Implementing Factor Analysis and PCA in R

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Assemble a data set of returns from correlated stocks

Use R to calculate principal components of the financial data

Eliminate low-value principal components using a Scree plot

Relate the principal components to underlying latent factors

PCA in R

Explain Google's returns
Yahoo finance

Using returns of correlated stocks

Covariance and Correlation
Correlation matrix signals trouble

Multicollinearity problems

Eigen Decomposition
Built-in R function

On covariance matrix

Scree Plot
Number of dimensions

Discard low-value dimensions

Principal Components
From eigen vectors

Uncorrelated components

Interpret and Regress
Beta, bonds, sectors

Now regress Google

Demo

Implement Eigen analysis and PCA in R

PCA should always be applied on the covariance matrix of standardised vectors

Data Frame: Data in Rows and Columns

Each column represents 1 variable (a list or vector)

DATE	OPEN	...	ADJUSTED CLOSE
2016-12-01	772	...	779
2016-11-01	758	...	747
2006-01-01	302	...	309

Each row represents 1 observation

From File to Data Frame

DATE	OPEN	…	ADJUSTED CLOSE
2016-12-01	772	…	779
2016-11-01	758	…	747
2006-01-01	302	…	309

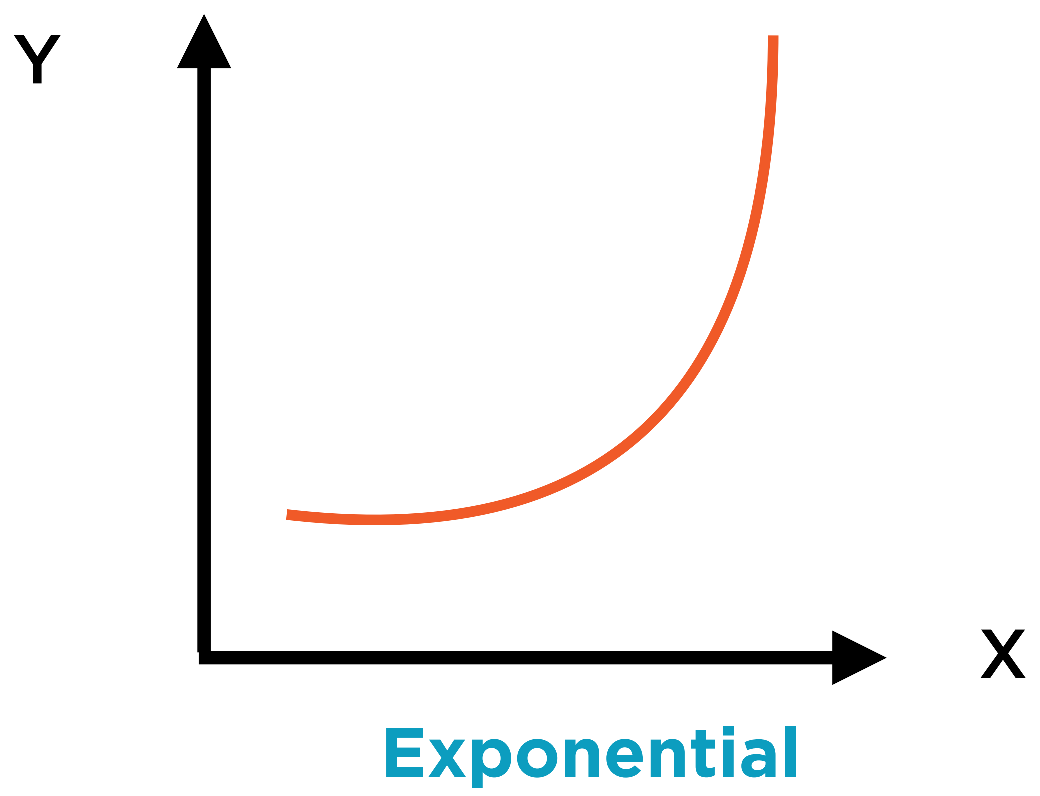
read.table

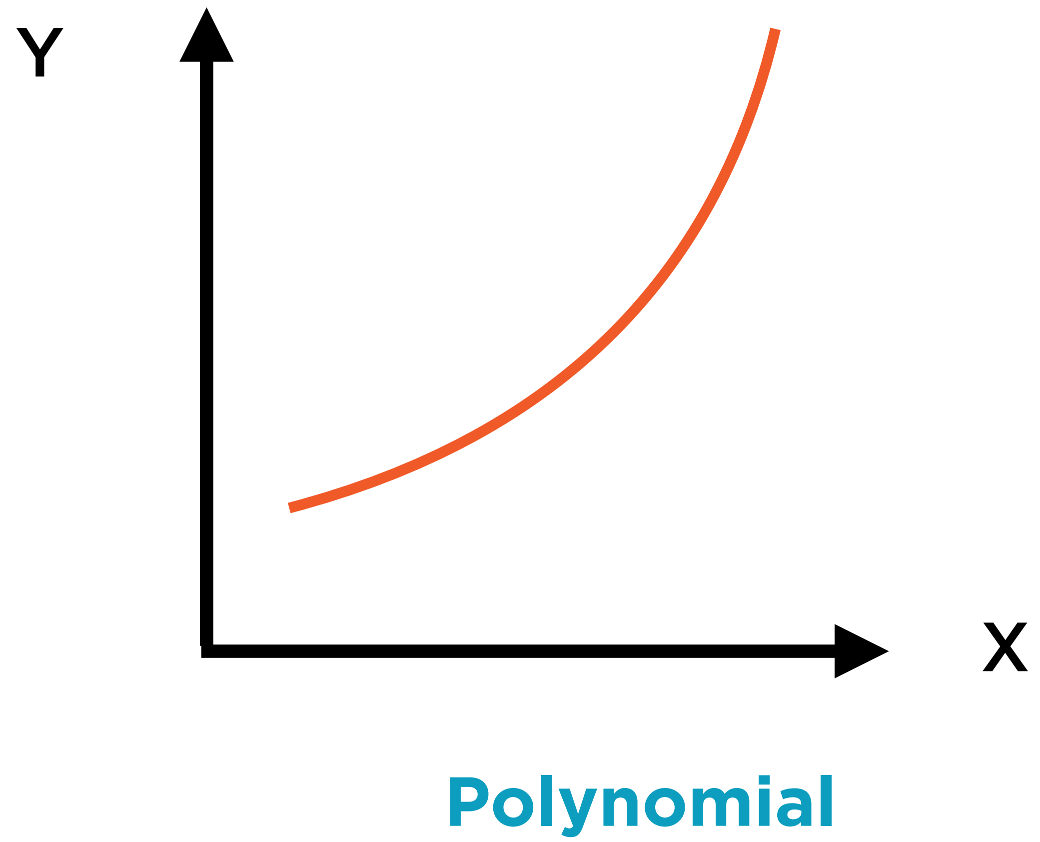
DATE	OPEN	…	ADJUSTED CLOSE
2016-12-01	772	…	779
2016-11-01	758	…	747
2006-01-01	302	…	309

File

Data Frame

Never Regress Non-Stationary Data





Smoothly trending data will lead to poor quality regression models

First Differences

$y'_{12} = \log y_2 - \log y_1$

$x'_{12} = \log x_2 - \log x_1$

Regress y' and x'

Log Differences

$y'_{12} = (y_2 - y_1)/y_1$

$x'_{12} = (x_2 - x_1)/x_1$

Regress y' and x'

Returns

Take first differences of smooth data converting either to log differences or returns

Negative Indices => Exclude Data

goog

DATE	GOOG. PRICE	NASDAQ. PRICE	
2016-12-01	779	5550	Row 1
2016-11-01	747	5324	
2006-01-01	309	1900	Row nrow(goog)
Column 1			

```
goog[-nrow(goog),-1]
```

Negative Indices => Exclude Data

DATE	GOOG. PRICE	NASDAQ. PRICE	
2016-12-01	779	5550	Row 1
2016-11-01	747	5324	
2006-01-01	309	1900	Row nrow(goog)

goog

Exclude

Column 1

goog[-nrow(goog),-1]

Negative Indices => Exclude Data

goog

Exclude

DATE	GOOG. PRICE	NASDAQ. PRICE	
2016-12-01	779	5550	Row 1
2016-11-01	747	5324	
2006-01-01	309	1900	Row nrow(goog)

Column 1

```
goog[-nrow(goog),-1]
```

Negative Indices => Exclude Data

goog

DATE	GOOG. PRICE	NASDAQ. PRICE	
2016-12-01	779	5550	Row 1
2016-11-01	747	5324	
2006-01-01	309	1900	Row nrow(goog)

Exclude

Column 1

goog[-nrow(goog),-1]

Element-wise Operations

779	5550

/

747	5324

=

779/747	5550/5324
...	...
...	...

```
goog[-nrow(goog),-1]/
goog[-1,-1]
```

Prices to Returns

779/747	5550/5324
...	...
...	...

-

1	1
1	1
1	1
1	1
1	1

=

779/747 - 1	5550/5324 -1
...	...
...	...

goog[-nrow(goog),-1]/
goog[-1,-1] - 1

This converts prices to returns

Standardising Data

$$\begin{bmatrix} X_{11} & & X_{1k} \\ X_{21} & & X_{2k} \\ X_{31} & \cdots & X_{3k} \\ \cdots & & \cdots \\ X_{n1} & & X_{nk} \end{bmatrix}$$

$avg(X_1)$ ... $avg(X_k)$

$stdev(X_1)$ ... $stdev(X_k)$

Standardising Data

$$\begin{bmatrix} \dfrac{x_{11} - avg(X_1)}{stdev(X_1)} & \cdots & \dfrac{x_{1k} - avg(X_k)}{stdev(X_k)} \\ \cdots & & \cdots \\ \dfrac{x_{n1} - avg(X_1)}{stdev(X_1)} & & \dfrac{x_{nk} - avg(X_k)}{stdev(X_k)} \end{bmatrix}$$

Each column of the standardised data has mean 0 and variance 1

Principal Components Analysis

$[\ X_1\ X_2\ X_3 \ldots X_k\ ]$ →

Eigenvalue
Decomposition

→

Principal Components:

$[\ F_1\ F_2\ F_3 \ldots F_k\ ]$ n rows

k columns

Eigenvectors:

$[\ v_1\ v_2\ v_3 \ldots v_k\ ]$ k rows

k columns

Eigenvalues:

$[\ e_1\ e_2\ e_3 \ldots e_k\ ]$ 1 row

k columns

Interpreting Eigenvalues

$$[\ F_1 \quad F_2 \quad F_3 \quad \cdots \quad F_k\ ]$$

$$var(F_1) > var(F_2) > var(F_3) \quad > var(F_k)$$

These vectors F_i are arranged in order of decreasing variance

The greater the variance of a principal component, the more important it is

Interpreting Eigenvalues

$$[\ F_1 \quad F_2 \quad F_3 \quad \ldots \quad F_k\ ]$$

$$\mathrm{var}(F_1) > \mathrm{var}(F_2) > \mathrm{var}(F_3) \quad > \mathrm{var}(F_k)$$

Eigenvalue 1

Eigenvalue 2

Eigenvalue 3

Eigenvalue k

The greater the eigenvalue of a principal component, the more important it is

Use the Scree plot to determine how many principal components to discard

Exploratory Factor Analysis: Experts trace back principal components to observable factors

PCA for Latent Factor Identification

$$[\ F_1 \quad F_2 \quad F_3 \quad \dots \quad F_k\ ]$$

$$[\ L_1 \quad L_2 \quad L_3 \quad \dots \quad L_k\ ]$$

3 Latent Factors in Stock Returns

Market Movements

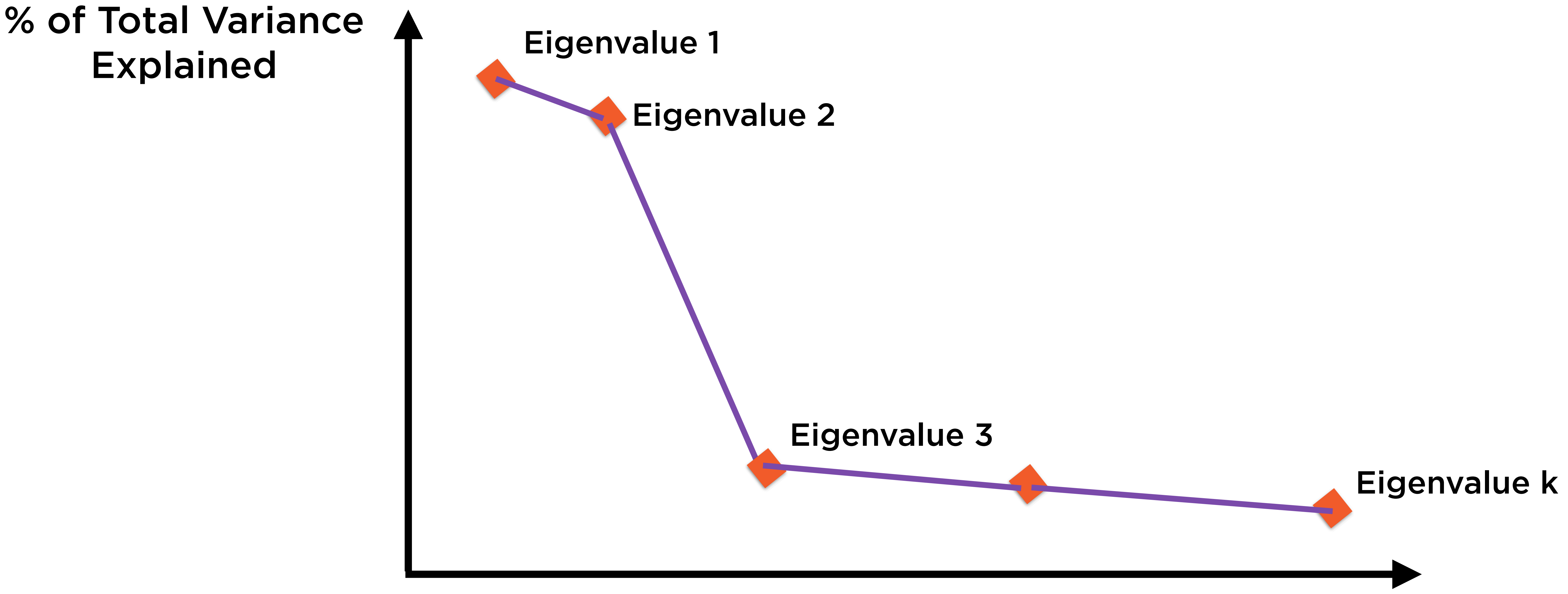
Interest Rates

Industry Sectors

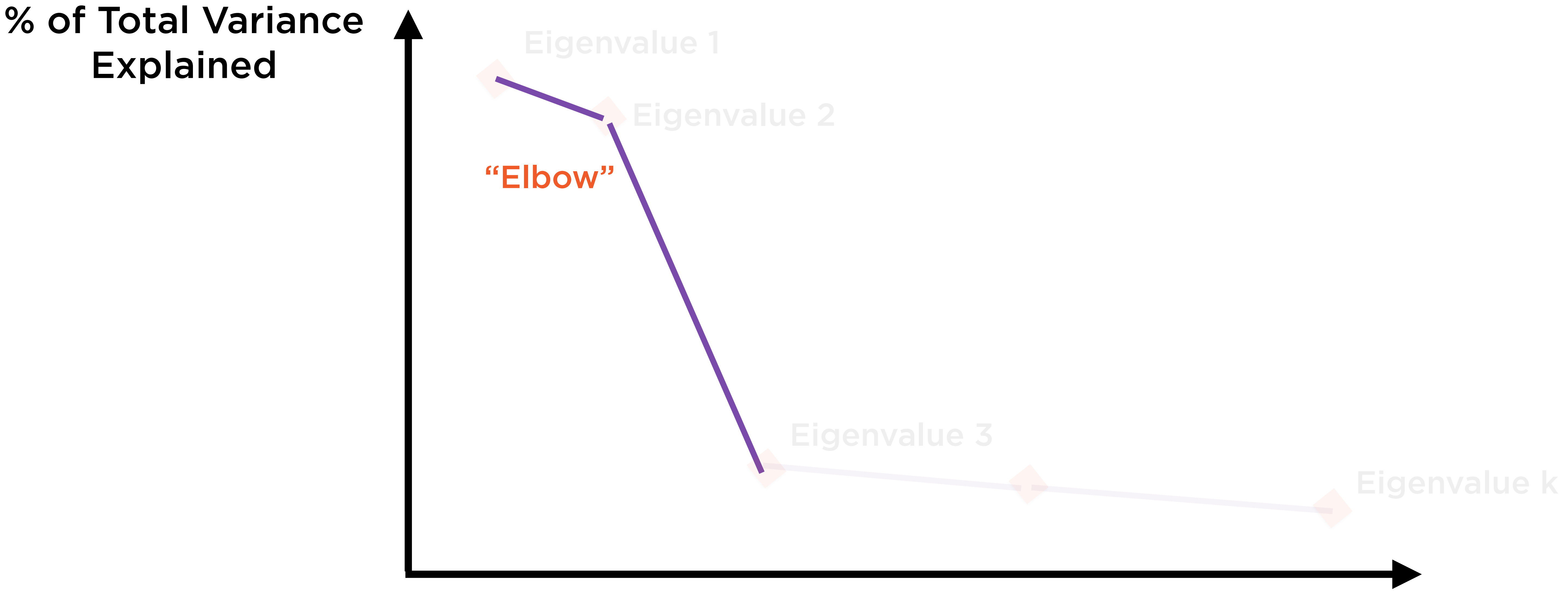
Matrix Multiplication

F_i	=	X	v_i
n rows, 1 column		n rows, k columns	k rows, 1 column

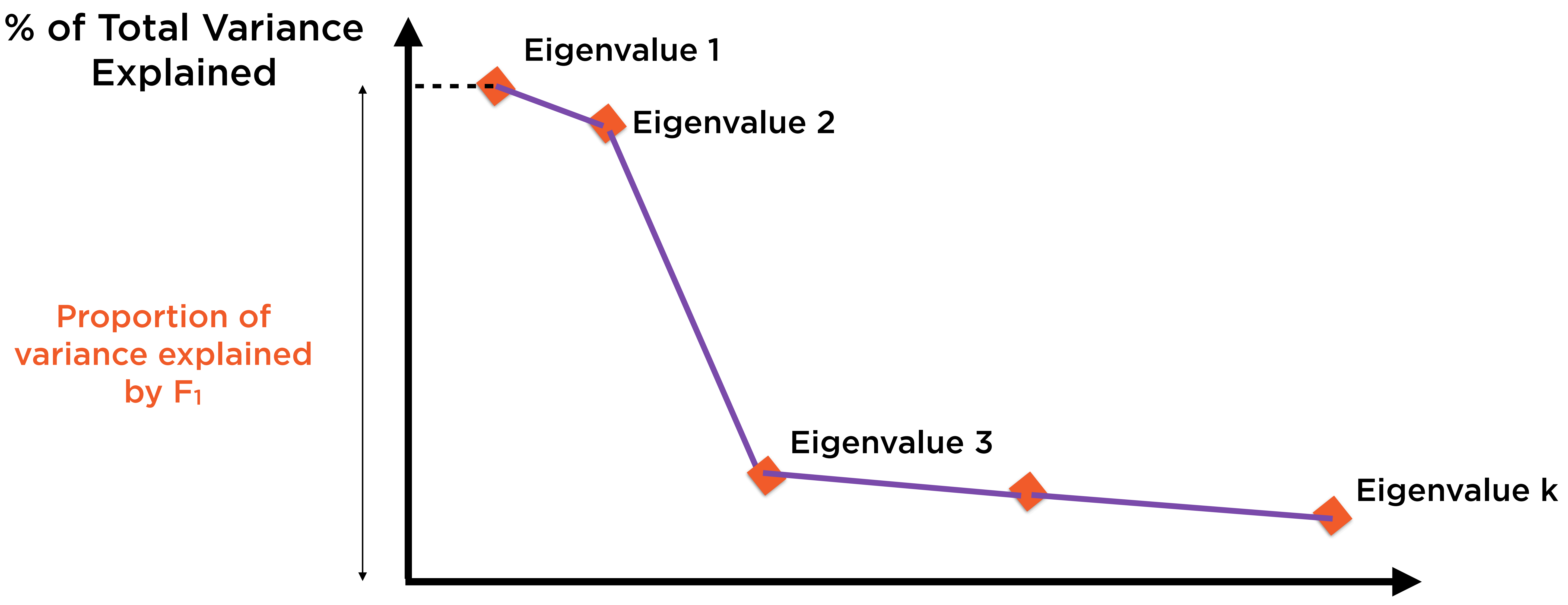
Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k

Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
"Elbow"
Eigenvalue 3
Eigenvalue k

Scree Plots


% of Total Variance Explained
Eigenvalue 1
Eigenvalue 2
Eigenvalue 3
Eigenvalue k
Proportion of variance explained by F_1

Summary

R makes PCA very simple and easy-to-use

PCA of equity returns reveals three important principal components

These closely correlate with underlying economic factors