GENERAL ASSEMBLY

Principal Component Analysis (PCA)

Patrick Smith

LEARNING OBJECTIVES

- Conduct a full PCA analysis manually and using scikit-learn
- Explain the mathematical process behind PCA

PRE-WORK

- Understand how to calculate principal components without using scikit-learn
- Have a basic understanding of linear algebra

OPENING

Intro to PCA

PCA is a very popular technique for performing "dimensionality reduction" on your data.

Intro to PCA

PCA is a very popular technique for performing "dimensionality reduction" on your data.

Dimensionality reduction is the process of combining or collapsing your existing features (columns in X) into new features that not only retain the original information but also ideally reduce noise.

Intro to PCA

Technically speaking, PCA finds the linear combinations of your current predictor variables that will create new "principal components" that explain, in order, the maximum possible amount of variance in your predictors.

Intro to PCA

The more intuitive way of thinking about PCA is that it transforms the coordinate system so that the axes become the most concise, informative descriptors of our data as a whole.

The new axes are the principal components.

Intro to PCA

A Brief Mathematical Introduction to Principal Component Analysis

Intro to PCA

Say we have a matrix $\boldsymbol{X}$ of predictor variables. PCA will give us the ability to transform our $\boldsymbol{X}$ matrix into a new matrix $\boldsymbol{Z}$.

1. First we will derive a weighting matrix $\boldsymbol{W}$ from the correlational/covariance structure of $\boldsymbol{X}$ that allows us to perform the transformation.

Intro to PCA

Say we have a matrix $\boldsymbol{X}$ of predictor variables. PCA will give us the ability to transform our $\boldsymbol{X}$ matrix into a new matrix $\boldsymbol{Z}$.

1. First we will derive a weighting matrix $\boldsymbol{W}$ from the correlational/covariance structure of $\boldsymbol{X}$ that allows us to perform the transformation.
2. Each successive dimension (column) in Z will be rank-ordered according to variance in it's values!

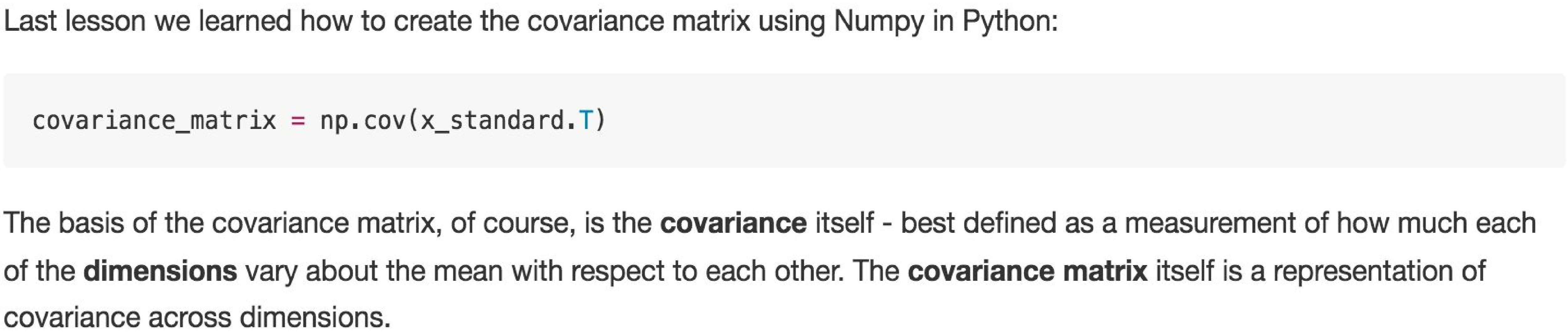
Intro to PCA

There are 3 assumptions that PCA makes:

1. **Linearity:** Our data does not hold nonlinear relationships.
2. **Large variances define importance:** our dimensions are constructed to maximize remaining variance.
3. **Principal components are orthogonal:** each component (columns of Z) is completely un-correlated with the others.

Intro to PCA

The Covariance Matrix + Correlation Matrix



Last lesson we learned how to create the covariance matrix using Numpy in Python:

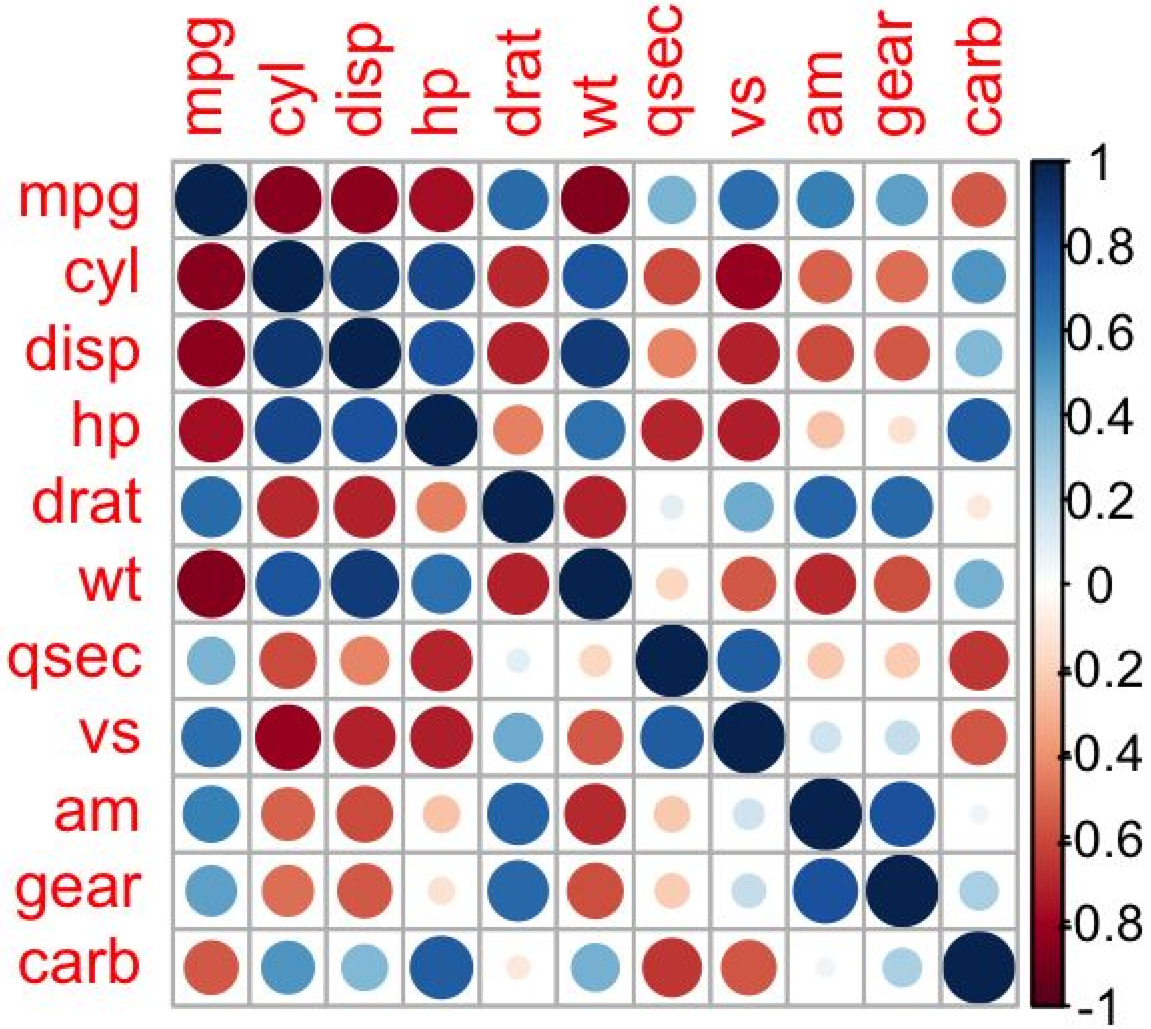
```
covariance_matrix = np.cov(x_standard.T)
```

The basis of the covariance matrix, of course, is the **covariance** itself - best defined as a measurement of how much each of the **dimensions** vary about the mean with respect to each other. The **covariance matrix** itself is a representation of covariance across dimensions.

Intro to PCA

The Covariance Matrix + Correlation Matrix

Likewise, the correlation matrix is used to show relationships between variables and can be used in place of the covariance matrix



Intro to PCA

The Covariance Matrix + Correlation Matrix

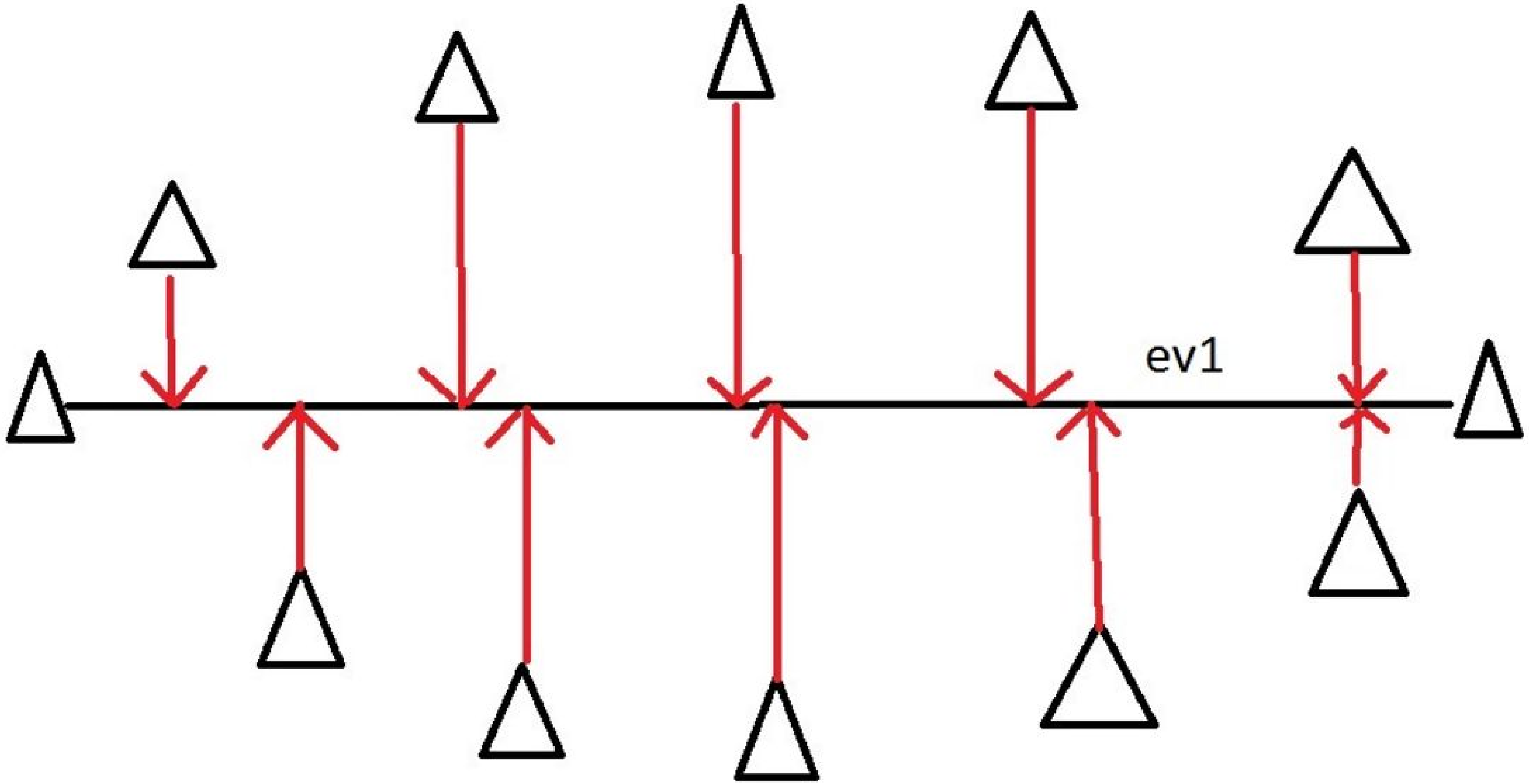
We use a covariance matrix when the scales or our dimensions are roughly similar

On the contrary, we use the correlation matrix when our data has different scales

The correlation matrix will standardize the data!

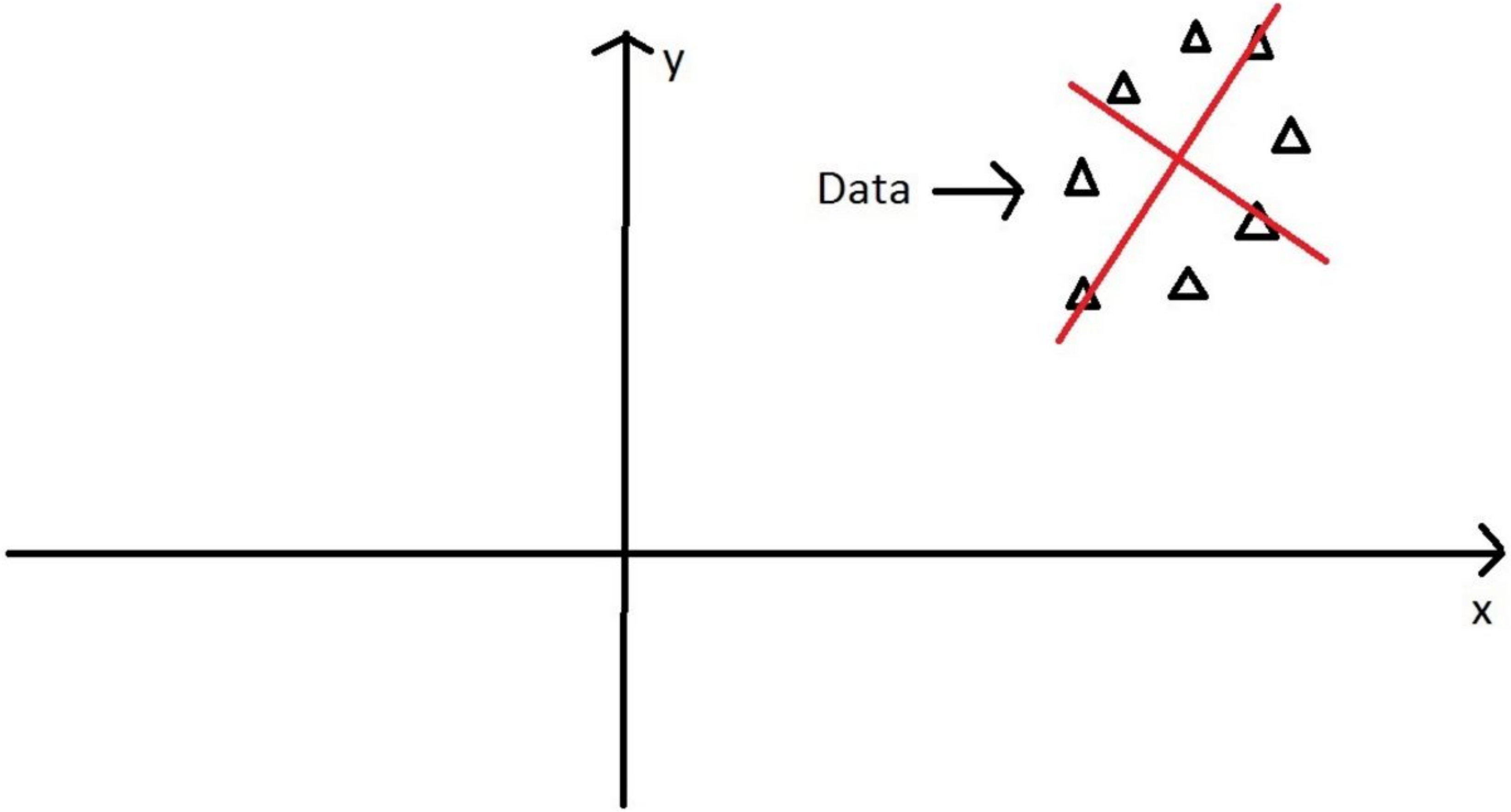
Intro to PCA

Eigenvalues and Eigenvectors



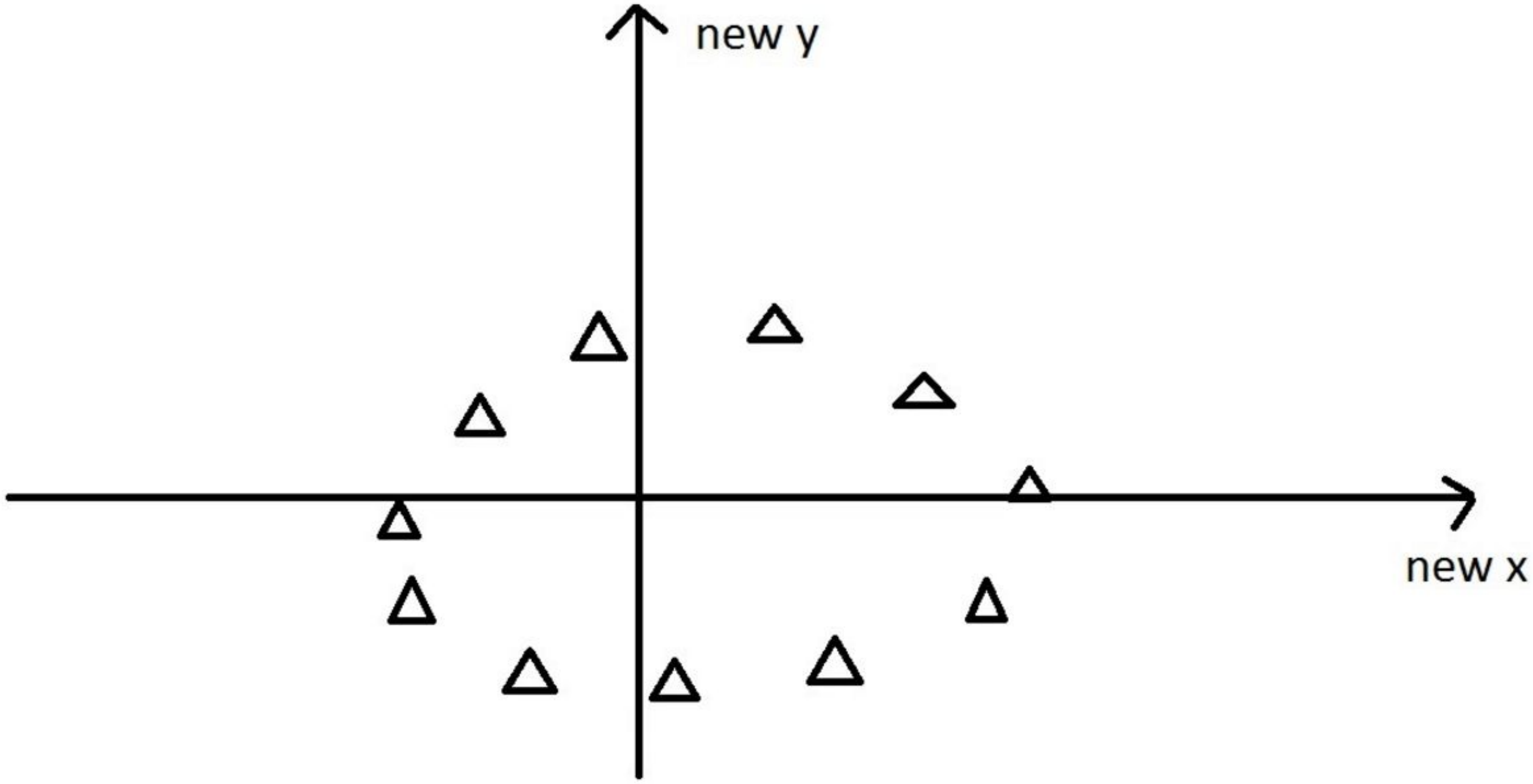
Intro to PCA

Eigenvalues and Eigenvectors



Intro to PCA

Eigenvalues and Eigenvectors



Intro to PCA

Eigenvalues and Eigenvectors

EIGENVECTORS

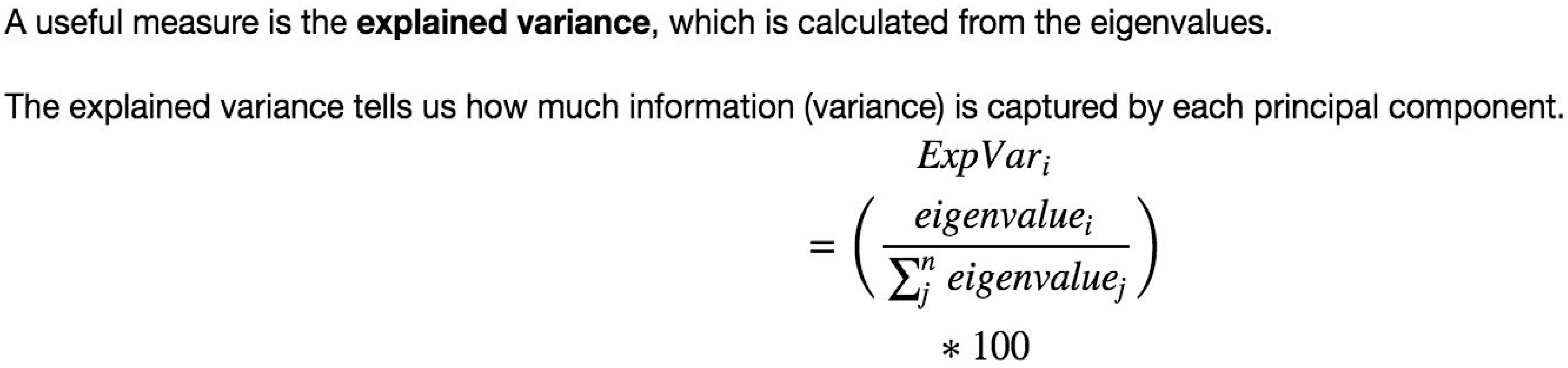
An eigenvector specifies a direction through the original coordinate space. The eigenvector with the highest correspoding eigenvalue is the first principal component.

EIGENVALUES

Eigenvalues indicate the amount of variance in the direction of it's corresponding eigenvector.

Intro to PCA

Explained Variance



A useful measure is the **explained variance**, which is calculated from the eigenvalues.

The explained variance tells us how much information (variance) is captured by each principal component.

$$ExpVar_i = \left(\frac{eigenvalue_i}{\sum_j^n eigenvalue_j} \right) * 100$$

Intro to PCA

What are Principal Components?

What is a principal component?

What are Principal Components?

What is a principal component?

- Principal components are the vectors that define the new coordinate system for your data.

Intro to PCA

What are Principal Components?

What is a principal component?

- **Principal components are the vectors that define the new coordinate system for your data.**
- Transforming your original data columns onto the principal component axes constructs new variables that are optimized to explain as much variance as possible and to be independent (uncorrelated).

What are Principal Components?

What is a principal component?

Creating these variables is a well-defined mathematical process, but in essence each component is created as a weighted sum of your original columns, such that all components are orthogonal (perpendicular) to each other.

Intro to PCA

Why would we want to do PCA?

- We can reduce the number of dimensions (remove bottom number of components) and lose the least possible amount of variance information in our data.

Intro to PCA

Why would we want to do PCA?

- We can reduce the number of dimensions (remove bottom number of components) and lose the least possible amount of variance information in our data.
- Since we are assuming our variables are interrelated (at least in the sense that they together explain a dependent variable), the information of interest should exist along directions with largest variance.

Why would we want to do PCA?

- We can reduce the number of dimensions (remove bottom number of components) and lose the least possible amount of variance information in our data.
- Since we are assuming our variables are interrelated (at least in the sense that they together explain a dependent variable), the information of interest should exist along directions with largest variance.
- Correlated predictor variables (also referred to as "redundancy" of information) are combined into independent variables. Our predictors from PCA are guaranteed to be independent

Intro to PCA

Manual PCA Codealong

Intro to PCA

1. Standardize data: centering is required, but full normalization is nice for visuals later.
2. Calculate eigenvectors and eigenvalues from correlation or covariance matrix.
3. Sort eigenvalues and choose eigenvectors that correspond to the largest eigenvalues. The number you choose is up to you, but we will take 2 for the sake of visualization here.
4. Construct the projection weighting matrix W from the eigenvectors.
5. Transform the original dataset X with W to obtain the new 2-dimensional transformed matrix Z.

Independent Practice: Automated PCA

Intro to PCA

Conclusion

Q & A