthogonal

- Eigenvectors are ordered by magnitude of their eigenvalues.

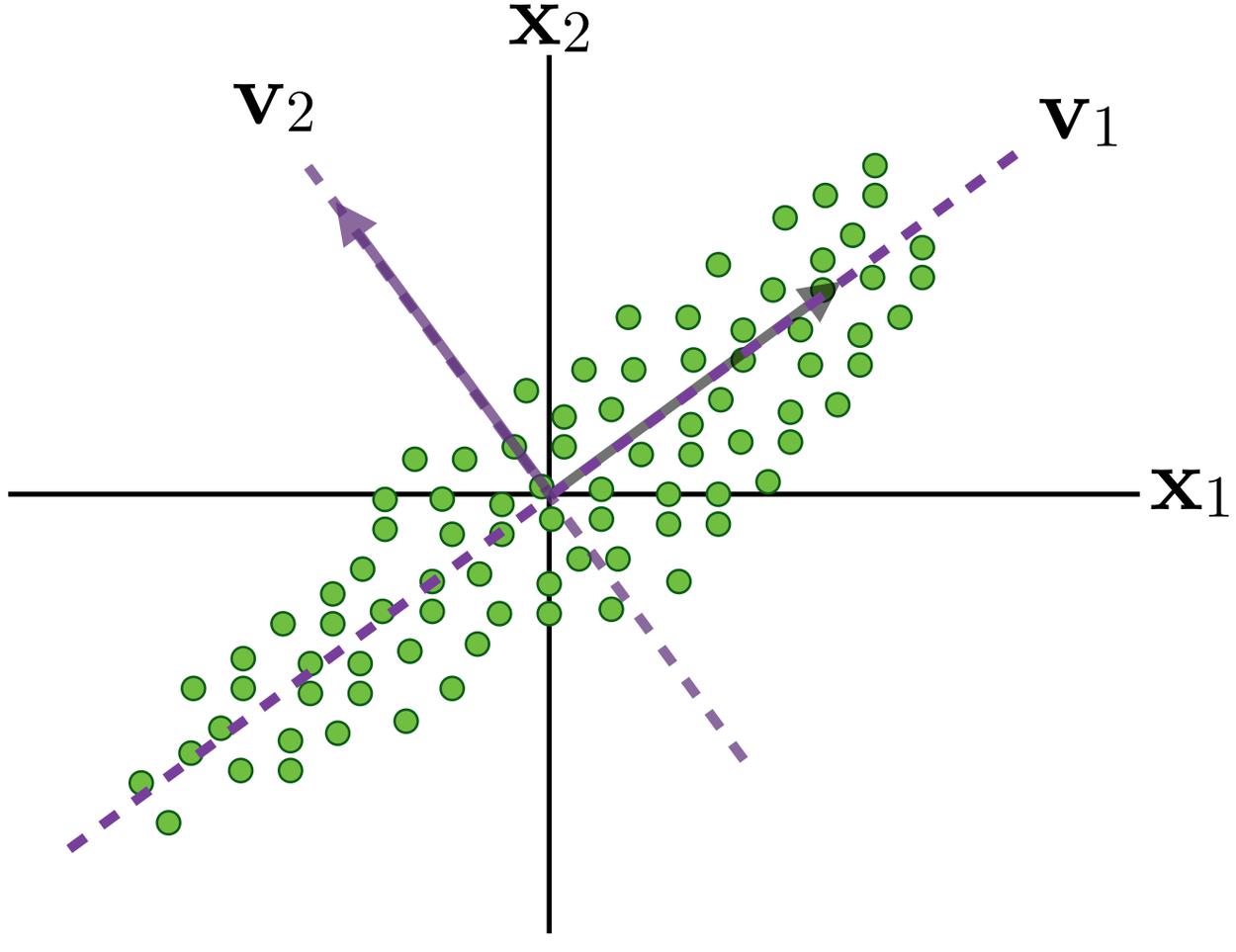
Principal Components

The **first principal component** is the first eigenvector of the covariance matrix and points in the direction of maximal variance



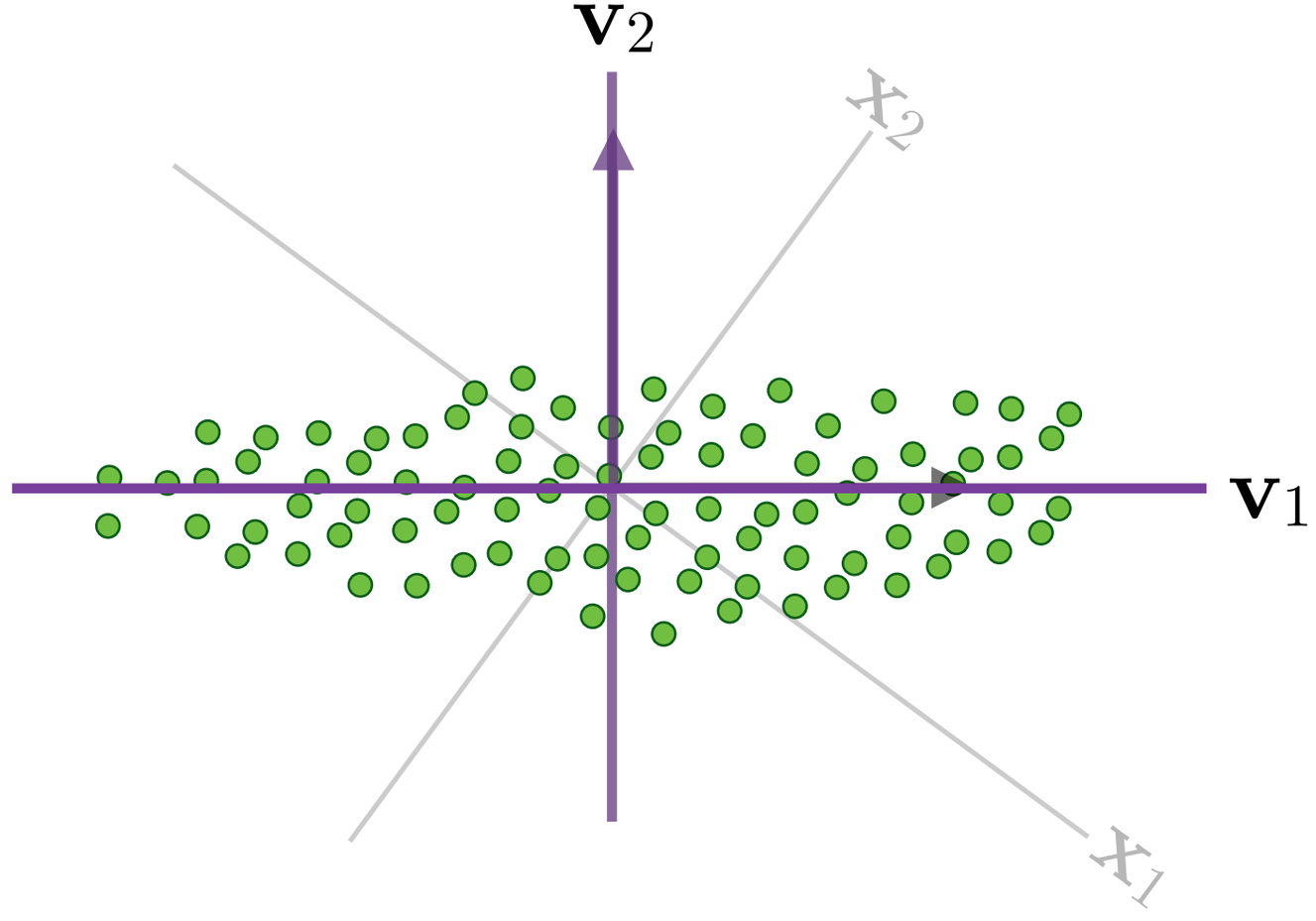
Principal Components

The **second principal component** is the second eigenvector of the covariance matrix and points in the direction, orthogonal to the first, that has maximal variance



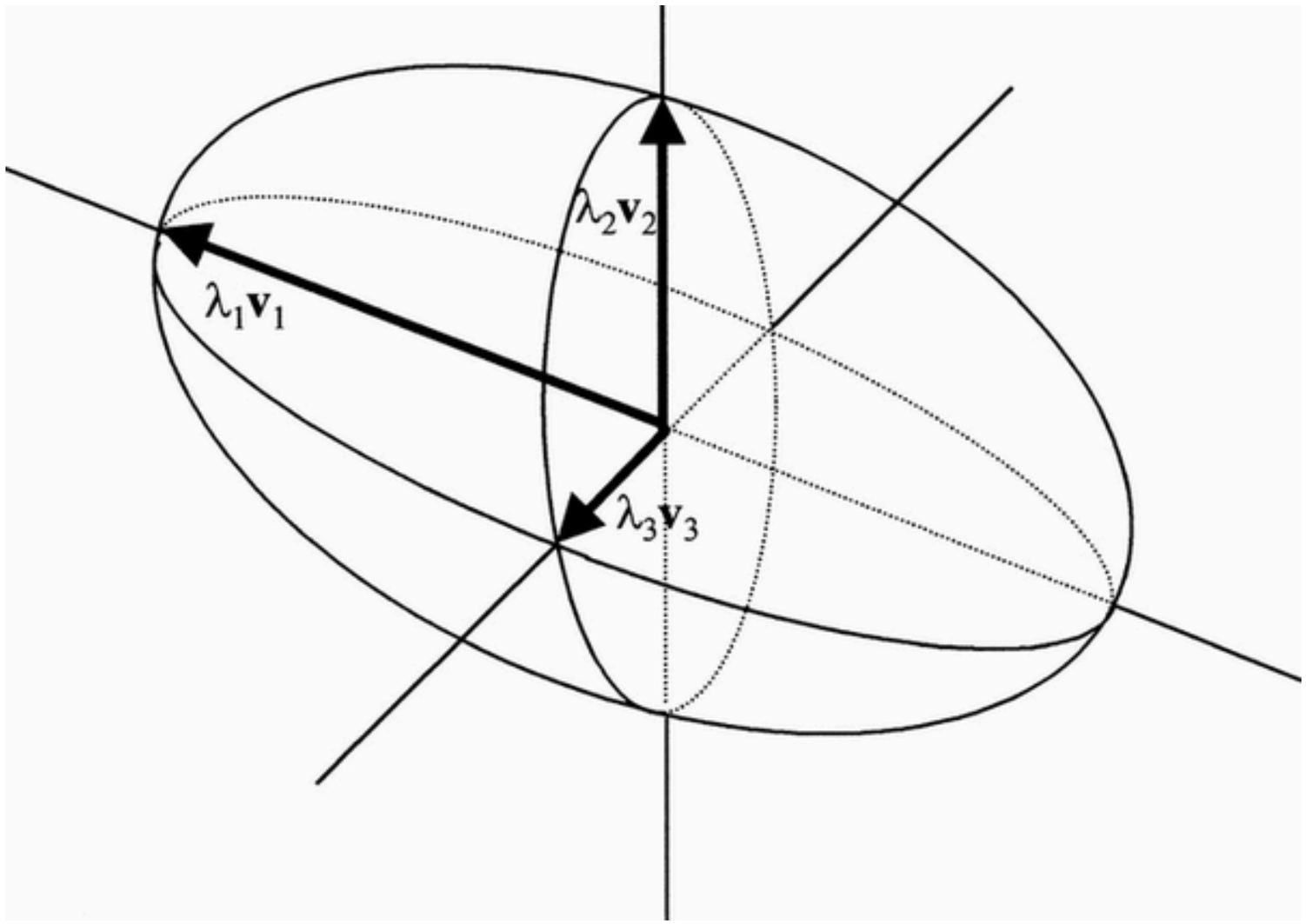
Principal Components

Principal components provide us with a new orthogonal basis where the coordinates of the data points are uncorrelated.



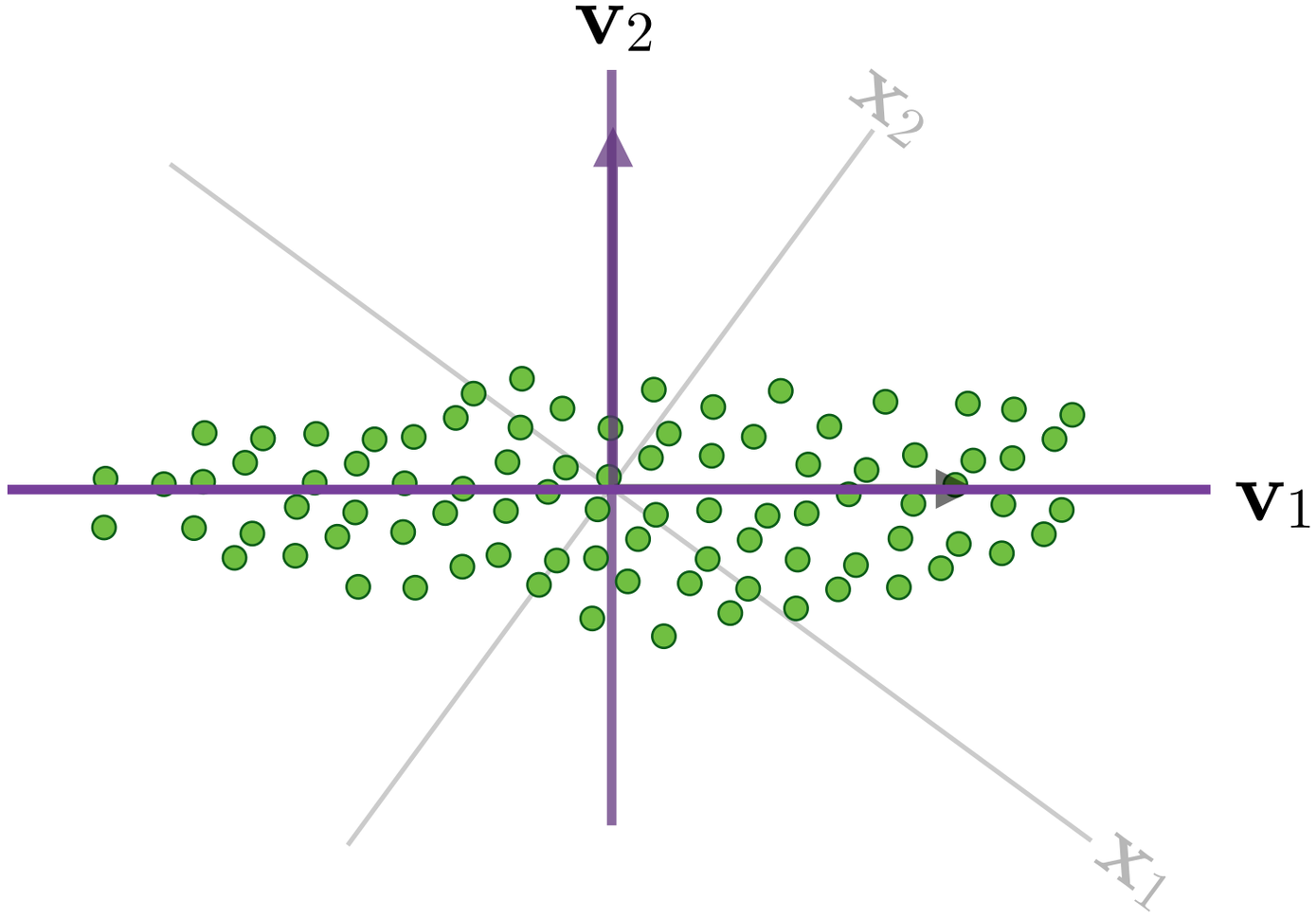
Eigenvalues give Variance

The corresponding eigenvalues, λ_1 and λ_2, tell us the amount of variance in each direction.



Eigenvalues give Variance

The corresponding eigenvalues, λ_1 and λ_2, tell us the amount of variance in each new direction. Same as saying the variance of the new variables $\mathbf{v}_1$ and $\mathbf{v}_2$.



Computed in the same old way, using the coordinates of the data in new basis.

Total Variance

We'll define the total amount of variance to be the sum of the variances of each variable:

$$\text{Total Variance} = var(\mathbf{x}_1) + var(\mathbf{x}_2)$$

Once we change to the principal component basis, we have new variables, $\mathbf{v}_1$ and $\mathbf{v}_2$, which would provide

$$\text{Total Variance} = \lambda_1 + \lambda_2$$

Total Variance

The cool fact is, the total amount of variance is the same no matter what orthogonal basis you consider!

$$var(\mathbf{x}_1) + var(\mathbf{x}_2) \;=\; \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2$$

In general, the total amount of variance will be the **trace** of the covariance matrix. (Sum of diagonal elements)

Proportion of Variance

Therefore, the proportion of total variance directed along (or explained by) the first principal component would be:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Likewise, for the second component:

$$\frac{\lambda_2}{\lambda_1 + \lambda_2}$$

Proportion of Variance

In general, when we have many components:

- **Eigenvalues give variance** of each component
- **Sum of eigenvalues** gives the **total** variance
- Total variance is equal to the sum of variances of each original variable

Proportion of Variance

The proportion of variance explained by the i^{th} component is

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

The cumulative proportion of variance explained by the first k components is

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

Practice

1

Suppose we have a dataset with 8 variables and we used standardized data. (*Note, this would amount to running a PCA on the correlation matrix)*

a) How many eigenvalues would we have in this case?

b) What would be the sum of the eigenvalues?

2

Suppose I have a dataset with 3 variables and the eigenvalues of the covariance matrix are:

$$\lambda_1=3 \quad \lambda_2=2 \quad \lambda_3=1$$

a) What proportion of variance explained by 1st PC?

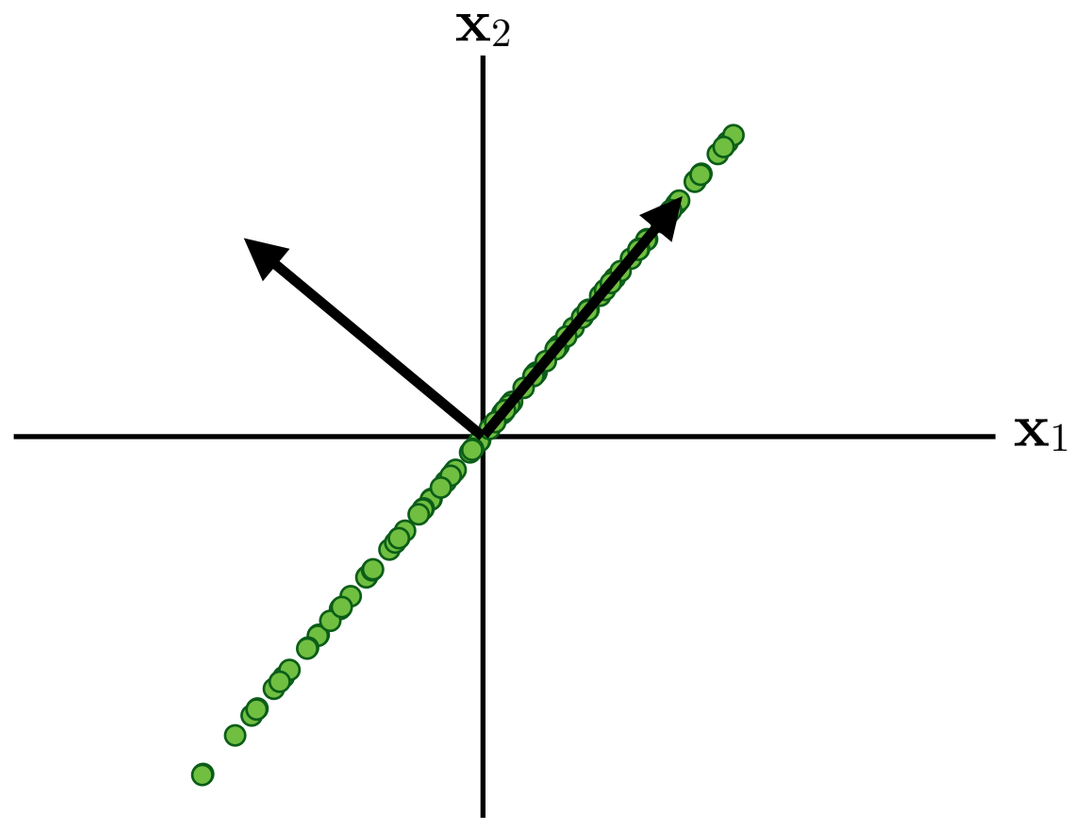
b) What is the variance of the second PC?

c) What proportion of variance is captured by using both the first and second principal components?

Another Look at Zero Eigenvalues

What would it mean if λ_2=0?

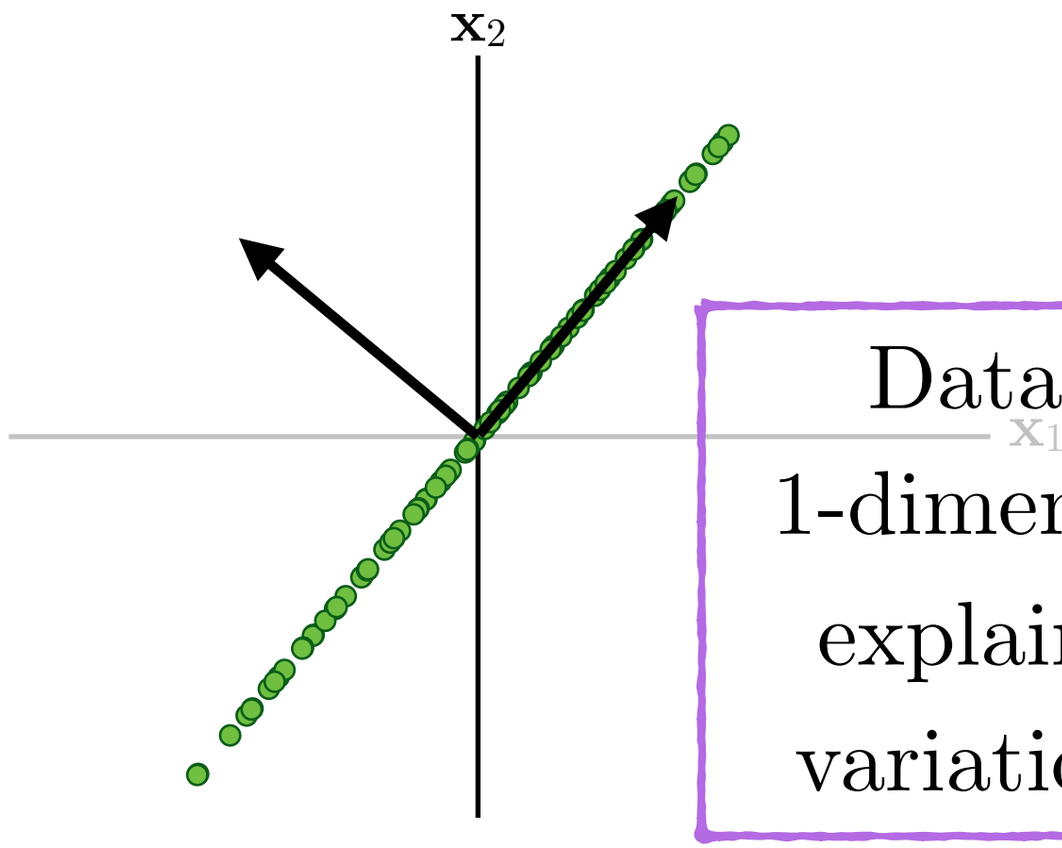
Variance along that direction is exactly zero.



Another Look at Zero Eigenvalues

All data points have exactly the same coordinate along $\mathbf{v}_2$

Original variables $\mathbf{x}_1$ and $\mathbf{x}_2$ are perfectly correlated.



Data is essentially 1-dimensional! $\mathbf{v}_1$ alone explains 100% of the variation in $\mathbf{x}_1$ and $\mathbf{x}_2$

Small Eigenvalues

When Eigenvalues are *close* to zero...

- Not much variance in this direction
- Won't lose much by ignoring or dropping this component

Dropping components ⇒ Orthogonal Projection

...onto a subspace

...the span of the principal components

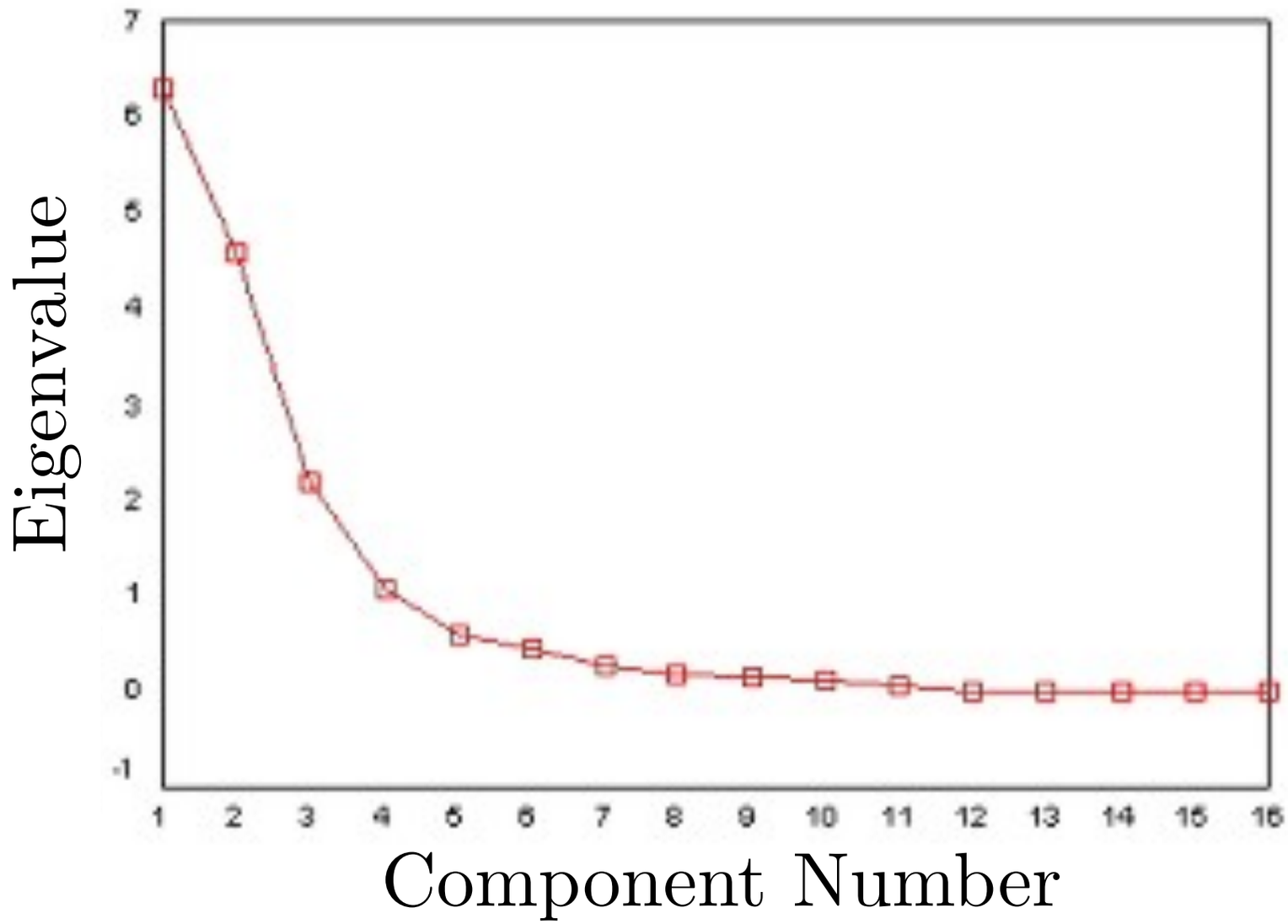
#DimensionReduction

Multicollinearity (Geometrically)

in R!

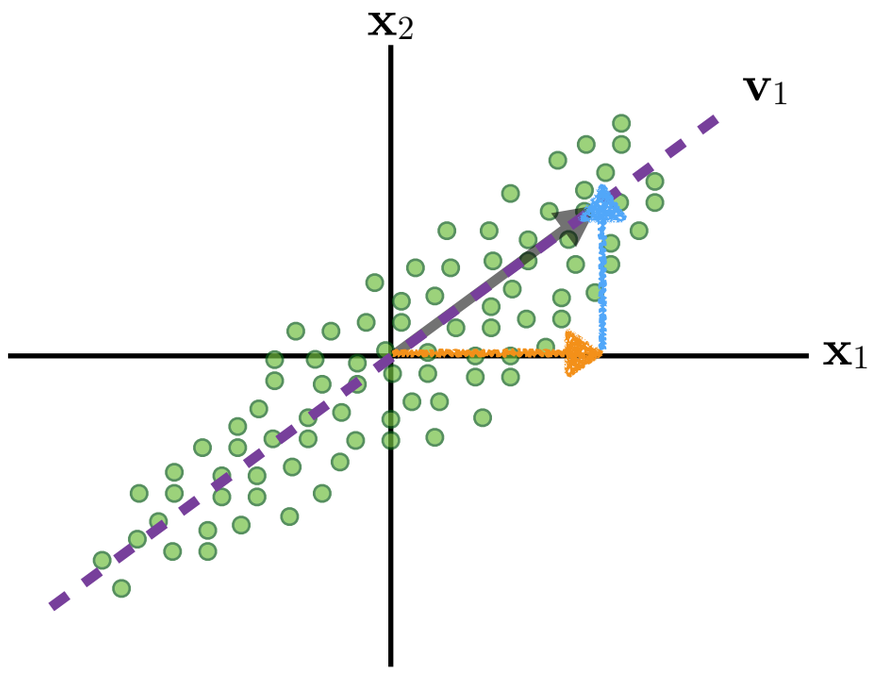
Scree plot

Plot of the eigenvalues. Sometimes used to *guess* the number of latent components by finding an 'elbow' in the curve.



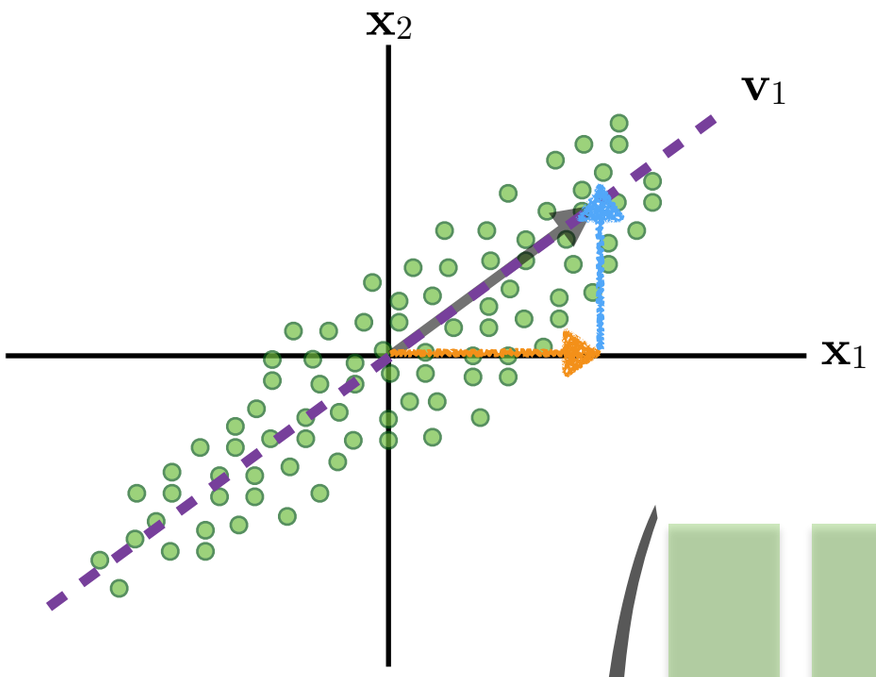
Coordinates in the New Basis

To find the new coordinates (called the scores) of the data in the new basis of principal components, we use the fact that principal components are linear combinations of original variables with weights given by the loadings.



$$\mathbf{v}_1 = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} = 0.7\mathbf{x}_1 + 0.7\mathbf{x}_2$$

Coordinates in the New Basis



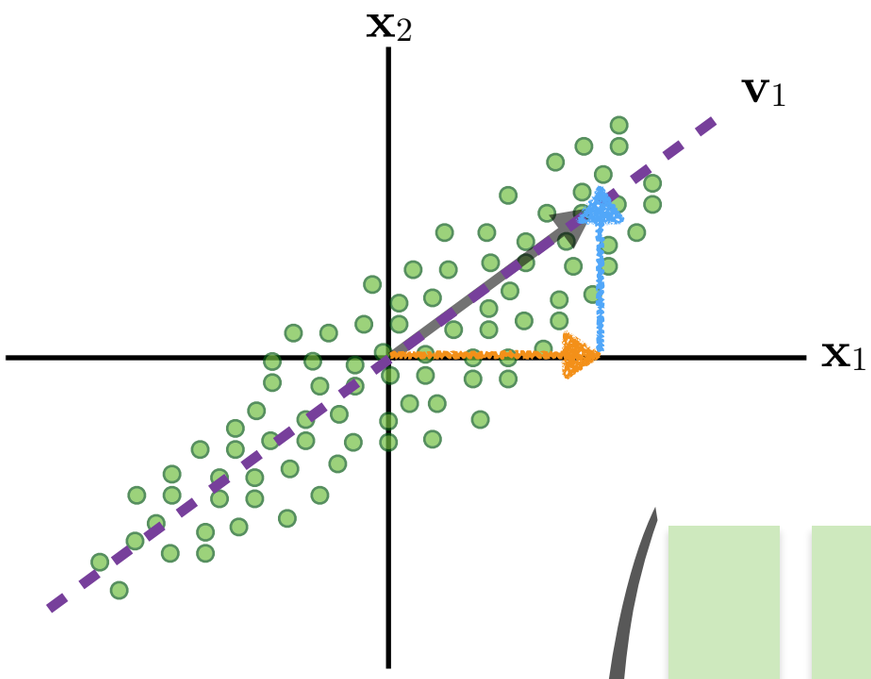
Loadings

$$\mathbf{v}_1 = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} = 0.7\mathbf{x}_1 + 0.7\mathbf{x}_2$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{pmatrix}$$

Your original data, after centering

Coordinates in the New Basis



Loadings

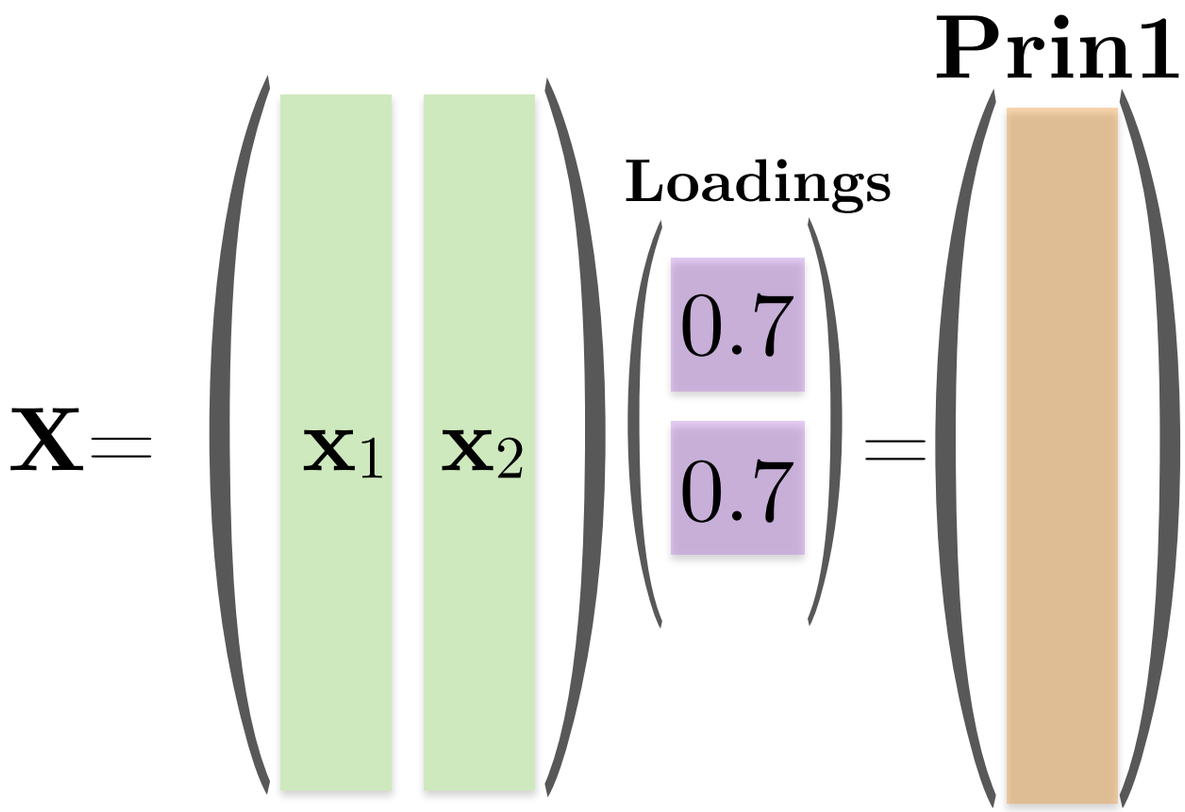
$$\mathbf{v}_1 = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} = 0.7\mathbf{x}_1 + 0.7\mathbf{x}_2$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{pmatrix} \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} = \begin{pmatrix} \quad \end{pmatrix}$$

new variable

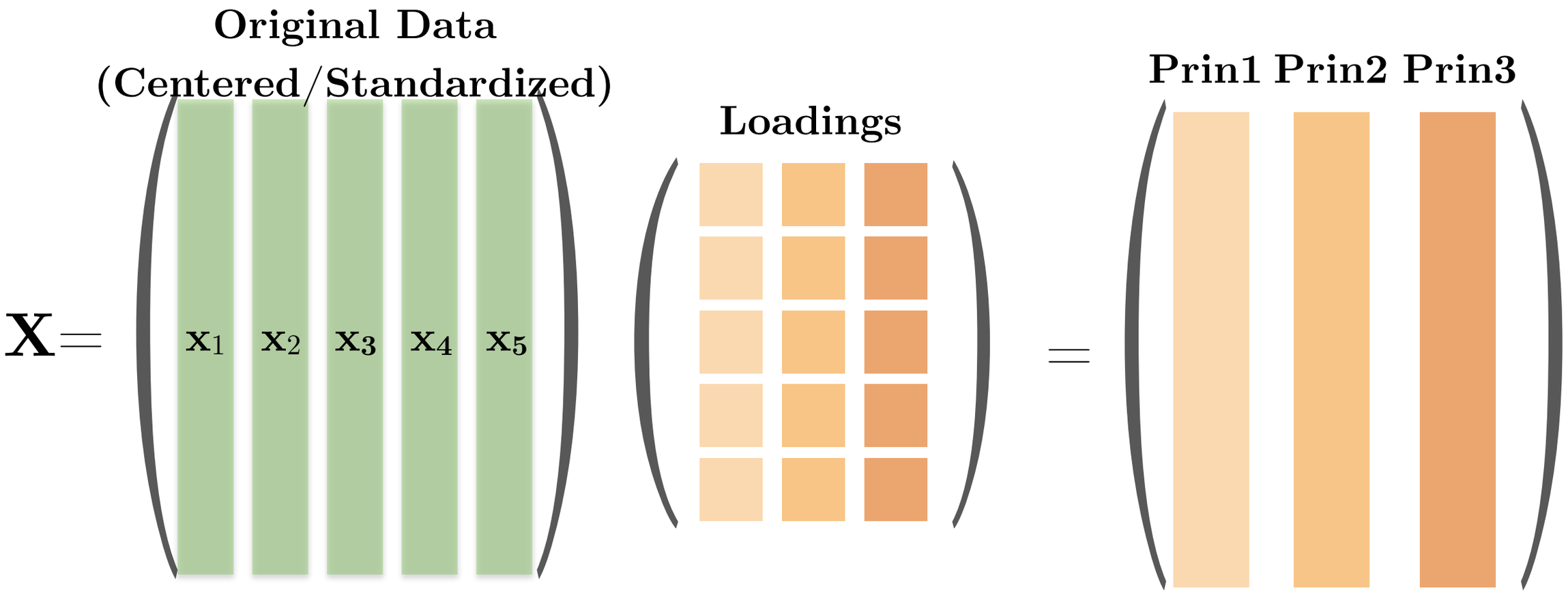
Prin1

Coordinates in the New Basis



Prin1: New variable in your spreadsheet. *The coordinates of the data projected onto the direction of the first principal component*

Coordinates in the New Basis



Can do this for **all components at once**, which amounts to matrix multiplication. **XV** where **V** is matrix of eigenvectors (i.e. loadings).

Summary of Output

3 Major Pieces of Output:

1. **Eigenvectors** (**Principal Components**, Variable Loadings, sometimes called *Rotation Matrix*)
    - output window in SAS
2. **Eigenvalues** (**Variances** of the new variables)
3. **Coordinates** of Data in the new basis (often called **scores** or scored data)
    - output dataset in SAS

Correlation matrix vs. Covariance matrix

PCA can be done using eigenvectors of either the covariance matrix or the correlation matrix.

- Default in SAS is correlation
- Default in R is covariance (for most packages - check!)

Recall: The correlation matrix is simply the covariance matrix of the standardized data.

Correlation matrix vs. Covariance matrix

Covariance PCA

- Data is centered (essentially unchanged) and directions of maximal variance are drawn
- Use when scales are not very different

Correlation PCA

- Data is centered *and* normalized/standardized before directions of maximal variance are drawn.
- Use when scales of variables are very different

Why not *always* use correlation matrix?

What could it hurt?

SAS Example: Test Scores

- Scores for 100 students on 5 tests
- An illustrative example, not a practical one (5 variables not likely to be reduced via PCA).

```
proc princomp data=Testscores
              out=testPCs;
var _all_;
run;
```

```
proc sgplot data=testPCs;
scatter x=Prin1 y=Prin2;
run;
```

```
proc princomp data=Testscores
              out=covTestPCs
              cov;
var _all_;
run;
```

```
proc sgplot data=covTestPCs;
scatter x=Prin1 y=Prin2;
run;
```

SAS's terrible problem

SAS will not perform a typical PCA for datasets with fewer observations than variables.

For our examples like this we will use R