Eigen Manual

v1.0

Chapter 1

Basics

Introduction

*"**Principal component analysis** (**PCA**) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible **variance** (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components."*

(from Wikipedia - Principal Component Analysis)

Section 2

Data & File Requirements

File Types

It is possible to import **tab-delimited** or comma-separated **text files**. For international users it is important to notice that comma-separated text-files may interfere with the decimal separator (,). The decimal separator is set automatically or can be set manually (cf. Options)

The files may contain a header row and row names. The matrix size (columns x rows) influences the speed and performance of the calculation:

Table 1.1 Performance

Matrix Size (columns * rows)	Duration (Calculation) in seconds
1000 (25*40)	< 1
10000 (25*400)	~ 2
100000 (50*2000)	~ 8
1000000 (50*20000)	~ 46

The table depicts the duration of a calculation in relation to a given matrix size. The values were determined using the new iPad (March 2012).

Transposing data

It may be necessary to **transpose** a given dataset (matrix). If a data data file is arranged as depicted in Table 1.2 transposing of data is necessary. For calculation of the principal components the samples (observations) should be grouped in rows and all variables in columns. Eigen is capable to transpose the data and this option should be selected (cf. Figure) if transposing of data is necessary.

Transposing required

Table 1.2 Matrix A

Sample Name	Sample 1	Sample 2	Sample 3
Variable 1	1	2	3
Variable 2	4	5	6
Variable 3	7	8	9
Variable 4	10	11	12
Variable 5	13	14	15

Transposing not required

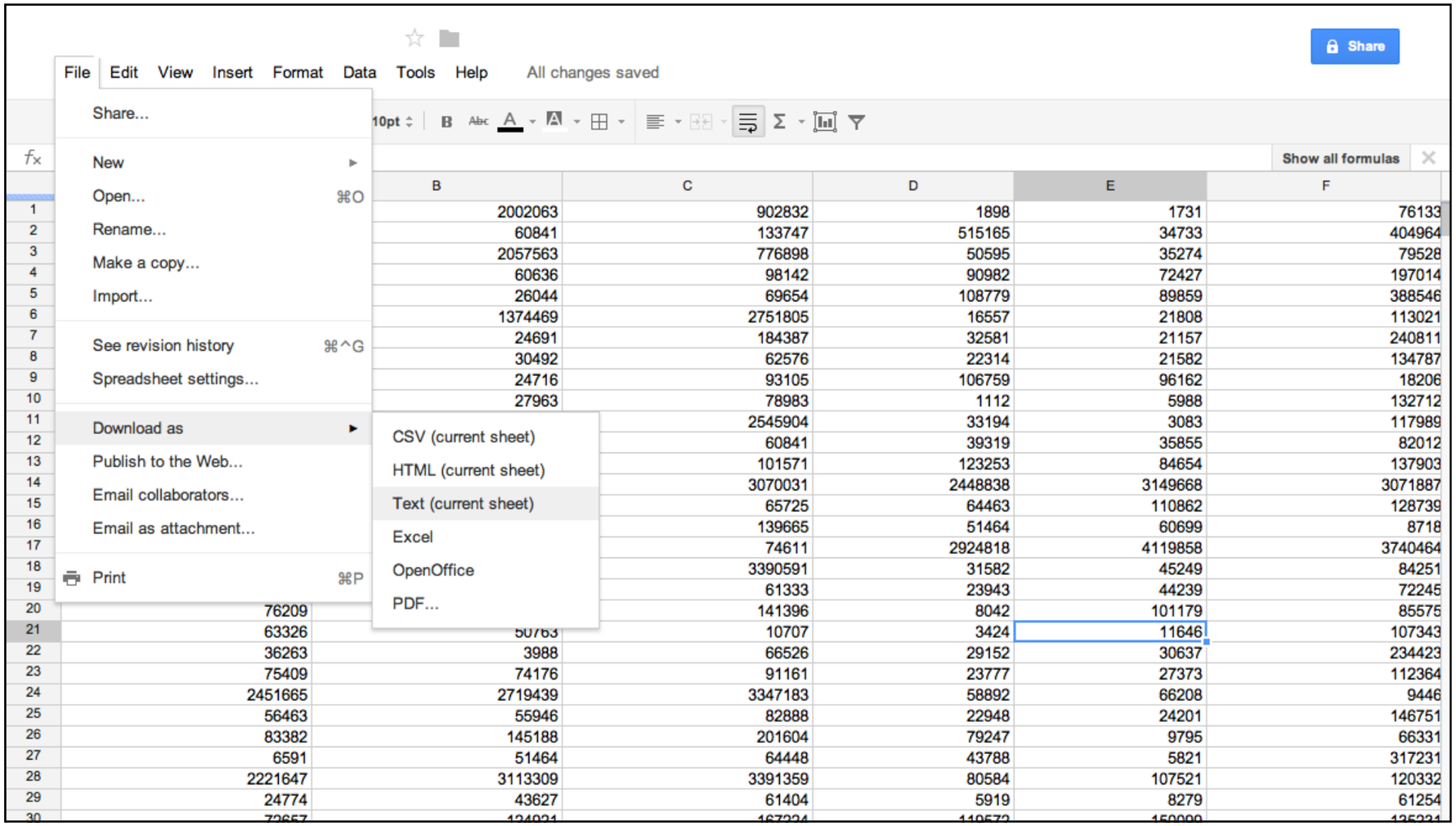
Table 1.3 Matrix A'

Sample Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Sample 1	1	4	7	10	13
Sample 2	2	5	8	11	14
Sample 3	3	6	9	12	15

Generation of a text-delimited text-file

Data from Excel or other software can be exported as tab-delimited text-files. The file extension should be **.txt, .text, .csv or .tsv** to be recognized by Eigen. Data can be transferred to Eigen via iTunes or Dropbox.

Figure 1.1 Export from Google Docs



The figure depicts the export of data from Google Docs.

Raw data and descriptive data

It is possible to import two kinds of data:

1. Raw data (Data from measurements) [required]

2. descriptive Data (Metadata, Description of samples) [optional]

The first kind of data (required) is used to calculate the scores of the principal components. It is mandatory to provide this data. The second kind of data (optional) is used to determine the coloring of the plot and to calculate correlations (based on **regression** analysis or a **students t-test**). If non-numerical data is provided, Eigen will try to convert the provided textual information to numerical information. It is important to notice that this classification might misinterpret qualitative data (cf. Table 1.4, left columns). To avoid this phenomenon change the order of samples or provide numerical information.

Table 1.4 Conversion of provided descriptive data

FALSE	0		Low	0
TRUE	1		High	1
TRUE	1		Normal	2
FALSE	0		Heigh	1

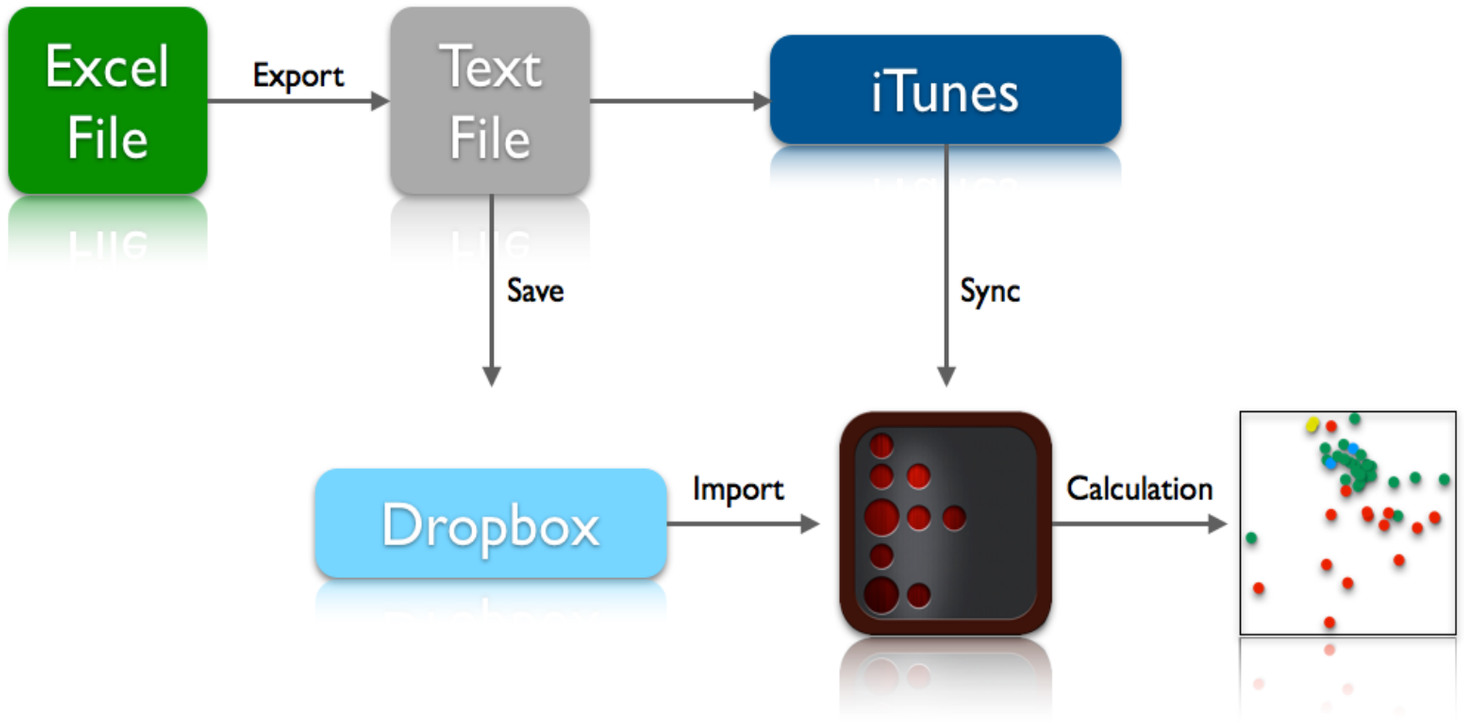
The table demonstrate the conversion from textual information (e.g. “FALSE/TRUE” to numerical information “0/1”).

The alignment of raw data and descriptive data is based on the sample names. Therefore it is mandatory to avoid duplicate sample names. If no sample names are provided the order within the tables will be used.

Exemplary Data

To demonstrate the function of Eigen an exemplary dataset and corresponding descriptive data was generated and subjected to Eigen according to the flowchart. For details to sync data via iTunes please refer to the online guide. A free Dropbox account is available here.

Figure 1.2 Flowchart Data processing



The figure shows the data export and import to Eigen.

Table 1.5 Dataset

No	Model	CPU	Chip	Storage	RAM	Weight	Size	Pixel	Resolution	Camera
1	Apple iPhone	412	1	16	128	135	3.5	153600	43885.7143	1.92
2	Apple iPhone 3G	412	1	16	128	133	3.5	153600	43885.7143	1.92
3	Apple iPhone 3GS	600	1	32	256	135	3.5	153600	43885.7143	3
4	Apple iPhone 4	800	1	32	512	137	3.5	614400	175542.857	5.0
5	Apple iPhone 4S	800	1	64	512	140	3.5	614400	175542.857	8.0
6	HTC 7 Mozart	1000	2	8.5	576	130	3.7	384000	103783.784	8
7	HTC 7 Trophy	1001	2	8.5	576	140	3.8	384000	101052.632	5
8	HTC Desire	1002	2	0.5	576	135	3.7	384000	103783.784	5
9	HTC Desire HD	1003	2	1.5	768	164	4.3	384000	89302.3256	8
10	HTC Desire S	1004	2	1.1	768	130	3.7	384000	103783.784	5
11	HTC Desire Z	800	2	1.5	512	180	3.7	384000	103783.784	5
...	...	...	...	...	...	...	...	...	...	...
45	Sony Xperia Play	1000	2	0.4	512	175	4	409920	102480	5
46	Sony Xperia X10	1000	2	1	384	135	4.0	409920	102480	8

The table depicts data of smartphones.

Source: Comparison of smartphones

Table 1.6 Descriptive Data

No	Model	Manufacturer	Chip	OS
1	Apple iPhone	Apple	ARM	iOS
2	Apple iPhone 3G	Apple	ARM	iOS
3	Apple iPhone 3GS	Apple	ARM	iOS
4	Apple iPhone 4	Apple	ARM	iOS
5	Apple iPhone 4S	Apple	ARM	iOS
6	HTC 7 Mozart	HTC	Qualcomm	WP7
7	HTC 7 Trophy	HTC	Qualcomm	WP7
8	HTC Desire	HTC	Qualcomm	Android 2
9	HTC Desire HD	HTC	Qualcomm	Android 2
10	HTC Desire S	HTC	Qualcomm	Android 2
11	HTC Desire Z	HTC	Qualcomm	Android 2
...	...	...	...	
45	Sony Xperia Play	Sony	Qualcomm	Android 2
46	Sony Xperia X10	Sony	Qualcomm	Android 2

The table depicts the descriptive data of Table 1.5.

Chapter 2

User Interface

Section 1

Projects

Projects

All successful calculations will be automatically listed in the project table and (if enabled) sync to iCloud. Projects can be selected from this list to display the corresponding graph or directly deleted via a swipe gesture.

New projects can be started by tapping the “New Calculation” button. Options can be assessed via the information button.

Figure 2.1 Project List

Data Import

Data import

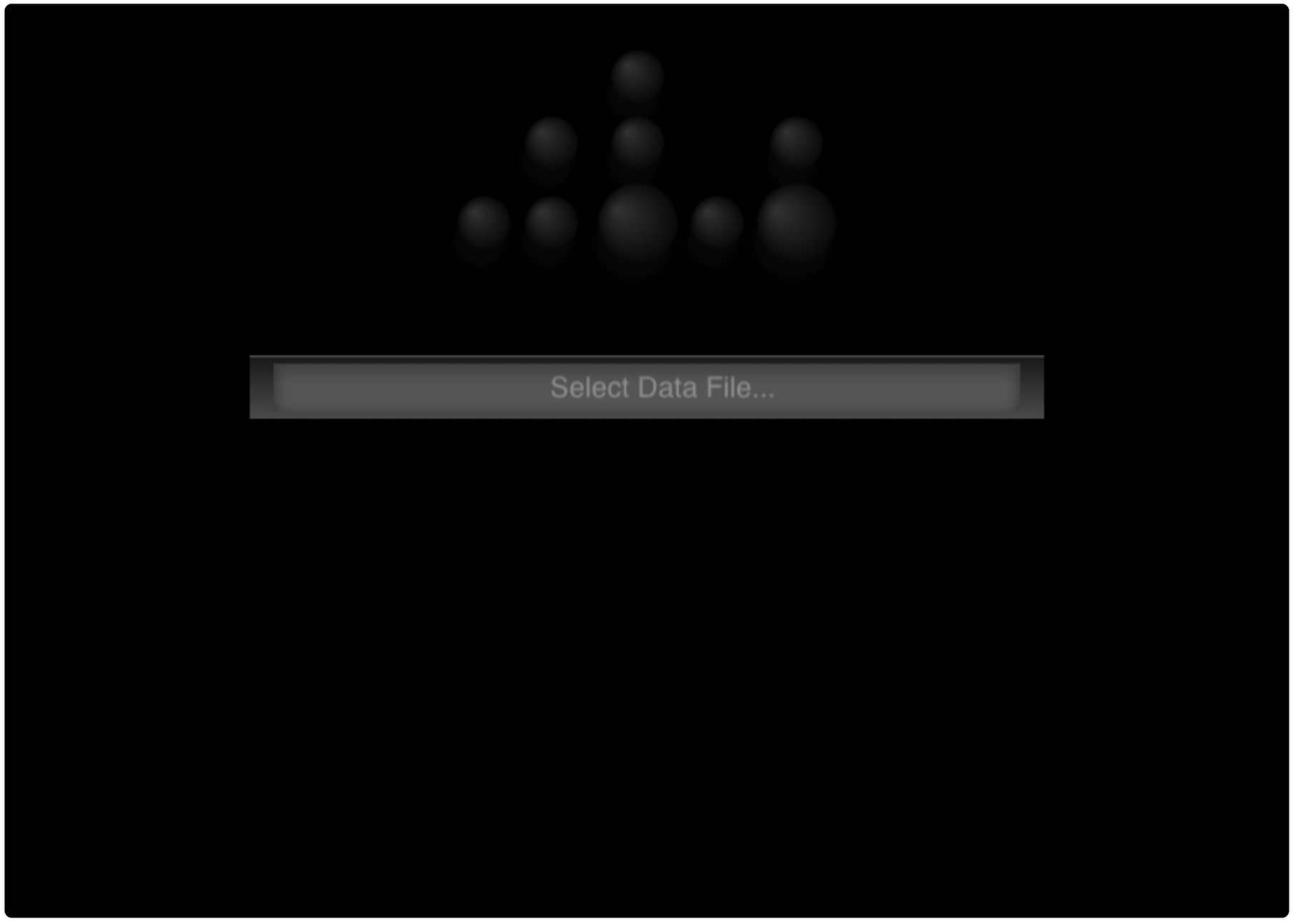
Data can be imported from the local directory or from Dropbox (*authorization required*). A special folder (Apps) and a subfolder (Eigen) will be generated. Files stored in the Eigen folder can be assessed.

Selected files from Dropbox will be downloaded and stored locally. Local files can be deleted via a swipe gesture. Local files can be assessed via iTunes (Guide).

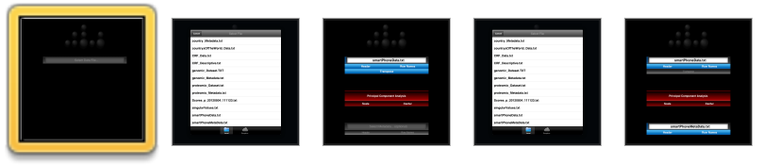
The gallery shows the sequence of data import.

Figure 2.2 Dropbox navigation

Gallery 2.1 Data Import



Basic Interface. Tap to select a data file.



File Parameter & Preprocessing

After selection of files Eigen will try to determine if header rows and row names are present. Additionally Eigen will try to determine if a transposing of data is required. This determination (transposing) is solely based on the geometry of the provided matrix (number of columns and rows) and should be adjusted by the user as required.

Before starting the calculation by tapping the “Principal Component Analysis” button, it is possible to determine the preprocessing of data (scale and/or center). The table shows the impact on the original data.

Centering is done by subtracting the column means from the corresponding columns.

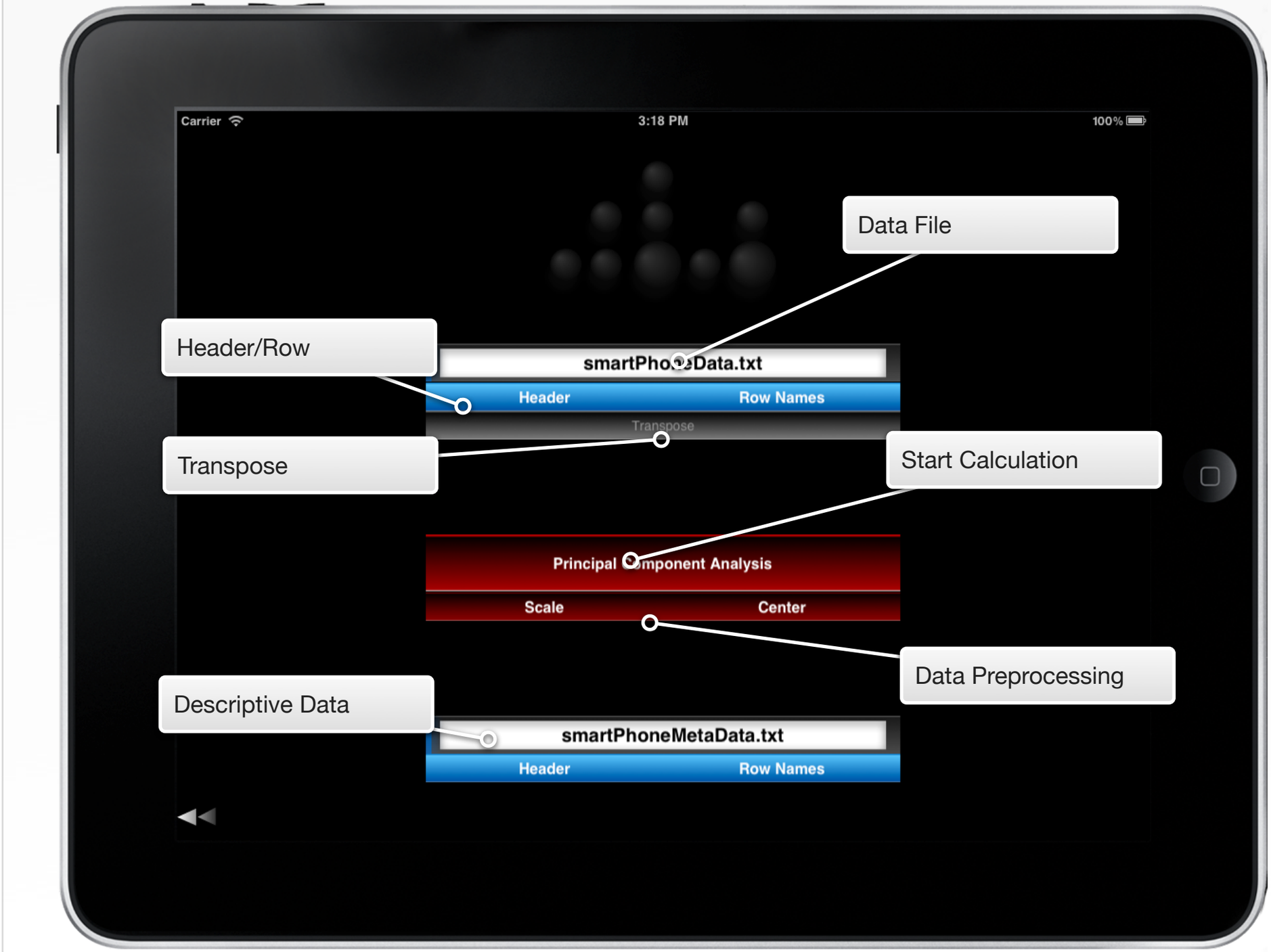
Scaling is done by dividing the (centered) columns by their standard deviations or the (non-centered) columns by the root mean square otherwise.

Table 2.1 Center and Scale

Original Data	Center	Scale	Scale and Center
1	-4	0.1675416	-1.4605935
2	-3	0.3350831	-1.0954451
3	-2	0.5026247	-0.7302967
4	-1	0.6701663	-0.3651484
5	0	0.8377078	0.0000000
6	1	1.0052494	0.3651484
7	2	1.1727909	0.7302967
8	3	1.3403325	1.0954451
9	4	1.5078741	1.4605935

The table demonstrates the effects of centering and scaling

Figure 2.3 User Interface - Data Import



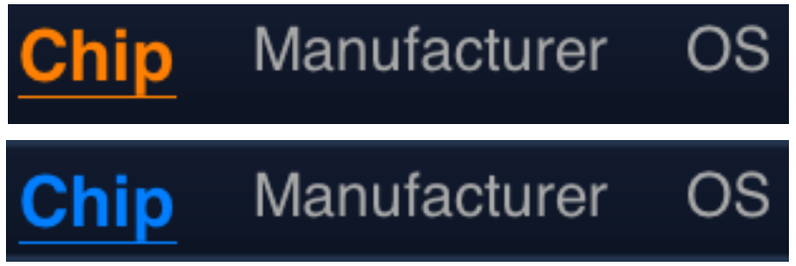
Section 3

Graph

Graphical Representation

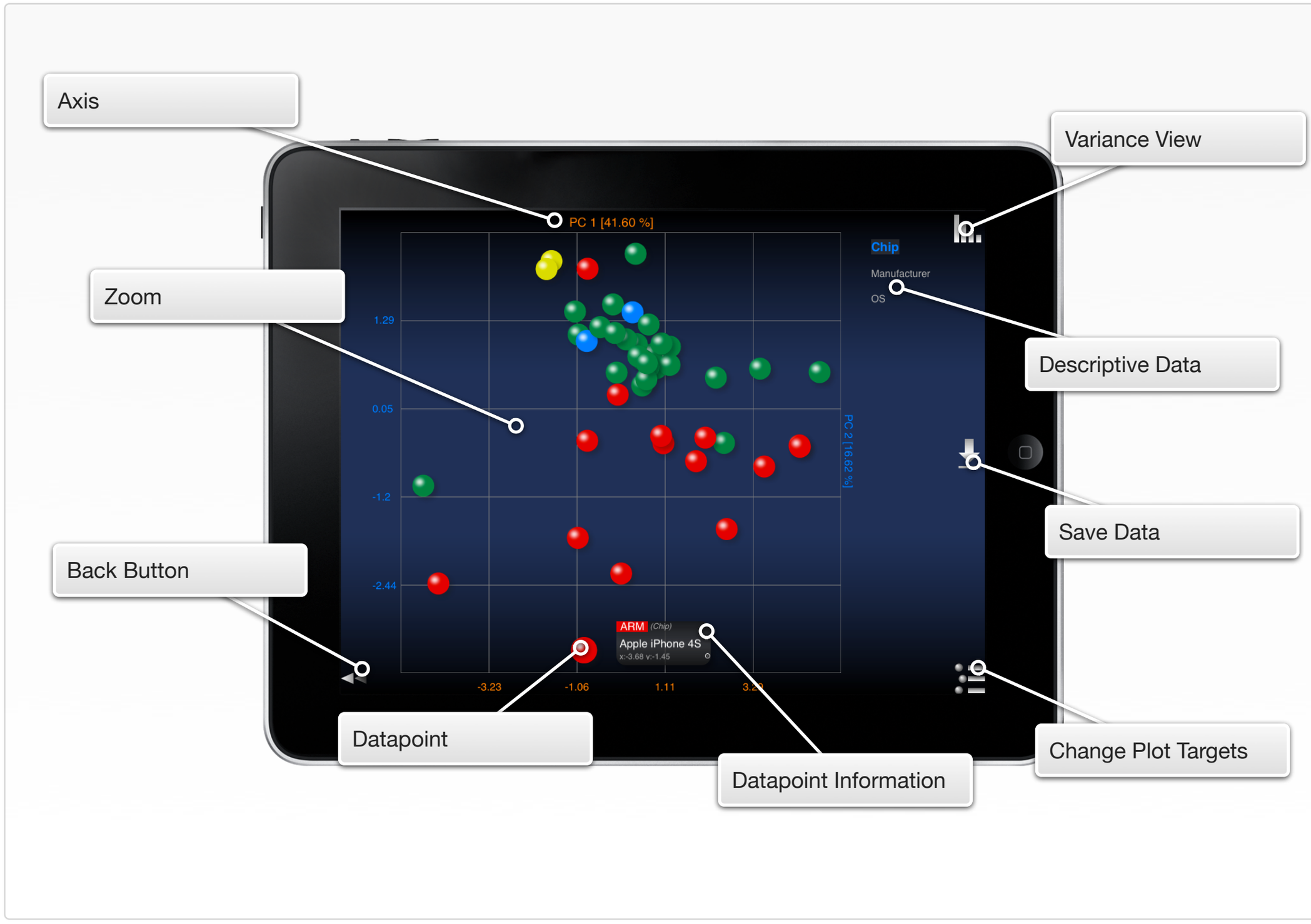
The results of the principal component analysis (Scores) are presented as a 2D scatter plot. Each sample is represented by a dot. Tapping a dot will reveal the corresponding information. The color of the dots are dependent from the provided descriptive data.

The provided descriptive data will be statistically evaluated to reveal significant correlations to specific scores. Correlating variables will be depicted in bold and in specific colors: Orange (correlation to the PC selected for the x-axis) or Blue (correlation to the PC selected for the y-axis). It is possible to plot variables from the descriptive data to calculated scores (Change Plot Targets). If no descriptive data is provided all dots will be displayed with the same color.



The results (scores) can be saved locally (.eigen) as a tab-delimited text-file.

Figure 2.4 Elements of the graphical representation of the scores of the principal components



Chapter 3

Applications

Examples

Application Examples

The presentation shows three application examples of principal component analyses to elucidate structures (cluster) within complex data. The statistical evaluation of corresponding descriptive data helps to explain the results of the PCA.

Examples from three different areas were selected:

- Distribution of wealth in the world

  This example illustrates the outcome of a PCA to reveal a basic phenomenon. It shows the differentiation between industrialized, emerging and less-developed countries.

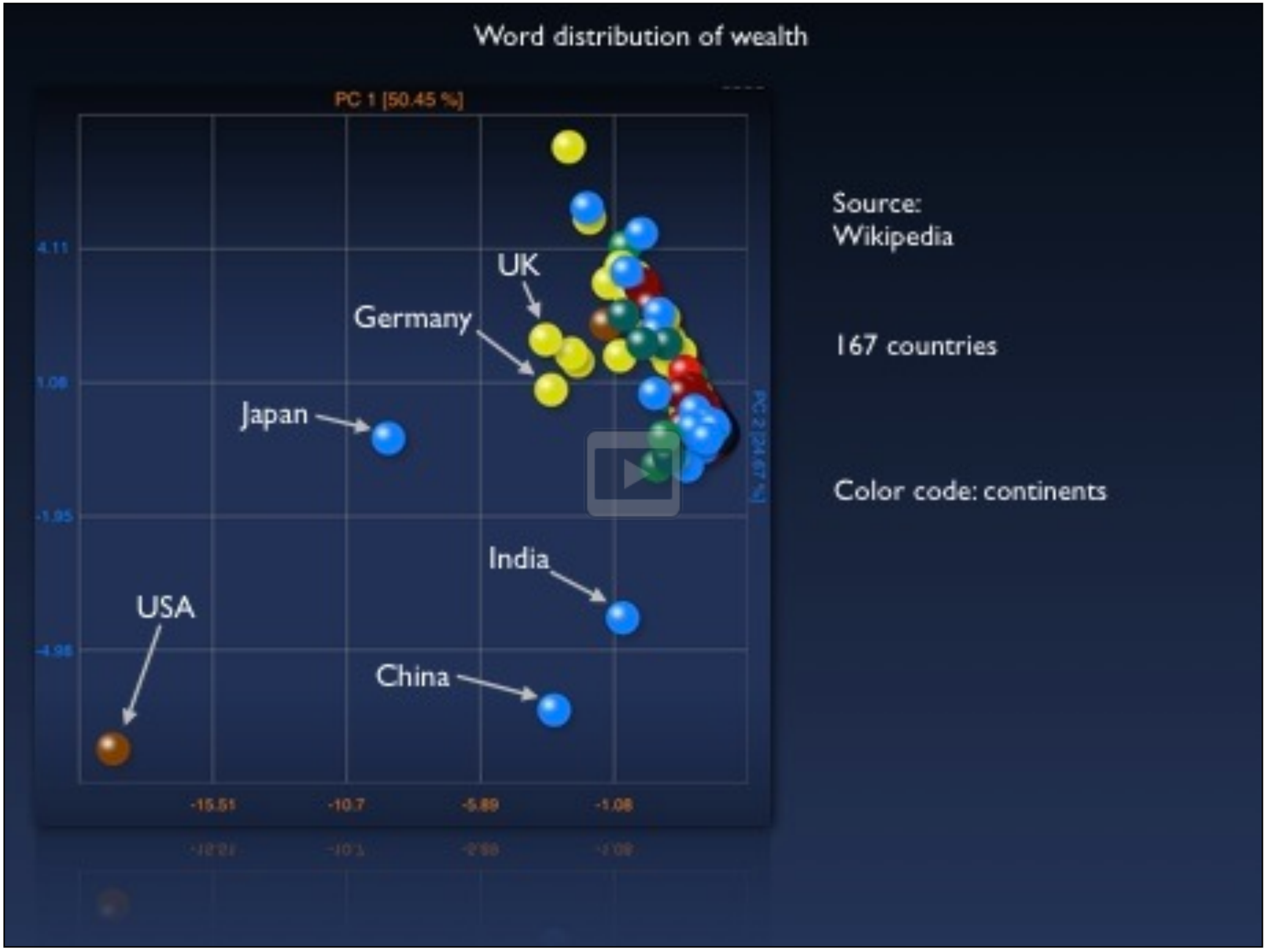
- **Microarray** Data

  The dataset derived from a study of expression profiles (mRNA) in primary breast cancer by Wang et al. The data shows the clustering according to the type of breast cancer.

- Data from a clinical study in chronic renal disease

  The data shows the clustering of samples according to the stage of the disease.

Interactive 3.1 Application Examples



The slides demonstrate three different applications to analyze complex data by principal component analysis.

Matrix

A rectangular array of quantities or expressions in rows and columns that is treated as a single entity and manipulated according to particular rules.

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 1 - Data & File Requirements

Microarray

A microarray is a multiplex lab-on-a-chip. It is a 2D array on a solid substrate (usually a glass slide or silicon thin-film cell) that assays large amounts of biological material using high-throughput screening methods.

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 3 - Examples

PCA

Principal Component Analysis

Related Glossary Terms

Principal component analysis

Index

Find Term

Chapter 1 - Introduction

Principal component analysis

A method of analysis which involves finding the linear combination of a set of variables that has maximum variance and removing its effect, repeating this successively.

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 1 - Introduction

Regression

A measure of the relation between the mean value of one variable (e.g., output) and corresponding values of other variables (e.g., time and cost).

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 1 - Data & File Requirements

Students t-test

A test for statistical significance that uses tables of a statistical distribution called Student's t-distribution, which is that of a fraction (t) whose numerator is drawn from a normal distribution with a mean of zero, and whose denominator is the root mean square of k terms drawn from the same normal distribution (where k is the number of degrees of freedom).

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 1 - Data & File Requirements

Tab-delimited

A tab-separated values file is a simple text format for a database table. Each record in the table is one line of the text file. Each field value of a record is separated from the next by a tab stop character – it is a form of the more general delimiter-separated values format.

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 1 - Data & File Requirements

Text files

Text files (sometimes spelled "textfiles": an old alternate name is "flatfile") are a kind of computer files that are structured as a sequence of lines of electronic text.

A text file exists within a computer file system. The end of a text file is often denoted by placing one or more special characters, known as an end-of-file marker, after the last line in a text file.

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 1 - Data & File Requirements

Transpose

A matrix obtained from a given matrix by interchanging each row and the corresponding column.

Related Glossary Terms

Drag related terms here

Index

Find Term

Chapter 1 - Data & File Requirements

Variance

In probability theory and statistics, the variance is a measure of how far a set of numbers is spread out. It is one of several descriptors of a probability distribution, describing how far the numbers lie from the mean (expected value). In particular, the variance is one of the moments of a distribution. In that context, it forms part of a systematic approach to distinguishing between probability distributions. While other such approaches have been developed, those based on moments are advantageous in terms of mathematical and computational simplicity.

Related Glossary Terms

Principal component analysis

Index Find Term

Chapter 1 - Introduction