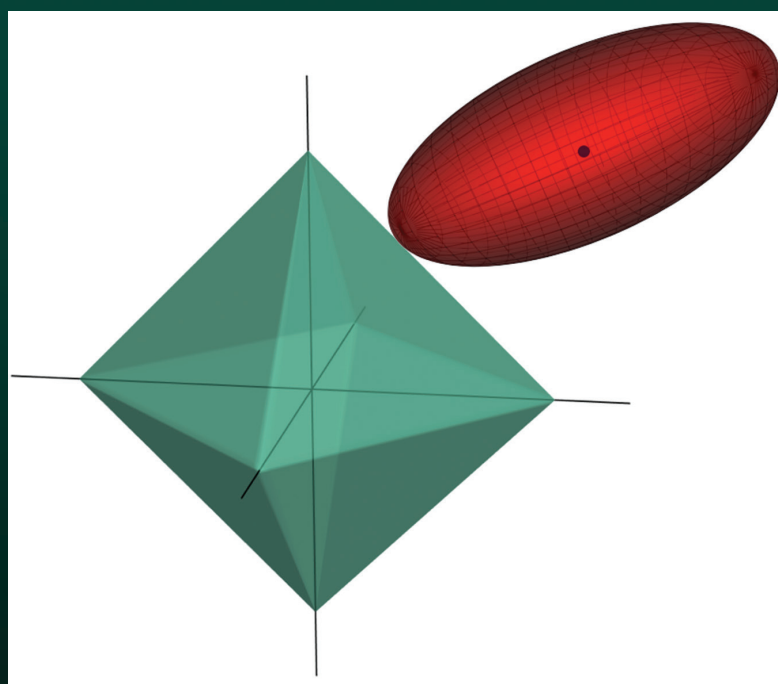


Statistical Learning with Sparsity

The Lasso and Generalizations



Trevor Hastie
Robert Tibshirani
Martin Wainwright



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Contents

Preface	xv
1 Introduction	1
2 The Lasso for Linear Models	7
2.1 Introduction	7
2.2 The Lasso Estimator	8
2.3 Cross-Validation and Inference	13
2.4 Computation of the Lasso Solution	14
2.4.1 Single Predictor: Soft Thresholding	15
2.4.2 Multiple Predictors: Cyclic Coordinate Descent	15
2.4.3 Soft-Thresholding and Orthogonal Bases	17
2.5 Degrees of Freedom	17
2.6 Uniqueness of the Lasso Solutions	19
2.7 A Glimpse at the Theory	20
2.8 The Nonnegative Garrote	20
2.9 ℓ_q Penalties and Bayes Estimates	22
2.10 Some Perspective	23
Exercises	24
3 Generalized Linear Models	29
3.1 Introduction	29
3.2 Logistic Regression	31
3.2.1 Example: Document Classification	32
3.2.2 Algorithms	35
3.3 Multiclass Logistic Regression	36
3.3.1 Example: Handwritten Digits	37
3.3.2 Algorithms	39
3.3.3 Grouped-Lasso Multinomial	39
3.4 Log-Linear Models and the Poisson GLM	40
3.4.1 Example: Distribution Smoothing	40
3.5 Cox Proportional Hazards Models	42
3.5.1 Cross-Validation	43
3.5.2 Pre-Validation	45
3.6 Support Vector Machines	46
3.6.1 Logistic Regression with Separable Data	49

3.7	Computational Details and <code>glmnet</code>	50
	Bibliographic Notes	52
	Exercises	53
4	Generalizations of the Lasso Penalty	55
4.1	Introduction	55
4.2	The Elastic Net	56
4.3	The Group Lasso	58
4.3.1	Computation for the Group Lasso	62
4.3.2	Sparse Group Lasso	64
4.3.3	The Overlap Group Lasso	65
4.4	Sparse Additive Models and the Group Lasso	69
4.4.1	Additive Models and Backfitting	69
4.4.2	Sparse Additive Models and Backfitting	70
4.4.3	Approaches Using Optimization and the Group Lasso	72
4.4.4	Multiple Penalization for Sparse Additive Models	74
4.5	The Fused Lasso	76
4.5.1	Fitting the Fused Lasso	77
4.5.1.1	Reparametrization	78
4.5.1.2	A Path Algorithm	79
4.5.1.3	A Dual Path Algorithm	79
4.5.1.4	Dynamic Programming for the Fused Lasso	80
4.5.2	Trend Filtering	81
4.5.3	Nearly Isotonic Regression	83
4.6	Nonconvex Penalties	84
	Bibliographic Notes	86
	Exercises	88
5	Optimization Methods	95
5.1	Introduction	95
5.2	Convex Optimality Conditions	95
5.2.1	Optimality for Differentiable Problems	95
5.2.2	Nondifferentiable Functions and Subgradients	98
5.3	Gradient Descent	100
5.3.1	Unconstrained Gradient Descent	101
5.3.2	Projected Gradient Methods	102
5.3.3	Proximal Gradient Methods	103
5.3.4	Accelerated Gradient Methods	107
5.4	Coordinate Descent	109
5.4.1	Separability and Coordinate Descent	110
5.4.2	Linear Regression and the Lasso	112
5.4.3	Logistic Regression and Generalized Linear Models	115
5.5	A Simulation Study	117
5.6	Least Angle Regression	118
5.7	Alternating Direction Method of Multipliers	121

5.8	Minorization-Maximization Algorithms	123
5.9	Biconvexity and Alternating Minimization	124
5.10	Screening Rules	127
	Bibliographic Notes	131
	Appendix	132
	Exercises	134
6	Statistical Inference	139
6.1	The Bayesian Lasso	139
6.2	The Bootstrap	142
6.3	Post-Selection Inference for the Lasso	147
6.3.1	The Covariance Test	147
6.3.2	A General Scheme for Post-Selection Inference	150
6.3.2.1	Fixed- λ Inference for the Lasso	154
6.3.2.2	The Spacing Test for LAR	156
6.3.3	What Hypothesis Is Being Tested?	157
6.3.4	Back to Forward Stepwise Regression	158
6.4	Inference via a Debiased Lasso	158
6.5	Other Proposals for Post-Selection Inference	160
	Bibliographic Notes	161
	Exercises	162
7	Matrix Decompositions, Approximations, and Completion	167
7.1	Introduction	167
7.2	The Singular Value Decomposition	169
7.3	Missing Data and Matrix Completion	169
7.3.1	The Netflix Movie Challenge	170
7.3.2	Matrix Completion Using Nuclear Norm	174
7.3.3	Theoretical Results for Matrix Completion	177
7.3.4	Maximum Margin Factorization and Related Methods	181
7.4	Reduced-Rank Regression	184
7.5	A General Matrix Regression Framework	185
7.6	Penalized Matrix Decomposition	187
7.7	Additive Matrix Decomposition	190
	Bibliographic Notes	195
	Exercises	196
8	Sparse Multivariate Methods	201
8.1	Introduction	201
8.2	Sparse Principal Components Analysis	202
8.2.1	Some Background	202
8.2.2	Sparse Principal Components	204
8.2.2.1	Sparsity from Maximum Variance	204
8.2.2.2	Methods Based on Reconstruction	206
8.2.3	Higher-Rank Solutions	207

8.2.3.1	Illustrative Application of Sparse PCA	209
8.2.4	Sparse PCA via Fantope Projection	210
8.2.5	Sparse Autoencoders and Deep Learning	210
8.2.6	Some Theory for Sparse PCA	212
8.3	Sparse Canonical Correlation Analysis	213
8.3.1	Example: Netflix Movie Rating Data	215
8.4	Sparse Linear Discriminant Analysis	217
8.4.1	Normal Theory and Bayes' Rule	217
8.4.2	Nearest Shrunk Centroids	218
8.4.3	Fisher's Linear Discriminant Analysis	221
8.4.3.1	Example: Simulated Data with Five Classes	222
8.4.4	Optimal Scoring	225
8.4.4.1	Example: Face Silhouettes	226
8.5	Sparse Clustering	227
8.5.1	Some Background on Clustering	227
8.5.1.1	Example: Simulated Data with Six Classes	228
8.5.2	Sparse Hierarchical Clustering	228
8.5.3	Sparse K -Means Clustering	230
8.5.4	Convex Clustering	231
	Bibliographic Notes	232
	Exercises	234
9	Graphs and Model Selection	241
9.1	Introduction	241
9.2	Basics of Graphical Models	241
9.2.1	Factorization and Markov Properties	241
9.2.1.1	Factorization Property	242
9.2.1.2	Markov Property	243
9.2.1.3	Equivalence of Factorization and Markov Properties	243
9.2.2	Some Examples	244
9.2.2.1	Discrete Graphical Models	244
9.2.2.2	Gaussian Graphical Models	245
9.3	Graph Selection via Penalized Likelihood	246
9.3.1	Global Likelihoods for Gaussian Models	247
9.3.2	Graphical Lasso Algorithm	248
9.3.3	Exploiting Block-Diagonal Structure	251
9.3.4	Theoretical Guarantees for the Graphical Lasso	252
9.3.5	Global Likelihood for Discrete Models	253
9.4	Graph Selection via Conditional Inference	254
9.4.1	Neighborhood-Based Likelihood for Gaussians	255
9.4.2	Neighborhood-Based Likelihood for Discrete Models	256
9.4.3	Pseudo-Likelihood for Mixed Models	259
9.5	Graphical Models with Hidden Variables	261
	Bibliographic Notes	261

Exercises	263
10 Signal Approximation and Compressed Sensing	269
10.1 Introduction	269
10.2 Signals and Sparse Representations	269
10.2.1 Orthogonal Bases	269
10.2.2 Approximation in Orthogonal Bases	271
10.2.3 Reconstruction in Overcomplete Bases	274
10.3 Random Projection and Approximation	276
10.3.1 Johnson–Lindenstrauss Approximation	277
10.3.2 Compressed Sensing	278
10.4 Equivalence between ℓ_0 and ℓ_1 Recovery	280
10.4.1 Restricted Nullspace Property	281
10.4.2 Sufficient Conditions for Restricted Nullspace	282
10.4.3 Proofs	284
10.4.3.1 Proof of Theorem 10.1	284
10.4.3.2 Proof of Proposition 10.1	284
Bibliographic Notes	285
Exercises	286
11 Theoretical Results for the Lasso	289
11.1 Introduction	289
11.1.1 Types of Loss Functions	289
11.1.2 Types of Sparsity Models	290
11.2 Bounds on Lasso ℓ_2 -Error	291
11.2.1 Strong Convexity in the Classical Setting	291
11.2.2 Restricted Eigenvalues for Regression	293
11.2.3 A Basic Consistency Result	294
11.3 Bounds on Prediction Error	299
11.4 Support Recovery in Linear Regression	301
11.4.1 Variable-Selection Consistency for the Lasso	301
11.4.1.1 Some Numerical Studies	303
11.4.2 Proof of Theorem 11.3	305
11.5 Beyond the Basic Lasso	310
Bibliographic Notes	311
Exercises	312
Bibliography	315
Author Index	337
Index	343

