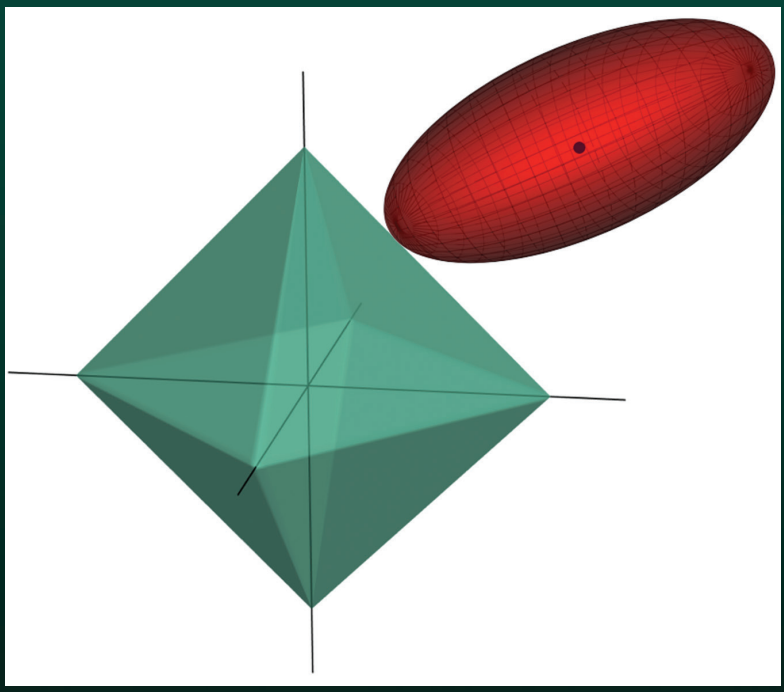
Monographs on Statistics and Applied Probability 143

# Statistical Learning with Sparsity

## The Lasso and Generalizations



Trevor Hastie
Robert Tibshirani
Martin Wainwright

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Monographs on Statistics and Applied Probability 143

# Statistical Learning with Sparsity
## The Lasso and Generalizations

**Trevor Hastie**
Stanford University
USA

**Robert Tibshirani**
Stanford University
USA

**Martin Wainwright**
University of California, Berkeley
USA


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

**F. Bunea, V. Isham, N. Keiding, T. Louis, R. L. Smith, and H. Tong**

1. Stochastic Population Models in Ecology and Epidemiology *M.S. Barlett* (1960)
2. Queues *D.R. Cox and W.L. Smith* (1961)
3. Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4. The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5. Population Genetics *W.J. Ewens* (1969)
6. Probability, Statistics and Time *M.S. Barlett* (1975)
7. Statistical Inference *S.D. Silvey* (1975)
8. The Analysis of Contingency Tables *B.S. Everitt* (1977)
9. Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10. Stochastic Abundance Models *S. Engen* (1978)
11. Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12. Point Processes *D.R. Cox and V. Isham* (1980)
13. Identification of Outliers *D.M. Hawkins* (1980)
14. Optimal Design *S.D. Silvey* (1980)
15. Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16. Classification *A.D. Gordon* (1981)
17. Distribution-Free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18. Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19. Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20. Risk Theory, 3rd edition *R.E. Beard, T. Pentikäinen and E. Pesonen* (1984)
21. Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22. An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23. Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24. Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25. The Statistical Analysis of Composition Data *J. Aitchison* (1986)
26. Density Estimation for Statistics and Data Analysis *B.W. Silverman* (1986)
27. Regression Analysis with Applications *G.B. Wetherill* (1986)
28. Sequential Methods in Statistics, 3rd edition *G.B. Wetherill and K.D. Glazebrook* (1986)
29. Tensor Methods in Statistics *P. McCullagh* (1987)
30. Transformation and Weighting in Regression *R.J. Carroll and D. Ruppert* (1988)
31. Asymptotic Techniques for Use in Statistics *O.E. Bandorff-Nielsen and D.R. Cox* (1989)
32. Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33. Analysis of Infectious Disease Data *N.G. Becker* (1989)
34. Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35. Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36. Symmetric Multivariate and Related Distributions *K.T. Fang, S. Kotz and K.W. Ng* (1990)
37. Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38. Cyclic and Computer Generated Designs, 2nd edition *J.A. John and E.R. Williams* (1995)
39. Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40. Subset Selection in Regression *A.J. Miller* (1990)
41. Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42. Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43. Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
44. Inspection Errors for Attributes in Quality Control *N.L. Johnson, S. Kotz and X. Wu* (1991)
45. The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)
46. The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)
47. Longitudinal Data with Serial Correlation—A State-Space Approach *R.H. Jones* (1993)

48. Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)
49. Markov Models and Optimization *M.H.A. Davis* (1993)
50. Networks and Chaos—Statistical and Probabilistic Aspects
*O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall* (1993)
51. Number-Theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)
52. Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)
53. Practical Risk Theory for Actuaries *C.D. Daykin, T. Pentikäinen and M. Pesonen* (1994)
54. Biplots *J.C. Gower and D.J. Hand* (1996)
55. Predictive Inference—An Introduction *S. Geisser* (1993)
56. Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)
57. An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)
58. Nonparametric Regression and Generalized Linear Models *P.J. Green and B.W. Silverman* (1994)
59. Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)
60. Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)
61. Statistics for Long Memory Processes *J. Beran* (1995)
62. Nonlinear Models for Repeated Measurement Data *M. Davidian and D.M. Giltinan* (1995)
63. Measurement Error in Nonlinear Models *R.J. Carroll, D. Rupert and L.A. Stefanski* (1995)
64. Analyzing and Modeling Rank Data *J.J. Marden* (1995)
65. Time Series Models—In Econometrics, Finance and Other Fields
*D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen* (1996)
66. Local Polynomial Modeling and its Applications *J. Fan and I. Gijbels* (1996)
67. Multivariate Dependencies—Models, Analysis and Interpretation *D.R. Cox and N. Wermuth* (1996)
68. Statistical Inference—Based on the Likelihood *A. Azzalini* (1996)
69. Bayes and Empirical Bayes Methods for Data Analysis *B.P. Carlin and T.A Louis* (1996)
70. Hidden Markov and Other Models for Discrete-Valued Time Series *I.L. MacDonald and W. Zucchini* (1997)
71. Statistical Evidence—A Likelihood Paradigm *R. Royall* (1997)
72. Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)
73. Multivariate Models and Dependence Concepts *H. Joe* (1997)
74. Theory of Sample Surveys *M.E. Thompson* (1997)
75. Retrial Queues *G. Falin and J.G.C. Templeton* (1997)
76. Theory of Dispersion Models *B. Jørgensen* (1997)
77. Mixed Poisson Processes *J. Grandell* (1997)
78. Variance Components Estimation—Mixed Models, Methodologies and Applications *P.S.R.S. Rao* (1997)
79. Bayesian Methods for Finite Population Sampling *G. Meeden and M. Ghosh* (1997)
80. Stochastic Geometry—Likelihood and computation
*O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout* (1998)
81. Computer-Assisted Analysis of Mixtures and Applications—Meta-Analysis, Disease Mapping and Others
*D. Böhning* (1999)
82. Classification, 2nd edition *A.D. Gordon* (1999)
83. Semimartingales and their Statistical Inference *B.L.S. Prakasa Rao* (1999)
84. Statistical Aspects of BSE and vCJD—Models for Epidemics *C.A. Donnelly and N.M. Ferguson* (1999)
85. Set-Indexed Martingales *G. Ivanoff and E. Merzbach* (2000)
86. The Theory of the Design of Experiments *D.R. Cox and N. Reid* (2000)
87. Complex Stochastic Systems *O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg* (2001)
88. Multidimensional Scaling, 2nd edition *T.F. Cox and M.A.A. Cox* (2001)
89. Algebraic Statistics—Computational Commutative Algebra in Statistics
*G. Pistone, E. Riccomagno and H.P. Wynn* (2001)
90. Analysis of Time Series Structure—SSA and Related Techniques
*N. Golyandina, V. Nekrutkin and A.A. Zhigljavsky* (2001)
91. Subjective Probability Models for Lifetimes *Fabio Spizzichino* (2001)
92. Empirical Likelihood *Art B. Owen* (2001)
93. Statistics in the 21st Century *Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells* (2001)
94. Accelerated Life Models: Modeling and Statistical Analysis
*Vilijandas Bagdonavicius and Mikhail Nikulin* (2001)

95. Subset Selection in Regression, Second Edition *Alan Miller* (2002)
96. Topics in Modelling of Clustered Data *Marc Aerts, Helena Geys, Geert Molenberghs, and Louise M. Ryan* (2002)
97. Components of Variance *D.R. Cox and P.J. Solomon* (2002)
98. Design and Analysis of Cross-Over Trials, 2nd Edition *Byron Jones and Michael G. Kenward* (2003)
99. Extreme Values in Finance, Telecommunications, and the Environment *Bärbel Finkenstädt and Holger Rootzén* (2003)
100. Statistical Inference and Simulation for Spatial Point Processes *Jesper Møller and Rasmus Plenge Waagepetersen* (2004)
101. Hierarchical Modeling and Analysis for Spatial Data *Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand* (2004)
102. Diagnostic Checks in Time Series *Wai Keung Li* (2004)
103. Stereology for Statisticians *Adrian Baddeley and Eva B. Vedel Jensen* (2004)
104. Gaussian Markov Random Fields: Theory and Applications *Håvard Rue and Leonhard Held* (2005)
105. Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition *Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu* (2006)
106. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood *Youngjo Lee, John A. Nelder, and Yudi Pawitan* (2006)
107. Statistical Methods for Spatio-Temporal Systems *Bärbel Finkenstädt, Leonhard Held, and Valerie Isham* (2007)
108. Nonlinear Time Series: Semiparametric and Nonparametric Methods *Jiti Gao* (2007)
109. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis *Michael J. Daniels and Joseph W. Hogan* (2008)
110. Hidden Markov Models for Time Series: An Introduction Using R *Walter Zucchini and Iain L. MacDonald* (2009)
111. ROC Curves for Continuous Data *Wojtek J. Krzanowski and David J. Hand* (2009)
112. Antedependence Models for Longitudinal Data *Dale L. Zimmerman and Vicente A. Núñez-Antón* (2009)
113. Mixed Effects Models for Complex Data *Lang Wu* (2010)
114. Intoduction to Time Series Modeling *Genshiro Kitagawa* (2010)
115. Expansions and Asymptotics for Statistics *Christopher G. Small* (2010)
116. Statistical Inference: An Integrated Bayesian/Likelihood Approach *Murray Aitkin* (2010)
117. Circular and Linear Regression: Fitting Circles and Lines by Least Squares *Nikolai Chernov* (2010)
118. Simultaneous Inference in Regression *Wei Liu* (2010)
119. Robust Nonparametric Statistical Methods, Second Edition *Thomas P. Hettmansperger and Joseph W. McKean* (2011)
120. Statistical Inference: The Minimum Distance Approach *Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park* (2011)
121. Smoothing Splines: Methods and Applications *Yuedong Wang* (2011)
122. Extreme Value Methods with Applications to Finance *Serguei Y. Novak* (2012)
123. Dynamic Prediction in Clinical Survival Analysis *Hans C. van Houwelingen and Hein Putter* (2012)
124. Statistical Methods for Stochastic Differential Equations *Mathieu Kessler, Alexander Lindner, and Michael Sørensen* (2012)
125. Maximum Likelihood Estimation for Sample Surveys *R. L. Chambers, D. G. Steel, Suojin Wang, and A. H. Welsh* (2012)
126. Mean Field Simulation for Monte Carlo Integration *Pierre Del Moral* (2013)
127. Analysis of Variance for Functional Data *Jin-Ting Zhang* (2013)
128. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition *Peter J. Diggle* (2013)
129. Constrained Principal Component Analysis and Related Techniques *Yoshio Takane* (2014)
130. Randomised Response-Adaptive Designs in Clinical Trials *Anthony C. Atkinson and Atanu Biswas* (2014)
131. Theory of Factorial Design: Single- and Multi-Stratum Experiments *Ching-Shui Cheng* (2014)
132. Quasi-Least Squares Regression *Justine Shults and Joseph M. Hilbe* (2014)
133. Data Analysis and Approximate Models: Model Choice, Location-Scale, Analysis of Variance, Nonparametric Regression and Image Analysis *Laurie Davies* (2014)
134. Dependence Modeling with Copulas *Harry Joe* (2014)
135. Hierarchical Modeling and Analysis for Spatial Data, Second Edition *Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand* (2014)

136. Sequential Analysis: Hypothesis Testing and Changepoint Detection *Alexander Tartakovsky, Igor Nikiforov, and Michèle Basseville* (2015)
137. Robust Cluster Analysis and Variable Selection *Gunter Ritter* (2015)
138. Design and Analysis of Cross-Over Trials, Third Edition *Byron Jones and Michael G. Kenward* (2015)
139. Introduction to High-Dimensional Statistics *Christophe Giraud* (2015)
140. Pareto Distributions: Second Edition *Barry C. Arnold* (2015)
141. Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data *Paul Gustafson* (2015)
142. Models for Dependent Time Series *Granville Tunnicliffe Wilson, Marco Reale, John Haywood* (2015)
143. Statistical Learning with Sparsity: The Lasso and Generalizations *Trevor Hastie, Robert Tibshirani, and Martin Wainwright* (2015)

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742


Version Date: 20150316

International Standard Book Number-13: 978-1-4987-1217-0 (eBook - PDF)

*To our parents:*

*Valerie and Patrick Hastie*

*Vera and Sami Tibshirani*

*Patricia and John Wainwright*

*and to our families:*

*Samantha, Timothy, and Lynda*

*Charlie, Ryan, Jess, Julie, and Cheryl*

*Haruko and Hana*

# Contents

# Preface

In this monograph, we have attempted to summarize the actively developing field of statistical learning with sparsity. A sparse statistical model is one having only a small number of nonzero parameters or weights. It represents a classic case of "*less is more*": a sparse model can be much easier to estimate and interpret than a dense model. In this age of big data, the number of features measured on a person or object can be large, and might be larger than the number of observations. The sparsity assumption allows us to tackle such problems and extract useful and reproducible patterns from big datasets.

The ideas described here represent the work of an entire community of researchers in statistics and machine learning, and we thank everyone for their continuing contributions to this exciting area. We particularly thank our colleagues at Stanford, Berkeley and elsewhere; our collaborators, and our past and current students working in this area. These include Alekh Agarwal, Arash Amini, Francis Bach, Jacob Bien, Stephen Boyd, Andreas Buja, Emmanuel Candes, Alexandra Chouldechova, David Donoho, John Duchi, Brad Efron, Will Fithian, Jerome Friedman, Max G'Sell, Iain Johnstone, Michael Jordan, Ping Li, Po-Ling Loh, Michael Lim, Jason Lee, Richard Lockhart, Rahul Mazumder, Balasubramanian Narashimhan, Sahand Negahban, Guillaume Obozinski, Mee-Young Park, Junyang Qian, Garvesh Raskutti, Pradeep Ravikumar, Saharon Rosset, Prasad Santhanam, Noah Simon, Dennis Sun, Yukai Sun, Jonathan Taylor, Ryan Tibshirani,[1] Stefan Wager, Daniela Witten, Bin Yu, Yuchen Zhang, Ji Zhou, and Hui Zou. We also thank our editor John Kimmel for his advice and support.

Stanford University
and
University of California, Berkeley

Trevor Hastie
Robert Tibshirani
Martin Wainwright

[1]Some of the bibliographic references, for example in Chapters 4 and 6, are to Tibshirani$_2$, R.J., rather than Tibshirani, R.; the former is Ryan Tibshirani, the latter is Robert (son and father).

Chapter 1

# Introduction

*"I never keep a scorecard or the batting averages. I hate statistics. What I got to know, I keep in my head."*

This is a quote from baseball pitcher Dizzy Dean, who played in the major leagues from 1930 to 1947.

How the world has changed in the 75 or so years since that time! Now large quantities of data are collected and mined in nearly every area of science, entertainment, business, and industry. Medical scientists study the genomes of patients to choose the best treatments, to learn the underlying causes of their disease. Online movie and book stores study customer ratings to recommend or sell them new movies or books. Social networks mine information about members and their friends to try to enhance their online experience. And yes, most major league baseball teams have statisticians who collect and analyze detailed information on batters and pitchers to help team managers and players make better decisions.

Thus the world is awash with data. But as Rutherford D. Roger (and others) has said:

*"We are drowning in information and starving for knowledge."*

There is a crucial need to sort through this mass of information, and pare it down to its bare essentials. For this process to be successful, we need to hope that the world is not as complex as it might be. For example, we hope that not all of the $30,000$ or so genes in the human body are directly involved in the process that leads to the development of cancer. Or that the ratings by a customer on perhaps 50 or 100 different movies are enough to give us a good idea of their tastes. Or that the success of a left-handed pitcher against left-handed batters will be fairly consistent for different batters.

This points to an underlying assumption of simplicity. One form of simplicity is *sparsity*, the central theme of this book. Loosely speaking, a sparse statistical model is one in which only a relatively small number of parameters (or predictors) play an important role. In this book we study methods that exploit sparsity to help recover the underlying signal in a set of data.

The leading example is linear regression, in which we observe $N$ observations of an outcome variable $y_i$ and $p$ associated predictor variables (or features) $x_i = (x_{i1}, \ldots x_{ip})^T$. The goal is to predict the outcome from the

predictors, both for actual prediction with future data and also to discover which predictors play an important role. A linear regression model assumes that

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \tag{1.1}$$

where $\beta_0$ and $\beta = (\beta_1, \beta_2, \ldots \beta_p)$ are unknown parameters and $e_i$ is an error term. The method of least squares provides estimates of the parameters by minimization of the least-squares objective function

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_{ij})^2. \tag{1.2}$$

Typically all of the least-squares estimates from (1.2) will be nonzero. This will make interpretation of the final model challenging if $p$ is large. In fact, if $p > N$, the least-squares estimates are not unique. There is an infinite set of solutions that make the objective function equal to zero, and these solutions almost surely overfit the data as well.

Thus there is a need to constrain, or *regularize* the estimation process. In the *lasso* or $\ell_1$*-regularized regression*, we estimate the parameters by solving the problem

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_{ij})^2 \text{ subject to } \|\beta\|_1 \leq t \tag{1.3}$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the $\ell_1$ norm of $\beta$, and $t$ is a user-specified parameter. We can think of $t$ as a budget on the total $\ell_1$ norm of the parameter vector, and the lasso finds the best fit within this budget.

Why do we use the $\ell_1$ norm? Why not use the $\ell_2$ norm or any $\ell_q$ norm? It turns out that the $\ell_1$ norm is special. If the budget $t$ is small enough, the lasso yields sparse solution vectors, having only some coordinates that are nonzero. This does not occur for $\ell_q$ norms with $q > 1$; for $q < 1$, the solutions are sparse but the problem is not convex and this makes the minimization very challenging computationally. The value $q = 1$ is the smallest value that yields a convex problem. Convexity greatly simplifies the computation, as does the sparsity assumption itself. They allow for scalable algorithms that can handle problems with even millions of parameters.

Thus the advantages of sparsity are interpretation of the fitted model and computational convenience. But a third advantage has emerged in the last few years from some deep mathematical analyses of this area. This has been termed the "bet on sparsity" principle:

> *Use a procedure that does well in sparse problems, since no procedure does well in dense problems.*

We can think of this in terms of the amount of information $N/p$ per parameter. If $p \gg N$ and the true model is not sparse, then the number of samples $N$ is too small to allow for accurate estimation of the parameters. But if the true model is sparse, so that only $k < N$ parameters are actually nonzero in the true underlying model, then it turns out that we can estimate the parameters effectively, using the lasso and related methods that we discuss in this book. This may come as somewhat of a surprise, because we are able to do this even though we are not told *which* $k$ of the $p$ parameters are actually nonzero. Of course we cannot do as well as we could if we had that information, but it turns out that we can still do reasonably well.

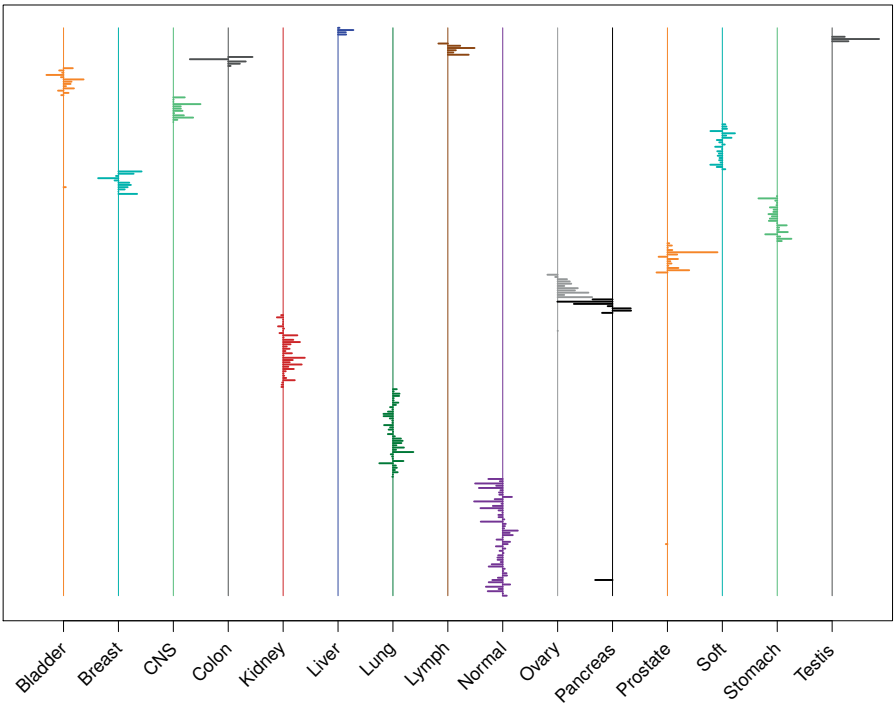



**Figure 1.1** *15-class gene expression cancer data: estimated nonzero feature weights from a lasso-regularized multinomial classifier. Shown are the 254 genes (out of 4718) with at least one nonzero weight among the 15 classes. The genes (unlabelled) run from top to bottom. Line segments pointing to the right indicate positive weights, and to the left, negative weights. We see that only a handful of genes are needed to characterize each class.*

For all of these reasons, the area of sparse statistical modelling is exciting—for data analysts, computer scientists, and theorists—and practically useful. Figure 1.1 shows an example. The data consists of quantitative gene expression measurements of 4718 genes on samples from 349 cancer patients. The cancers have been categorized into 15 different types such as "Bladder," "Breast",

"CNS," etc. The goal is to build a classifier to predict cancer class based on some or all of the 4718 features. We want the classifier to have a low error rate on independent samples and would prefer that it depend only on a subset of the genes, to aid in our understanding of the underlying biology.

For this purpose we applied a lasso-regularized multinomial classifier to these data, as described in Chapter 3. This produces a set of 4718 weights or coefficients for each of the 15 classes, for discriminating each class from the rest. Because of the $\ell_1$ penalty, only some of these weights may be nonzero (depending on the choice of the regularization parameter). We used cross-validation to estimate the optimal choice of regularization parameter, and display the resulting weights in Figure 1.1. Only 254 genes have at least one nonzero weight, and these are displayed in the figure. The cross-validated error rate for this classifier is about 10%, so the procedure correctly predicts the class of about 90% of the samples. By comparison, a standard support vector classifier had a slightly higher error rate (13%) using all of the features. Using sparsity, the lasso procedure has dramatically reduced the number of features without sacrificing accuracy. Sparsity has also brought computational efficiency: although there are potentially $4718 \times 15 \approx 70,000$ parameters to estimate, the entire calculation for Figure 1.1 was done on a standard laptop computer in less than a minute. For this computation we used the `glmnet` procedure described in Chapters 3 and 5.

Figure 1.2 shows another example taken from an article by Candès and Wakin (2008) in the field of *compressed sensing*. On the left is a megapixel image. In order to reduce the amount of space needed to store the image, we represent it in a wavelet basis, whose coefficients are shown in the middle panel. The largest $25,000$ coefficients are then retained and the rest zeroed out, yielding the excellent reconstruction in the right image. This all works because of sparsity: although the image seems complex, in the wavelet basis it is simple and hence only a relatively small number of coefficients are nonzero. The original image can be perfectly recovered from just $96,000$ incoherent measurements. Compressed sensing is a powerful tool for image analysis, and is described in Chapter 10.

In this book we have tried to summarize the hot and rapidly evolving field of sparse statistical modelling. In Chapter 2 we describe and illustrate the lasso for linear regression, and a simple coordinate descent algorithm for its computation. Chapter 3 covers the application of $\ell_1$ penalties to generalized linear models such as multinomial and survival models, as well as support vector machines. Generalized penalties such as the elastic net and group lasso are discussed in Chapter 4. Chapter 5 reviews numerical methods for optimization, with an emphasis on first-order methods that are useful for the large-scale problems that are discussed in this book. In Chapter 6, we discuss methods for statistical inference for fitted (lasso) models, including the bootstrap, Bayesian methods and some more recently developed approaches. Sparse matrix decomposition is the topic of Chapter 7, and we apply these methods in the context of sparse multivariate analysis in Chapter 8. Graph-

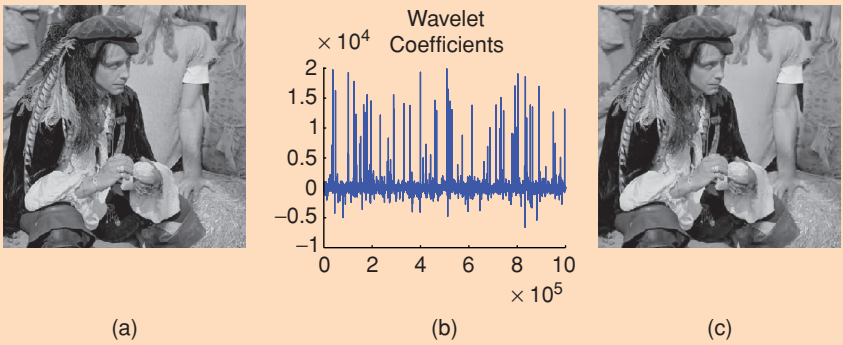



**Figure 1.2** *(a) Original megapixel image with pixel values in the range* $[0, 255]$ *and (b) its wavelet transform coefficients (arranged in random order for enhanced visibility). Relatively few wavelet coefficients capture most of the signal energy; many such images are highly compressible. (c) The reconstruction obtained by zeroing out all the coefficients in the wavelet expansion but the* $25,000$ *largest (pixel values are thresholded to the range* $[0, 255]$*). The differences from the original picture are hardly noticeable.*

ical models and their selection are discussed in Chapter 9 while compressed sensing is the topic of Chapter 10. Finally, a survey of theoretical results for the lasso is given in Chapter 11.

We note that both *supervised* and *unsupervised* learning problems are discussed in this book, the former in Chapters 2, 3, 4, and 10, and the latter in Chapters 7 and 8.

### *Notation*

We have adopted a notation to reduce mathematical clutter. Vectors are column vectors by default; hence $\beta \in \mathbb{R}^p$ is a column vector, and its transpose $\beta^T$ is a row vector. All vectors are lower case and non-bold, except $N$-vectors which are bold, where $N$ is the sample size. For example $\mathbf{x}_j$ might be the $N$-vector of observed values for the $j^{th}$ variable, and $\mathbf{y}$ the response $N$-vector. All matrices are bold; hence $\mathbf{X}$ might represent the $N \times p$ matrix of observed predictors, and $\mathbf{\Theta}$ a $p \times p$ precision matrix. This allows us to use $x_i \in \mathbb{R}^p$ to represent the vector of $p$ features for observation $i$ (i.e., $x_i^T$ is the $i^{th}$ row of $\mathbf{X}$), while $\mathbf{x}_k$ is the $k^{th}$ column of $\mathbf{X}$, without ambiguity.

Chapter 2

# The Lasso for Linear Models

In this chapter, we introduce the lasso estimator for linear regression. We describe the basic lasso method, and outline a simple approach for its implementation. We relate the lasso to ridge regression, and also view it as a Bayesian estimator.

## 2.1 Introduction

In the linear regression setting, we are given $N$ samples $\{(x_i, y_i)\}_{i=1}^N$, where each $x_i = (x_{i1}, \ldots, x_{ip})$ is a $p$-dimensional vector of features or predictors, and each $y_i \in \mathbb{R}$ is the associated response variable. Our goal is to approximate the response variable $y_i$ using a linear combination of the predictors

$$\eta(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j. \tag{2.1}$$

The model is parametrized by the vector of regression weights $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ and an intercept (or "bias") term $\beta_0 \in \mathbb{R}$.

The usual "least-squares" estimator for the pair $(\beta_0, \beta)$ is based on minimizing squared-error loss:

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \right\}. \tag{2.2}$$

There are two reasons why we might consider an alternative to the least-squares estimate. The first reason is *prediction accuracy*: the least-squares estimate often has low bias but large variance, and prediction accuracy can sometimes be improved by shrinking the values of the regression coefficients, or setting some coefficients to zero. By doing so, we introduce some bias but reduce the variance of the predicted values, and hence may improve the overall prediction accuracy (as measured in terms of the mean-squared error). The second reason is for the purposes of *interpretation*. With a large number of predictors, we often would like to identify a smaller subset of these predictors that exhibit the strongest effects.

This chapter is devoted to discussion of the *lasso*, a method that combines the least-squares loss (2.2) with an $\ell_1$-constraint, or bound on the sum of the absolute values of the coefficients. Relative to the least-squares solution, this constraint has the effect of shrinking the coefficients, and even setting some to zero.[1] In this way it provides an automatic way for doing model selection in linear regression. Moreover, unlike some other criteria for model selection, the resulting optimization problem is convex, and can be solved efficiently for large problems.

## 2.2 The Lasso Estimator

Given a collection of $N$ predictor-response pairs $\{(x_i, y_i)\}_{i=1}^N$, the lasso finds the solution $(\widehat{\beta}_0, \widehat{\beta})$ to the optimization problem

$$\begin{aligned} &\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t. \end{aligned} \tag{2.3}$$

The constraint $\sum_{j=1}^p |\beta_j| \le t$ can be written more compactly as the $\ell_1$-norm constraint $\|\beta\|_1 \le t$. Furthermore, (2.3) is often represented using matrix-vector notation. Let $\mathbf{y} = (y_1, \ldots, y_N)$ denote the $N$-vector of responses, and $\mathbf{X}$ be an $N \times p$ matrix with $x_i \in \mathbb{R}^p$ in its $i^{th}$ row, then the optimization problem (2.3) can be re-expressed as

$$\begin{aligned} &\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{2N} \|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\beta\|_2^2 \right\} \\ &\text{subject to } \|\beta\|_1 \le t, \end{aligned} \tag{2.4}$$

where $\mathbf{1}$ is the vector of $N$ ones, and $\|\cdot\|_2$ denotes the usual Euclidean norm on vectors. The bound $t$ is a kind of "budget": it limits the sum of the absolute values of the parameter estimates. Since a shrunken parameter estimate corresponds to a more heavily-constrained model, this budget limits how well we can fit the data. It must be specified by an external procedure such as cross-validation, which we discuss later in the chapter.

Typically, we first standardize the predictors $\mathbf{X}$ so that each column is centered ($\frac{1}{N}\sum_{i=1}^N x_{ij} = 0$) and has unit variance ($\frac{1}{N}\sum_{i=1}^N x_{ij}^2 = 1$). Without

[1] A *lasso* is a long rope with a noose at one end, used to catch horses and cattle. In a figurative sense, the method "lassos" the coefficients of the model. In the original lasso paper (Tibshirani 1996), the name "lasso" was also introduced as an acronym for "Least Absolute Selection and Shrinkage Operator."
Pronunciation: in the US "lasso" tends to be pronounced "lass-oh" (oh as in goat), while in the UK "lass-oo." In the OED (2nd edition, 1965): "lasso is pronounced lăsoo by those who use it, and by most English people too."

standardization, the lasso solutions would depend on the units (e.g., feet versus meters) used to measure the predictors. On the other hand, we typically would not standardize if the features were measured in the same units. For convenience, we also assume that the outcome values $y_i$ have been centered, meaning that $\frac{1}{N}\sum_{i=1}^{N} y_i = 0$. These centering conditions are convenient, since they mean that we can omit the intercept term $\beta_0$ in the lasso optimization. Given an optimal lasso solution $\widehat{\beta}$ on the centered data, we can recover the optimal solutions for the uncentered data: $\widehat{\beta}$ is the same, and the intercept $\widehat{\beta}_0$ is given by

$$\widehat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \widehat{\beta}_j,$$

where $\bar{y}$ and $\{\bar{x}_j\}_1^p$ are the original means.[2] For this reason, we omit the intercept $\beta_0$ from the lasso for the remainder of this chapter.

It is often convenient to rewrite the lasso problem in the so-called Lagrangian form

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{2.5}$$

for some $\lambda \geq 0$. By Lagrangian duality, there is a one-to-one correspondence between the constrained problem (2.3) and the Lagrangian form (2.5): for each value of $t$ in the range where the constraint $\|\beta\|_1 \leq t$ is active, there is a corresponding value of $\lambda$ that yields the same solution from the Lagrangian form (2.5). Conversely, the solution $\widehat{\beta}_\lambda$ to problem (2.5) solves the bound problem with $t = \|\widehat{\beta}_\lambda\|_1$.

We note that in many descriptions of the lasso, the factor $1/2N$ appearing in (2.3) and (2.5) is replaced by $1/2$ or simply 1. Although this makes no difference in (2.3), and corresponds to a simple reparametrization of $\lambda$ in (2.5), this kind of standardization makes $\lambda$ values comparable for different sample sizes (useful for cross-validation).

The theory of convex analysis tells us that necessary and sufficient conditions for a solution to problem (2.5) take the form

$$-\frac{1}{N}\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\beta \rangle + \lambda s_j = 0, \ j = 1, \ldots, p. \tag{2.6}$$

Here each $s_j$ is an unknown quantity equal to $\text{sign}(\beta_j)$ if $\beta_j \neq 0$ and some value lying in $[-1, 1]$ otherwise—that is, it is a subgradient for the absolute value function (see Chapter 5 for details). In other words, the solutions $\hat{\beta}$ to problem (2.5) are the same as solutions $(\hat{\beta}, \hat{s})$ to (2.6). This system is a form of the so-called Karush–Kuhn–Tucker (KKT) conditions for problem (2.5). Expressing a problem in subgradient form can be useful for designing

[2]This is typically only true for linear regression with squared-error loss; it's not true, for example, for lasso logistic regression.

algorithms for finding its solutions. More details are given in Exercises (2.3) and (2.4).

As an example of the lasso, let us consider the data given in Table 2.1, taken from Thomas (1990). The outcome is the total overall reported crime rate per

**Table 2.1** *Crime data: Crime rate and five predictors, for $N = 50$ U.S. cities.*

| city | funding | hs | not-hs | college | college4 | crime rate |
|---|---|---|---|---|---|---|
| 1 | 40 | 74 | 11 | 31 | 20 | 478 |
| 2 | 32 | 72 | 11 | 43 | 18 | 494 |
| 3 | 57 | 70 | 18 | 16 | 16 | 643 |
| 4 | 31 | 71 | 11 | 25 | 19 | 341 |
| 5 | 67 | 72 | 9 | 29 | 24 | 773 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | | |
| 50 | 66 | 67 | 26 | 18 | 16 | 940 |

one million residents in 50 U.S cities. There are five predictors: annual police funding in dollars per resident, percent of people 25 years and older with four years of high school, percent of 16- to 19-year olds not in high school and not high school graduates, percent of 18- to 24-year olds in college, and percent of people 25 years and older with at least four years of college. This small example is for illustration only, but helps to demonstrate the nature of the lasso solutions. Typically the lasso is most useful for much larger problems, including "wide" data for which $p \gg N$.

The left panel of Figure 2.1 shows the result of applying the lasso with the bound $t$ varying from zero on the left, all the way to a large value on the right, where it has no effect. The horizontal axis has been scaled so that the maximal bound, corresponding to the least-squares estimates $\tilde{\beta}$, is one. We see that for much of the range of the bound, many of the estimates are exactly zero and hence the corresponding predictor(s) would be excluded from the model. Why does the lasso have this model selection property? It is due to the geometry that underlies the $\ell_1$ constraint $\|\beta\|_1 \leq t$. To understand this better, the right panel shows the estimates from *ridge regression*, a technique that predates the lasso. It solves a criterion very similar to (2.3):

$$\begin{aligned} &\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \\ &\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t^2. \end{aligned} \tag{2.7}$$

The ridge profiles in the right panel have roughly the same shape as the lasso profiles, but are not equal to zero except at the left end. Figure 2.2 contrasts the two constraints used in the lasso and ridge regression. The residual sum

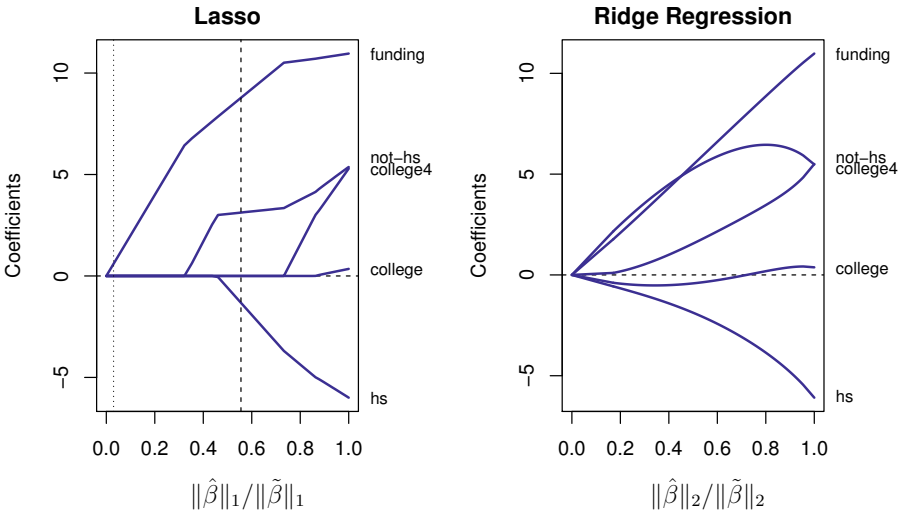



**Figure 2.1** *Left: Coefficient path for the lasso, plotted versus the $\ell_1$ norm of the coefficient vector, relative to the norm of the unrestricted least-squares estimate $\tilde{\beta}$. Right: Same for ridge regression, plotted against the relative $\ell_2$ norm.*

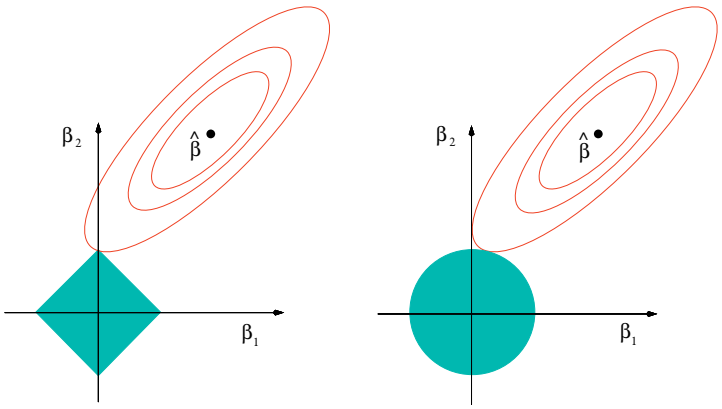



**Figure 2.2** *Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions $|\beta_1|+|\beta_2| \leq t$ and $\beta_1^2+\beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point $\widehat{\beta}$ depicts the usual (unconstrained) least-squares estimate.*

**Table 2.2** *Results from analysis of the crime data. Left panel shows the least-squares estimates, standard errors, and their ratio (Z-score). Middle and right panels show the corresponding results for the lasso, and the least-squares estimates applied to the subset of predictors chosen by the lasso.*

| | LS coef | SE | Z | Lasso | SE | Z | LS | SE | Z |
|---|---|---|---|---|---|---|---|---|---|
| `funding` | 10.98 | 3.08 | 3.6 | 8.84 | 3.55 | 2.5 | 11.29 | 2.90 | 3.9 |
| `hs` | -6.09 | 6.54 | -0.9 | -1.41 | 3.73 | -0.4 | -4.76 | 4.53 | -1.1 |
| `not-hs` | 5.48 | 10.05 | 0.5 | 3.12 | 5.05 | 0.6 | 3.44 | 7.83 | 0.4 |
| `college` | 0.38 | 4.42 | 0.1 | 0.0 | - | - | 0.0 | - | - |
| `college4` | 5.50 | 13.75 | 0.4 | 0.0 | - | - | 0.0 | - | - |

of squares has elliptical contours, centered at the full least-squares estimates. The constraint region for ridge regression is the disk $\beta_1^2 + \beta_2^2 \le t^2$, while that for lasso is the diamond $|\beta_1| + |\beta_2| \le t$. Both methods find the first point where the elliptical contours hit the constraint region. Unlike the disk, the diamond has corners; if the solution occurs at a corner, then it has one parameter $\beta_j$ equal to zero. When $p > 2$, the diamond becomes a rhomboid, and has many corners, flat edges, and faces; there are many more opportunities for the estimated parameters to be zero (see Figure 4.2 on page 58.)

We use the term *sparse* for a model with few nonzero coefficients. Hence a key property of the $\ell_1$-constraint is its ability to yield sparse solutions. This idea can be applied in many different statistical models, and is the central theme of this book.

Table 2.2 shows the results of applying three fitting procedures to the crime data. The lasso bound $t$ was chosen by cross-validation, as described in Section 2.3. The left panel corresponds to the full least-squares fit, while the middle panel shows the lasso fit. On the right, we have applied least-squares estimation to the subset of three predictors with nonzero coefficients in the lasso. The standard errors for the least-squares estimates come from the usual formulas. No such simple formula exists for the lasso, so we have used the bootstrap to obtain the estimate of standard errors in the middle panel (see Exercise 2.6; Chapter 6 discusses some promising new approaches for post-selection inference). Overall it appears that `funding` has a large effect, probably indicating that police resources have been focused on higher crime areas. The other predictors have small to moderate effects.

Note that the lasso sets two of the five coefficients to zero, and tends to shrink the coefficients of the others toward zero relative to the full least-squares estimate. In turn, the least-squares fit on the subset of the three predictors tends to expand the lasso estimates away from zero. The nonzero estimates from the lasso tend to be biased toward zero, so the debiasing in the right panel can often improve the prediction error of the model. This two-stage process is also known as the *relaxed lasso* (Meinshausen 2007).

## 2.3 Cross-Validation and Inference

The bound $t$ in the lasso criterion (2.3) controls the complexity of the model; larger values of $t$ free up more parameters and allow the model to adapt more closely to the training data. Conversely, smaller values of $t$ restrict the parameters more, leading to sparser, more interpretable models that fit the data less closely. Forgetting about interpretability, we can ask for the value of $t$ that gives the most accurate model for predicting independent test data from the same population. Such accuracy is called the *generalization* ability of the model. A value of $t$ that is too small can prevent the lasso from capturing the main signal in the data, while too large a value can lead to overfitting. In this latter case, the model adapts to the noise as well as the signal that is present in the training data. In both cases, the prediction error on a test set will be inflated. There is usually an intermediate value of $t$ that strikes a good balance between these two extremes, and in the process, produces a model with some coefficients equal to zero.

In order to estimate this best value for $t$, we can create artificial training and test sets by splitting up the given dataset at random, and estimating performance on the test data, using a procedure known as *cross-validation*. In more detail, we first randomly divide the full dataset into some number of groups $K > 1$. Typical choices of $K$ might be 5 or 10, and sometimes $N$. We fix one group as the test set, and designate the remaining $K - 1$ groups as the training set. We then apply the lasso to the training data for a range of different $t$ values, and we use each fitted model to predict the responses in the test set, recording the mean-squared prediction errors for each value of $t$. This process is repeated a total of $K$ times, with each of the $K$ groups getting the chance to play the role of the test data, with the remaining $K-1$ groups used as training data. In this way, we obtain $K$ different estimates of the prediction error over a range of values of $t$. These $K$ estimates of prediction error are averaged for each value of $t$, thereby producing a *cross-validation error curve*.

Figure 2.3 shows the cross-validation error curve for the crime-data example, obtained using $K = 10$ splits. We plot the estimated mean-squared prediction error versus the relative bound $\tilde{t} = \|\widehat{\beta}(t)\|_1/\|\tilde{\beta}\|_1$, where the estimate $\widehat{\beta}(t)$ correspond to the lasso solution for bound $t$ and $\tilde{\beta}$ is the ordinary least-squares solution. The error bars in Figure 2.3 indicate plus and minus one standard error in the cross-validated estimates of the prediction error. A vertical dashed line is drawn at the position of the minimum ($\tilde{t} = 0.56$) while a dotted line is drawn at the "one-standard-error rule" choice ($\tilde{t} = 0.03$). This is the smallest value of $t$ yielding a CV error no more than one standard error above its minimum value. The number of nonzero coefficients in each model is shown along the top. Hence the model that minimizes the CV error has three predictors, while the one-standard-error-rule model has just one.

We note that the cross-validation process above focused on the bound parameter $t$. One can just as well carry out cross-validation in the Lagrangian

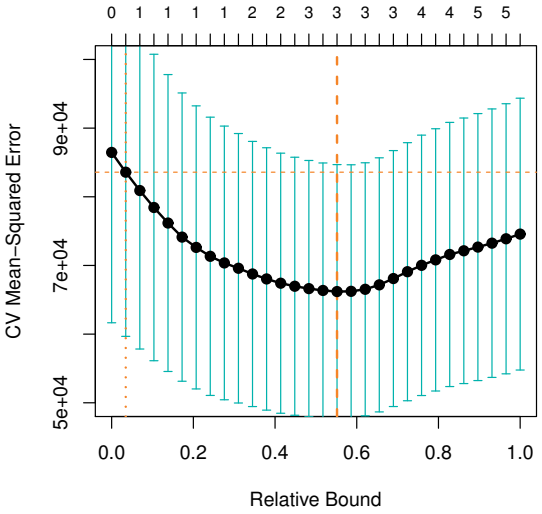



**Figure 2.3** *Cross-validated estimate of mean-squared prediction error, as a function of the relative $\ell_1$ bound $\tilde{t} = \|\widehat{\beta}(t)\|_1/\|\tilde{\beta}\|_1$. Here $\widehat{\beta}(t)$ is the lasso estimate corresponding to the $\ell_1$ bound $t$ and $\tilde{\beta}$ is the ordinary least-squares solution. Included are the location of the minimum, pointwise standard-error bands, and the "one-standard-error" location. The standard errors are large since the sample size $N$ is only 50.*

form (2.5), focusing on the parameter $\lambda$. The two methods will give similar but not identical results, since the mapping between $t$ and $\lambda$ is data-dependent.

## 2.4 Computation of the Lasso Solution

The lasso problem is a convex program, specifically a quadratic program (QP) with a convex constraint. As such, there are many sophisticated QP methods for solving the lasso. However there is a particularly simple and effective computational algorithm, that gives insight into how the lasso works. For convenience, we rewrite the criterion in Lagrangian form:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \right\}. \tag{2.8}$$

As before, we will assume that both $y_i$ and the features $x_{ij}$ have been standardized so that $\frac{1}{N}\sum_i y_i = 0$, $\frac{1}{N}\sum_i x_{ij} = 0$, and $\frac{1}{N}\sum_i x_{ij}^2 = 1$. In this case, the intercept term $\beta_0$ can be omitted. The Lagrangian form is especially convenient for numerical computation of the solution by a simple procedure known as *coordinate descent*.

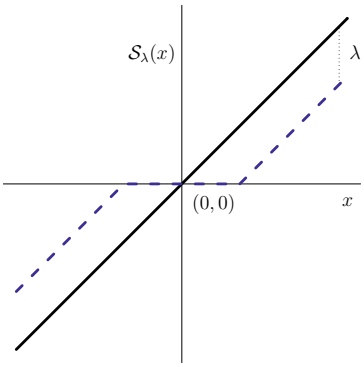



**Figure 2.4** *Soft thresholding function* $\mathcal{S}_\lambda(x) = \text{sign}(x)\,(|x| - \lambda)_+$ *is shown in blue (broken lines), along with the* 45° *line in black.*

### *2.4.1 Single Predictor: Soft Thresholding*

Let's first consider a single predictor setting, based on samples $\{(z_i, y_i)\}_{i=1}^N$ (for convenience we have renamed $z_i$ to be one the $x_{ij}$). The problem then is to solve

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - z_i\beta)^2 + \lambda|\beta| \right\}. \tag{2.9}$$

The standard approach to this univariate minimization problem would be to take the gradient (first derivative) with respect to $\beta$, and set it to zero. There is a complication, however, because the absolute value function $|\beta|$ does not have a derivative at $\beta = 0$. However we can proceed by direct inspection of the function (2.9), and find that

$$\widehat{\beta} = \begin{cases} \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle - \lambda & \text{if } \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle > \lambda, \\ 0 & \text{if } \frac{1}{N}|\langle \mathbf{z}, \mathbf{y}\rangle| \leq \lambda, \\ \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle + \lambda & \text{if } \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle < -\lambda. \end{cases} \tag{2.10}$$

(Exercise 2.2), which we can write succinctly as

$$\widehat{\beta} = \mathcal{S}_\lambda\big(\tfrac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle\big). \tag{2.11}$$

Here the *soft-thresholding operator*

$$\mathcal{S}_\lambda(x) = \text{sign}(x)\,\big(|x| - \lambda\big)_+ \tag{2.12}$$

translates its argument $x$ toward zero by the amount $\lambda$, and sets it to zero if $|x| \leq \lambda$.[3] See Figure 2.4 for an illustration. Notice that for standardized data with $\frac{1}{N}\sum_i z_i^2 = 1$, (2.11) is just a soft-thresholded version of the usual least-squares estimate $\tilde{\beta} = \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle$. One can also derive these results using the notion of subgradients (Exercise 2.3).

[3] $t_+$ denotes the positive part of $t \in \mathbb{R}$, equal to $t$ if $t > 0$ and 0 otherwise.

### 2.4.2 Multiple Predictors: Cyclic Coordinate Descent

Using this intuition from the univariate case, we can now develop a simple coordinatewise scheme for solving the full lasso problem (2.5). More precisely, we repeatedly cycle through the predictors in some fixed (but arbitrary) order (say $j = 1, 2, \ldots, p$), where at the $j^{th}$ step, we update the coefficient $\beta_j$ by minimizing the objective function in this coordinate while holding fixed all other coefficients $\{\widehat{\beta}_k, k \neq j\}$ at their current values.

Writing the objective in (2.5) as

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|, \tag{2.13}$$

we see that solution for each $\beta_j$ can be expressed succinctly in terms of the *partial residual* $r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\widehat{\beta}_k$, which removes from the outcome the current fit from all but the $j^{th}$ predictor. In terms of this partial residual, the $j^{th}$ coefficient is updated as

$$\widehat{\beta}_j = \mathcal{S}_\lambda\Big(\tfrac{1}{N}\langle \mathbf{x}_j, \boldsymbol{r}^{(j)}\rangle\Big). \tag{2.14}$$

Equivalently, the update can be written as

$$\widehat{\beta}_j \leftarrow \mathcal{S}_\lambda\Big(\widehat{\beta}_j + \tfrac{1}{N}\langle \mathbf{x}_j, \boldsymbol{r}\rangle\Big), \tag{2.15}$$

where $r_i = y_i - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j$ are the full residuals (Exercise 2.4). The overall algorithm operates by applying this soft-thresholding update (2.14) repeatedly in a cyclical manner, updating the coordinates of $\widehat{\beta}$ (and hence the residual vectors) along the way.

Why does this algorithm work? The criterion (2.5) is a convex function of $\beta$ and so has no local minima. The algorithm just described corresponds to the method of *cyclical coordinate descent*, which minimizes this convex objective along each coordinate at a time. Under relatively mild conditions (which apply here), such coordinate-wise minimization schemes applied to a convex function converge to a global optimum. It is important to note that some conditions are required, because there are instances, involving nonseparable penalty functions, in which coordinate descent schemes can become "jammed." Further details are in given in Chapter 5.

Note that the choice $\lambda = 0$ in (2.5) delivers the solution to the ordinary least-squares problem. From the update (2.14), we see that the algorithm does a univariate regression of the partial residual onto each predictor, cycling through the predictors until convergence. When the data matrix $\mathbf{X}$ is of full rank, this point of convergence is the least-squares solution. However, it is not a particularly efficient method for computing it.

In practice, one is often interested in finding the lasso solution not just for a single fixed value of $\lambda$, but rather the entire path of solutions over a range

of possible $\lambda$ values (as in Figure 2.1). A reasonable method for doing so is to begin with a value of $\lambda$ just large enough so that the only optimal solution is the all-zeroes vector. As shown in Exercise 2.1, this value is equal to $\lambda_{max} = \max_j |\frac{1}{N}\langle \mathbf{x}_j, \mathbf{y}\rangle|$. Then we decrease $\lambda$ by a small amount and run coordinate descent until convergence. Decreasing $\lambda$ again and using the previous solution as a "warm start," we then run coordinate descent until convergence. In this way we can efficiently compute the solutions over a grid of $\lambda$ values. We refer to this method as *pathwise coordinate descent*.

Coordinate descent is especially fast for the lasso because the coordinate-wise minimizers are explicitly available (Equation (2.14)), and thus an iterative search along each coordinate is not needed. Secondly, it exploits the sparsity of the problem: for large enough values of $\lambda$ most coefficients will be zero and will not be moved from zero. In Section 5.4, we discuss computational hedges for guessing the active set, which speed up the algorithm dramatically.

*Homotopy methods* are another class of techniques for solving the lasso. They produce the entire path of solutions in a sequential fashion, starting at zero. This path is actually piecewise linear, as can be seen in Figure 2.1 (as a function of $t$ or $\lambda$). The *least angle regression* (LARS) algorithm is a homotopy method that efficiently constructs the piecewise linear path, and is described in Chapter 5.

### *2.4.3 Soft-Thresholding and Orthogonal Bases*

The soft-thresholding operator plays a central role in the lasso and also in signal denoising. To see this, notice that the coordinate minimization scheme above takes an especially simple form if the predictors are orthogonal, meaning that $\frac{1}{N}\langle \mathbf{x}_j, \mathbf{x}_k\rangle = 0$ for each $j \neq k$. In this case, the update (2.14) simplifies dramatically, since $\frac{1}{N}\langle \mathbf{x}_j, \boldsymbol{r}^{(j)}\rangle = \frac{1}{N}\langle \mathbf{x}_j, \mathbf{y}\rangle$ so that $\widehat{\beta}_j$ is simply the soft-thresholded version of the univariate least-squares estimate of $\mathbf{y}$ regressed against $\mathbf{x}_j$. Thus, in the special case of an orthogonal design, the lasso has an explicit closed-form solution, and no iterations are required.

Wavelets are a popular form of orthogonal bases, used for smoothing and compression of signals and images. In wavelet smoothing one represents the data in a wavelet basis, and then denoises by soft-thresholding the wavelet coefficients. We discuss this further in Section 2.10 and in Chapter 10.

## 2.5 Degrees of Freedom

Suppose we have $p$ predictors, and fit a linear regression model using only a subset of $k$ of these predictors. Then if these $k$ predictors were chosen without regard to the response variable, the fitting procedure "spends" $k$ degrees of freedom. This is a loose way of saying that the standard test statistic for testing the hypothesis that all $k$ coefficients are zero has a Chi-squared distribution with $k$ degrees of freedom (with the error variance $\sigma^2$ assumed to be known)

However if the $k$ predictors were chosen using knowledge of the response variable, for example to yield the smallest training error among all subsets of size $k$, then we would expect that the fitting procedure spends more than $k$ degrees of freedom. We call such a fitting procedure *adaptive*, and clearly the lasso is an example of one.

Similarly, a forward-stepwise procedure in which we sequentially add the predictor that most decreases the training error is adaptive, and we would expect that the resulting model uses more than $k$ degrees of freedom after $k$ steps. For these reasons and in general, one cannot simply count as degrees of freedom the number of nonzero coefficients in the fitted model. However, it turns out that for the lasso, one *can* count degrees of freedom by the number of nonzero coefficients, as we now describe.

First we need to define precisely what we mean by the degrees of freedom of an adaptively fitted model. Suppose we have an additive-error model, with

$$y_i = f(x_i) + \epsilon_i,\ i = 1, \ldots, N, \tag{2.16}$$

for some unknown $f$ and with the errors $\epsilon_i$ iid $(0, \sigma^2)$. If the $N$ sample predictions are denoted by $\widehat{\mathbf{y}}$, then we define

$$\mathrm{df}(\widehat{\mathbf{y}}) := \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathrm{Cov}\left(\widehat{y}_i, y_i\right). \tag{2.17}$$

The covariance here is taken over the randomness in the response variables $\{y_i\}_{i=1}^N$ with the predictors held fixed. Thus, the degrees of freedom corresponds to the total amount of *self-influence* that each response measurement has on its prediction. The more the model fits—that is, adapts—to the data, the larger the degrees of freedom. In the case of a fixed linear model, using $k$ predictors chosen independently of the response variable, it is easy to show that $\mathrm{df}(\widehat{\mathbf{y}}) = k$ (Exercise 2.7). However, under adaptive fitting, it is typically the case that the degrees of freedom is larger than $k$.

Somewhat miraculously, one can show that for the lasso, with a fixed penalty parameter $\lambda$, the number of nonzero coefficients $k_\lambda$ is an unbiased estimate of the degrees of freedom[4] (Zou, Hastie and Tibshirani 2007, Tibshirani$_2$ and Taylor 2012). As discussed earlier, a variable-selection method like forward-stepwise regression uses more than $k$ degrees of freedom after $k$ steps. Given the apparent similarity between forward-stepwise regression and the lasso, how can the lasso have this simple degrees of freedom property? The reason is that the lasso not only selects predictors (which inflates the degrees of freedom), but also shrinks their coefficients toward zero, relative to the usual least-squares estimates. This shrinkage turns out to be just the right

[4] An even stronger statement holds for the LAR path, where the degrees of freedom after $k$ steps is exactly $k$, under some conditions on $\mathbf{X}$. The LAR path relates closely to the lasso, and is described in Section 5.6.

amount to bring the degrees of freedom down to $k$. This result is useful because it gives us a qualitative measure of the amount of fitting that we have done at any point along the lasso path.

In the general setting, a proof of this result is quite difficult. In the special case of an orthogonal design, it is relatively easy to prove, using the fact that the lasso estimates are simply soft-thresholded versions of the univariate regression coefficients for the orthogonal design. We explore the details of this argument in Exercise 2.8. This idea is taken one step further in Section 6.3.1 where we describe the *covariance test* for testing the significance of predictors in the context of the lasso.

## 2.6 Uniqueness of the Lasso Solutions

We first note that the theory of convex duality can be used to show that when the columns of $\mathbf{X}$ are in general position, then for $\lambda > 0$ the solution to the lasso problem (2.5) is unique. This holds even when $p \geq N$, although then the number of nonzero coefficients in any lasso solution is at most $N$ (Rosset, Zhu and Hastie 2004, Tibshirani$_2$ 2013). Now when the predictor matrix $\mathbf{X}$ is not of full column rank, the least squares fitted values are unique, but the parameter estimates themselves are not. The non-full-rank case can occur when $p \leq N$ due to collinearity, and always occurs when $p > N$. In the latter scenario, there are an infinite number of solutions $\widehat{\beta}$ that yield a perfect fit with zero training error. Now consider the lasso problem in Lagrange form (2.5) for $\lambda > 0$. As shown in Exercise 2.5, the fitted values $\mathbf{X}\widehat{\beta}$ are unique. But it turns out that the solution $\widehat{\beta}$ may not be unique. Consider a simple example with two predictors $\mathbf{x}_1$ and $\mathbf{x}_2$ and response $\mathbf{y}$, and suppose the lasso solution coefficients $\widehat{\beta}$ at $\lambda$ are $(\widehat{\beta}_1,\ \widehat{\beta}_2)$. If we now include a third predictor $\mathbf{x}_3 = \mathbf{x}_2$ into the mix, an identical copy of the second, then for any $\alpha \in [0, 1]$, the vector $\tilde{\beta}(\alpha) = (\widehat{\beta}_1,\ \alpha \cdot \widehat{\beta}_2,\ (1-\alpha) \cdot \widehat{\beta}_2)$ produces an identical fit, and has $\ell_1$ norm $\|\tilde{\beta}(\alpha)\|_1 = \|\widehat{\beta}\|_1$. Consequently, for this model (in which we might have either $p \leq N$ or $p > N$), there is an infinite family of solutions.

In general, when $\lambda > 0$, one can show that if the columns of the model matrix $\mathbf{X}$ are in *general position*, then the lasso solutions are unique. To be precise, we say the columns $\{\mathbf{x}_j\}_{j=1}^p$ are in general position if any affine subspace $\mathbb{L} \subset \mathbb{R}^N$ of dimension $k < N$ contains at most $k + 1$ elements of the set $\{\pm\mathbf{x}_1, \pm\mathbf{x}_2, \ldots \pm \mathbf{x}_p\}$, excluding antipodal pairs of points (that is, points differing only by a sign flip). We note that the data in the example in the previous paragraph are not in general position. If the $X$ data are drawn from a continuous probability distribution, then with probability one the data are in general position and hence the lasso solutions will be unique. As a result, non-uniqueness of the lasso solutions can only occur with discrete-valued data, such as those arising from dummy-value coding of categorical predictors. These results have appeared in various forms in the literature, with a summary given by Tibshirani$_2$ (2013).

We note that numerical algorithms for computing solutions to the lasso will