
Contents

Preface	xiii
Acknowledgments	xv
1 Introduction	1
1.1 High-Dimensional Data	1
1.2 Curse of Dimensionality	3
1.2.1 Lost in the Immensity of High-Dimensional Spaces	3
1.2.2 Fluctuations Cumulate	6
1.2.3 An Accumulation of Rare Events May Not Be Rare	10
1.2.4 Computational Complexity	13
1.3 High-Dimensional Statistics	13
1.3.1 Circumventing the Curse of Dimensionality	13
1.3.2 A Paradigm Shift	16
1.3.3 Mathematics of High-Dimensional Statistics	17
1.4 About This Book	18
1.4.1 Statistics and Data Analysis	18
1.4.2 Purpose of This Book	19
1.4.3 Overview	19
1.5 Discussion and References	21
1.5.1 Take-Home Message	21
1.5.2 References	21
1.6 Exercises	21
1.6.1 Strange Geometry of High-Dimensional Spaces	21
1.6.2 Volume of a p -Dimensional Ball	22
1.6.3 Tails of a Standard Gaussian Distribution	22
1.6.4 Principal Component Analysis	23
1.6.5 Basics of Linear Regression	24
1.6.6 Concentration of Square Norm of Gaussian Random Variable	24
2 Model Selection	27
2.1 Statistical Setting	27
2.2 To Select among a Collection of Models	30
2.2.1 Models and Oracle	30
2.2.2 Model Selection Procedures	33

2.3	Risk Bound for Model Selection	37
2.3.1	Oracle Risk Bound	37
2.4	Optimality	41
2.4.1	Minimax Optimality	41
2.4.2	Frontier of Estimation in High Dimensions	45
2.4.3	Minimal Penalties	45
2.5	Computational Issues	46
2.6	Illustration	47
2.7	An Alternative Point of View on Model Selection	49
2.8	Discussion and References	50
2.8.1	Take-Home Message	50
2.8.2	References	50
2.9	Exercises	51
2.9.1	Orthogonal Design	51
2.9.2	Risk Bounds for the Different Sparsity Settings	53
2.9.3	Collections of Nested Models	54
2.9.4	Segmentation with Dynamic Programming	55
2.9.5	Goldenshluger–Lepski Method	56
2.9.6	Minimax Lower Bounds	58
3	Aggregation of Estimators	61
3.1	Introduction	61
3.2	Gibbs Mixing of Estimators	61
3.3	Oracle Risk Bound	62
3.4	Numerical Approximation by Metropolis–Hastings	65
3.5	Numerical Illustration	68
3.6	Discussion and References	69
3.6.1	Take-Home Message	69
3.6.2	References	69
3.7	Exercises	69
3.7.1	Gibbs Distribution	69
3.7.2	Orthonormal Setting with Power Law Prior	70
3.7.3	Group-Sparse Setting	70
3.7.4	Gain of Combining	70
3.7.5	Online Aggregation	71
4	Convex Criteria	73
4.1	Reminder on Convex Multivariate Functions	73
4.1.1	Subdifferentials	73
4.1.2	Two Useful Properties	74
4.2	Lasso Estimator	74
4.2.1	Geometric Insights	76
4.2.2	Analytic Insights	76
4.2.3	Oracle Risk Bound	79
4.2.4	Computing the Lasso Estimator	82

4.2.5	Removing the Bias of the Lasso Estimator	86
4.3	Convex Criteria for Various Sparsity Patterns	87
4.3.1	Group-Lasso (Group Sparsity)	87
4.3.2	Sparse-Group Lasso (Sparse-Group Sparsity)	90
4.3.3	Fused-Lasso (Variation Sparsity)	91
4.4	Discussion and References	91
4.4.1	Take-Home Message	91
4.4.2	References	91
4.5	Exercises	92
4.5.1	When Is the Lasso Solution Unique?	92
4.5.2	Support Recovery via the Witness Approach	93
4.5.3	Lower Bound on the Compatibility Constant	94
4.5.4	On the Group-Lasso	94
4.5.5	Dantzig Selector	96
4.5.6	Projection on the ℓ^1 -Ball	97
4.5.7	Ridge and Elastic-Net	98
5	Estimator Selection	101
5.1	Estimator Selection	102
5.2	Cross-Validation Techniques	103
5.3	Complexity Selection Techniques	105
5.3.1	Coordinate-Sparse Regression	106
5.3.2	Group-Sparse Regression	107
5.3.3	Multiple Structures	108
5.4	Scaled-Invariant Criteria	108
5.5	References and Discussion	113
5.5.1	Take-Home Message	113
5.5.2	References	115
5.6	Exercises	116
5.6.1	Expected V -Fold CV ℓ^2 -Risk	116
5.6.2	Proof of Corollary 5.5	117
5.6.3	Some Properties of Penalty (5.4)	117
5.6.4	Selecting the Number of Steps for the Forward Algorithm	119
6	Multivariate Regression	121
6.1	Statistical Setting	121
6.2	A Reminder on Singular Values	122
6.3	Low-Rank Estimation	123
6.3.1	If We Knew the Rank of A^*	123
6.3.2	When the Rank of A^* Is Unknown	126
6.4	Low Rank and Sparsity	129
6.4.1	Row-Sparse matrices	129
6.4.2	Criterion for Row-Sparse and Low-Rank Matrices	130
6.4.3	Convex Criterion for Low Rank Matrices	134
6.4.4	Convex Criterion for Sparse and Low-Rank Matrices	137

6.5	Discussion and References	137
6.5.1	Take-Home Message	137
6.5.2	References	138
6.6	Exercises	138
6.6.1	Hard-Thresholding of the Singular Values	138
6.6.2	Exact Rank Recovery	138
6.6.3	Rank Selection with Unknown Variance	139
7	Graphical Models	141
7.1	Reminder on Conditional Independence	142
7.2	Graphical Models	142
7.2.1	Directed Acyclic Graphical Models	142
7.2.2	Nondirected Models	144
7.3	Gaussian Graphical Models (GGM)	147
7.3.1	Connection with the Precision Matrix and the Linear Regression	147
7.3.2	Estimating g by Multiple Testing	148
7.3.3	Sparse Estimation of the Precision Matrix	149
7.3.4	Estimation of g by Regression	150
7.4	Practical Issues	154
7.5	Discussion and References	155
7.5.1	Take-Home Message	155
7.5.2	References	155
7.6	Exercises	156
7.6.1	Factorization in Directed Models	156
7.6.2	Moralization of a Directed Graph	156
7.6.3	Convexity of $-\log(\det(K))$	157
7.6.4	Block Gradient Descent with the ℓ^1/ℓ^2 Penalty	157
7.6.5	Gaussian Graphical Models with Hidden Variables	157
7.6.6	Dantzig Estimation of Sparse Gaussian Graphical Models	158
7.6.7	Gaussian Copula Graphical Models	160
7.6.8	Restricted Isometry Constant for Gaussian Matrices	162
8	Multiple Testing	165
8.1	An Introductory Example	165
8.1.1	Differential Expression of a Single Gene	165
8.1.2	Differential Expression of Multiple Genes	166
8.2	Statistical Setting	167
8.2.1	p -Values	167
8.2.2	Multiple Testing Setting	168
8.2.3	Bonferroni Correction	169
8.3	Controlling the False Discovery Rate	169
8.3.1	Heuristics	169
8.3.2	Step-Up Procedures	170
8.3.3	FDR Control under the WPRDS Property	173

8.4	Illustration	175
8.5	Discussion and References	177
8.5.1	Take-Home Message	177
8.5.2	References	177
8.6	Exercises	177
8.6.1	FDR versus FWER	177
8.6.2	WPRDS Property	178
8.6.3	Positively Correlated Normal Test Statistics	178
9	Supervised Classification	181
9.1	Statistical Modeling	181
9.1.1	Bayes Classifier	181
9.1.2	Parametric Modeling	182
9.1.3	Semi-Parametric Modeling	183
9.1.4	Nonparametric Modeling	185
9.2	Empirical Risk Minimization	185
9.2.1	Misclassification Probability of the Empirical Risk Minimizer	186
9.2.2	Vapnik–Chervonenkis Dimension	190
9.2.3	Dictionary Selection	193
9.3	From Theoretical to Practical Classifiers	195
9.3.1	Empirical Risk Convexification	195
9.3.2	Statistical Properties	198
9.3.3	Support Vector Machines	202
9.3.4	AdaBoost	206
9.3.5	Classifier Selection	207
9.4	Discussion and References	208
9.4.1	Take-Home Message	208
9.4.2	References	208
9.5	Exercises	208
9.5.1	Linear Discriminant Analysis	208
9.5.2	VC Dimension of Linear Classifiers in $\mathbb{R}^d$	209
9.5.3	Linear Classifiers with Margin Constraints	209
9.5.4	Spectral Kernel	210
9.5.5	Computation of the SVM Classifier	210
9.5.6	Kernel Principal Component Analysis (KPCA)	211
Appendix A	Gaussian Distribution	213
A.1	Gaussian Random Vectors	213
A.2	Chi-Square Distribution	214
A.3	Gaussian Conditioning	215

Appendix B Probabilistic Inequalities	217
B.1 Basic Inequalities	217
B.2 Concentration Inequalities	218
B.2.1 McDiarmid Inequality	218
B.2.2 Gaussian Concentration Inequality	221
B.3 Symmetrization and Contraction Lemmas	223
B.3.1 Symmetrization Lemma	223
B.3.2 Contraction Principle	224
B.4 Birgé's Inequality	227
Appendix C Linear Algebra	229
C.1 Singular Value Decomposition (SVD)	229
C.2 Moore–Penrose Pseudo-Inverse	230
C.3 Matrix Norms	231
C.4 Matrix Analysis	232
Appendix D Subdifferentials of Convex Functions	235
D.1 Subdifferentials and Subgradients	235
D.2 Examples of Subdifferentials	237
Appendix E Reproducing Kernel Hilbert Spaces	239
Notations	243
Bibliography	245
Index	253