Alexander N. Gorban Balázs Kégl
Donald C. Wunsch Andrei Zinovyev
(Eds.)

Principal Manifolds for Data Visualization and Dimension Reduction

With 82 Figures and 22 Tables

Springer

Contents

1 Developments and Applications of Nonlinear Principal Component Analysis – a Review
Uwe Kruger, Junping Zhang, and Lei Xie ... 1
1.1 Introduction ... 1
1.2 PCA Preliminaries ... 3
1.3 Nonlinearity Test for PCA Models ... 6
1.3.1 Assumptions ... 7
1.3.2 Disjunct Regions ... 7
1.3.3 Confidence Limits for Correlation Matrix ... 8
1.3.4 Accuracy Bounds ... 10
1.3.5 Summary of the Nonlinearity Test ... 11
1.3.6 Example Studies ... 12
1.4 Nonlinear PCA Extensions ... 15
1.4.1 Principal Curves and Manifolds ... 16
1.4.2 Neural Network Approaches ... 24
1.4.3 Kernel PCA ... 29
1.5 Analysis of Existing Work ... 31
1.5.1 Computational Issues ... 31
1.5.2 Generalization of Linear PCA? ... 33
1.5.3 Roadmap for Future Developments (Basics and Beyond) ... 37
1.6 Concluding Summary ... 38
References ... 39

2 Nonlinear Principal Component Analysis: Neural Network Models and Applications
Matthias Scholz, Martin Fraunholz, and Joachim Selbig ... 44
2.1 Introduction ... 44
2.2 Standard Nonlinear PCA ... 47
2.3 Hierarchical Nonlinear PCA ... 48
2.3.1 The Hierarchical Error Function ... 49
2.4 Circular PCA ... 51

2.5 Inverse Model of Nonlinear PCA 52
2.5.1 The Inverse Network Model 53
2.5.2 NLPCA Models Applied to Circular Data 55
2.5.3 Inverse NLPCA for Missing Data 56
2.5.4 Missing Data Estimation 57
2.6 Applications 58
2.6.1 Application of Hierarchical NLPCA 59
2.6.2 Metabolite Data Analysis 60
2.6.3 Gene Expression Analysis 62
2.7 Summary 64
References 65

3 Learning Nonlinear Principal Manifolds by Self-Organising Maps
Hujun Yin 68
3.1 Introduction 68
3.2 Biological Background 69
3.2.1 Lateral Inhibition and Hebbian Learning 69
3.2.2 From Von Marsburg and Willshaw's Model to Kohonen's SOM 72
3.2.3 The SOM Algorithm 75
3.3 Theories 76
3.3.1 Convergence and Cost Functions 76
3.3.2 Topological Ordering Measures 79
3.4 SOMs, Multidimensional Scaling and Principal Manifolds 80
3.4.1 Multidimensional Scaling 80
3.4.2 Principal Manifolds 82
3.4.3 Visualisation Induced SOM (ViSOM) 84
3.5 Examples 86
3.5.1 Data Visualisation 87
3.5.2 Document Organisation and Content Management 88
References 91

4 Elastic Maps and Nets for Approximating Principal Manifolds and Their Application to Microarray Data Visualization
Alexander N. Gorban and Andrei Y. Zinovyev 96
4.1 Introduction and Overview 96
4.1.1 Fréchet Mean and Principal Objects: K-Means, PCA, what else? 96
4.1.2 Principal Manifolds 98
4.1.3 Elastic Functional and Elastic Nets 100
4.2 Optimization of Elastic Nets for Data Approximation 103
4.2.1 Basic Optimization Algorithm 103

4.2.2 Missing Data Values 105
4.2.3 Adaptive Strategies 106
4.3 Elastic Maps 109
4.3.1 Piecewise Linear Manifolds and Data Projectors 109
4.3.2 Iterative Data Approximation 109
4.4 Principal Manifold as Elastic Membrane 110
4.5 Method Implementation 112
4.6 Examples 112
4.6.1 Test Examples 112
4.6.2 Modeling Molecular Surfaces 113
4.6.3 Visualization of Microarray Data 114
4.7 Discussion 125
References 127

5 Topology-Preserving Mappings for Data Visualisation
Marian Peña, Wesam Barbakh, and Colin Fyfe 131
5.1 Introduction 131
5.2 Clustering Techniques 132
5.2.1 K-Means 132
5.2.2 K-Harmonic Means 133
5.2.3 Neural Gas 135
5.2.4 Weighted K-Means 136
5.2.5 The Inverse Weighted K-Means 137
5.3 Topology Preserving Mappings 138
5.3.1 Generative Topographic Map 138
5.3.2 Topographic Product of Experts ToPoE 140
5.3.3 The Harmonic Topograpic Map 141
5.3.4 Topographic Neural Gas 143
5.3.5 Inverse-Weighted K-Means Topology-Preserving Map 143
5.4 Experiments 144
5.4.1 Projections in Latent Space 144
5.4.2 Responsibilities 144
5.4.3 U-matrix, Hit Histograms and Distance Matrix 145
5.4.4 The Quality of The Map 147
5.5 Conclusions 149
References 149

6 The Iterative Extraction Approach to Clustering
Boris Mirkin 151
6.1 Introduction 151
6.2 Clustering Entity-to-feature Data 152
6.2.1 Principal Component Analysis 152
6.2.2 Additive Clustering Model and ITEX 154
6.2.3 Overlapping and Fuzzy Clustering Case 156
6.2.4 K-Means and iK-Means Clustering 157

6.3 ITEX Structuring and Clustering for Similarity Data 162
6.3.1 Similarity Clustering: a Review 162
6.3.2 The Additive Structuring Model and ITEX 163
6.3.3 Additive Clustering Model.............................. 165
6.3.4 Approximate Partitioning 166
6.3.5 One Cluster Clustering.................................. 168
6.3.6 Some Applications....................................... 171
References 174

7 Representing Complex Data Using Localized Principal Components with Application to Astronomical Data
Jochen Einbeck, Ludger Evers, and Coryn Bailer-Jones 178
7.1 Introduction 178
7.2 Localized Principal Component Analysis 181
7.2.1 Cluster-wise PCA 181
7.2.2 Principal Curves 185
7.2.3 Further Approaches 188
7.3 Combining Principal Curves and Regression 189
7.3.1 Principal Component Regression and its Shortcomings 189
7.3.2 The Generalization to Principal Curves 190
7.3.3 Using Directions Other than the Local Principal Components 192
7.3.4 A Simple Example 193
7.4 Application to the Gaia Survey Mission 194
7.4.1 The Astrophysical Data.................................. 194
7.4.2 Principal Manifold Based Approach 196
7.5 Conclusion 198
References 199

8 Auto-Associative Models, Nonlinear Principal Component Analysis, Manifolds and Projection Pursuit
Stéphane Girard and Serge Iovleff.................................. 202
8.1 Introduction 202
8.2 Auto-Associative Models 203
8.2.1 Approximation by Manifolds.............................. 203
8.2.2 A Projection Pursuit Algorithm.......................... 205
8.2.3 Theoretical Results..................................... 206
8.3 Examples.. 207
8.3.1 Linear Auto-Associative Models and PCA 207
8.3.2 Additive Auto-Associative Models and Neural Networks 208
8.4 Implementation Aspects.. 209
8.4.1 Estimation of the Regression Functions 209
8.4.2 Computation of Principal Directions..................... 211
8.5 Illustration on Real and Simulated Data....................... 213
References 216

9 Beyond The Concept of Manifolds: Principal Trees, Metro Maps, and Elastic Cubic Complexes
Alexander N. Gorban, Neil R. Sumner, and Andrei Y. Zinovyev 219
9.1 Introduction and Overview 219
9.1.1 Elastic Principal Graphs 221
9.2 Optimization of Elastic Graphs for Data Approximation 222
9.2.1 Elastic Functional Optimization 222
9.2.2 Optimal Application of Graph Grammars 223
9.2.3 Factorization and Transformation of Factors 224
9.3 Principal Trees (Branching Principal Curves) 225
9.3.1 Simple Graph Grammar ("Add a Node", "Bisect an Edge") 225
9.3.2 Visualization of Data Using "Metro Map" Two-Dimensional Tree Layout 226
9.3.3 Example of Principal Cubic Complex: Product of Principal Trees 227
9.4 Analysis of the Universal 7-Cluster Structure of Bacterial Genomes 229
9.4.1 Brief Introduction 230
9.4.2 Visualization of the 7-Cluster Structure 232
9.5 Visualization of Microarray Data 232
9.5.1 Dataset Used 232
9.5.2 Principal Tree of Human Tissues 234
9.6 Discussion 235
References 235

10 Diffusion Maps - a Probabilistic Interpretation for Spectral Embedding and Clustering Algorithms
Boaz Nadler, Stephane Lafon, Ronald Coifman, and Ioannis G. Kevrekidis 238
10.1 Introduction 238
10.2 Diffusion Distances and Diffusion Maps 240
10.2.1 Asymptotics of the Diffusion Map 245
10.3 Spectral Embedding of Low Dimensional Manifolds 246
10.4 Spectral Clustering of a Mixture of Gaussians 251
10.5 Summary and Discussion 258
References 258

11 On Bounds for Diffusion, Discrepancy and Fill Distance Metrics
Steven B. Damelin 261
11.1 Introduction 261
11.2 Energy, Discrepancy, Distance and Integration on Measurable Sets in Euclidean Space 262

11.3 Set Learning via Normalized Laplacian Dimension Reduction and Diffusion Distance 266
11.4 Main Result: Bounds for Discrepancy, Diffusion and Fill Distance Metrics 268
References 269

12 Geometric Optimization Methods for the Analysis of Gene Expression Data
Michel Journée, Andrew E. Teschendorff, Pierre-Antoine Absil, Simon Tavaré, and Rodolphe Sepulchre 271
12.1 Introduction 271
12.2 ICA as a Geometric Optimization Problem 272
12.3 Contrast Functions 274
12.3.1 Mutual Information 274
12.3.2 $\mathcal{F}$-Correlation 276
12.3.3 Non-Gaussianity 277
12.3.4 Joint Diagonalization of Cumulant Matrices 278
12.4 Matrix Manifolds for ICA 279
12.5 Optimization Algorithms 280
12.5.1 Line-Search Algorithms 280
12.5.2 FastICA 282
12.5.3 Jacobi Rotations 284
12.6 Analysis of Gene Expression Data by ICA 284
12.6.1 Some Issues About the Application of ICA 284
12.6.2 Evaluation of the Biological Relevance of the Expression Modes 287
12.6.3 Results Obtained on the Breast Cancer Microarray Data Set 288
12.7 Conclusion 290
References 290

13 Dimensionality Reduction and Microarray Data
David A. Elizondo, Benjamin N. Passow, Ralph Birkenhead, and Andreas Huemer 293
13.1 Introduction 293
13.2 Background 295
13.2.1 Microarray Data 295
13.2.2 Methods for Dimension Reduction 296
13.2.3 Linear Separability 297
13.3 Comparison Procedure 300
13.3.1 Data Sets 300
13.3.2 Dimensionality Reduction 301
13.3.3 Perceptron Models 303
13.4 Results 303
13.5 Conclusions 306
References 307

14 PCA and K-Means Decipher Genome
Alexander N. Gorban and Andrei Y. Zinovyev 309
14.1 Introduction 309
14.2 Required Materials 310
14.3 Genomic Sequence 311
14.3.1 Background 311
14.3.2 Sequences for the Analysis 312
14.4 Converting Text to a Numerical Table 312
14.5 Data Visualization 313
14.5.1 Visualization 313
14.5.2 Understanding Plots 314
14.6 Clustering and Visualizing Results 315
14.7 Task List and Further Information 317
14.8 Conclusion 318
References 318

Color Plates 325

Index 333