Springer Series in Statistics

Statistics for High-Dimensional Data

Methods, Theory and Applications

Springer

Springer Series in Statistics

Advisors:
P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

For other titles published in this series, go to
http://www.springer.com/series/692

Peter Bühlmann • Sara van de Geer

Statistics for High-Dimensional Data

Methods, Theory and Applications

Springer

Peter Bühlmann
Seminar for Statistics
ETH Zürich
CH-8092 Zürich
Switzerland
buhlmann@stat.math.ethz.ch

Sara van de Geer
Seminar for Statistics
ETH Zürich
CH-8092 Zürich
Switzerland
geer@stat.math.ethz.ch

ISSN 0172-7397
ISBN 978-3-642-20191-2 e-ISBN 978-3-642-20192-9
DOI 10.1007/978-3-642-20192-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930793

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To

Anthony and Sigi

and

Tabea, Anna, Sophia,
Simon and Lukas

Preface

High-dimensional data are nowadays rule rather than exception in areas like information technology, bioinformatics or astronomy, to name just a few. The word "high-dimensional" refers to the situation where the number of unknown parameters which are to be estimated is one or several orders of magnitude larger than the number of samples in the data. Classical statistical inference cannot be used for high-dimensional problems. For example, least-squares fitting of a linear model having many more unknown parameters than observations and assigning corresponding standard errors and measures of significance is ill-posed. It is rather obvious that without additional assumptions, or say restricting to a certain class of models, high-dimensional statistical inference is impossible. A well-established framework for fitting many parameters is based on assuming structural smoothness, enabling estimation of smooth functions. The last years have witnessed a revolution of methodological, computational and mathematical advances which allow for high-dimensional statistical inference based on assuming certain notions of sparsity. Shifting the focus from smoothness to sparsity constraints, or combining the two, opens the path for many more applications involving complex data. For example, the sparsity assumption that the health status of a person is depending only on a few among several thousands of biomarkers appears much more realistic than considering a model where all the thousands of variables would contribute in a smooth way to the state of health.

This book brings together methodological concepts, computational algorithms, a few applications and mathematical theory for high-dimensional statistics. The mathematical underpinning of methodology and computing has implications on exploring exciting possibilities and understanding fundamental limitations. In this sense, the combination of methodology and theory builds the foundation of the book. We present the methods and their potential for data analysis with a view on the underlying mathematical assumptions and properties and vice-versa, the theoretical derivations are motivated by applicability and implications to real data problems. The mathematical results yield additional insights and allow to categorize different methods and algorithms in terms of what they can achieve and what not. The book

is not meant as an overview of the state-of-the-art, but rather as a selective treatment with emphasis on our own work.

It is possible to read the book with more emphasis on methods and applications or on theory; but of course, one can also focus on all aspects with equal intensity. As such, we hope that the book will be useful and appealing to statisticians, data analysts and other researchers who appreciate the possibilities to learn about methods and algorithms, mathematical theory and the combination of both of them.

This book emerged from a very nice collaboration between the authors. We acknowledge many people who have contributed in various ways to its completion. Wolfgang Härdle proposed to write a book on high-dimensional statistics while hiking in the black forest at Oberwolfach, and we are thankful for it. Alain Hauser, Mohamed Hebiri, Markus Kalisch, Johannes Lederer, Lukas Meier, Nicolai Meinshausen, Patric Müller, Jürg Schelldorfer and Nicolas Städler have contributed with many original ideas and concepts as collaborators of joint research projects or making some thoughtful suggestions for the book. Finally, we would like to express our gratitude to our families for providing a different, interesting, supportive and beautiful environment.

Zürich, December 2010 *Peter Bühlmann and Sara van de Geer*

Contents

1 Introduction 1
1.1 The framework 1
1.2 The possibilities and challenges 2
1.3 About the book 3
1.3.1 Organization of the book 3
1.4 Some examples 4
1.4.1 Prediction and biomarker discovery in genomics 5

2 Lasso for linear models 7
2.1 Organization of the chapter 7
2.2 Introduction and preliminaries 8
2.2.1 The Lasso estimator 9
2.3 Orthonormal design 10
2.4 Prediction 11
2.4.1 Practical aspects about the Lasso for prediction 12
2.4.2 Some results from asymptotic theory 13
2.5 Variable screening and $\|\hat{\beta} - \beta^0\|_q$-norms 14
2.5.1 Tuning parameter selection for variable screening 17
2.5.2 Motif regression for DNA binding sites 18
2.6 Variable selection 19
2.6.1 Neighborhood stability and irrepresentable condition 22
2.7 Key properties and corresponding assumptions: a summary 23
2.8 The adaptive Lasso: a two-stage procedure 25
2.8.1 An illustration: simulated data and motif regression 25
2.8.2 Orthonormal design 27
2.8.3 The adaptive Lasso: variable selection under weak conditions 28
2.8.4 Computation 29
2.8.5 Multi-step adaptive Lasso 30
2.8.6 Non-convex penalty functions 32
2.9 Thresholding the Lasso 33
2.10 The relaxed Lasso 34

2.11 Degrees of freedom of the Lasso 34
2.12 Path-following algorithms 36
2.12.1 Coordinatewise optimization and shooting algorithms 38
2.13 Elastic net: an extension 41
Problems 42

3 Generalized linear models and the Lasso 45
3.1 Organization of the chapter 45
3.2 Introduction and preliminaries 45
3.2.1 The Lasso estimator: penalizing the negative log-likelihood . 46
3.3 Important examples of generalized linear models 47
3.3.1 Binary response variable and logistic regression 47
3.3.2 Poisson regression 49
3.3.3 Multi-category response variable and multinomial distribution 50
Problems 53

4 The group Lasso 55
4.1 Organization of the chapter 55
4.2 Introduction and preliminaries 56
4.2.1 The group Lasso penalty 56
4.3 Factor variables as covariates 58
4.3.1 Prediction of splice sites in DNA sequences 59
4.4 Properties of the group Lasso for generalized linear models 61
4.5 The generalized group Lasso penalty 64
4.5.1 Groupwise prediction penalty and parametrization invariance 65
4.6 The adaptive group Lasso 66
4.7 Algorithms for the group Lasso 67
4.7.1 Block coordinate descent 68
4.7.2 Block coordinate gradient descent 72
Problems 75

5 Additive models and many smooth univariate functions 77
5.1 Organization of the chapter 77
5.2 Introduction and preliminaries 78
5.2.1 Penalized maximum likelihood for additive models 78
5.3 The sparsity-smoothness penalty 79
5.3.1 Orthogonal basis and diagonal smoothing matrices 80
5.3.2 Natural cubic splines and Sobolev spaces 81
5.3.3 Computation 82
5.4 A sparsity-smoothness penalty of group Lasso type 85
5.4.1 Computational algorithm 86
5.4.2 Alternative approaches 88
5.5 Numerical examples 89
5.5.1 Simulated example 89

5.5.2 Motif regression . . . 90
5.6 Prediction and variable selection . . . 91
5.7 Generalized additive models . . . 92
5.8 Linear model with varying coefficients . . . 93
5.8.1 Properties for prediction . . . 95
5.8.2 Multivariate linear model . . . 95
5.9 Multitask learning . . . 95
Problems . . . 97

6 Theory for the Lasso . . . 99
6.1 Organization of this chapter . . . 99
6.2 Least squares and the Lasso . . . 101
6.2.1 Introduction . . . 101
6.2.2 The result assuming the truth is linear . . . 102
6.2.3 Linear approximation of the truth . . . 108
6.2.4 A further refinement: handling smallish coefficients . . . 112
6.3 The setup for general convex loss . . . 114
6.4 The margin condition . . . 119
6.5 Generalized linear model without penalty . . . 122
6.6 Consistency of the Lasso for general loss . . . 126
6.7 An oracle inequality . . . 128
6.8 The ℓ_q-error for $1 \leq q \leq 2$ . . . 135
6.8.1 Application to least squares assuming the truth is linear . . . 136
6.8.2 Application to general loss and a sparse approximation of the truth . . . 137
6.9 The weighted Lasso . . . 139
6.10 The adaptively weighted Lasso . . . 141
6.11 Concave penalties . . . 144
6.11.1 Sparsity oracle inequalities for least squares with ℓ_r-penalty 146
6.11.2 Proofs for this section (Section 6.11) . . . 147
6.12 Compatibility and (random) matrices . . . 150
6.13 On the compatibility condition . . . 156
6.13.1 Direct bounds for the compatibility constant . . . 158
6.13.2 Bounds using $\|\beta_S\|_1^2 \leq s\|\beta_S\|_2^2$ . . . 161
6.13.3 Sets $\mathcal{N}$ containing S . . . 167
6.13.4 Restricted isometry . . . 169
6.13.5 Sparse eigenvalues . . . 170
6.13.6 Further coherence notions . . . 172
6.13.7 An overview of the various eigenvalue flavored constants . . . 174
Problems . . . 178

7 Variable selection with the Lasso . . . 183
7.1 Introduction . . . 183
7.2 Some results from literature . . . 184
7.3 Organization of this chapter . . . 185

7.4 The beta-min condition . . . 187
7.5 The irrepresentable condition in the noiseless case . . . 189
7.5.1 Definition of the irrepresentable condition . . . 190
7.5.2 The KKT conditions . . . 190
7.5.3 Necessity and sufficiency for variable selection . . . 191
7.5.4 The irrepresentable condition implies the compatibility condition . . . 195
7.5.5 The irrepresentable condition and restricted regression . . . 197
7.5.6 Selecting a superset of the true active set . . . 199
7.5.7 The weighted irrepresentable condition . . . 200
7.5.8 The weighted irrepresentable condition and restricted regression . . . 201
7.5.9 The weighted Lasso with "ideal" weights . . . 203
7.6 Definition of the adaptive and thresholded Lasso . . . 204
7.6.1 Definition of adaptive Lasso . . . 204
7.6.2 Definition of the thresholded Lasso . . . 205
7.6.3 Order symbols . . . 206
7.7 A recollection of the results obtained in Chapter 6 . . . 206
7.8 The adaptive Lasso and thresholding: invoking sparse eigenvalues . . . 210
7.8.1 The conditions on the tuning parameters . . . 210
7.8.2 The results . . . 211
7.8.3 Comparison with the Lasso . . . 213
7.8.4 Comparison between adaptive and thresholded Lasso . . . 214
7.8.5 Bounds for the number of false negatives . . . 215
7.8.6 Imposing beta-min conditions . . . 216
7.9 The adaptive Lasso without invoking sparse eigenvalues . . . 218
7.9.1 The condition on the tuning parameter . . . 219
7.9.2 The results . . . 219
7.10 Some concluding remarks . . . 221
7.11 Technical complements for the noiseless case without sparse eigenvalues . . . 222
7.11.1 Prediction error for the noiseless (weighted) Lasso . . . 222
7.11.2 The number of false positives of the noiseless (weighted) Lasso . . . 224
7.11.3 Thresholding the noiseless initial estimator . . . 225
7.11.4 The noiseless adaptive Lasso . . . 227
7.12 Technical complements for the noisy case without sparse eigenvalues . . . 232
7.13 Selection with concave penalties . . . 237
Problems . . . 241

8 Theory for ℓ_1/ℓ_2-penalty procedures . . . 249
8.1 Introduction . . . 249
8.2 Organization and notation of this chapter . . . 250
8.3 Regression with group structure . . . 252
8.3.1 The loss function and penalty . . . 253

8.3.2 The empirical process . . . 254
8.3.3 The group Lasso compatibility condition . . . 255
8.3.4 A group Lasso sparsity oracle inequality . . . 256
8.3.5 Extensions . . . 258
8.4 High-dimensional additive model . . . 258
8.4.1 The loss function and penalty . . . 258
8.4.2 The empirical process . . . 260
8.4.3 The smoothed Lasso compatibility condition . . . 264
8.4.4 A smoothed group Lasso sparsity oracle inequality . . . 265
8.4.5 On the choice of the penalty . . . 270
8.5 Linear model with time-varying coefficients . . . 275
8.5.1 The loss function and penalty . . . 275
8.5.2 The empirical process . . . 277
8.5.3 The compatibility condition for the time-varying coefficients model . . . 278
8.5.4 A sparsity oracle inequality for the time-varying coefficients model . . . 279
8.6 Multivariate linear model and multitask learning . . . 281
8.6.1 The loss function and penalty . . . 281
8.6.2 The empirical process . . . 282
8.6.3 The multitask compatibility condition . . . 283
8.6.4 A multitask sparsity oracle inequality . . . 284
8.7 The approximation condition for the smoothed group Lasso . . . 286
8.7.1 Sobolev smoothness . . . 286
8.7.2 Diagonalized smoothness . . . 287
Problems . . . 288

9 Non-convex loss functions and ℓ_1-regularization . . . 293
9.1 Organization of the chapter . . . 293
9.2 Finite mixture of regressions model . . . 294
9.2.1 Finite mixture of Gaussian regressions model . . . 294
9.2.2 ℓ_1-penalized maximum likelihood estimator . . . 295
9.2.3 Properties of the ℓ_1-penalized maximum likelihood estimator . . . 299
9.2.4 Selection of the tuning parameters . . . 300
9.2.5 Adaptive ℓ_1-penalization . . . 301
9.2.6 Riboflavin production with bacillus subtilis . . . 301
9.2.7 Simulated example . . . 303
9.2.8 Numerical optimization . . . 304
9.2.9 GEM algorithm for optimization . . . 304
9.2.10 Proof of Proposition 9.2 . . . 308
9.3 Linear mixed effects models . . . 310
9.3.1 The model and ℓ_1-penalized estimation . . . 311
9.3.2 The Lasso in linear mixed effects models . . . 312
9.3.3 Estimation of the random effects coefficients . . . 312
9.3.4 Selection of the regularization parameter . . . 313

9.3.5 Properties of the Lasso in linear mixed effects models 313
9.3.6 Adaptive ℓ_1-penalized maximum likelihood estimator 314
9.3.7 Computational algorithm 314
9.3.8 Numerical results 317
9.4 Theory for ℓ_1-penalization with non-convex negative log-likelihood 320
9.4.1 The setting and notation 320
9.4.2 Oracle inequality for the Lasso for non-convex loss functions 323
9.4.3 Theory for finite mixture of regressions models 326
9.4.4 Theory for linear mixed effects models 329
9.5 Proofs for Section 9.4 332
9.5.1 Proof of Lemma 9.1 332
9.5.2 Proof of Lemma 9.2 333
9.5.3 Proof of Theorem 9.1 335
9.5.4 Proof of Lemma 9.3 337
Problems 337

10 Stable solutions 339
10.1 Organization of the chapter 339
10.2 Introduction, stability and subsampling 340
10.2.1 Stability paths for linear models 341
10.3 Stability selection 346
10.3.1 Choice of regularization and error control 346
10.4 Numerical results 351
10.5 Extensions 352
10.5.1 Randomized Lasso 352
10.6 Improvements from a theoretical perspective 354
10.7 Proofs 355
10.7.1 Sample splitting 355
10.7.2 Proof of Theorem 10.1 356
Problems 358

11 P-values for linear models and beyond 359
11.1 Organization of the chapter 359
11.2 Introduction, sample splitting and high-dimensional variable selection 360
11.3 Multi sample splitting and familywise error control 363
11.3.1 Aggregation over multiple p-values 364
11.3.2 Control of familywise error 365
11.4 Multi sample splitting and false discovery rate 367
11.4.1 Control of false discovery rate 368
11.5 Numerical results 369
11.5.1 Simulations and familywise error control 369
11.5.2 Familywise error control for motif regression in computational biology 372
11.5.3 Simulations and false discovery rate control 372

11.6 Consistent variable selection . . . 374
11.6.1 Single sample split method . . . 374
11.6.2 Multi sample split method . . . 377
11.7 Extensions . . . 377
11.7.1 Other models . . . 378
11.7.2 Control of expected false positive selections . . . 378
11.8 Proofs . . . 379
11.8.1 Proof of Proposition 11.1 . . . 379
11.8.2 Proof of Theorem 11.1 . . . 380
11.8.3 Proof of Theorem 11.2 . . . 382
11.8.4 Proof of Proposition 11.2 . . . 384
11.8.5 Proof of Lemma 11.3 . . . 384
Problems . . . 386

12 Boosting and greedy algorithms . . . 387
12.1 Organization of the chapter . . . 387
12.2 Introduction and preliminaries . . . 388
12.2.1 Ensemble methods: multiple prediction and aggregation . . . 388
12.2.2 AdaBoost . . . 389
12.3 Gradient boosting: a functional gradient descent algorithm . . . 389
12.3.1 The generic FGD algorithm . . . 390
12.4 Some loss functions and boosting algorithms . . . 392
12.4.1 Regression . . . 392
12.4.2 Binary classification . . . 393
12.4.3 Poisson regression . . . 396
12.4.4 Two important boosting algorithms . . . 396
12.4.5 Other data structures and models . . . 398
12.5 Choosing the base procedure . . . 398
12.5.1 Componentwise linear least squares for generalized linear models . . . 399
12.5.2 Componentwise smoothing spline for additive models . . . 400
12.5.3 Trees . . . 403
12.5.4 The low-variance principle . . . 404
12.5.5 Initialization of boosting . . . 404
12.6 L_2Boosting . . . 405
12.6.1 Nonparametric curve estimation: some basic insights about boosting . . . 405
12.6.2 L_2Boosting for high-dimensional linear models . . . 409
12.7 Forward selection and orthogonal matching pursuit . . . 413
12.7.1 Linear models and squared error loss . . . 414
12.8 Proofs . . . 418
12.8.1 Proof of Theorem 12.1 . . . 418
12.8.2 Proof of Theorem 12.2 . . . 420
12.8.3 Proof of Theorem 12.3 . . . 426
Problems . . . 430

13 Graphical modeling 433
13.1 Organization of the chapter 433
13.2 Preliminaries about graphical models 434
13.3 Undirected graphical models 434
13.3.1 Markov properties for undirected graphs 434
13.4 Gaussian graphical models 435
13.4.1 Penalized estimation for covariance matrix and edge set 436
13.4.2 Nodewise regression 440
13.4.3 Covariance estimation based on undirected graph 442
13.5 Ising model for binary random variables 444
13.6 Faithfulness assumption 445
13.6.1 Failure of faithfulness 446
13.6.2 Faithfulness and Gaussian graphical models 448
13.7 The PC-algorithm: an iterative estimation method 449
13.7.1 Population version of the PC-algorithm 449
13.7.2 Sample version for the PC-algorithm 451
13.8 Consistency for high-dimensional data 453
13.8.1 An illustration 455
13.8.2 Theoretical analysis of the PC-algorithm 456
13.9 Back to linear models 462
13.9.1 Partial faithfulness 463
13.9.2 The PC-simple algorithm 465
13.9.3 Numerical results 468
13.9.4 Asymptotic results in high dimensions 471
13.9.5 Correlation screening (sure independence screening) 474
13.9.6 Proofs 475
Problems 480

14 Probability and moment inequalities 481
14.1 Organization of this chapter 481
14.2 Some simple results for a single random variable 482
14.2.1 Sub-exponential random variables 482
14.2.2 Sub-Gaussian random variables 483
14.2.3 Jensen's inequality for partly concave functions 485
14.3 Bernstein's inequality 486
14.4 Hoeffding's inequality 487
14.5 The maximum of p averages 489
14.5.1 Using Bernstein's inequality 489
14.5.2 Using Hoeffding's inequality 491
14.5.3 Having sub-Gaussian random variables 493
14.6 Concentration inequalities 494
14.6.1 Bousquet's inequality 494
14.6.2 Massart's inequality 496
14.6.3 Sub-Gaussian random variables 496
14.7 Symmetrization and contraction 497

14.8 Concentration inequalities for Lipschitz loss functions 500
14.9 Concentration for squared error loss with random design 504
14.9.1 The inner product of noise and linear functions 505
14.9.2 Squared linear functions 505
14.9.3 Squared error loss 508
14.10 Assuming only lower order moments 508
14.10.1 Nemirovski moment inequality 509
14.10.2 A uniform inequality for quadratic forms 510
14.11 Using entropy for concentration in the sub-Gaussian case 511
14.12 Some entropy results 516
14.12.1 Entropy of finite-dimensional spaces and general convex hulls 518
14.12.2 Sets with restrictions on the coefficients 518
14.12.3 Convex hulls of small sets: entropy with log-term 519
14.12.4 Convex hulls of small sets: entropy without log-term 520
14.12.5 Further refinements 523
14.12.6 An example: functions with $(m-1)$-th derivative of bounded variation 523
14.12.7 Proofs for this section (Section 14.12) 525
Problems 535

Author Index 539

Index 543

References 547

Chapter 1
Introduction

Abstract High-dimensional statistics refers to statistical inference when the number of unknown parameters is of much larger order than sample size. We present some introductory motivation and a rough picture about high-dimensional statistics.

1.1 The framework

High-dimensional statistics refers to statistical inference when the number of unknown parameters p is of much larger order than sample size n, that is: $p \gg n$. This encompasses supervised regression and classification models where the number of covariates is of much larger order than n, unsupervised settings such as clustering or graphical modeling with more variables than observations or multiple testing where the number of considered testing hypotheses is larger than sample size. Among the mentioned examples, we discuss in this book regression and classification, graphical modeling and a few aspects of multiple testing.

High-dimensional statistics has relations to other areas. The methodological concepts share some common aspects with nonparametric statistics and machine learning, all of them involving a high degree of complexity making regularization necessary. An early and important book about statistics for complex data is Breiman et al. (1984) with a strong emphasis placed on the CART algorithm. The influential book by Hastie et al. (2001) covers a very broad range of methods and techniques at the interface between statistics and machine learning, also called "statistical learning" and "data mining". From an algorithmic point of view, convex optimization is a key ingredient for regularized likelihood problems which are a central focus of our book, and such optimization arises also in the area of kernel methods from machine learning, cf. Schölkopf and Smola (2002). We include also some deviations where non-convex optimization or iterative algorithms are used. Regarding many aspects of optimization, the book by Bertsekas (1995) has been an important

P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, DOI 10.1007/978-3-642-20192-9_1,

source for our use and understanding. Furthermore, the mathematical analysis of high-dimensional statistical inference has important connections to approximation theory, cf. Temlyakov (2008), in particular in the context of sparse approximations.

1.2 The possibilities and challenges

A simple yet very useful model for high-dimensional data is a linear model

$$Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n), \tag{1.1}$$

with $p \gg n$. It is intuitively clear that the unknown intercept μ and parameter vector $\beta = (\beta_1, \ldots, \beta_p)^T$ can only be estimated reasonably well, based on n observations, if β is sparse in some sense. Sparsity can be quantified in terms of the ℓ_q-norm for $1 \le q \le \infty$, the analogue (which is not a norm) with $0 < q < 1$, or the ℓ_0-analogue (which is not a norm) $\|\beta\|_0^0 = |\{j;\ \beta_j \neq 0\}|$ which counts the number of non-zero entries of the parameter. Note that the notation $\|\beta\|_0^0 = \sum_{j=1}^p |\beta_j|^0$ (where $0^0 = 0$) is in analogy to $\|\beta\|_q^q = \sum_{j=1}^p |\beta_j|^q$ for $0 < q < \infty$. In contrast to ℓ_0, the ℓ_1-norm $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ measures sparsity in a different way and has a computational advantage of being a convex function in β.

Roughly speaking, high-dimensional statistical inference is possible, in the sense of leading to reasonable accuracy or asymptotic consistency, if

$$\log(p) \cdot (\text{sparsity}(\beta)) \ll n,$$

depending on how we define sparsity and the setting under consideration.

Early progress of high-dimensional statistical inference has been achieved a while ago: Donoho and Johnstone (1994) present beautiful and clean results for the case of orthogonal design in a linear model where $p = n$. A lot of work has been done to analyze much more general designs in linear or generalized linear models where $p \gg n$, as occurring in many applications nowadays, cf. Donoho and Huo (2001), Donoho and Elad (2003), Fuchs (2004) and many other references given later. We present in this book a detailed treatment for high-dimensional linear and generalized linear models. Much of the methodology and techniques relies on the idea of ℓ_1-penalization for the negative log-likelihood, including versions of such regularization methods. Such ℓ_1-penalization has become tremendously popular due to its computational attractiveness and its statistical properties which reach optimality under certain conditions. Other problems involve more complicated models with e.g. some nonparametric components or some more demanding likelihood functions as occurring in e.g. mixture models. We also describe results and aspects when going beyond generalized linear models.

For sound statistical inference, we would like to quantify uncertainty of estimates or predictions. In particular, if statistical results cannot be validated with a scientific experiment, as for example in bio-medicine where say biomarkers of patients cannot be manipulated, the scientific conclusions hinge on statistical results only. In such cases, high-dimensional statistical inference must be equipped with measures of uncertainty, stability or significance. Our book presents some early ideas in this direction but more refined answers need to be developed in the future.

1.3 About the book

The book is intended for graduate students and researchers in statistics or related fields who are interested in methodological themes and/or detailed mathematical theory for high-dimensional statistics. It is possible to read the methodology and theory parts of the book separately.

Besides methodology and theory, the book touches on applications, as suggested by its title. Regarding the latter, we present illustrations largely without detailed scientific interpretation. Thus, the main emphasis is clearly on methodology and theory. We believe that the theory has its implications on using methods in practice and the book interweaves these aspects. For example, when using the so-called Lasso (ℓ_1-penalization) method for high-dimensional regression, the theory gives some important insights about variable selection and more particularly about false positive and false negative selections.

The book presents important advances in high-dimensional statistical inference. Some of them, like the Lasso and some of its versions, are treated comprehensively with details on practical methodology, computation and mathematical theory. Other themes, like boosting algorithms and graphical modeling with covariance estimation, are discussed from a more practical view point and with less detailed mathematical theory. However, all chapters include a supporting mathematical argumentation.

1.3.1 Organization of the book

The book combines practical methodology and mathematical theory. For the so-called Lasso and group Lasso and versions thereof in linear, generalized linear and additive models, there are separate theory and methods chapters with cross-references to each other.

Other chapters on non-convex negative likelihood problems, stable solutions, p-values for high-dimensional inference, boosting algorithms or graphical modeling

with covariance estimation are presenting in each chapter the methods and some mathematical theory. The last chapter on probability inequalities presents mathematical results and theory which are used at various places in the book. Figure 1.1 gives an overview which parts belong closely to each other.

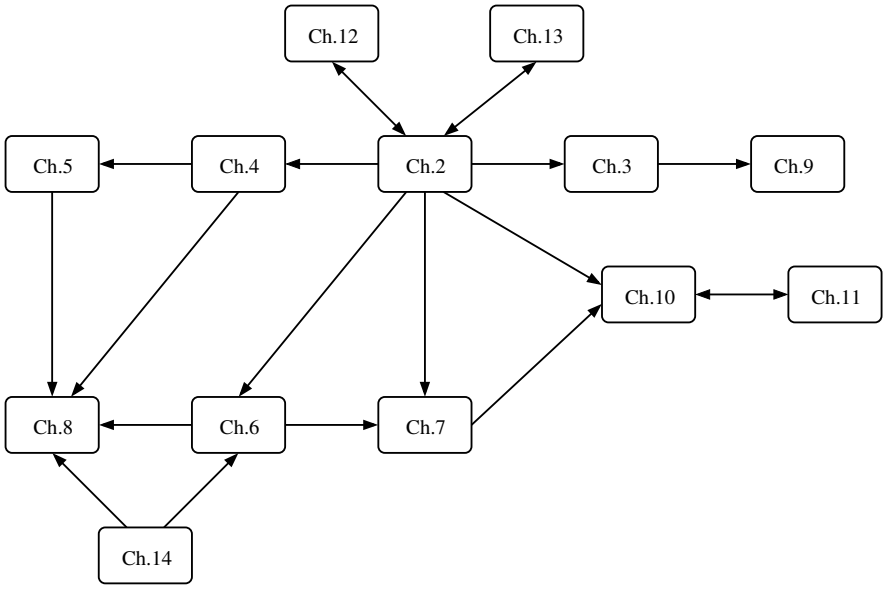

Fig. 1.1 Organization of the book. The arrowheads indicate the directions in which the chapters relate to each other. Chapters 2, 3, 4 and 5 describe statistical methodology and computation, Chapters 6, 7, 8 and 14 present detailed mathematical theory, and the remaining Chapters 9, 10, 11, 12 and 13 each contain methodological, theoretical and computational aspects.

1.4 Some examples

High-dimensional data arises nowadays in a wide variety of applications. The book contains illustrations and applications to problems from biology, a field of our own interest. However, the presented material includes models, methods, algorithms and theory whose relevance is very generic. In particular, we consider high-dimensional linear and generalized linear models as well as the more flexible generalized additive models, and both of them cover a very broad range of applications. Other areas of high-dimensional data applications include text mining, pattern recognition in imaging, astronomy and climate research.

1.4.1 Prediction and biomarker discovery in genomics

In genomics with high-throughput measurements, thousands of variables such as expressions of genes and abundances of proteins can be measured for each person in a (pre-)clinical study. A typical goal is to classify the health status of a person, e.g. healthy or diseased, based on its bio-molecular profile, i.e., the thousands of bio-molecular variables measured for the person.

1.4.1.1 Further biology applications treated in the book

We briefly describe now examples from genomics which will be considered in the book.

We consider motif regression in Chapters 2, 5, 10 and 11. The goal is to infer short DNA-words of approximate length $8-16$ base pairs, e.g., "ACCGTTAC", where a certain protein or transcription factor binds to the DNA. We have supervised data available with a continuous response variable Y_i and p-dimensional covariates X_i with continuous values. Thereby Y_i measures e.g. binding intensity of the protein of interest in the ith region of the whole DNA sequence and X_i contains abundance scores of p candidate motifs (or DNA words) in the ith region of the DNA. We relate the response Y_i and the covariates X_i with a linear model as in (1.1) (or an additive model as in Chapter 5), where $X_i^{(j)}$ denotes the abundance score of candidate word j in DNA region i. The task is to infer which candidate words are relevant for explaining the response Y. Statistically, we want to find the variables $X^{(j)}$ whose corresponding regression coefficients β_j are substantial in absolute value or significantly different from zero. That is, motif regression is concerned about variable or feature selection. The typical sizes for motif regression are $n \approx 50-1'000$ and $p \approx 100-2'000$ and hence, the number of variables or the dimensionality p is about of the same order as sample size n. In this sense, motif regression is a fairly but not truly high-dimensional problem.

Another example is the prediction of DNA splice sites which are the regions between coding and non-coding DNA segments. The problem is discussed in Chapter 4. We have binary response variables $Y_i \in \{0,1\}$, encoding whether there is a splice site or not at a certain position i of the DNA sequence, and categorical p-dimensional covariates $X_i \in \{A,C,G,T\}^p$ with four categories corresponding to the letters of the DNA alphabet. The p categorical variables correspond to p neighboring values of a certain position i of the DNA sequence: for example, 3 positions to the left and 4 positions to the right from i, corresponding to $p=7$ and e.g. $X_i = (A,A,T,G,G,C,G)$. We model the data as a binary logistic regression whose covariates consist of 7 factors each having 4 levels. The primary goal here is prediction or classification of a new, unknown splice site. The typical sizes for DNA splice site prediction is $n \approx 10'000-50'000$ and $p \approx 5-20$. When allowing for all interactions, the num-

ber of parameters in the logistic model is 4^p which can be huge in comparison to n, e.g., $4^{10} \approx 1.05 \cdot 10^6$. Depending on how many interactions we allow, the problem may involve a million unknown parameters which is of larger order than the typical sample size.

In Chapters 9 and 10 we illustrate some methods for a problem about riboflavin production with bacillus subtilis. The data consists of continuous response variables Y_i, measuring the log-concentration of riboflavin, and p-dimensional covariates X_i containing the log-expressions from essentially all genes from bacillus subtilis, for the ith individual. The goal is primarily variable selection to increase understanding which genes are relevant for the riboflavin production rate. A linear model as in (1.1) is often a reasonable approximation but we will also discuss in Chapter 9 a mixture model which is an attempt to model inhomogeneity of the data. The size of the data is about $n \approx 70 - 150$ and $p = 4088$, and hence it is a real high-dimensional problem.

Finally, we consider in Chapter 13 an unsupervised problem about genes in two biosynthesis pathways in arabidopsis thaliana. The data consists of continuous gene expressions from 39 genes for $n = 118$ samples of different arabidopsis plants. We illustrate covariance estimation and aspects of graphical modeling which involve $39 \cdot 40/2 = 780$ covariance parameters, i.e., more parameters than sample size.

Chapter 2
Lasso for linear models

Abstract The Lasso, proposed by Tibshirani (1996), is an acronym for Least Absolute Shrinkage and Selection Operator. Among the main reasons why it has become very popular for high-dimensional estimation problems are its statistical accuracy for prediction and variable selection coupled with its computational feasibility. Furthermore, since the Lasso is a penalized likelihood approach, the method is rather general and can be used in a broad variety of models. In the simple case of a linear model with orthonormal design, the Lasso equals the soft thresholding estimator introduced and analyzed by Donoho and Johnstone (1994). The Lasso for linear models is the core example to develop the methodology for ℓ_1-penalization in high-dimensional settings. We discuss in this chapter some fundamental methodological and computational aspects of the Lasso. We also present the adaptive Lasso, an important two-stage procedure which addresses some bias problems of the Lasso. The methodological steps are supported by describing various theoretical results which will be fully developed in Chapters 6 and 7.

2.1 Organization of the chapter

We present in this chapter the Lasso for linear models from a methodological point of view. Theoretical results are loosely described to support methodology and practical steps for analyzing high-dimensional data. After an introduction in Section 2.2 with the definition of the Lasso for linear models, we focus in Section 2.4 on prediction of a new response when given a new covariate. Afterwards, we discuss in Section 2.5 the Lasso for estimating the regression coefficients which is rather different from prediction. An important implication will be that under certain conditions, the Lasso will have the screening property for variable selection saying that it will include all relevant variables whose regression coefficients are sufficiently large (besides potentially false positive selections). In Section 2.6 we discuss the more ambitious goal of variable selection in terms of exact recovery of all the rele-

P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, DOI 10.1007/978-3-642-20192-9_2,

vant variables. Some of the drawbacks of the Lasso can be addressed by two-stage or multi-stage procedures. Among them are the adaptive Lasso (Zou, 2006) and the relaxed Lasso (Meinshausen, 2007), discussed in Sections 2.8 and 2.10, respectively. Finally, we present concepts and ideas for computational algorithms in Section 2.12.

2.2 Introduction and preliminaries

We consider here the setting where the observed data are realizations of

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

with p-dimensional covariates $X_i \in \mathscr{X} \subset \mathbb{R}^p$ and univariate response variables $Y_i \in \mathscr{Y} \subset \mathbb{R}$. The covariates are either deterministic fixed values or random variables: regarding the methodology, there is no difference between these two cases. Typically, we assume that the samples are independent but the generalization to stationary processes poses no essential methodological or theoretical problems.

Modeling high-dimensional data is challenging. For a continuous response variable $Y \in \mathbb{R}$, a simple yet very useful approach is given by a linear model

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n), \tag{2.1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d., independent of $\{X_i;\ i = 1, \ldots, n\}$ and with $\mathbb{E}[\varepsilon_i] = 0$.

For simplicity and without loss of generality, we usually assume that the intercept is zero and that all covariates are centered and measured on the same scale. Both of these assumptions can be approximately achieved by empirical mean centering and scaling with the standard deviation, and the standardized data then satisfies $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i = 0$ and $\hat{\sigma}_j^2 := n^{-1} \sum_{i=1}^{n} (X_i^{(j)} - \overline{X}^{(j)})^2 = 1$ for all j. The only unusual aspect of the linear model in (2.1) is the fact that $p \gg n$.

We often use for (2.1) the matrix- and vector-notation

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

with response vector $\mathbf{Y}_{n \times 1}$, design matrix $\mathbf{X}_{n \times p}$, parameter vector $\beta_{p \times 1}$ and error vector $\varepsilon_{n \times 1}$. If the model is correct, we denote the true underlying parameter by β^0. We denote the best approximating parameter, in a sense to be specified, by β^*: this case will be discussed from a theory point of view in Chapter 6 in Section 6.2.3.

2.2.1 The Lasso estimator

If $p > n$, the ordinary least squares estimator is not unique and will heavily overfit the data. Thus, a form of complexity regularization will be necessary. We focus here on regularization with the ℓ_1-penalty. The parameters in model (2.1) are estimated with the Lasso (Tibshirani, 1996):

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \Big(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_1 \Big), \tag{2.2}$$

where $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \sum_{i=1}^n (Y_i - (\mathbf{X}\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ and where $\lambda \geq 0$ is a penalty parameter. The estimator has the property that it does variable selection in the sense that $\hat{\beta}_j(\lambda) = 0$ for some j's (depending on the choice of λ) and $\hat{\beta}_j(\lambda)$ can be thought as a shrunken least squares estimator; hence, the name Least Absolute Shrinkage and Selection Operator (LASSO). An intuitive explanation for the variable selection property is given below.

The optimization in (2.2) is convex, enabling efficient computation of the estimator, see Section 2.12. In addition, the optimization problem in (2.2) is equivalent to

$$\hat{\beta}_{\text{primal}}(R) = \arg\min_{\beta; \|\beta\|_1 \leq R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n, \tag{2.3}$$

with a one-to-one correspondence between λ in (2.2) and R in (2.3), depending on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Such an equivalence holds since $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ is convex in β with convex constraint $\|\beta\|_1 \leq R$, see for example Bertsekas (1995, Ch. 5.3).

Because of the ℓ_1-geometry, the Lasso is performing variable selection in the sense that an estimated component can be exactly zero. To see this, we consider the representation in (2.3) and Figure 2.1: the residual sum of squares reaches a minimal value (for certain constellations of the data) if its contour lines hit the ℓ_1-ball in its corner which corresponds to the first component $\hat{\beta}_{\text{primal},1}$ being equal to zero. Figure 2.1 indicates that such a phenomenon does not occur with say Ridge regression,

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \arg\min_{\beta} \Big(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_2^2 \Big),$$

with its equivalent primal solution

$$\hat{\beta}_{\text{Ridge;primal}}(R) = \arg\min_{\beta; \|\beta\|_2 \leq R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n, \tag{2.4}$$

with again a data-dependent one-to-one correspondence between λ and R.

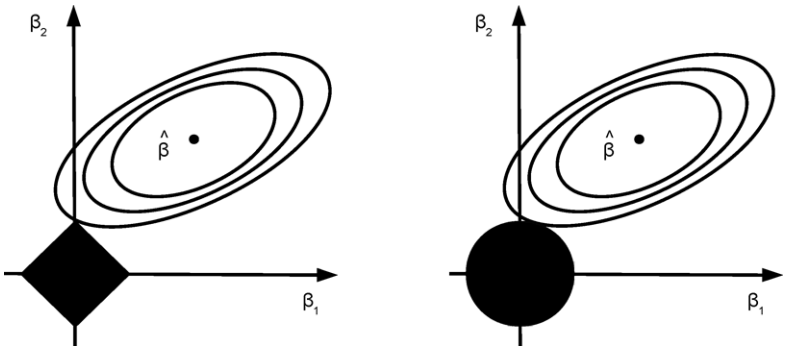

Fig. 2.1 Left: Contour lines of residual sum of squares, with $\hat{\beta}$ being the least squares estimator, and ℓ_1-ball corresponding to the Lasso problem in (2.3). Right: Analogous to left panel but with ℓ_2-ball corresponding to Ridge regression in (2.4). The figure is as in Tibshirani (1996).

2.2.1.1 Estimation of the error variance

The estimator in (2.2) does not directly provide an estimate for the error variance σ^2. One can construct an estimator using the residual sum of squares and the degrees of freedom of the Lasso (Section 2.11). Alternatively, and rigorously developed, we can estimate β and σ^2 simultaneously using a reparametrization: this is discussed in detail in Section 9.2.2.1 from Chapter 9.

2.3 Orthonormal design

It is instructive to consider the orthonormal design where $p = n$ and the design matrix satisfies $n^{-1}\mathbf{X}^T\mathbf{X} = I_{p\times p}$. In this case, the Lasso estimator is the soft-threshold estimator

$$\hat{\beta}_j(\lambda) = \text{sign}(Z_j)(|Z_j| - \lambda/2)_+,\ Z_j = (\mathbf{X}^T\mathbf{Y})_j/n \ \ (j = 1,\ldots,p = n), \quad (2.5)$$

where $(x)_+ = \max(x, 0)$ denotes the positive part and Z_j equals the ordinary least squares estimator for β_j. This follows from the general characterization in Lemma 2.1 below and we leave a direct derivation (without using Lemma 2.1) as Problem 2.1. Thus, the estimator can be written as

$$\hat{\beta}_j(\lambda) = g_{\text{soft},\lambda/2}(Z_j),$$

where $g_{\text{soft},\lambda}(z) = \text{sign}(z)(|z| - \lambda)_+$, is the soft-threshold function depicted in Figure 2.2. There, we also show for comparison the hard-threshold and the adaptive Lasso

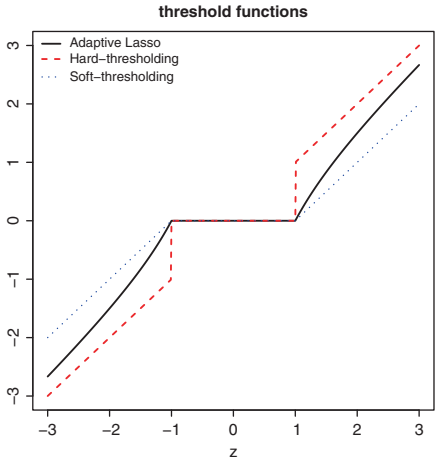

Fig. 2.2 Various threshold functions $g(\cdot)$ for orthonormal design: soft-threshold (dashed line), hard-threshold (dotted line), Adaptive Lasso (solid line). The estimators are of the form $\hat{\beta}_j = g(Z_j)$ with Z_j as in (2.5).

estimator (see Section 2.8) for β_j defined by

$$\hat{\beta}_{\text{hard},j}(\lambda) = g_{\text{hard},\lambda/2}(Z_j), \quad g_{\text{hard},\lambda}(z) = z\mathbb{1}(|z| \le \lambda),$$
$$\hat{\beta}_{\text{adapt},j}(\lambda) = g_{\text{adapt},\lambda/2}(Z_j), \quad g_{\text{adapt},\lambda}(z) = z(1 - \lambda/|z|^2)_+ = \text{sign}(z)(|z| - \lambda/|z|)_+.$$

2.4 Prediction

We refer to prediction whenever the goal is estimation of the regression function $\mathbb{E}[Y|X = x] = \sum_{j=1}^{p} \beta_j x^{(j)}$ in model (2.1). This is also the relevant quantity for predicting a new response.