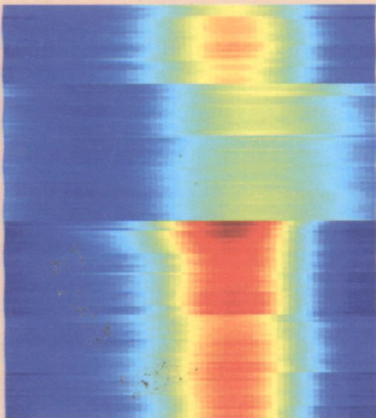
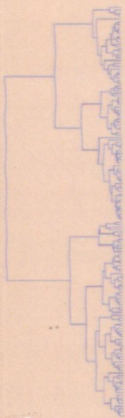


Computer Science and Data Analysis Series

Exploratory Data Analysis with MATLAB®

Third Edition



Wendy L. Martinez
Angel R. Martinez
Jeffrey L. Solka



CRC Press
Taylor & Francis Group

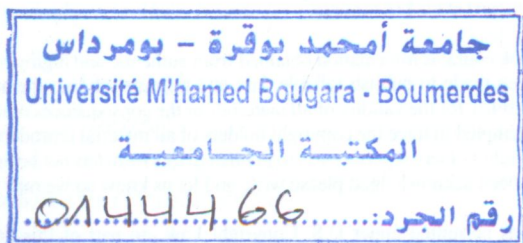
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Computer Science and Data Analysis Series

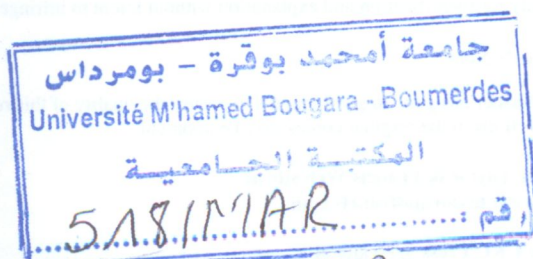


Exploratory Data Analysis with MATLAB®

Third Edition



Wendy L. Martinez
Angel R. Martinez
Jeffrey L. Solka



CRC Press
Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an *informa* business
A CHAPMAN & HALL BOOK

Table of Contents

Preface to the Third Edition.....	xvii
Preface to the Second Edition.....	xix
Preface to the First Edition.....	xxiii

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 Nonnegative Matrix Factorization	47
2.5 Factor Analysis	51
2.6 Fisher's Linear Discriminant	56
2.7 Random Projections	61
2.8 Intrinsic Dimensionality	65
2.8.1 Nearest Neighbor Approach	67
2.8.2 Correlation Dimension	71
2.8.3 Maximum Likelihood Approach	72
2.8.4 Estimation Using Packing Numbers	74
2.8.5 Estimation of Local Dimension	76
2.9 Summary and Further Reading	79
Exercises	81

Chapter 3

Dimensionality Reduction — Nonlinear Methods

3.1 Multidimensional Scaling — MDS	85
3.1.1 Metric MDS	87
3.1.2 Nonmetric MDS	97
3.2 Manifold Learning	105
3.2.1 Locally Linear Embedding	105
3.2.2 Isometric Feature Mapping — ISOMAP	107
3.2.3 Hessian Eigenmaps	109
3.3 Artificial Neural Network Approaches	114
3.3.1 Self-Organizing Maps	114
3.3.2 Generative Topographic Maps	117
3.3.3 Curvilinear Component Analysis	122
3.3.4 Autoencoders	127
3.4 Stochastic Neighbor Embedding	131
3.5 Summary and Further Reading	135
Exercises	136

Chapter 4

Data Tours

4.1 Grand Tour	140
4.1.1 Torus Winding Method	141
4.1.2 Pseudo Grand Tour	143
4.2 Interpolation Tours	146
4.3 Projection Pursuit	148
4.4 Projection Pursuit Indexes	156
4.4.1 Posse Chi-Square Index	156
4.4.2 Moment Index	159
4.5 Independent Component Analysis	161
4.6 Summary and Further Reading	165
Exercises	166

Chapter 5**Finding Clusters**

5.1 Introduction	169
5.2 Hierarchical Methods	171
5.3 Optimization Methods — <i>k</i> -Means	177
5.4 Spectral Clustering	181
5.5 Document Clustering	185
5.5.1 Nonnegative Matrix Factorization — Revisited	187
5.5.2 Probabilistic Latent Semantic Analysis	191
5.6 Minimum Spanning Trees and Clustering	196
5.6.1 Definitions	196
5.6.2 Minimum Spanning Tree Clustering	199
5.7 Evaluating the Clusters	204
5.7.1 Rand Index	205
5.7.2 Cophenetic Correlation	207
5.7.3 Upper Tail Rule	208
5.7.4 Silhouette Plot	211
5.7.5 Gap Statistic	213
5.7.6 Cluster Validity Indices	219
5.8 Summary and Further Reading	230
Exercises	232

Chapter 6**Model-Based Clustering**

6.1 Overview of Model-Based Clustering	237
6.2 Finite Mixtures	240
6.2.1 Multivariate Finite Mixtures	242
6.2.2 Component Models — Constraining the Covariances	243
6.3 Expectation-Maximization Algorithm	249
6.4 Hierarchical Agglomerative Model-Based Clustering	254
6.5 Model-Based Clustering	256
6.6 MBC for Density Estimation and Discriminant Analysis	263
6.6.1 Introduction to Pattern Recognition	263
6.6.2 Bayes Decision Theory	264
6.6.3 Estimating Probability Densities with MBC	267
6.7 Generating Random Variables from a Mixture Model	271
6.8 Summary and Further Reading	273
Exercises	276

Chapter 7**Smoothing Scatterplots**

7.1 Introduction	279
7.2 Loess	280
7.3 Robust Loess	291
7.4 Residuals and Diagnostics with Loess	293

7.4.1 Residual Plots	293
7.4.2 Spread Smooth	297
7.4.3 Loess Envelopes — Upper and Lower Smooths	300
7.5 Smoothing Splines	301
7.5.1 Regression with Splines	302
7.5.2 Smoothing Splines	304
7.5.3 Smoothing Splines for Uniformly Spaced Data	310
7.6 Choosing the Smoothing Parameter	313
7.7 Bivariate Distribution Smooths	317
7.7.1 Pairs of Middle Smoothings	317
7.7.2 Polar Smoothing	319
7.8 Curve Fitting Toolbox	323
7.9 Summary and Further Reading	325
Exercises	326

Part III

Graphical Methods for EDA

Chapter 8

Visualizing Clusters

8.1 Dendrogram	333
8.2 Treemaps	335
8.3 Rectangle Plots	338
8.4 ReClus Plots	344
8.5 Data Image	349
8.6 Summary and Further Reading	355
Exercises	356

Chapter 9

Distribution Shapes

9.1 Histograms	359
9.1.1 Univariate Histograms	359
9.1.2 Bivariate Histograms	366
9.2 Kernel Density	368
9.2.1 Univariate Kernel Density Estimation	369
9.2.2 Multivariate Kernel Density Estimation	371
9.3 Boxplots	374
9.3.1 The Basic Boxplot	374
9.3.2 Variations of the Basic Boxplot	380
9.3.3 Violin Plots	383
9.3.4 Beeswarm Plot	385
9.3.5 Beanplot	388
9.4 Quantile Plots	390
9.4.1 Probability Plots	392

11.3.1 Background	481
11.3.2 Bar Plots	483
11.3.3 Spine Plots	486
11.3.4 Mosaic Plots	489
11.3.5 Sieve Plots	490
11.3.6 Log Odds Plot	493
11.4 Summary and Further Reading	498
Exercises	500

Appendix A

Proximity Measures

A.1 Definitions	503
A.1.1 Dissimilarities	504
A.1.2 Similarity Measures	506
A.1.3 Similarity Measures for Binary Data	506
A.1.4 Dissimilarities for Probability Density Functions	507
A.2 Transformations	508
A.3 Further Reading	509

Appendix B

Software Resources for EDA

B.1 MATLAB Programs	511
B.2 Other Programs for EDA	515
B.3 EDA Toolbox	516

Appendix C

Description of Data Sets	517
--------------------------------	-----

Appendix D

MATLAB® Basics

D.1 Desktop Environment	523
D.2 Getting Help and Other Documentation	525
D.3 Data Import and Export	526
D.3.1 Data Import and Export in Base MATLAB®	526
D.3.2 Data Import and Export with the Statistics Toolbox	528
D.4 Data in MATLAB®	529
D.4.1 Data Objects in Base MATLAB®	529
D.4.2 Accessing Data Elements	532
D.4.3 Object-Oriented Programming	535
D.5 Workspace and Syntax	535
D.5.1 File and Workspace Management	536
D.5.2 Syntax in MATLAB®	537
D.5.3 Functions in MATLAB®	539
D.6 Basic Plot Functions	540

D.6.1 Plotting 2D Data 540

D.6.2 Plotting 3D Data 543

D.6.3 Scatterplots 544

D.6.4 Scatterplot Matrix 545

D.6.5 GUIs for Graphics 545

D.7 Summary and Further Reading 547

References 551

Author Index 575

Subject Index 583

Exploratory Data Analysis with MATLAB®

Third Edition

Praise for the Second Edition

"The authors present an intuitive and easy-to-read book. ... accompanied by many examples, proposed exercises, good references, and comprehensive appendices that initiate the reader unfamiliar with MATLAB®."

—Adolfo Alvarez Pinto, *International Statistical Review*

"Practitioners of EDA who use MATLAB will want a copy of this book. ... The authors have done a great service by bringing together so many EDA routines, but their main accomplishment in this dynamic text is providing the understanding and tools to do EDA."

—David A Huckaby, *MAA Reviews*

Exploratory Data Analysis (EDA) is an important part of the data analysis process. The methods presented in this text are ones that should be in the toolkit of every data scientist. As computational sophistication has increased and data sets have grown in size and complexity, EDA has become an even more important process for visualizing and summarizing data before making assumptions to generate hypotheses and models.

Exploratory Data Analysis with MATLAB®, Third Edition presents EDA methods from a computational perspective and uses numerous examples and applications to show how the methods are used in practice. The authors use MATLAB code, pseudo-code, and algorithm descriptions to illustrate the concepts. The MATLAB code for examples, data sets, and the EDA Toolbox are available for download on the book's website.

New to the Third Edition

- Random projections and estimating local intrinsic dimensionality
- Deep learning autoencoders and stochastic neighbor embedding
- Minimum spanning tree and additional cluster validity indices
- Kernel density estimation
- Plots for visualizing data distributions, such as beanplots and violin plots
- A chapter on visualizing categorical data



CRC Press

Taylor & Francis Group
an informa business

www.crcpress.com

ISBN 978-1-03-217905-6



9 781032 179056

www.crcpress.com

