# Contents

*Preface* *xvii*

1 *Introduction* *1*
*1.1 Linear representation of multivariate data* *1*
*1.1.1 The general statistical setting* *1*
*1.1.2 Dimension reduction methods* *2*
*1.1.3 Independence as a guiding principle* *3*
*1.2 Blind source separation* *3*
*1.2.1 Observing mixtures of unknown signals* *4*
*1.2.2 Source separation based on independence* *5*
*1.3 Independent component analysis* *6*
*1.3.1 Definition* *6*
*1.3.2 Applications* *7*
*1.3.3 How to find the independent components* *7*
*1.4 History of ICA* *11*

*Part I MATHEMATICAL PRELIMINARIES*

*2 Random Vectors and Independence* 15
*2.1 Probability distributions and densities* 15
*2.1.1 Distribution of a random variable* 15
*2.1.2 Distribution of a random vector* 17
*2.1.3 Joint and marginal distributions* 18
*2.2 Expectations and moments* 19
*2.2.1 Definition and general properties* 19
*2.2.2 Mean vector and correlation matrix* 20
*2.2.3 Covariances and joint moments* 22
*2.2.4 Estimation of expectations* 24
*2.3 Uncorrelatedness and independence* 24
*2.3.1 Uncorrelatedness and whiteness* 24
*2.3.2 Statistical independence* 27
*2.4 Conditional densities and Bayes' rule* 28
*2.5 The multivariate gaussian density* 31
*2.5.1 Properties of the gaussian density* 32
*2.5.2 Central limit theorem* 34
*2.6 Density of a transformation* 35
*2.7 Higher-order statistics* 36
*2.7.1 Kurtosis and classification of densities* 37
*2.7.2 Cumulants, moments, and their properties* 40
*2.8 Stochastic processes ** 43
*2.8.1 Introduction and definition* 43
*2.8.2 Stationarity, mean, and autocorrelation* 45
*2.8.3 Wide-sense stationary processes* 46
*2.8.4 Time averages and ergodicity* 48
*2.8.5 Power spectrum* 49
*2.8.6 Stochastic signal models* 50
*2.9 Concluding remarks and references* 51
*Problems* 52

*3 Gradients and Optimization Methods* 57
*3.1 Vector and matrix gradients* 57
*3.1.1 Vector gradient* 57
*3.1.2 Matrix gradient* 59
*3.1.3 Examples of gradients* 59

3.1.4 Taylor series expansions 62
3.2 Learning rules for unconstrained optimization 63
3.2.1 Gradient descent 63
3.2.2 Second-order learning 65
3.2.3 The natural gradient and relative gradient 67
3.2.4 Stochastic gradient descent 68
3.2.5 Convergence of stochastic on-line algorithms * 71
3.3 Learning rules for constrained optimization 73
3.3.1 The Lagrange method 73
3.3.2 Projection methods 73
3.4 Concluding remarks and references 75
Problems 75

4 Estimation Theory 77
4.1 Basic concepts 78
4.2 Properties of estimators 80
4.3 Method of moments 84
4.4 Least-squares estimation 86
4.4.1 Linear least-squares method 86
4.4.2 Nonlinear and generalized least squares * 88
4.5 Maximum likelihood method 90
4.6 Bayesian estimation * 94
4.6.1 Minimum mean-square error estimator 94
4.6.2 Wiener filtering 96
4.6.3 Maximum a posteriori (MAP) estimator 97
4.7 Concluding remarks and references 99
Problems 101

5 Information Theory 105
5.1 Entropy 105
5.1.1 Definition of entropy 105
5.1.2 Entropy and coding length 107
5.1.3 Differential entropy 108
5.1.4 Entropy of a transformation 109
5.2 Mutual information 110
5.2.1 Definition using entropy 110
5.2.2 Definition using Kullback-Leibler divergence 110

5.3 Maximum entropy 111
5.3.1 Maximum entropy distributions 111
5.3.2 Maximality property of gaussian distribution 112
5.4 Negentropy 112
5.5 Approximation of entropy by cumulants 113
5.5.1 Polynomial density expansions 113
5.5.2 Using expansions for entropy approximation 114
5.6 Approximation of entropy by nonpolynomial functions 115
5.6.1 Approximating the maximum entropy 116
5.6.2 Choosing the nonpolynomial functions 117
5.6.3 Simple special cases 118
5.6.4 Illustration 119
5.7 Concluding remarks and references 120
Problems 121
Appendix proofs 122

6 Principal Component Analysis and Whitening 125
6.1 Principal components 125
6.1.1 PCA by variance maximization 127
6.1.2 PCA by minimum MSE compression 128
6.1.3 Choosing the number of principal components 129
6.1.4 Closed-form computation of PCA 131
6.2 PCA by on-line learning 132
6.2.1 The stochastic gradient ascent algorithm 133
6.2.2 The subspace learning algorithm 134
6.2.3 The PAST algorithm * 135
6.2.4 PCA and back-propagation learning * 136
6.2.5 Extensions of PCA to nonquadratic criteria * 137
6.3 Factor analysis 138
6.4 Whitening 140
6.5 Orthogonalization 141
6.6 Concluding remarks and references 143
Problems 144

*Part II BASIC INDEPENDENT COMPONENT ANALYSIS*

*7 What is Independent Component Analysis?* 147
*7.1 Motivation* 147
*7.2 Definition of independent component analysis* 151
*7.2.1 ICA as estimation of a generative model* 151
*7.2.2 Restrictions in ICA* 152
*7.2.3 Ambiguities of ICA* 154
*7.2.4 Centering the variables* 154
*7.3 Illustration of ICA* 155
*7.4 ICA is stronger that whitening* 158
*7.4.1 Uncorrelatedness and whitening* 158
*7.4.2 Whitening is only half ICA* 160
*7.5 Why gaussian variables are forbidden* 161
*7.6 Concluding remarks and references* 163
*Problems* 164

*8 ICA by Maximization of Nongaussianity* 165
*8.1 "Nongaussian is independent"* 166
*8.2 Measuring nongaussianity by kurtosis* 171
*8.2.1 Extrema give independent components* 171
*8.2.2 Gradient algorithm using kurtosis* 175
*8.2.3 A fast fixed-point algorithm using kurtosis* 178
*8.2.4 Examples* 179
*8.3 Measuring nongaussianity by negentropy* 182
*8.3.1 Critique of kurtosis* 182
*8.3.2 Negentropy as nongaussianity measure* 182
*8.3.3 Approximating negentropy* 183
*8.3.4 Gradient algorithm using negentropy* 185
*8.3.5 A fast fixed-point algorithm using negentropy* 188
*8.4 Estimating several independent components* 192
*8.4.1 Constraint of uncorrelatedness* 192
*8.4.2 Deflationary orthogonalization* 194
*8.4.3 Symmetric orthogonalization* 194
*8.5 ICA and projection pursuit* 197
*8.5.1 Searching for interesting directions* 197
*8.5.2 Nongaussian is interesting* 197
*8.6 Concluding remarks and references* 198

*Problems* 199
*Appendix proofs* 201

*9 ICA by Maximum Likelihood Estimation* 203
*9.1 The likelihood of the ICA model* 203
*9.1.1 Deriving the likelihood* 203
*9.1.2 Estimation of the densities* 204
*9.2 Algorithms for maximum likelihood estimation* 207
*9.2.1 Gradient algorithms* 207
*9.2.2 A fast fixed-point algorithm* 209
*9.3 The infomax principle* 211
*9.4 Examples* 213
*9.5 Concluding remarks and references* 214
*Problems* 218
*Appendix proofs* 219

*10 ICA by Minimization of Mutual Information* 221
*10.1 Defining ICA by mutual information* 221
*10.1.1 Information-theoretic concepts* 221
*10.1.2 Mutual information as measure of dependence* 222
*10.2 Mutual information and nongaussianity* 223
*10.3 Mutual information and likelihood* 224
*10.4 Algorithms for minimization of mutual information* 224
*10.5 Examples* 225
*10.6 Concluding remarks and references* 225
*Problems* 227

*11 ICA by Tensorial Methods* 229
*11.1 Definition of cumulant tensor* 229
*11.2 Tensor eigenvalues give independent components* 230
*11.3 Tensor decomposition by a power method* 232
*11.4 Joint approximate diagonalization of eigenmatrices* 234
*11.5 Weighted correlation matrix approach* 235
*11.5.1 The FOBI algorithm* 235
*11.5.2 From FOBI to JADE* 235
*11.6 Concluding remarks and references* 236
*Problems* 237

*12 ICA by Nonlinear Decorrelation and Nonlinear PCA* 239
*12.1 Nonlinear correlations and independence* 240
*12.2 The Hérault-Jutten algorithm* 242
*12.3 The Cichocki-Unbehauen algorithm* 243
*12.4 The estimating functions approach ** 245
*12.5 Equivariant adaptive separation via independence* 247
*12.6 Nonlinear principal components* 249
*12.7 The nonlinear PCA criterion and ICA* 251
*12.8 Learning rules for the nonlinear PCA criterion* 254
*12.8.1 The nonlinear subspace rule* 254
*12.8.2 Convergence of the nonlinear subspace rule ** 255
*12.8.3 Nonlinear recursive least-squares rule* 258
*12.9 Concluding remarks and references* 261
*Problems* 262

*13 Practical Considerations* 263
*13.1 Preprocessing by time filtering* 263
*13.1.1 Why time filtering is possible* 264
*13.1.2 Low-pass filtering* 265
*13.1.3 High-pass filtering and innovations* 265
*13.1.4 Optimal filtering* 266
*13.2 Preprocessing by PCA* 267
*13.2.1 Making the mixing matrix square* 267
*13.2.2 Reducing noise and preventing overlearning* 268
*13.3 How many components should be estimated?* 269
*13.4 Choice of algorithm* 271
*13.5 Concluding remarks and references* 272
*Problems* 272

*14 Overview and Comparison of Basic ICA Methods* 273
*14.1 Objective functions vs. algorithms* 273
*14.2 Connections between ICA estimation principles* 274
*14.2.1 Similarities between estimation principles* 274
*14.2.2 Differences between estimation principles* 275
*14.3 Statistically optimal nonlinearities* 276
*14.3.1 Comparison of asymptotic variance ** 276
*14.3.2 Comparison of robustness ** 277
*14.3.3 Practical choice of nonlinearity* 279

14.4 Experimental comparison of ICA algorithms 280
14.4.1 Experimental set-up and algorithms 281
14.4.2 Results for simulated data 282
14.4.3 Comparisons with real-world data 286
14.5 References 287
14.6 Summary of basic ICA 287
Appendix Proofs 289

Part III EXTENSIONS AND RELATED METHODS

15 Noisy ICA 293
15.1 Definition 293
15.2 Sensor noise vs. source noise 294
15.3 Few noise sources 295
15.4 Estimation of the mixing matrix 295
15.4.1 Bias removal techniques 296
15.4.2 Higher-order cumulant methods 298
15.4.3 Maximum likelihood methods 299
15.5 Estimation of the noise-free independent components 299
15.5.1 Maximum a posteriori estimation 299
15.5.2 Special case of shrinkage estimation 300
15.6 Denoising by sparse code shrinkage 303
15.7 Concluding remarks 304

16 ICA with Overcomplete Bases 305
16.1 Estimation of the independent components 306
16.1.1 Maximum likelihood estimation 306
16.1.2 The case of supergaussian components 307
16.2 Estimation of the mixing matrix 307
16.2.1 Maximizing joint likelihood 307
16.2.2 Maximizing likelihood approximations 308
16.2.3 Approximate estimation by quasiorthogonality 309
16.2.4 Other approaches 311
16.3 Concluding remarks 313

*17 Nonlinear ICA* 315
*17.1 Nonlinear ICA and BSS* 315
*17.1.1 The nonlinear ICA and BSS problems* 315
*17.1.2 Existence and uniqueness of nonlinear ICA* 317
*17.2 Separation of post-nonlinear mixtures* 319
*17.3 Nonlinear BSS using self-organizing maps* 320
*17.4 A generative topographic mapping approach ** 322
*17.4.1 Background* 322
*17.4.2 The modified GTM method* 323
*17.4.3 An experiment* 326
*17.5 An ensemble learning approach to nonlinear BSS* 328
*17.5.1 Ensemble learning* 328
*17.5.2 Model structure* 329
*17.5.3 Computing Kullback-Leibler cost function ** 330
*17.5.4 Learning procedure ** 332
*17.5.5 Experimental results* 333
*17.6 Other approaches* 337
*17.7 Concluding remarks* 339

*18 Methods using Time Structure* 341
*18.1 Separation by autocovariances* 342
*18.1.1 An alternative to nongaussianity* 342
*18.1.2 Using one time lag* 343
*18.1.3 Extension to several time lags* 344
*18.2 Separation by nonstationarity of variances* 346
*18.2.1 Using local autocorrelations* 347
*18.2.2 Using cross-cumulants* 349
*18.3 Separation principles unified* 351
*18.3.1 Comparison of separation principles* 351
*18.3.2 Kolmogoroff complexity as unifying framework* 352
*18.4 Concluding remarks* 354

19 Convolutive Mixtures and Blind Deconvolution 355
19.1 Blind deconvolution 356
19.1.1 Problem definition 356
19.1.2 Bussgang methods 357
19.1.3 Cumulant-based methods 358
19.1.4 Blind deconvolution using linear ICA 360
19.2 Blind separation of convolutive mixtures 361
19.2.1 The convolutive BSS problem 361
19.2.2 Reformulation as ordinary ICA 363
19.2.3 Natural gradient methods 364
19.2.4 Fourier transform methods 365
19.2.5 Spatiotemporal decorrelation methods 367
19.2.6 Other methods for convolutive mixtures 367
19.3 Concluding remarks 368
Appendix Discrete-time filters and the z-transform 369

20 Other Extensions 371
20.1 Priors on the mixing matrix 371
20.1.1 Motivation for prior information 371
20.1.2 Classic priors 372
20.1.3 Sparse priors 374
20.1.4 Spatiotemporal ICA 377
20.2 Relaxing the independence assumption 378
20.2.1 Multidimensional ICA 379
20.2.2 Independent subspace analysis 380
20.2.3 Topographic ICA 382
20.3 Complex-valued data 383
20.3.1 Basic concepts of complex random variables 383
20.3.2 Indeterminacy of the independent components 384
20.3.3 Choice of the nongaussianity measure 385
20.3.4 Consistency of estimator 386
20.3.5 Fixed-point algorithm 386
20.3.6 Relation to independent subspaces 387
20.4 Concluding remarks 387

*Part IV APPLICATIONS OF ICA*

*21 Feature Extraction by ICA* 391
*21.1 Linear representations* 392
*21.1.1 Definition* 392
*21.1.2 Gabor analysis* 392
*21.1.3 Wavelets* 394
*21.2 ICA and Sparse Coding* 396
*21.3 Estimating ICA bases from images* 398
*21.4 Image denoising by sparse code shrinkage* 398
*21.4.1 Component statistics* 399
*21.4.2 Remarks on windowing* 400
*21.4.3 Denoising results* 401
*21.5 Independent subspaces and topographic ICA* 401
*21.6 Neurophysiological connections* 403
*21.7 Concluding remarks* 405

*22 Brain Imaging Applications* 407
*22.1 Electro- and magnetoencephalography* 407
*22.1.1 Classes of brain imaging techniques* 407
*22.1.2 Measuring electric activity in the brain* 408
*22.1.3 Validity of the basic ICA model* 409
*22.2 Artifact identification from EEG and MEG* 410
*22.3 Analysis of evoked magnetic fields* 411
*22.4 ICA applied on other measurement techniques* 413
*22.5 Concluding remarks* 414

*23 Telecommunications* 417
*23.1 Multiuser detection and CDMA communications* 417
*23.2 CDMA signal model and ICA* 422
*23.3 Estimating fading channels* 424
*23.3.1 Minimization of complexity* 424
*23.3.2 Channel estimation ** 426
*23.3.3 Comparisons and discussion* 428
*23.4 Blind separation of convolved CDMA mixtures ** 430
*23.4.1 Feedback architecture* 430
*23.4.2 Semiblind separation method* 431
*23.4.3 Simulations and discussion* 432

*23.5 Improving multiuser detection using complex ICA ** 434
*23.5.1 Data model* 435
*23.5.2 ICA based receivers* 436
*23.5.3 Simulation results* 438
*23.6 Concluding remarks and references* 439

*24 Other Applications* 441
*24.1 Financial applications* 441
*24.1.1 Finding hidden factors in financial data* 441
*24.1.2 Time series prediction by ICA* 443
*24.2 Audio separation* 446
*24.3 Further applications* 448

*References* 449

*Index* 476