Statistical Foundations of Data Science

Jianqing Fan

Runze Li

Cun-Hui Zhang

Hui Zou

Preface

Big data are ubiquitous. They come in varying volume, velocity, and variety. They have a deep impact on systems such as storages, communications and computing architectures and analysis such as statistics, computation, optimization, and privacy. Engulfed by a multitude of applications, data science aims to address the large-scale challenges of data analysis, turning big data into smart data for decision making and knowledge discoveries. Data science integrates theories and methods from statistics, optimization, mathematical science, computer science, and information science to extract knowledge, make decisions, discover new insights, and reveal new phenomena from data. The concept of data science has appeared in the literature for several decades and has been interpreted differently by different researchers. It has nowadays become a multi-disciplinary field that distills knowledge in various disciplines to develop new methods, processes, algorithms and systems for knowledge discovery from various kinds of data, which can be either low or high dimensional, and either structured, unstructured or semi-structured. Statistical modeling plays critical roles in the analysis of complex and heterogeneous data and quantifies uncertainties of scientific hypotheses and statistical results.

This book introduces commonly-used statistical models, contemporary statistical machine learning techniques and algorithms, along with their mathematical insights and statistical theories. It aims to serve as a graduate-level textbook on the statistical foundations of data science as well as a research monograph on sparsity, covariance learning, machine learning and statistical inference. For a one-semester graduate level course, it may cover Chapters 2, 3, 9, 10, 12, 13 and some topics selected from the remaining chapters. This gives a comprehensive view on statistical machine learning models, theories and methods. Alternatively, one-semester graduate course may cover Chapters 2, 3, 5, 7, 8 and selected topics from the remaining chapters. This track focuses more on high-dimensional statistics, model selection and inferences but both paths emphasize a great deal on sparsity and variable selections.

Frontiers of scientific research rely on the collection and processing of massive complex data. Information and technology allow us to collect big data of unprecedented size and complexity. Accompanying big data is the rise of dimensionality and high dimensionality characterizes many contemporary statistical problems, from sciences and engineering to social science and humanities. Many traditional statistical procedures for finite or low-dimensional data are still useful in data science, but they become infeasible or ineffective for dealing with high-dimensional data. Hence, new statistical methods are indispensable. The authors have worked on high-dimensional statistics for two decades, and started to write the book on the topics of high-dimensional data analysis over a decade ago. Over the last decide, there have been surges in interest and exciting developments in high-dimensional and big data. This led us to concentrate mainly on statistical aspects of data science.

We aim to introduce commonly-used statistical models, methods and procedures in data science and provide readers with sufficient and sound theoretical justifications. It has been a challenge for us to balance statistical theories and methods and to choose the topics and works to cover since the amount of publications in this emerging area is enormous. Thus, we focus on the foundational aspects that are related to sparsity, covariance learning, machine learning, and statistical inference.

Sparsity is a common assumption in the analysis of high-dimensional data. By sparsity, we mean that only a handful of features embedded in a huge pool suffice for certain scientific questions or predictions. This book introduces various regularization methods to deal with sparsity, including how to determine penalties and how to choose tuning parameters in regularization methods and numerical optimization algorithms for various statistical models. They can be found in Chapters 3–6 and 8.

High-dimensional measurements are frequently dependent, since these variables often measure similar things, such as aspects of economics or personal health. Many of these variables have heavy tails due to big number of collected variables. To model the dependence, factor models are frequently employed, which exhibit low-rank plus sparse structures in data matrices and can be solved by robust principal component analysis from high-dimensional covariance. Robust covariance learning, principal component analysis, as well as their applications to community detection, topic modeling, recommender systems, ect. are also a feature of this book. They can be found in Chapters 9–11. Note that factor learning or more generally latent structure learning can also be regarded as unsupervised statistical machine learning.

Machine learning is critical in analyzing high-dimensional and complex data. This book also provides readers with a comprehensive account on statistical machine learning methods and algorithms in data science. We introduce statistical procedures for supervised learning in which the response variable (often categorical) is available and the goal is to predict the response based on input variables. This book also provides readers with statistical procedures for unsupervised learning, in which the responsible variable is missing and the goal concentrates on learning the association and patterns among a set of input variables. Feature creations and sparsity learning also arise in these problems. See Chapters 2, 12–14 for details.

Statistical inferences on high-dimensional data are another focus of this book. Statistical inferences require one to characterize the uncertainty, estimate the standard errors of the estimated parameters of primary interest and derive the asymptotic distributions of the resulting estimates. This is very challenging under the high-dimensional regime. See Chapter 7.

Fueled by the surging demands on processing high-dimensional and big data, there have been rapid and vast developments in high-dimensional statistics and machine learning over the last decade, contributed by data scientists from various fields such as statistics, computer science, information theory, applied and computational mathematics, among others. Even though we have narrowed the scope of the book to the statistical aspects of data science, the field is still too broad for us to cover. Many important contributions that do not fit our presentation have been omitted. Conscientious effort was made in the composition of the reference list and bibliographical notes, but they merely reflect our immediate interests. Omissions and discrepancies are inevitable. We apologize for their occurrence.

Although we all contribute to various chapters and share the responsibility for the whole book, Jianqing Fan was the lead author for Chapters 1, 3 and 9–11, 14 and some sections in other chapters, Runze Li for Chapters 5, and 8 and part of Chapters 6–7, Cun-Hui Zhang for Chapters 4 and 7, and Hui Zou for Chapters 2, 6, 11 and 12 and part of Chapter 5. In addition, Jianqing Fan and Runze Li oversaw the whole book project.

Many people have contributed importantly to the completion of this book. In particular, we would like to thank the editor, John Kimmel, who has been extremely helpful and patient with us for over 10 years! We greatly appreciate a set of around 10 anonymous reviewers for valuable comments that lead to the improvement of the book. We are particularly grateful to Cong Ma and Yiqiao Zhong for preparing a draft of Chapter 14, to Zhao Chen for helping us with putting our unsorted and non-uniform references into the present form, to Tracy Ke, Bryan Kelly, Dacheng Xiu and Jia Wang for helping us with constructing Figure 1.3, and to Boxiang Wang, Yi Yang for helping produce some figures in Chapter 12. Various people have carefully proof-read certain chapters of the book and made useful suggestions. They include Krishna Balasubramanian, Pierre Bayle, Elynn Chen, Wenyan Gong, Yongyi Guo, Cong Ma, Igor Silin, Qiang Sun, Francesca Tang, Bingyan Wang, Kaizheng Wang, Weichen Wang, Yuling Yan, Zhuoran Yang, Mengxin Yu, Wenxin Zhou, Yifeng Zhou, and Ziwei Zhu. We owe them many thanks.

In the spring semester of 2019, we used a draft of this book as a textbook for a first-year graduate course at Princeton University and a senior graduate topic course at the Pennsylvania State University. We would like to thank the graduate students in the classes for their careful readings. In particular, we are indebted to Cong Ma, Kaizheng Wang and Zongjun Tan for assisting in preparing the homework problems at Princeton, most of which are now a part of our exercise at the end of each chapter. At Princeton, we covered chapters 2-3, 5, 8.1, 8.3, 9–14.

We are very grateful for grant supports from National Science Foundation and National Institutes of Health on our research. Finally, we would like to thank our families and our parents for their love and support.

Jianqing Fan
Runze Li
Cun-Hui Zhang
Hui Zou

January 2020.

Contents

1 Introduction **3**

1.1 Rise of Big Data and Dimensionality 3

1.1.1 Biological Sciences 4

1.1.2 Health Sciences 6

1.1.3 Computer and Information Sciences 7

1.1.4 Economics and Finance 9

1.1.5 Business and Program Evaluation 11

1.1.6 Earth Sciences and Astronomy 11

1.2 Impact of Big Data 11

1.3 Impact of Dimensionality 13

1.3.1 Computation 13

1.3.2 Noise Accumulation 14

1.3.3 Spurious Correlation 16

1.3.4 Statistical theory 19

1.4 Aim of High-dimensional Statistical Learning 20

1.5 What big data can do 21

1.6 Scope of the book 21

2 Multiple and Nonparametric Regression **23**

2.1 Introduction 23

2.2 Multiple Linear Regression 23

2.2.1 The Gauss-Markov Theorem 25

2.2.2 Statistical Tests 28

2.3 Weighted Least-Squares 29

2.4 Box-Cox Transformation 31

2.5 Model Building and Basis Expansions 32

2.5.1 Polynomial Regression 33

2.5.2 Spline Regression 34

2.5.3 Multiple Covariates 37

2.6 Ridge Regression 38

2.6.1 Bias-Variance Tradeoff 39

2.6.2 ℓ_2 Penalized Least Squares 39

2.6.3 Bayesian Interpretation 40

2.6.4 Ridge Regression Solution Path 41

2.6.5 Kernel Ridge Regression 42

2.7 Regression in Reproducing Kernel Hilbert Space 44
2.8 Leave-one-out and Generalized Cross-validation 49
2.9 Exercises 51

3 Introduction to Penalized Least-Squares 57
3.1 Classical Variable Selection Criteria 57
3.1.1 Subset selection 57
3.1.2 Relation with penalized regression 58
3.1.3 Selection of regularization parameters 59
3.2 Folded-concave Penalized Least Squares 61
3.2.1 Orthonormal designs 63
3.2.2 Penalty functions 64
3.2.3 Thresholding by SCAD and MCP 65
3.2.4 Risk properties 66
3.2.5 Characterization of folded-concave PLS 67
3.3 Lasso and L_1 Regularization 68
3.3.1 Nonnegative garrote 68
3.3.2 Lasso 70
3.3.3 Adaptive Lasso 73
3.3.4 Elastic Net 74
3.3.5 Dantzig selector 76
3.3.6 SLOPE and Sorted Penalties 79
3.3.7 Concentration inequalities and uniform convergence 80
3.3.8 A brief history of model selection 82
3.4 Bayesian Variable Selection 83
3.4.1 Bayesian view of the PLS 83
3.4.2 A Bayesian framework for selection 85
3.5 Numerical Algorithms 86
3.5.1 Quadratic programs 86
3.5.2 Least angle regression* 88
3.5.3 Local quadratic approximations 91
3.5.4 Local linear algorithm 92
3.5.5 Penalized linear unbiased selection* 93
3.5.6 Cyclic coordinate descent algorithms 95
3.5.7 Iterative shrinkage-thresholding algorithms 96
3.5.8 Projected proximal gradient method 98
3.5.9 ADMM 98
3.5.10 Iterative Local Adaptive Majorization and Minimization 99
3.5.11 Other Methods and Timeline 100
3.6 Regularization parameters for PLS 101
3.6.1 Degrees of freedom 102
3.6.2 Extension of information criteria 103
3.6.3 Application to PLS estimators 104
3.7 Residual variance and refitted cross-validation 105

3.7.1 Residual variance of Lasso 105
3.7.2 Refitted cross-validation 106
3.8 Extensions to Nonparametric Modeling 108
3.8.1 Structured nonparametric models 108
3.8.2 Group penalty 109
3.9 Applications 111
3.10 Bibliographical notes 116
3.11 Exercises 117

4 Penalized Least Squares: Properties 125
4.1 Performance Benchmarks 125
4.1.1 Performance measures 126
4.1.2 Impact of model uncertainty 129
4.1.2.1 Bayes lower bounds for orthogonal design 130
4.1.2.2 Minimax lower bounds for general design 134
4.1.3 Performance goals, sparsity and sub-Gaussian noise 140
4.2 Penalized L_0 Selection 143
4.3 Lasso and Dantzig Selector 149
4.3.1 Selection consistency 150
4.3.2 Prediction and coefficient estimation errors 154
4.3.3 Model size and least squares after selection 165
4.3.4 Properties of the Dantzig selector 171
4.3.5 Regularity conditions on the design matrix 179
4.4 Properties of Concave PLS. 187
4.4.1 Properties of penalty functions 189
4.4.2 Local and oracle solutions 194
4.4.3 Properties of local solutions 199
4.4.4 Global and approximate global solutions 204
4.5 Smaller and Sorted Penalties 210
4.5.1 Sorted concave penalties and its local approximation 211
4.5.2 Approximate PLS with smaller and sorted penalties 215
4.5.3 Properties of LLA and LCA 224
4.6 Bibliographical notes 228
4.7 Exercises 229

5 Generalized Linear Models and Penalized Likelihood 231
5.1 Generalized Linear Models 231
5.1.1 Exponential family 231
5.1.2 Elements of generalized linear models 234
5.1.3 Maximum likelihood 235
5.1.4 Computing MLE: Iteratively reweighed least squares 236
5.1.5 Deviance and Analysis of Deviance 238
5.1.6 Residuals 240
5.2 Examples 242
5.2.1 Bernoulli and binomial models 242

5.2.2 Models for count responses 245
5.2.3 Models for nonnegative continuous responses 246
5.2.4 Normal error models 247
5.3 Sparest solution in high confidence set 247
5.3.1 A general setup 247
5.3.2 Examples 248
5.3.3 Properties 249
5.4 Variable Selection via Penalized Likelihood 250
5.5 Algorithms 253
5.5.1 Local quadratic approximation 253
5.5.2 Local linear approximation 254
5.5.3 Coordinate descent 255
5.5.4 Iterative Local Adaptive Majorization and Minimization 256
5.6 Tuning parameter selection 256
5.7 An Application 258
5.8 Sampling Properties in low-dimension 260
5.8.1 Notation and regularity conditions 261
5.8.2 The oracle property 262
5.8.3 Sampling Properties with Diverging Dimensions 264
5.8.4 Asymptotic properties of GIC selectors 266
5.9 Properties under Ultrahigh Dimensions 268
5.9.1 The Lasso penalized estimator and its risk property 268
5.9.2 Strong oracle property 272
5.9.3 Numeric studies 277
5.10 Risk properties 278
5.11 Bibliographical notes 282
5.12 Exercises 283

6 Penalized M-estimators 291
6.1 Penalized quantile regression 291
6.1.1 Quantile regression 291
6.1.2 Variable selection in quantile regression 293
6.1.3 A fast algorithm for penalized quantile regression 295
6.2 Penalized composite quantile regression 298
6.3 Variable selection in robust regression 301
6.3.1 Robust regression 301
6.3.2 Variable selection in Huber regression 303
6.4 Rank regression and its variable selection 305
6.4.1 Rank regression 306
6.4.2 Penalized weighted rank regression 306
6.5 Variable Selection for Survival Data 307
6.5.1 Partial likelihood 308
6.5.2 Variable selection via penalized partial likelihood and its properties 310

6.6 Theory of folded-concave penalized M-estimator 312
6.6.1 Conditions on penalty and restricted strong convexity 313
6.6.2 Statistical accuracy of penalized M-estimator with folded concave penalties 314
6.6.3 Computational accuracy 318
6.7 Bibliographical notes 321
6.8 Exercises 323

7 High Dimensional Inference 327
7.1 Inference in linear regression 328
7.1.1 Debias of regularized regression estimators 329
7.1.2 Choices of weights 331
7.1.3 Inference for the noise level 333
7.2 Inference in generalized linear models 336
7.2.1 Desparsified Lasso 337
7.2.2 Decorrelated score estimator 338
7.2.3 Test of linear hypotheses 341
7.2.4 Numerical comparison 343
7.2.5 An application 344
7.3 Asymptotic efficiency 345
7.3.1 Statistical efficiency and Fisher information 345
7.3.2 Linear regression with random design 351
7.3.3 Partial linear regression 357
7.4 Gaussian graphical models 361
7.4.1 Inference via penalized least squares 361
7.4.2 Sample size in regression and graphical models 367
7.5 General solutions 373
7.5.1 Local semi-LD decomposition 374
7.5.2 Data swap 375
7.5.3 Gradient approximation 380
7.6 Bibliographical notes 382
7.7 Exercises 383

8 Feature Screening 387
8.1 Correlation Screening 387
8.1.1 Sure screening property 388
8.1.2 Connection to multiple comparison 390
8.1.3 Iterative SIS 391
8.2 Generalized and Rank Correlation Screening 392
8.3 Feature Screening for Parametric Models 395
8.3.1 Generalized linear models 395
8.3.2 A unified strategy for parametric feature screening 397
8.3.3 Conditional sure independence screening 400
8.4 Nonparametric Screening 401
8.4.1 Additive models 401

8.4.2 Varying coefficient models 402
8.4.3 Heterogeneous nonparametric models 406
8.5 Model-free Feature Screening 407
8.5.1 Sure independent ranking screening procedure 407
8.5.2 Feature screening via distance correlation 409
8.5.3 Feature screening for high-dimensional categorial data 412
8.6 Screening and Selection 415
8.6.1 Feature screening via forward regression 415
8.6.2 Sparse maximum likelihood estimate 416
8.6.3 Feature screening via partial correlation 418
8.7 Refitted Cross-Validation 423
8.7.1 RCV algorithm 423
8.7.2 RCV in linear models 424
8.7.3 RCV in nonparametric regression 426
8.8 An Illustration 428
8.9 Bibliographical notes 432
8.10 Exercises 434

9 Covariance Regularization and Graphical Models 437
9.1 Basic facts about matrix 437
9.2 Sparse Covariance Matrix Estimation 441
9.2.1 Covariance regularization by thresholding and banding 441
9.2.2 Asymptotic properties 444
9.2.3 Nearest positive definite matrices 447
9.3 Robust covariance inputs 449
9.4 Sparse Precision Matrix and Graphical Models 452
9.4.1 Gaussian graphical models 452
9.4.2 Penalized likelihood and M-estimation 453
9.4.3 Penalized least-squares 454
9.4.4 CLIME and its adaptive version 457
9.5 Latent Gaussian Graphical Models 462
9.6 Technical Proofs 465
9.6.1 Proof of Theorem 9.1 465
9.6.2 Proof of Theorem 9.3 467
9.6.3 Proof of Theorem 9.4 468
9.6.4 Proof of Theorem 9.6 468
9.7 Bibliographical notes 470
9.8 Exercises 472

10 Covariance Learning and Factor Models 477
10.1 Principal Component Analysis 477
10.1.1 Introduction to PCA 477
10.1.2 Power Method 479
10.2 Factor Models and Structured Covariance Learning 480
10.2.1 Factor model and high-dimensional PCA 481

10.2.2 Extracting latent factors and POET 484
10.2.3 Methods for selecting number of factors 486
10.3 Covariance and Precision Learning with Known Factors 489
10.3.1 Factor model with observable factors 489
10.3.2 Robust initial estimation of covariance matrix 491
10.4 Augmented factor models and projected PCA 494
10.5 Asymptotic Properties 497
10.5.1 Properties for estimating loading matrix 497
10.5.2 Properties for estimating covariance matrices 499
10.5.3 Properties for estimating realized latent factors 499
10.5.4 Properties for estimating idiosyncratic components 501
10.6 Technical Proofs 501
10.6.1 Proof of Theorem 10.1 501
10.6.2 Proof of Theorem 10.2 506
10.6.3 Proof of Theorem 10.3 507
10.6.4 Proof of Theorem 10.4 510
10.7 Bibliographical Notes 512
10.8 Exercises 513

11 Applications of Factor Models and PCA 519
11.1 Factor-adjusted Regularized Model Selection 519
11.1.1 Importance of factor adjustments 519
11.1.2 FarmSelect 521
11.1.3 Application to forecasting bond risk premia 522
11.1.4 Application to a neuroblastoma data 524
11.1.5 Asymptotic theory for FarmSelect 526
11.2 Factor-adjusted robust multiple testing 526
11.2.1 False discovery rate control 527
11.2.2 Multiple testing under dependence measurements 529
11.2.3 Power of factor adjustments 530
11.2.4 FarmTest 532
11.2.5 Application to neuroblastoma data 534
11.3 Factor Augmented Regression Methods 536
11.3.1 Principal Component Regression 536
11.3.2 Augmented Principal Component Regression 538
11.3.3 Application to Forecast Bond Risk Premia 539
11.4 Applications to Statistical Machine Learning 540
11.4.1 Community detection 541
11.4.2 Topic model 547
11.4.3 Matrix completion 548
11.4.4 Item ranking 550
11.4.5 Gaussian Mixture models 553
11.5 Bibliographical Notes 556
11.6 Exercises 557

12 Supervised Learning 563
12.1 Model-based Classifiers 563
12.1.1 Linear and quadratic discriminant analysis 563
12.1.2 Logistic regression 567
12.2 Kernel Density Classifiers and Naive Bayes 569
12.3 Nearest Neighbor Classifiers 573
12.4 Classification Trees and Ensemble Classifiers 574
12.4.1 Classification trees 574
12.4.2 Bagging 577
12.4.3 Random forests 578
12.4.4 Boosting 580
12.5 Support Vector Machines 584
12.5.1 The standard support vector machine 584
12.5.2 Generalizations of SVMs 587
12.6 Sparse Classifiers via Penalized Empirical Loss 590
12.6.1 The importance of sparsity under high-dimensionality 590
12.6.2 Sparse support vector machines 592
12.6.3 Sparse large margin classifiers 593
12.7 Sparse Discriminant Analysis 595
12.7.1 Nearest shrunken centroids classifier 597
12.7.2 Features annealed independent rule 598
12.7.3 Selection bias of sparse independence rules 600
12.7.4 Regularized optimal affine discriminant 600
12.7.5 Linear programming discriminant 602
12.7.6 Direct sparse discriminant analysis 603
12.7.7 Solution path equivalence between ROAD and DSDA 605
12.8 Feature Augmention and Sparse Additive Classifiers 606
12.8.1 Feature augmentation 606
12.8.2 Penalized additive logistic regression 607
12.8.3 Semiparametric sparse discriminant analysis 609
12.9 Bibliographical notes 611
12.10Exercises 611

13 Unsupervised Learning 619
13.1 Cluster Analysis 619
13.1.1 K-means clustering 620
13.1.2 Hierarchical clustering 621
13.1.3 Model-based clustering 623
13.1.4 Spectral clustering 627
13.2 Data-driven choices of the number of clusters 629
13.3 Variable Selection in Clustering 632
13.3.1 Sparse K-means clustering 632
13.3.2 Sparse model-based clustering 634
13.3.3 Sparse Mixture of Experts Model 636
13.4 An introduction of Sparse PCA 639

13.4.1 Inconsistency of the regular PCA 639
13.4.2 Consistency under sparse eigenvector model 640
13.5 Sparse Principal Component Analysis 642
13.5.1 Sparse PCA 642
13.5.2 An iterative SVD thresholding approach 647
13.5.3 A penalized matrix decomposition approach 648
13.5.4 A semidefinite programming approach 649
13.5.5 A generalized power method 650
13.6 Bibliographical notes 652
13.7 Exercises 653

14 An Introduction to Deep Learning 657
14.1 Rise of Deep Learning 657
14.2 Feed-forward neural networks 660
14.2.1 Model setup 660
14.2.2 Back-propagation in computational graphs 662
14.3 Popular models 664
14.3.1 Convolutional neural networks 664
14.3.2 Recurrent neural networks 668
14.3.2.1 Vanilla RNNs 668
14.3.2.2 GRUs and LSTM 669
14.3.2.3 Multilayer RNNs 670
14.3.3 Modules 671
14.4 Deep unsupervised learning 672
14.4.1 Autoencoders 673
14.4.2 Generative adversarial networks 675
14.4.2.1 Sampling view of GANs 676
14.4.2.2 Minimum distance view of GANs 677
14.5 Training deep neural nets 678
14.5.1 Stochastic gradient descent 679
14.5.1.1 Mini-batch SGD 680
14.5.1.2 Momentum-based SGD 681
14.5.1.3 SGD with adaptive learning rates 681
14.5.2 Easing numerical instability 682
14.5.2.1 ReLU activation function 682
14.5.2.2 Skip connections 683
14.5.2.3 Batch normalization 683
14.5.3 Regularization techniques 684
14.5.3.1 Weight decay 684
14.5.3.2 Dropout 684
14.5.3.3 Data augmentation 685
14.6 Example: image classification 685
14.7 Bibliography notes 686

References 691

Author Index **739**

Index **751**