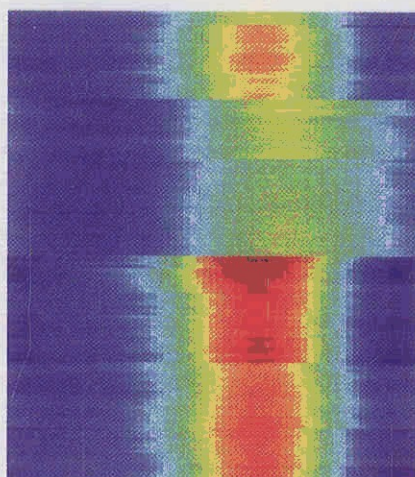


Computer Science and Data Analysis Series

Exploratory Data Analysis with MATLAB®

Second Edition



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



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Table of Contents

Preface to the Second Edition.....xiii
Preface to the First Edition.....xvii

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 Intrinsic Dimensionality	61
2.7.1 Nearest Neighbor Approach	63
2.7.2 Correlation Dimension	67
2.7.3 Maximum Likelihood Approach	68
2.7.4 Estimation Using Packing Numbers	70
2.8 Summary and Further Reading	72
Exercises	73

Chapter 3

Dimensionality Reduction — Nonlinear Methods

3.1 Multidimensional Scaling — MDS	79
3.1.1 Metric MDS	81
3.1.2 Nonmetric MDS	91
3.2 Manifold Learning	99
3.2.1 Locally Linear Embedding	99
3.2.2 Isometric Feature Mapping — ISOMAP	101
3.2.3 Hessian Eigenmaps	103
3.3 Artificial Neural Network Approaches	108
3.3.1 Self-Organizing Maps	108
3.3.2 Generative Topographic Maps	111
3.3.3 Curvilinear Component Analysis	116
3.4 Summary and Further Reading	121
Exercises	122

Chapter 4

Data Tours

4.1 Grand Tour	126
4.1.1 Torus Winding Method	127
4.1.2 Pseudo Grand Tour	129
4.2 Interpolation Tours	132
4.3 Projection Pursuit	134
4.4 Projection Pursuit Indexes	142
4.4.1 Posse Chi-Square Index	142
4.4.2 Moment Index	145
4.5 Independent Component Analysis	147
4.6 Summary and Further Reading	151
Exercises	152

Chapter 5

Finding Clusters

5.1 Introduction	155
5.2 Hierarchical Methods	157

5.3 Optimization Methods — k -Means	163
5.4 Spectral Clustering	167
5.5 Document Clustering	171
5.5.1 Nonnegative Matrix Factorization — Revisited	173
5.5.2 Probabilistic Latent Semantic Analysis	177
5.6 Evaluating the Clusters	182
5.6.1 Rand Index	182
5.6.2 Cophenetic Correlation	185
5.6.3 Upper Tail Rule	186
5.6.4 Silhouette Plot	189
5.6.5 Gap Statistic	191
5.7 Summary and Further Reading	197
Exercises	200

Chapter 6

Model-Based Clustering

6.1 Overview of Model-Based Clustering	205
6.2 Finite Mixtures	207
6.2.1 Multivariate Finite Mixtures	210
6.2.2 Component Models — Constraining the Covariances	211
6.3 Expectation-Maximization Algorithm	217
6.4 Hierarchical Agglomerative Model-Based Clustering	222
6.5 Model-Based Clustering	224
6.6 MBC for Density Estimation and Discriminant Analysis	231
6.6.1 Introduction to Pattern Recognition	231
6.6.2 Bayes Decision Theory	232
6.6.3 Estimating Probability Densities with MBC	235
6.7 Generating Random Variables from a Mixture Model	239
6.8 Summary and Further Reading	241
Exercises	244

Chapter 7

Smoothing Scatterplots

7.1 Introduction	247
7.2 Loess	248
7.3 Robust Loess	259
7.4 Residuals and Diagnostics with Loess	261
7.4.1 Residual Plots	261
7.4.2 Spread Smooth	265
7.4.3 Loess Envelopes — Upper and Lower Smooths	268
7.5 Smoothing Splines	269
7.5.1 Regression with Splines	270
7.5.2 Smoothing Splines	272
7.5.3 Smoothing Splines for Uniformly Spaced Data	278
7.6 Choosing the Smoothing Parameter	281

7.7 Bivariate Distribution Smooths	285
7.7.1 Pairs of Middle Smoothings	285
7.7.2 Polar Smoothing	287
7.8 Curve Fitting Toolbox	291
7.9 Summary and Further Reading	293
Exercises	294

Part III

Graphical Methods for EDA

Chapter 8

Visualizing Clusters

8.1 Dendrogram	301
8.2 Treemaps	303
8.3 Rectangle Plots	306
8.4 ReClus Plots	312
8.5 Data Image	317
8.6 Summary and Further Reading	323
Exercises	324

Chapter 9

Distribution Shapes

9.1 Histograms	327
9.1.1 Univariate Histograms	327
9.1.2 Bivariate Histograms	334
9.2 Boxplots	336
9.2.1 The Basic Boxplot	337
9.2.2 Variations of the Basic Boxplot	342
9.3 Quantile Plots	347
9.3.1 Probability Plots	347
9.3.2 Quantile-Quantile Plot	349
9.3.3 Quantile Plot	352
9.4 Bagplots	354
9.5 Rangefinder Boxplot	356
9.6 Summary and Further Reading	359
Exercises	361

Chapter 10

Multivariate Visualization

10.1 Glyph Plots	365
10.2 Scatterplots	366
10.2.1 2-D and 3-D Scatterplots	368
10.2.2 Scatterplot Matrices	371
10.2.3 Scatterplots with Hexagonal Binning	372

10.3 Dynamic Graphics	374
10.3.1 Identification of Data	376
10.3.2 Linking	378
10.3.3 Brushing	381
10.4 Coplots	384
10.5 Dot Charts	387
10.5.1 Basic Dot Chart	387
10.5.2 Multiway Dot Chart	388
10.6 Plotting Points as Curves	392
10.6.1 Parallel Coordinate Plots	393
10.6.2 Andrews' Curves	395
10.6.3 Andrews' Images	399
10.6.4 More Plot Matrices	400
10.7 Data Tours Revisited	403
10.7.1 Grand Tour	404
10.7.2 Permutation Tour	405
10.8 Biplots	408
10.9 Summary and Further Reading	411
Exercises	413

Appendix A

Proximity Measures

A.1 Definitions	417
A.1.1 Dissimilarities	418
A.1.2 Similarity Measures	420
A.1.3 Similarity Measures for Binary Data	420
A.1.4 Dissimilarities for Probability Density Functions	421
A.2 Transformations	422
A.3 Further Reading	423

Appendix B

Software Resources for EDA

B.1 MATLAB Programs	425
B.2 Other Programs for EDA	429
B.3 EDA Toolbox	431

Appendix C

Description of Data Sets	433
--------------------------------	-----

Appendix D

Introduction to MATLAB

D.1 What Is MATLAB?	439
D.2 Getting Help in MATLAB	440
D.3 File and Workspace Management	440

D.4 Punctuation in MATLAB	443
D.5 Arithmetic Operators	444
D.6 Data Constructs in MATLAB	444
Basic Data Constructs	444
Building Arrays	445
Cell Arrays	445
Structures	447
D.7 Script Files and Functions	447
D.8 Control Flow	448
for Loop	448
while Loop	449
if-else Statements	449
switch Statement	449
D.9 Simple Plotting	450
D.10 Where to get MATLAB Information	452

Appendix E

MATLAB Functions

E.1 MATLAB	455
E.2 Statistics Toolbox	457
E.3 Exploratory Data Analysis Toolbox	458
E.4 EDA GUI Toolbox	459
References	475
Author Index	497
Subject Index	503