Contents

1 Introduction
   1.1 Rise of Big Data and Dimensionality
      1.1.1 Biological Sciences
      1.1.2 Health Sciences
      1.1.3 Computer and Information Sciences
      1.1.4 Economics and Finance
      1.1.5 Business and Program Evaluation
      1.1.6 Earth Sciences and Astronomy
   1.2 Impact of Big Data
   1.3 Impact of Dimensionality
      1.3.1 Computation
      1.3.2 Noise Accumulation
      1.3.3 Spurious Correlation
      1.3.4 Statistical theory
   1.4 Aim of High-dimensional Statistical Learning
   1.5 What big data can do
   1.6 Scope of the book

2 Multiple and Nonparametric Regression
   2.1 Introduction
   2.2 Multiple Linear Regression
      2.2.1 The Gauss-Markov Theorem
      2.2.2 Statistical Tests
   2.3 Weighted Least-Squares
   2.4 Box-Cox Transformation
   2.5 Model Building and Basis expansions
      2.5.1 Polynomial Regression
      2.5.2 Spline Regression
      2.5.3 Multiple Covariates
   2.6 Ridge Regression
      2.6.1 Bias-Variance Tradeoff
      2.6.2 $\ell_2$ Penalized Least Squares
      2.6.3 Bayesian Interpretation
      2.6.4 Ridge Regression Solution Path
      2.6.5 Kernel Ridge Regression
3 Introduction to Penalized Least Squares

3.1 Classical Variable Selection Criteria
   3.1.1 Subset selection
   3.1.2 Relation with penalized regression
   3.1.3 Selection of regularization parameters

3.2 Folded-concave Penalized Least Squares
   3.2.1 Orthonormal designs
   3.2.2 Penalty functions
   3.2.3 Thresholding by SCAD and MCP
   3.2.4 Risk properties
   3.2.5 Characterization of folded-concave PLS

3.3 Lasso and $L_1$ Regularization
   3.3.1 Nonnegative garrote
   3.3.2 Lasso
   3.3.3 Adaptive Lasso
   3.3.4 Elastic Net
   3.3.5 Dantzig selector
   3.3.6 SLOPE and Sorted Penalties
   3.3.7 Concentration inequalities and uniform convergence
   3.3.8 A brief history of model selection

3.4 Bayesian Variable Selection
   3.4.1 Bayesian view of the PLS
   3.4.2 A Bayesian framework for selection

3.5 Numerical Algorithms
   3.5.1 Quadratic programs
   3.5.2 Least angle regression
   3.5.3 Local quadratic approximations
   3.5.4 Local linear algorithm
   3.5.5 Penalized linear unbiased selection
   3.5.6 Cyclic coordinate descent algorithms
   3.5.7 Iterative shrinkage-thresholding algorithms
   3.5.8 Projected proximal gradient method
   3.5.9 ADMM
   3.5.10 Iterative Local Adaptive Majorization and Minimization
   3.5.11 Other Methods and Timeline

3.6 Regularization parameters for PLS
   3.6.1 Degrees of freedom
   3.6.2 Extension of information criteria
   3.6.3 Application to PLS estimators

3.7 Residual variance and refitted cross-validation
CONTENTS

3.7.1 Residual variance of Lasso 103
3.7.2 Refitted cross-validation 104
3.8 Extensions to Nonparametric Modeling 106
3.8.1 Structured nonparametric models 106
3.8.2 Group penalty 107
3.9 Applications 109
3.10 Bibliographical notes 113
3.11 Exercises 115

4 Penalized Least Squares: Properties 119
4.1 Performance Benchmarks 119
4.1.1 Performance measures 120
4.1.2 Impact of model uncertainty 123
4.1.2.1 Bayes lower bounds for orthogonal design 124
4.1.2.2 Minimax lower bounds for general design 128
4.1.3 Performance goals, sparsity and sub-Gaussian noise 134
4.2 Penalized \( L_0 \) Selection 137
4.3 Lasso and Dantzig Selector 143
4.3.1 Selection consistency 144
4.3.2 Prediction and coefficient estimation errors 148
4.3.3 Model size and least squares after selection 159
4.3.4 Properties of the Dantzig selector 165
4.3.5 Regularity conditions on the design matrix 173
4.4 Properties of Concave PLS. 181
4.4.1 Properties of penalty functions 183
4.4.2 Local and oracle solutions 188
4.4.3 Properties of local solutions 193
4.4.4 Global and approximate global solutions 198
4.5 Smaller and Sorted Penalties 204
4.5.1 Sorted concave penalties and its local approximation 205
4.5.2 Approximate PLS with smaller and sorted penalties 209
4.5.3 Properties of LLA and LCA 218
4.6 Bibliographical notes 222
4.7 Exercises 223

5 Generalized Linear Models and Penalized Likelihood 225
5.1 Generalized Linear Models 225
5.1.1 Exponential family 225
5.1.2 Elements of generalized linear models 228
5.1.3 Maximum likelihood 229
5.1.4 Computing MLE: Iteratively reweighed least squares 230
5.1.5 Deviance and Analysis of Deviance 232
5.1.6 Residuals 234
5.2 Examples 236
5.2.1 Bernoulli and binomial models 236
5.2.2 Models for count responses
5.2.3 Models for nonnegative continuous responses
5.2.4 Normal error models
5.3 Variable Selection via Penalized Likelihood
5.4 Algorithms
5.4.1 Local quadratic approximation
5.4.2 Local linear approximation
5.4.3 Coordinate descent
5.4.4 Iterative Local Adaptive Majorization and Minimization
5.5 Tuning parameter selection
5.6 An Application
5.7 Sampling Properties in low-dimension
5.7.1 Notation and regularity conditions
5.7.2 The oracle property
5.7.3 Sampling Properties with Diverging Dimensions
5.7.4 Asymptotic properties of GIC selectors
5.8 Properties under Ultrahigh Dimensions
5.8.1 The Lasso penalized estimator and its risk property
5.8.2 Strong oracle property
5.8.3 Numeric studies
5.9 Risk properties
5.10 Bibliographical notes
5.11 Exercises

6 Penalized M-estimators
6.1 Penalized quantile regression
6.1.1 Quantile regression
6.1.2 Variable selection in quantile regression
6.1.3 A fast algorithm for penalized quantile regression
6.2 Penalized composite quantile regression
6.3 Variable selection in robust regression
6.3.1 Robust regression
6.3.2 Variable selection in Huber regression
6.4 Rank regression and its variable selection
6.4.1 Rank regression
6.4.2 Penalized weighted rank regression
6.5 Variable Selection for Survival Data
6.5.1 Variable selection in proportional hazard models
6.6 Theory of penalized M-estimator
6.6.1 Conditions on penalty and restricted strong convexity
6.6.2 Statistical accuracy
6.6.3 Computational accuracy
6.7 Bibliographical notes
6.8 Exercises
CONTENTS

8 Feature Screening 313
  8.1 Correlation Screening 313
    8.1.1 Sure screening property 314
    8.1.2 Connection to multiple comparison 315
    8.1.3 Iterative SIS 316
  8.2 Generalized and Rank Correlation Screening 317
  8.3 Parametric Feature Screening 320
    8.3.1 Generalized linear models 320
    8.3.2 A unified strategy for parametric feature screening 322
  8.4 Nonparametric Screening 324
    8.4.1 Additive models 324
    8.4.2 Varying coefficient models 326
    8.4.3 Heterogeneous nonparametric models 328
  8.5 Model-free Feature Screening 329
    8.5.1 Sure independent ranking screening procedure 329
    8.5.2 Feature screening via distance correlation 332
    8.5.3 Feature screening for high-dimensional categorial data 335
  8.6 Screening and Selection 337
    8.6.1 Feature screening via forward regression 338
    8.6.2 Sparse maximum likelihood estimate 338
    8.6.3 Feature screening via partial correlation 340
  8.7 Refitted Cross-Validation 344
    8.7.1 RCV algorithm 345
    8.7.2 RCV in linear models 345
    8.7.3 RCV in nonparametric regression 349
  8.8 An Illustration 350
  8.9 Bibliographical notes 353
  8.10 Exercises 355

9 Covariance Regularization and Graphical Models 359
  9.1 Basic facts about matrix 359
  9.2 Sparse Covariance Matrix Estimation 363
    9.2.1 Covariance regularization by thresholding and banding 363
    9.2.2 Asymptotic properties 366
    9.2.3 Nearest positive definite matrices 369
  9.3 Robust covariance inputs 371
  9.4 Sparse Precision Matrix and Graphical Models 374
    9.4.1 Gaussian graphical models 374
    9.4.2 Penalized likelihood and M-estimation 375
    9.4.3 Penalized least-squares 376
    9.4.4 CLIME and its adaptive version 379
  9.5 A Latent Gaussian Graphical Model 384
  9.6 Technical Proofs 387
    9.6.1 Proof of Theorem 9.1 387
    9.6.2 Proof of Theorem 9.3 389
CONTENTS

11.3.2 Augmented Principal Component Regression 458
11.3.3 Application to Forecast Bond Risk Premia 459
11.4 Applications to Statistical Machine Learning 460
  11.4.1 Community detection 461
  11.4.2 Matrix completion 465
  11.4.3 Item ranking 467
  11.4.4 Gaussian Mixture models 470
11.5 Exercises 473
11.6 Bibliographical Notes 476

12 Supervised Learning 479
  12.1 Model-based Classifiers 479
    12.1.1 Linear and quadratic discriminant analysis 479
    12.1.2 Logistic regression 483
  12.2 Kernel Density Classifiers and Naive Bayes 485
  12.3 Nearest Neighbor Classifiers 489
  12.4 Classification Trees and Ensemble Classifiers 490
    12.4.1 Classification trees 490
    12.4.2 Bagging 493
    12.4.3 Random forests 494
    12.4.4 Boosting 496
  12.5 Support Vector Machines 500
    12.5.1 The standard support vector machine 500
    12.5.2 Generalizations of SVMs 503
  12.6 Sparse classifiers via penalized empirical loss 505
    12.6.1 The importance of sparsity under high-dimensionality 506
    12.6.2 Sparse support vector machines 507
    12.6.3 Sparse large margin classifiers 509
  12.7 Sparse Discriminant Analysis 511
    12.7.1 Nearest shrunken centroids classifier 512
    12.7.2 Features annealed independent rule 514
    12.7.3 Selection bias of sparse independence rules 515
    12.7.4 Regularized optimal affine discriminant 516
    12.7.5 Linear programming discriminant 518
    12.7.6 Direct sparse discriminant analysis 519
    12.7.7 Solution path equivalence between ROAD and DSDA 520
  12.8 Sparse Additive Classifiers 521
    12.8.1 Penalized additive logistic regression 521
    12.8.2 Feature augmentation 522
    12.8.3 Semiparametric sparse discriminant analysis 523
  12.9 Bibliographical notes 526
  12.10 Exercises 527
13 Unsupervised Learning 531
13.1 Cluster Analysis 531
  13.1.1 K-means clustering 532
  13.1.2 Hierarchical clustering 533
  13.1.3 Model-based clustering 535
  13.1.4 Spectral clustering 539
13.2 Data-driven choices of the number of clusters 541
13.3 Variable Selection in Clustering 544
  13.3.1 Sparse K-means clustering 544
  13.3.2 Sparse model-based clustering 546
  13.3.3 Sparse Mixture of Experts Model 547
13.4 An introduction of Sparse PCA 550
  13.4.1 Inconsistency of the regular PCA 551
  13.4.2 Consistency under sparse eigenvector model 552
13.5 Sparse Principal Component Analysis 554
  13.5.1 Sparse PCA 554
  13.5.2 An iterative SVD thresholding approach 558
  13.5.3 A penalized matrix decomposition approach 560
  13.5.4 A semidefinite programming approach 561
  13.5.5 A generalized power method 562
13.6 Bibliographical notes 564
13.7 Exercises 565

14 Introduction to Deep Learning 569
  14.1 Convolution Network 569

Index 601