



Exploratory Data Analysis with MATLAB®

**Wendy L. Martinez
Angel R. Martinez**



CHAPMAN & HALL/CRC

A CRC Press Company

Boca Raton London New York Washington, D.C.

Table of Contents

Table of Contents vii

Prefacexiii

Part I

Introduction to Exploratory Data Analysis

Chapter 1

Introduction to Exploratory Data Analysis

1.1 What is Exploratory Data Analysis 3

1.2 Overview of the Text 6

1.3 A Few Words About Notation 8

1.4 Data Sets Used in the Book 9

 1.4.1 Unstructured Text Documents 9

 1.4.2 Gene Expression Data 12

 1.4.3 Oronsay Data Set 18

 1.4.4 Software Inspection 19

1.5 Transforming Data 20

 1.5.1 Power Transformations 21

 1.5.2 Standardization 22

 1.5.3 Sphering the Data 24

1.6 Further Reading 25

Exercises 27

Part II

EDA as Pattern Discovery

Chapter 2

Dimensionality Reduction - Linear Methods

2.1 Introduction 31

2.2 Principal Component Analysis - PCA 33

 2.2.1 PCA Using the Sample Covariance Matrix 34

 2.2.2 PCA Using the Sample Correlation Matrix 37

 2.2.3 How Many Dimensions Should We Keep? 38

2.3 Singular Value Decomposition - SVD 42

2.4 Factor Analysis 46

2.5 Intrinsic Dimensionality	52
2.6 Summary and Further Reading	57
Exercises	57

Chapter 3

Dimensionality Reduction - Nonlinear Methods

3.1 Multidimensional Scaling - MDS	61
3.1.1 Metric MDS	63
3.1.2 Nonmetric MDS	72
3.2 Manifold Learning	81
3.2.1 Locally Linear Embedding	81
3.2.2 Isometric Feature Mapping - ISOMAP	83
3.2.3 Hessian Eigenmaps	85
3.3 Artificial Neural Network Approaches	90
3.3.1 Self-Organizing Maps - SOM	90
3.3.2 Generative Topographic Maps - GTM	94
3.4 Summary and Further Reading	98
Exercises	100

Chapter 4

Data Tours

4.1 Grand Tour	104
4.1.1 Torus Winding Method	105
4.1.2 Pseudo Grand Tour	107
4.2 Interpolation Tours	110
4.3 Projection Pursuit	112
4.4 Projection Pursuit Indexes	120
4.4.1 Posse Chi-Square Index	120
4.4.2 Moment Index	124
4.5 Summary and Further Reading	125
Exercises	126

Chapter 5

Finding Clusters

5.1 Introduction	127
5.2 Hierarchical Methods	129
5.3 Optimization Methods - k -Means	135
5.4 Evaluating the Clusters	139
5.4.1 Rand Index	141
5.4.2 Cophenetic Correlation	143
5.5.3 Upper Tail Rule	144
5.5.4 Silhouette Plot	147
5.5.5 Gap Statistic	149
5.5 Summary and Further Reading	155

Exercises 158

Chapter 6

Model-Based Clustering

6.1 Overview of Model-Based Clustering 163

6.2 Finite Mixtures 166

 6.2.1 Multivariate Finite Mixtures 167

 6.2.2 Component Models - Constraining the Covariances 168

6.3 Expectation-Maximization Algorithm 176

6.4 Hierarchical Agglomerative Model-Based Clustering 181

6.5 Model-Based Clustering 182

6.6 Generating Random Variables from a Mixture Model 188

6.7 Summary and Further Reading 192

Exercises 193

Chapter 7

Smoothing Scatterplots

7.1 Introduction 197

7.2 Loess 198

7.3 Robust Loess 208

7.4 Residuals and Diagnostics 211

 7.4.1 Residual Plots 212

 7.4.2 Spread Smooth 216

 7.4.3 Loess Envelopes - Upper and Lower Smooths 218

7.5 Bivariate Distribution Smooths 219

 7.5.1 Pairs of Middle Smoothings 219

 7.5.2 Polar Smoothing 222

7.6 Curve Fitting Toolbox 226

7.7 Summary and Further Reading 228

Exercises 229

Part III

Graphical Methods for EDA

Chapter 8

Visualizing Clusters

8.1 Dendrogram 233

8.2 Treemaps 235

8.3 Rectangle Plots 238

8.4 ReClus Plots 244

8.5 Data Image 249

8.6 Summary and Further Reading 255

Exercises 256

Chapter 9

Distribution Shapes

9.1 Histograms	259
9.1.1 Univariate Histograms	259
9.1.2 Bivariate Histograms	266
9.2 Boxplots	268
9.2.1 The Basic Boxplot	269
9.2.2 Variations of the Basic Boxplot	274
9.3 Quantile Plots	279
9.3.1 Probability Plots	279
9.3.2 Quantile-quantile Plot	281
9.3.3 Quantile Plot	284
9.4 Bagplots	286
9.5 Summary and Further Reading	289
Exercises	289

Chapter 10

Multivariate Visualization

10.1 Glyph Plots	293
10.2 Scatterplots	294
10.2.1 2-D and 3-D Scatterplots	294
10.2.2 Scatterplot Matrices	298
10.2.3 Scatterplots with Hexagonal Binning	299
10.3 Dynamic Graphics	301
10.3.1 Identification of Data	301
10.3.2 Linking	305
10.3.3 Brushing	308
10.4 Coplots	309
10.5 Dot Charts	312
10.5.1 Basic Dot Chart	313
10.5.2 Multiway Dot Chart	314
10.6 Plotting Points as Curves	318
10.6.1 Parallel Coordinate Plots	318
10.6.2 Andrews' Curves	321
10.6.3 More Plot Matrices	325
10.7 Data Tours Revisited	326
10.7.1 Grand Tour	326
10.7.2 Permutation Tour	328
10.8 Summary and Further Reading	332
Exercises	333

Appendix A

Proximity Measures

A.1 Definitions	337
A.1.1 Dissimilarities	338

A.1.2 Similarity Measures	340
A.1.3 Similarity Measures for Binary Data	340
A.1.4 Dissimilarities for Probability Density Functions	341
A.2 Transformations	342
A.3 Further Reading	343

Appendix B

Software Resources for EDA

B.1 MATLAB Programs	345
B.2 Other Programs for EDA	348
B.3 EDA Toolbox	350

Appendix C

Description of Data Sets	351
--------------------------------	-----

Appendix D

Introduction to MATLAB

D.1 What Is MATLAB?	357
D.2 Getting Help in MATLAB	358
D.3 File and Workspace Management	358
D.4 Punctuation in MATLAB	360
D.5 Arithmetic Operators	361
D.6 Data Constructs in MATLAB	362
Basic Data Constructs	362
Building Arrays	363
Cell Arrays	363
Structures	364
D.7 Script Files and Functions	365
D.8 Control Flow	366
for Loop	366
while Loop	366
if-else Statements	367
switch Statement	367
D.9 Simple Plotting	367
D.10 Where to get MATLAB Information	370

Appendix E

MATLAB Functions

E.1 MATLAB	371
E.2 Statistics Toolbox - Versions 4 and 5	373
E.3 Exploratory Data Analysis Toolbox	374

References 377

Author Index 395

Subject Index 401