EAS Publication Series
Vol. 77 (2016)

Statistics for Astrophysics
Clustering and Classification

Didier Fraix-Burnet & Stéphane Girard (Editors)

Detailed Table of Contents

INTRODUCTION TO R

Didier Fraix-Burnet

1 Introduction . . . 3
1.1 Installation of R . . . 4
1.2 Graphical User Interface (GUI) or Integrated Development Environment . . . 4
1.3 Packages . . . 4
1.4 Files . . . 5
2 Basics . . . 5
2.1 Commands . . . 6
2.2 Objects . . . 6
2.3 Operators . . . 6
2.4 Functions . . . 7
3 Data types . . . 7
3.1 Variables . . . 7
3.2 Indices . . . 8
3.3 Objects classes . . . 9
4 Graphics . . . 9
4.1 Introduction . . . 9
4.2 An illustration with simple classification tools . . . 10
4.3 Other graphical examples . . . 11
5 Import/export . . . 12

ELEMENTS OF STATISTICS

Gérard Grégoire

1 Some probability . . . 13

1.1 Definitions . . . 13

1.2 Some basic distributions . . . 14

1.3 Moments of 1rst and 2nd order: mean, variance and covariance 16

1.4 Conditioning and Independency . . . 17

2 Some important univariate distributions . . . 18

2.1 The normal distribution . . . 18

2.2 The chi-squared distribution . . . 19

2.3 The Student distribution . . . 19

2.4 The Fisher distribution . . . 20

3 Random vectors . . . 21

3.1 Cumulative distribution function . . . 21

3.2 Density function . . . 22

3.3 Marginal and conditional distributions for a continuous random vector . . . 22

3.4 Distributions of a discrete random vector . . . 22

3.5 Mean vector and variance-covariance matrix . . . 23

4 Estimation and tests . . . 23

4.1 Estimation . . . 23

4.2 Confidence intervals . . . 25

4.3 Test of H_0 against H_1 . . . 25

5 Bayesian approach: basic elements for clustering . . . 26

5.1 The total probabilility rule . . . 26

5.2 The Bayes theorem . . . 26

5.3 The Bayes classifier . . . 27

6 Maximum likelihood method . . . 27

6.1 The likelihood function $L(\theta)$ . . . 27

6.2 Maximum likelihood estimate . . . 28

6.3 Score vector, information matrix, Fisher information, observed information . . . 28

6.4 Examples . . . 28

6.5 Asymptotic results . . . 29

7 Gaussian vectors . . . 30

7.1 Gaussian vectors: definition . . . 30

7.2 Some elementary properties . . . 31

7.3 MLE for parameters of Gaussian vectors . . . 31

7.4 Conditional distributions . . . 32

8 Mixture of two Gaussian distributions . . . 32

8.1 Mixture of Gaussian distributions . . . 32

8.2 The naïve Bayes classifier . . . 33

8.3 Classification and clustering . . . 34

A A table of usual distributions definitions . . . 35

Some basic elements in clustering and classification

Gérard Grégoire

1 A few basics in classification and clustering 39

1.1 Notation 39

1.2 Unsupervised Clustering 40

1.3 Classification or Supervised Clustering 41

1.4 Distances and dissimilarities 41

2 K-means clustering 44

2.1 Theory and algorithm 44

2.2 Step by step on a toy data set 46

2.3 Using *kmeans* `R` function. 48

3 Hierarchical clustering 48

4 The DBSCAN algorithm 51

4.1 Description of the algorithm 51

4.2 Example 52

5 Choosing the number of clusters 55

5.1 Criteria based on dispersion or variance decomposition formula 56

5.2 Criteria based on distances between clusters or compacity of the clusters 56

5.3 Choosing the number of clusters with `R` 58

6 Comparison of partitions: Rand index 59

6.1 Definition 59

6.2 Computing the Rand index with `R` 60

7 Classification methods (supervised clustering) 61

7.1 Training error, test error and cross-validation 61

7.2 The K-Nearest Neighbors method (KNN) 62

7.3 Logistic regression 63

7.4 Discriminating between stars and quasars with KNN and Logistic Regression 64

7.5 A comparison between KNN, Logistic Regression, LDA and QDA approaches 64

SUPERVISED AND UNSUPERVISED CLASSIFICATION USING MIXTURE MODELS

Stéphane Girard

1 Introduction 69
2 Supervised classification 70
2.1 Notations 70
2.2 Fisher Discriminant Analysis 71
2.3 Mixture model 73
2.4 Gaussian Mixture Model 77
3 Unsupervised classification 78
3.1 K-means algorithm 79
3.2 Maximum likelihood in the GMM 79
3.3 Selecting the number of clusters 84
4 Conclusion, recent developments 86
A Exercices 87
A.1 Gaussian Mixture Models 87
A.2 Sample drawn from a dataset 87
A.3 A simple EM algorithm 87
A.4 Use of the Rmixmod package 88

MODEL-BASED CLUSTERING OF HIGH-DIMENSIONAL DATA IN ASTROPHYSICS
Charles Bouveyron
1 Introduction 91
2 Curse and blessings of the dimensionality 93
2.1 Bellman’s curse of the dimensionality 93
2.2 The curse of dimensionality in model-based clustering 94
2.3 The blessing of dimensionality in clustering 94
3 Earliest approaches for high-dimensional clustering 95
3.1 Dimension reduction 96
3.2 Regularization 99
3.3 Parsimonious models 100
4 Subspace clustering methods 101
4.1 Mixture of high-dimensional Gaussian mixture models 101
4.2 The discriminative latent mixture models 104
5 Variable selection for model-based clustering 106
5.1 Variable selection as a model selection problem 107
5.2 Variable selection by penalization of the loadings 108
6 Conclusion 110
A Practical Session 110
A.1 The curse of dimensionality 110
A.2 PCA and LDA analyses 111
A.3 Gaussian Mixture Models 113
A.4 The High Dimensional Data Clustering algorithm 113
A.5 The FisherEM algorithm 115
A.6 SparseFEM 116

CLUSTERING OF VARIABLES FOR MIXED DATA

Jérôme Saracco

1 Introduction 121
2 The homogeneity criterion 123
2.1 Some notations 123
2.2 Definition of the synthetic variable of a cluster C_k 124
2.3 Homogeneity H of a cluster C_k 124
2.4 Homogeneity $\mathcal{H}$ of a partition $P_k = (C_1, \ldots, C_K)$ 125
3 The clustering algorithms 125
3.1 Hierarchical clustering of variables 126
3.2 k-means type clustering 127
4 A brief presentation of PCAmix method: a principal component analysis for mixed data 129
5 Illustration on a real data set: PCAmix and clustering of variables . . 131
5.1 Methodology 131
5.2 Description of the data 132
5.3 Clustering of individuals based on PCAmix principal components 133
5.4 Clustering of individuals based on ClustOfVar synthetic variables 139
5.5 Concluding remark 146
A The PCAmix algorithm. 146
B R practical work 148
B.1 Work 1: clustering on decathlon data 148
B.2 Work 2: Clustering of variables on Wine data (mixed data: quantitative variables and categorical variables) 162

Introduction to Kernel Methods: Classification of multivariate data

Mathieu Fauvel

1 Introductory example 171
1.1 Linear case 171
1.2 Non linear case 173
1.3 Conclusion 175
2 Kernel Function 175
2.1 Positive semi-definite kernels 175
2.2 Some kernels 176
2.3 Kernels on images 177
3 Introductory example continued: Kernel K-NN 177
4 Support Vectors Machines 178
4.1 Learn from data 178
4.2 Linear SVM 179
4.3 Non linear SVM 182
4.4 Fitting the hyperparameters 183
4.5 Multiclass SVM 183
5 SVM in practice 184
5.1 Simulated data 184
5.2 Load the data and extract few training samples for learning the model 184
5.3 Estimate the optimal hyperparameter of the model 186
5.4 Learn the model with optimal parameter and predict the whole validation samples 187
5.5 Speeding up the process ? 187
5.6 Results 188
A `R` codes 188

MODELLING STRUCTURED DATA WITH PROBABILISTIC GRAPHICAL MODELS

Florence Forbes

1 Introduction 195
2 Elements of probability theory 196
3 Directed graphs and Bayesian networks 197
4 Conditional independence and Markov properties 200
4.1 Reading conditional independence 201
4.2 d-separation 202
5 Undirected graphs and Markov Random Fields 203
5.1 Factorization 203
5.2 Trivial configurations 204
5.3 Separation and conditional independence 204
6 Inference and learning 206
6.1 Markov model based segmentation 207
6.2 Variational EM algorithm 208
A Practical work 210
A.1 Non spatial segmentation 210
A.2 Spatial segmentation. 210
B `R` commandes 212
B.1 Input of the `R` function 212
B.2 Output of the `R` function 212
B.3 `R` function 212
B.4 Example of use 215
C The SPACEM^3 software 217

CONCEPTS OF CLASSIFICATION AND TAXONOMY – PHYLOGENETIC CLASSIFICATION

Didier Fraix-Burnet

1 Why phylogenetic tools in astrophysics? 221
1.1 History of classification 221
1.2 Two routes for classification 222
1.3 Taxonomy 224
1.4 Stars, galaxies: multivariate objects in evolution 225
2 Distance-based approaches 225
2.1 Minimum Spanning Tree (MST) 225
2.2 Neighbor Joining Tree Estimation (NJ) 227
2.3 Difference between a hierarchical tree and a phylogenetic tree 227
3 Character-based approaches 229
3.1 Maximum Parsimony (cladistics) and Maximum Likelihood . 229
3.2 Cladistics: constructing a Tree 229
3.3 Parameters as Characters 233
3.4 Counting the steps: the cost of evolution 234
3.5 The most Parsimonious Tree 235
3.6 Robustness Assessment 236
3.7 Continuous Parameters 236
3.8 Interpretation 237
4 Clustering vs phylogenetic approaches 238
4.1 A simple example 238
5 Astrocladistics in Practice 243
5.1 Workflow 243
5.2 Softwares 243
5.3 An application of cladistics in astrophysics: the Globular Clusters of our Galaxy 245
5.4 To go further 248
6 Generalization: Networks 248
6.1 Split networks 248
7 Conclusion 249
A Exercices 251
A.1 Presentation of the session 251
A.2 Packages to install 251
A.3 MP analysis 251
A.4 Minimum Spanning Tree 254
A.5 Neighbor Joining Tree Estimation 255
A.6 Comparison of the three results 255
A.7 Interpretation of the tree 255