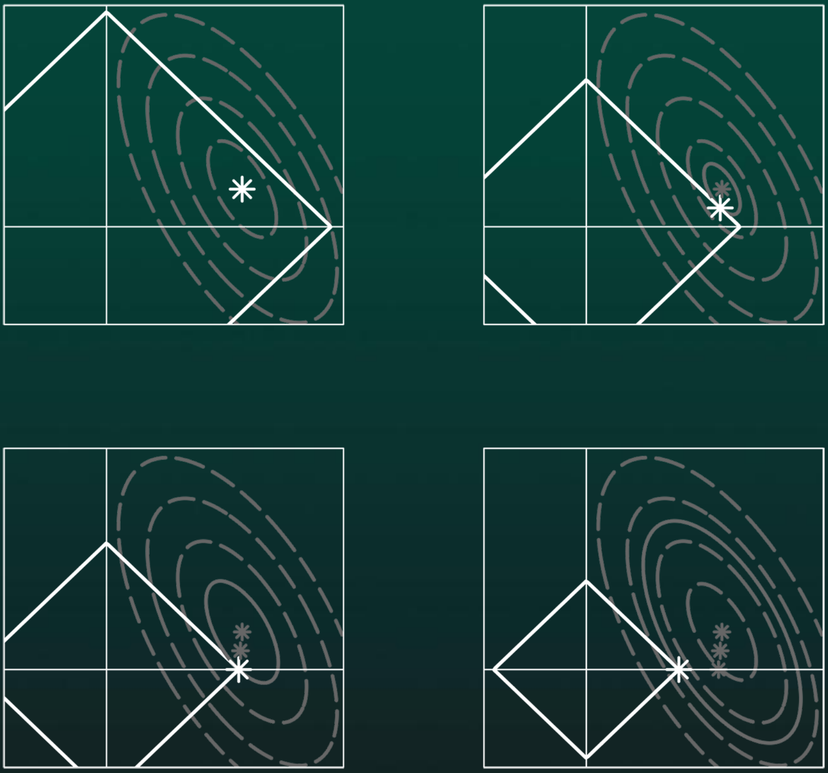
Monographs on Statistics and Applied Probability 139

Introduction to High-Dimensional Statistics



Christophe Giraud

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Introduction to High-Dimensional Statistics

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

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)

Monographs on Statistics and Applied Probability 139

Introduction to High-Dimensional Statistics

Christophe Giraud

Université Paris-Sud

France

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

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

Version Date: 20141107

International Standard Book Number-13: 978-1-4822-3795-5 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

Over the last twenty years (or so), the dramatic development of data acquisition technologies has enabled devices able to take thousands (up to millions) of measurements simultaneously. The data produced by such wide-scale devices are said to be *high dimensional*. They can be met in almost any branch of human activities, including medicine (biotech data, medical imaging, etc.), basic sciences (astrophysics, environmental sciences, etc.), e-commerce (tracking, loyalty programs, etc.), finance, cooperative activities (crowdsourcing data), etc. Having access to such massive data sounds like a blessing. Unfortunately, the analysis of high-dimensional data is extremely challenging. Indeed, separating the useful information from the noise is *generally* almost impossible in high-dimensional settings. This issue is often referred as the *curse of dimensionality*.

Most of the *classical* statistics developed during the 20th century focused on data where the number n of *experimental units* (number of individuals in a medical cohort, number of experiments in biology or physics, etc.) was large compared to the number p of *unknown features*. Accordingly, most of the *classical* statistical theory provides results for the asymptotic setting where p is fixed and n goes to infinity. This theory is very insightful for analyzing data where "n is large" and "p is small," but it can be seriously misleading for modern high-dimensional data. Analyzing "large p" data then requires some new statistics. It has given rise to a huge effort from the statistical and data analyst community for developing new tools able to circumvent the curse of dimensionality. In particular, building on the concept of *sparsity* has shown to be successful in this setting.

This book is an introduction to the mathematical foundations of high-dimensional statistics. It is intended to be a concise guide for students and researchers unfamiliar with the area, and interested in the mathematics involved. In particular, this book is not conceived as a comprehensive catalog of statistical methods for high-dimensional data. It is based on lectures given in the master programs "Mathematics for Life Sciences" and "Data Sciences," from Paris Sud University (Orsay), Ecole Polytechnique (Palaiseau), Ecole Normale Supérieure de Cachan (Cachan), and Telecom ParisTech (Paris). The primary goal is to explain, as simply as possible, the main concepts and ideas on some selected topics of high-dimensional statistics. The focus is mainly on some simple settings, avoiding, thereby, unessential technicalities that could blur the main arguments. To achieve this goal, the book includes significantly streamlined proofs issued from the recent research literature.

Each chapter of the book ends with some exercises, and the reader is invited to share his solutions on the wiki-site
`http://high-dimensional-statistics.wikidot.com`.

Finally, I apologize for any remaining typos and errors (despite my efforts to remove them). The reader is invited to point them out at
`high.dimensional.statistics@gmail.com`
for errata that will be published on the book's website,
`http://sites.google.com/site/highdimensionalstatistics`.

Enjoy your reading!

Christophe Giraud
Orsay, France

Acknowledgments

This book has strongly benefited from the contributions of many people.

My first thanks go to Sylvain Arlot, Marc Hoffmann, Sylvie Huet, Guillaume Lecué, Pascal Massart, Vincent Rivoirard, Etienne Roquain, François Roueff, and Nicolas Verzelen, who very kindly agreed to review one or several chapters of this book. Thank you very much to all of you for your suggestions and corrections, which have significantly improved the quality of the manuscript. I also very warmly thank Marten Wegkamp and Irina Gaynanova for their feedback, suggestions and corrections on an early version of my lecture notes. My warm thanks go also to all my students, who have pointed out some typos and errors in the first draft of my lecture notes.

I have been lucky to receive some advice from famous statisticians on the writing of a book. Thank you Florentina Bunea, Pascal Massart, and Sacha Tsybakov for your invaluable advice from which this book has strongly benefited. Thank you also Elisabeth Gassiat for your cheerful encouragement, especially during my final months of writing.

This book would not exist without the invitation of David Grubbs to convert my initial lecture notes into a book. I am very grateful to you and the team at Chapman & Hall for your support, enthusiasm, and kindness. I am also deeply grateful to the anonymous reviewers for their insightful suggestions, comments, and feedback.

I have learned statistics mostly from my collaborators, so this book is theirs in some sense (of course, typos, errors, and misuses of English language are mine). Thank you so much to all of you! In particular, I am happy to thank Yannick Baraud and Sylvie Huet for introducing me to statistics a few years ago. My deepest gratitude goes to Nicolas Verzelen. Thank you for all I have learned from you, for the many invaluable scientific discussions we had, for the uncountable references you have pointed out to me, and for your major contribution to the reviewing of this book.

Last but not least, I am extremely grateful to you, Dominique, for your patience during these last months, your beautiful illustrations, and all the nice time we have together.

I am really indebted to all of you.
Merci beaucoup!

Chapter 1

Introduction

1.1 High-Dimensional Data

The sustained development of technologies, data storage resources, and computing resources give rise to the production, storage, and processing of an exponentially growing volume of data. Data are ubiquitous and have a dramatic impact on almost every branch of human activities, including science, medicine, business, finance and administration. For example, wide-scale data enable to better understand the regulation mechanisms of living organisms, to create new therapies, to monitor climate and biodiversity changes, to optimize the resources in the industry and in administrations, to personalize the marketing for each individual consumer, etc.

A major characteristic of modern data is that they often record simultaneously thousands up to millions of *features* on each *object* or *individual.* Such data are said to be *high-dimensional.* Let us illustrate this characteristic with a few examples. These examples are relevant at the time of writing and may become outdated in a few years, yet we emphasize that the mathematical ideas conveyed in this book are independent of these examples and will remain relevant.

- **Biotech data:** Recent biotechnologies enable to acquire high-dimensional data on single individuals. For example, DNA microarrays measure the transcription level[1] of tens of thousands of genes simultaneously; see Figure 1.1. Next generation sequencing devices (NGS) improve on these microarrays by allowing to sense the "transcription level" of virtually any part of the genome. Similarly, in proteomics some technologies can gauge the abundance of thousands of proteins simultaneously. These data are crucial for investigating biological regulation mechanisms and creating new drugs. In such biotech data, the number p of "variables" that are sensed scales in thousands and is most of the time much larger than the number n of "individuals" involved in the experiment (number of repetitions, rarely exceeding a few hundreds).
- **Images (and videos):** Large databases of images are continuously collected all around the world. They include medical images, massive astrophysic images, video surveillance images, etc. Each image is made of thousands up to millions of

[1] The transcription level of a gene in a cell at a given time corresponds to the quantity of ARNm associated to this gene present at this time in the cell.

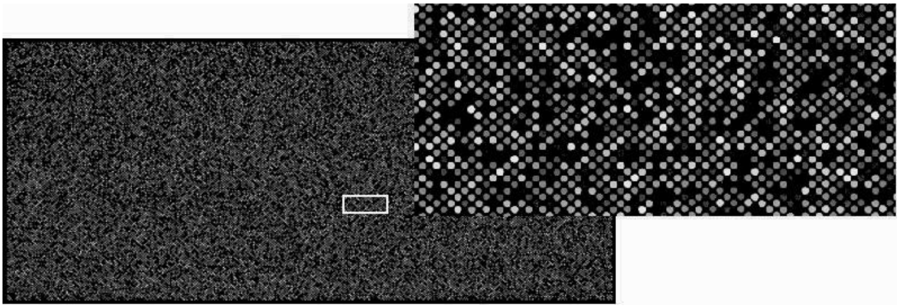


Figure 1.1 *Whole human genome microarray covering more than 41,000 human genes and transcripts on a standard* $1'' \times 3''$ *glass slide format. ©Agilent Technologies, Inc. 2004. Reproduced with permission, courtesy of Agilent Technologies, Inc.*

pixels or voxels. For medical images, as for biotech data, the number p of pixels can be much larger than the number n of patients in the cohort under study.

- **Consumers preferences data:** Websites and loyalty programs collect huge amounts of information on the preferences and the behaviors of customers. These data are processed for marketing purposes, for recommendations, and for fixing personalized prices. For example, recommendation systems (for movies, books, music, etc.) gather the customers' ratings on various products, together with some personal data (age, sex, location) and guess from them which products could be of interest for a given consumer.
- **Business data:** Every major company has its own chief data officer who supervises the optimal exploitation of internal and external data. For example, logistic and transportation companies intensively process internal and geo-economic data in order to optimize the allocation of their resources and to try to forecast precisely the future demand. Insurance companies widely rely on various sources of data in order to control their risk and allocate at best their financial resources. Many profitable activities of the financial industry are based on the intensive processing of transaction data from all over the world. Again, the dimensionality of the data processed in these examples can scale in thousands.
- **Crowdsourcing data:** The launch of websites dedicated to participative data recording together with the spreading of smartphones enable volunteers to record online massive data sets. For example, the Cornell Lab of Ornithology and the National Audubon Society have jointly launched a crowdsourcing program "eBird" `http://ebird.org` inviting all bird-watchers from North America to record via an online checklist all the birds they have seen and heard during their last birding session. The purpose of this program is to monitor birds abundances and their evolutions across North America. In 2014, eBird involved tens of thousands of participants, which had already recorded millions of observations.

Blessing?

Being able to sense simultaneously thousands of variables on each "individual" sounds like good news: Potentially we will be able to scan every variable that may influence the phenomenon under study. The statistical reality unfortunately clashes with this optimistic statement: Separating the signal from the noise is *in general* almost impossible in high-dimensional data. This phenomenon described in the next section is often called the "curse of dimensionality."

1.2 Curse of Dimensionality

The impact of high dimensionality on statistics is multiple. First, high-dimensional spaces are vast and data points are isolated in their immensity. Second, the accumulation of small fluctuations in many different directions can produce a large global fluctuation. Third, an event that is an accumulation of rare events may not be rare. Finally, numerical computations and optimizations in high-dimensional spaces can be overly intensive.

1.2.1 Lost in the Immensity of High-Dimensional Spaces

Let us illustrate this issue with an example. We consider a situation where we want to explain a response variable $Y \in \mathbb{R}$ by p variables $X_1,\ldots,X_p \in [0,1]$. Assume, for example, that each variable X_k follows a uniform distribution on $[0,1]$. If these variables are independent, then the variable $X = (X_1,\ldots,X_p) \in [0,1]^p$ follows a uniform distribution on the hypercube $[0,1]^p$. Our data consist of n i.i.d. observations $(Y_i, X^{(i)})_{i=1,\ldots,n}$ of the variables Y and X. We model them with the classical regression equation

$$Y_i = f(X^{(i)}) + \varepsilon_i, \quad i = 1,\ldots,n,$$
$$\text{with } f : [0,1]^p \to \mathbb{R} \text{ and } \varepsilon_1,\ldots,\varepsilon_n \text{ independent and centered.}$$

Assuming that the function f is smooth, it is natural to estimate $f(x)$ by some average of the Y_i associated to the $X^{(i)}$ in the vicinity of x. The most simple version of this idea is the k-Nearest Neighbors estimator, where $f(x)$ is estimated by the mean of the Y_i associated to the k points $X^{(i)}$, which are the nearest from x. Some more sophisticated versions of this idea use a weighted average of the Y_i with weights that are a decreasing function of the distance $\|X^{(i)} - x\|$ (like kernel smoothing). The basic idea being in all cases to use a local average of the data. This idea makes perfect sense in low-dimensional settings, as illustrated in Figure 1.2.

Unfortunately, when the dimension p increases, the notion of "nearest points" vanishes. This phenomenon is illustrated in Figure 1.3, where we have plotted the histograms of the distribution of the pairwise-distances $\{\|X^{(i)} - X^{(j)}\| : 1 \leq i < j \leq n\}$ for $n = 100$ and dimensions $p = 2, 10, 100$, and 1000. When the dimension p increases, we observe in Figure 1.3 that

- the minimal distance between two points increases,

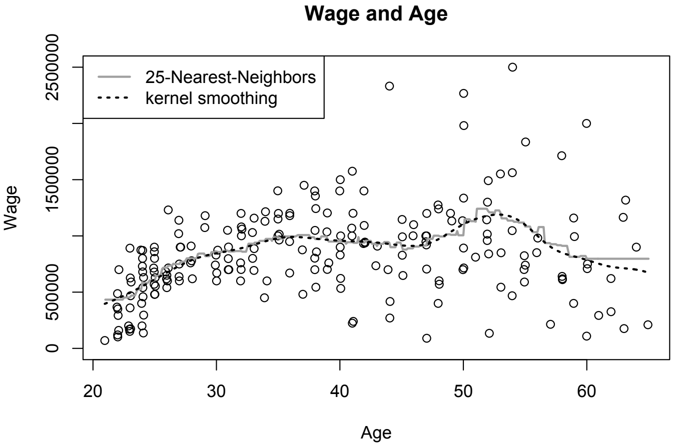


Figure 1.2 *Canadian high school graduate earnings from the 1971 Canadian Census Public Use Tapes. Black dots: records. Gray line: 25-Nearest Neighbors estimator. Dashed line: local averaging by kernel smoothing.*

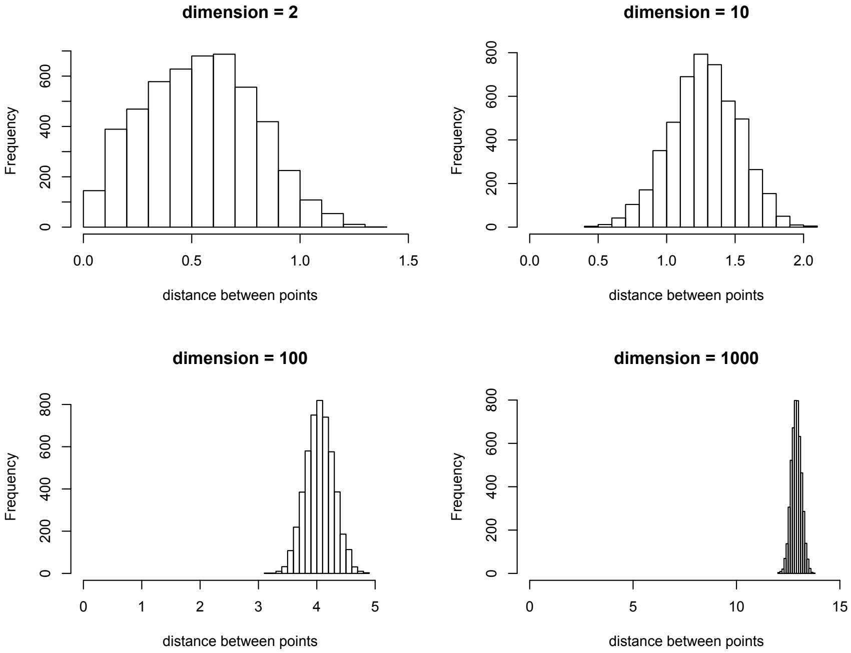


Figure 1.3 *Histograms of the pairwise-distances between* $n = 100$ *points sampled uniformly in the hypercube* $[0,1]^p$*, for* $p = 2, 10, 100,$ *and* 1000.

- all the points are at a similar distance one from the others, so the notion of “nearest-points” vanishes.

In particular, any estimator based on a local averaging will fail with such data.

Let us quantify roughly the above observations. Writing U and U' for two independent random variables with uniform distribution on $[0,1]$, the mean square distance between $X^{(i)}$ and $X^{(j)}$ is

$$\mathbb{E}\left[\|X^{(i)}-X^{(j)}\|^2\right]=\sum_{k=1}^{p}\mathbb{E}\left[\left(X_k^{(i)}-X_k^{(j)}\right)^2\right]=p\,\mathbb{E}\left[(U-U')^2\right]=p/6,$$

and the standard deviation of this square distance is

$$\mathrm{sdev}\left[\|X^{(i)}-X^{(j)}\|^2\right]=\sqrt{\sum_{k=1}^{p}\mathrm{var}\left[\left(X_k^{(i)}-X_k^{(j)}\right)^2\right]}=\sqrt{p\mathrm{var}\left[(U'-U)^2\right]}\approx 0.2\sqrt{p}.$$

In particular, we observe that the typical square distance between two points sampled uniformly in $[0,1]^p$ grows linearly with p, while the scaled deviation $\mathrm{sdev}\left[\|X^{(i)}-X^{(j)}\|^2\right]/\mathbb{E}\left[\|X^{(i)}-X^{(j)}\|^2\right]$ shrinks like $p^{-1/2}$.

How Many Observations Do We Need?

Figure 1.3 shows that if the number n of observations remains fixed while the dimension p of the observations increases, the observations $X^{(1)},\ldots,X^{(n)}$ get rapidly very isolated and local methods cannot work. If for any $x\in[0,1]^p$ we want to have at least one observation $X^{(i)}$ at distance less than one from x, then we must increase the number n of observations. How should this number n increase with the dimension p? We investigate below this issue by computing a lower bound on the number n of points needed in order to fill the hypercube $[0,1]^p$ in such a way that at any $x\in[0,1]^p$ there exists at least one point at distance less than 1 from x.

The volume $V_p(r)$ of a p-dimensional ball of radius $r>0$ is equal to (see Exercise 1.6.2)

$$V_p(r)=\frac{\pi^{p/2}}{\Gamma(p/2+1)}r^p \overset{p\to\infty}{\sim} \left(\frac{2\pi e r^2}{p}\right)^{p/2}(p\pi)^{-1/2}, \tag{1.1}$$

where Γ represents the Gamma function $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}\,dt$ for $x>0$.

If $x^{(1)},\ldots,x^{(n)}$ are such that for any $x\in[0,1]^p$ there exists a point $x^{(i)}$ fulfilling $\|x^{(i)}-x\|\leq 1$, then the hypercube is covered by the family of unit balls centered in $x^{(1)},\ldots,x^{(n)}$,

$$[0,1]^p\subset\bigcup_{i=1}^{n}B_p(x^{(i)},1).$$

As a consequence, the volume of the union of the n unit balls is larger than the volume of the hypercube, so $1\leq nV_p(1)$. According to Equation (1.1), we then need at least

$$n\geq\frac{\Gamma(p/2+1)}{\pi^{p/2}} \overset{p\to\infty}{\sim} \left(\frac{p}{2\pi e}\right)^{p/2}\sqrt{p\pi} \tag{1.2}$$

The Strange Geometry of High-Dimensional Spaces (I)

High-dimensional balls have a vanishing volume!

From the Formula (1.1), we observe that for any $r > 0$, the volume $V_p(r)$ of a ball of radius r goes to zero more than exponentially fast with the dimension p. We illustrate this phenomenon in Figure 1.4 by plotting $p \to V_p(1)$. We observe that for $p = 20$ the volume of the unit ball is already almost 0.

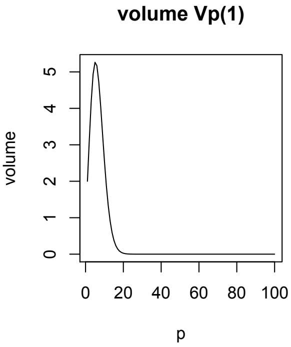


Figure 1.4

points in order to fill the hypercube $[0,1]^p$. This number of points then grows more than exponentially fast with p. If we come back to our above example in the regression setting, it means that if we want a local average estimator to work with observations uniformly distributed in $[0,1]^p$ with p larger than a few tens, then we would need a number n of observations, which is completely unrealistic (see Table 1.1).

p	20	30	50	100	150	200
n	39	45630	$5.7\,10^{12}$	$42\,10^{39}$	$1.28\,10^{72}$	larger than the estimated number of particles in the observable universe

Table 1.1 *Lower Bound (1.2) on the required number of points for filling the hypercube* $[0,1]^p$.

The moral of this example is that we have to be very careful with our geometric intuitions in high-dimensional spaces. These spaces have some counterintuitive geometric properties, as illustrated in Figures 1.4 and 1.5.

1.2.2 Fluctuations Cumulate

Assume that you want to evaluate some function $F(\theta_1)$ of some scalar value $\theta_1 \in \mathbb{R}$. Assume that you have only access to a noisy observation of θ_1, denoted by $X_1 = \theta_1 + \varepsilon_1$, with $\mathbb{E}[\varepsilon_1] = 0$ and var$(\varepsilon_1) = \sigma^2$. If the function F is 1-Lipschitz, then the mean square error is

$$\mathbb{E}\left[\|F(X_1) - F(\theta_1)\|^2\right] \leq \mathbb{E}\left[|\varepsilon_1|^2\right] = \sigma^2.$$

In particular, if the variance σ^2 of the noise is small, then this error is small.

Assume now that you need to evaluate a function $F(\theta_1, \ldots, \theta_p)$ from noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j. Assume that the noise variables $\varepsilon_1, \ldots, \varepsilon_p$ are all centered

The Strange Geometry of High-Dimensional Spaces (II)

The volume of a high-dimensional ball is concentrated in its crust!

Let us write $B_p(0,r)$ for the p-dimensional ball centered at 0 with radius $r > 0$, and $C_p(r)$ for the "crust" obtained by removing from $B_p(0,r)$ the sub-ball $B_p(0,0.99r)$. In other words, the "crust" gathers the points in $B_p(0,r)$, which are at a distance less than $0.01r$ from its surface.

In Figure 1.5, we plot as a function of p the ratio of the volume of $C_p(r)$ to the volume of $B_p(0,r)$

$$\frac{\text{volume}(C_p(r))}{\text{volume}(B_p(0,r))} = 1 - 0.99^p,$$

which goes exponentially fast to 1.

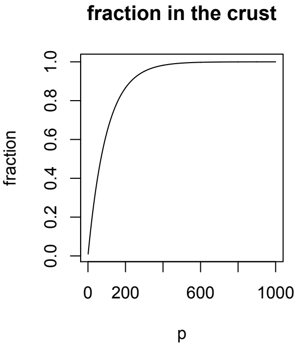


Figure 1.5

with variance σ^2. If as before F is 1-Lipschitz, we have

$$\mathbb{E}\left[\|F(X_1,\ldots,X_p) - F(\theta_1,\ldots,\theta_p)\|^2\right] \le \mathbb{E}\left[\|(\varepsilon_1,\ldots,\varepsilon_p)\|^2\right] = \sum_{j=1}^{p} \mathbb{E}\left[\varepsilon_j^2\right] = p\sigma^2.$$

Furthermore, if F fulfills $\|F(x+h) - F(x)\| \ge c\|h\|$ for some $c > 0$, then the mean square error $\mathbb{E}\left[\|F(X_1,\ldots,X_p) - F(\theta_1,\ldots,\theta_p)\|^2\right]$ scales like $p\sigma^2$. This error can be very large in high-dimensional settings, even if σ^2 is small. A central example where such a situation arises is in the linear regression model with high-dimensional covariates.

High-Dimensional Linear Regression

Assume that we have n observations $Y_i = \langle x^{(i)}, \beta^* \rangle + \varepsilon_i$ for $i = 1,\ldots,n$, with the response Y_i in $\mathbb{R}$ and the covariates $x^{(i)}$ in $\mathbb{R}^p$. We want to estimate $\beta^* \in \mathbb{R}^p$, and we assume that $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. centered, with variance σ^2. Writing

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(n)})^T \end{pmatrix} \quad \text{and} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

we have $Y = \mathbf{X}\beta^* + \varepsilon$. A classical estimator of β^* is the least-square estimator

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - \mathbf{X}\beta\|^2,$$

which is uniquely defined when the rank of $\mathbf{X}$ is p. Let us focus on this case. The solution of this minimization problem is $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$, which fulfills (see Exercise 1.6.5)

$$\mathbb{E}\left[\|\widehat{\beta}-\beta^*\|^2\right] = \mathbb{E}\left[\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\|^2\right] = \mathrm{Tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\sigma^2.$$

Assume for simplicity that the columns of $\mathbf{X}$ are orthonormal (i.e., orthogonal with norm 1). Then, the mean square error is

$$\mathbb{E}\left[\|\widehat{\beta}-\beta^*\|^2\right] = p\sigma^2.$$

So the more high dimensional is the covariate $x^{(i)}$, the larger is this estimation error.

We cannot give a direct picture of a linear regression in dimension p, with p larger than 2. Yet, we can illustrate the above phenomenon with the following example. Assume that the covariates $x^{(i)}$ are given by $x^{(i)} = \phi(i/n)$, where $\phi : [0,1] \to \mathbb{R}^p$ is defined by $\phi(t) = [\cos(\pi jt)]_{j=1,\dots,p}$. Then we observe

$$Y_i = \sum_{j=1}^{p} \beta_j^* \cos(\pi ji/n) + \varepsilon_i = f_{\beta^*}(i/n) + \varepsilon_i, \quad \text{for } i = 1,\dots,n, \tag{1.3}$$

with $f_\beta(t) = \sum_{j=1}^{p} \beta_j \cos(\pi jt)$. We can illustrate the increase of the error $\|\widehat{\beta}-\beta^*\|^2$ with p by plotting the function f_{β^*} and $f_{\widehat{\beta}}$ for increasing values of p. In Figure 1.6, the noise $\varepsilon_1,\dots,\varepsilon_n$ is i.i.d., with $\mathcal{N}(0,1)$ distribution and the function f_{β^*} has been generated by sampling the β_j^* independently with $\mathcal{N}(0,j^{-4})$ distribution. We choose $n = 100$ and the 4 figures correspond to $p = 10, 20, 50$ and 100, respectively. We observe that when p increases, the estimated function $f_{\widehat{\beta}}(t)$ becomes more and more wavy. These increasing oscillations are the direct consequence of the increasing error $\|\widehat{\beta}-\beta^*\|^2 \approx p\sigma^2$.

Tails of High-Dimensional Gaussian Distributions Are Thin but Concentrate the Mass

Gaussian distributions are known to have very thin tails. Actually, the density $g_p(x) = (2\pi)^{-p/2}\exp(-\|x\|^2/2)$ of a standard Gaussian distribution $\mathcal{N}(0,I_p)$ in $\mathbb{R}^p$ decreases exponentially fast with the square norm of x. Yet, when p is large, most of the mass of the standard Gaussian distribution lies in its tails!

First, we observe that the maximal value of $g_p(x)$ is $g_p(0) = (2\pi)^{-p/2}$, which decreases exponentially fast toward 0 when p increases, so the Gaussian distribution in high dimensions is much more flat than in dimension one or two. Let us compute the mass in its "bell" (central part of the distribution where the density is the largest). Let $\delta > 0$ be a small positive real number and write

$$B_{p,\delta} = \left\{x \in \mathbb{R}^p : g_p(x) \geq \delta g_p(0)\right\} = \left\{x \in \mathbb{R}^p : \|x\|^2 \leq 2\log(\delta^{-1})\right\}$$

for the ball gathering all the points $x \in \mathbb{R}^p$, such that the density $g_p(x)$ is larger or

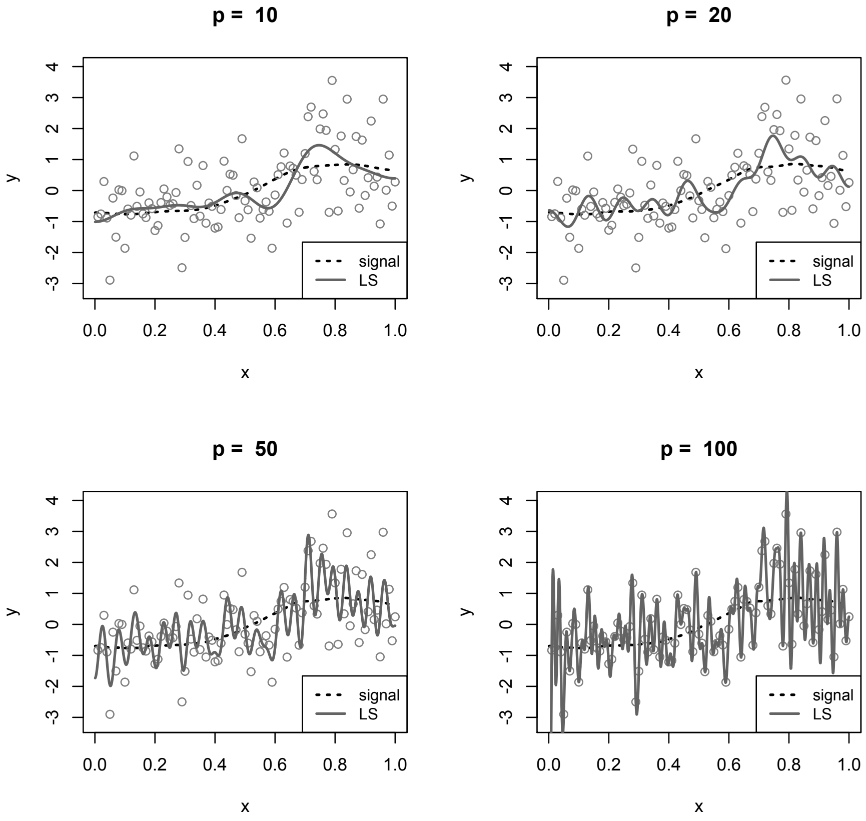


Figure 1.6 *Least-Square (LS) estimator* $f_{\widehat{\beta}}$ *in the setting (1.3), with* $n = 100$ *and for* $p = 10, 20, 50$, *and* 100. *Gray dots: the observations. Dashed line: the function* f_{β^*}. *Gray line: the LS estimator* $f_{\widehat{\beta}}$.

equal to δ times the density at 0. Our intuition from the one-dimensional case is that for δ small, say $\delta = 0.001$, the probability for a $\mathcal{N}(0, I_p)$ Gaussian random variable X to be in the "bell" $B_{p,\delta}$ is close to one. Yet, as illustrated in the Figure 1.7 it is the opposite!

Actually, from the Markov inequality (Lemma B.1 in Appendix B), we have

$$\begin{aligned}\mathbb{P}\left(X \in B_{p,\delta}\right) &= \mathbb{P}\left(e^{-\|X\|^2/2} \geq \delta\right)\\ &\leq \frac{1}{\delta}\,\mathbb{E}\left[e^{-\|X\|^2/2}\right] = \frac{1}{\delta}\int_{x\in\mathbb{R}^p} e^{-\|x\|^2}\,\frac{dx}{(2\pi)^{p/2}} = \frac{1}{\delta\, 2^{p/2}}.\end{aligned}$$

So most of the mass of the standard Gaussian distribution is in the tail

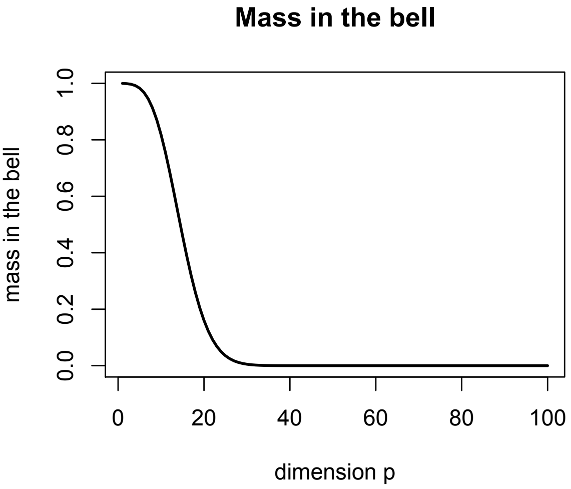


Figure 1.7 *Mass of the standard Gaussian distribution $g_p(x)\,dx$ in the "bell" $B_{p,0.001} = \left\{x \in \mathbb{R}^p : g_p(x) \geq 0.001 g_p(0)\right\}$ for increasing values of p.*

$\left\{x \in \mathbb{R}^p : g_p(x) < \delta g_0(x)\right\}$. If we want to have $\mathbb{P}\left(X \in B_{p,\delta}\right) \geq 1/2$, we must choose $\delta \leq 2^{-p/2+1}$ which is exponentially small.

How can we explain this counterintuitive phenomenon? It is related to the geometric properties of the high-dimensional spaces described above. We have seen that the volume of a ball of radius r grows like r^p. So when r increases, the density $g_p(x)$ on the ring $\{x : r < \|x\| < r + dr\}$ shrinks like $e^{-r^2/2}$, but at the same time the volume $V_p(r+dr) - V_p(r)$ of the thin ring $\{x : r < \|x\| < r + dr\}$ grows with r like r^{p-1}, so the probability for X to be in $\{x : r < \|x\| < r + dr\}$ evolves as $r^{p-1}e^{-r^2/2}$ with $r > 0$. In particular, this probability is maximal for $r^2 = p - 1$, and the Gaussian density at a point x with norm $\|x\|^2 = p - 1$ fulfills $g_p(x) = e^{-(p-1)^2/2} g_p(0) \ll g_p(0)$: most of the mass of a Gaussian distribution is located in areas where the density is extremely small compared to its maximum value.

1.2.3 An Accumulation of Rare Events May Not Be Rare

Assume that we have an observation Z_1 of a single quantity θ_1 blurred by some $\mathcal{N}(0,1)$ Gaussian noise ε_1. From Lemma B.3 in Appendix B we have $\mathbb{P}(|\varepsilon_1| \geq x) \leq e^{-x^2/2}$ for $x > 0$, so with probability at least $1 - \alpha$, the noise ε_1 has an absolute value smaller than $\left(2\log(1/\alpha)\right)^{1/2}$.

Assume that we observe now p quantities $\theta_1, \ldots, \theta_p$ blurred by $\varepsilon_1, \ldots, \varepsilon_p$ i.i.d. with

$\mathcal{N}(0,1)$ Gaussian distribution. We have

$$\mathbb{P}\left(\max_{j=1,\ldots,p}|\varepsilon_j|\geq x\right)=1-(1-\mathbb{P}(|\varepsilon_1|\geq x))^p \overset{x\to\infty}{\sim} p\,\mathbb{P}(|\varepsilon_1|\geq x)\,.$$

If we want to bound simultaneously the absolute values $|\varepsilon_1|,\ldots,|\varepsilon_p|$ with probability $1-\alpha$, then we can only guarantee that $\max_{j=1,\ldots,p}|\varepsilon_j|$ is smaller than $\left(2\log(p/\alpha)\right)^{1/2}$. This extra $\log(p)$ factor can be a serious issue in practice as illustrated below.

False Discoveries

Assume that for a given individual i, we can measure for p genes simultaneously the ratios between their expression levels in two different environments (e.g., with microarrays). For the individual i, let us denote by $Z_j^{(i)}$ the log-ratio of the expression levels of the gene j between the two environments. Assume that (after some normalization) these log-ratios can be modeled by

$$Z_j^{(i)}=\theta_j+\varepsilon_j^{(i)},\quad j=1,\ldots,p,\ \ i=1,\ldots,n,$$

with the $\varepsilon_j^{(i)}$ i.i.d. with $\mathcal{N}(0,1)$ Gaussian distribution. Our goal is to detect the genes j, such that $\theta_j\neq 0$, which means that they are involved in the response to the change of environment (such a gene is said to be “positive”).

We write $X_j=n^{-1}(Z_j^{(1)}+\ldots+Z_j^{(n)})$ for the mean of the observed log-ratios for the gene j. The random variables $n^{1/2}X_1,\ldots,n^{1/2}X_p$ are independent, and $n^{1/2}X_j$ follows a $\mathcal{N}(\sqrt{n}\theta_j,1)$ Gaussian distribution. For W with $\mathcal{N}(0,1)$ Gaussian distribution, we have $\mathbb{P}[|W|>1.96]\approx 5\%$, so a natural idea is to declare “positive” all the genes j, such that $n^{1/2}|X_j|$ is larger than 1.96. Nevertheless, this procedure would produce many false positives (genes declared “positive” while they are not) and thereby many false discoveries of genes responding to the change of environment. Let us illustrate this point. Assume that $p=5000$, and among them 200 genes are positive. Then, the average number of false positive genes is

$$\text{card}\left\{j:\theta_j=0\right\}\times 0.05=4800\times 0.05=240 \text{ false positive genes,}$$

which is larger than the number of positive genes (200). It means that, on average, more than half of the discoveries will be false discoveries. If we want to avoid false positives, we must choose a threshold larger than 1.96. From Exercise 1.6.3, for $W_1,\ldots,W_p$ i.i.d. with $\mathcal{N}(0,1)$ Gaussian distribution, we have

$$\mathbb{P}\left(\max_{j=1,\ldots,p}|W_j|\geq\sqrt{\alpha\log(p)}\right)=1-\exp\left(-\sqrt{\frac{2}{\alpha\pi}}\,\frac{p^{1-\alpha/2}}{(\log p)^{1/2}}+O\left(\frac{p^{1-\alpha/2}}{(\log p)^{3/2}}\right)\right)$$

$$\xrightarrow{p\to\infty}\begin{cases}0 & \text{if } \alpha\geq 2\\ 1 & \text{if } \alpha<2\,.\end{cases}$$

Therefore, in order to avoid false positives, it seems sensible to declare positive the genes j, such that $n^{1/2}|X_j| \geq \sqrt{2\log(p)}$. With this choice, we will roughly be able to detect the θ_j whose absolute value is larger than $\sqrt{2\log(p)/n}$. We then observe that the larger p, the less we are able to detect the nonzero θ_j. This can be a severe issue when p scales in thousands and n is only a few units. We refer to Chapter 8 for techniques suited to this setting.

Empirical Covariance Is Not Reliable In High-Dimensional Settings

Another important issue with high-dimensional data is that the empirical covariance matrix of a p-dimensional vector is not reliable when p scales like the sample size n. Let us illustrate briefly this point. Assume that we observe some i.i.d. random vectors $X^{(1)},\ldots,X^{(n)}$ in $\mathbb{R}^p$, which are centered with covariance matrix $\text{cov}(X^{(i)}) = I_p$. The empirical covariance matrix $\widehat{\Sigma}$ associated to the observations $X^{(1)},\ldots,X^{(n)}$ is given by

$$\widehat{\Sigma}_{ab} = \frac{1}{n}\sum_{i=1}^{n} X_a^{(i)}X_b^{(i)}, \quad \text{for } a,b = 1,\ldots,p.$$

We have $\mathbb{E}\left[X_a^{(i)}X_b^{(i)}\right] = \mathbf{1}_{\{a=b\}}$, so by the strong law of large numbers, $\widehat{\Sigma}_{a,b}$ tends to $\mathbf{1}_{\{a=b\}}$ almost surely when n goes to infinity. Therefore, the empirical covariance matrix $\widehat{\Sigma}$ converges almost surely to the identity matrix when n goes to infinity with p fixed. In particular, the spectrum of $\widehat{\Sigma}$ is concentrated around 1 when n is large and p small. This property is lost when p increases proportionally to n. Figure 1.8 displays three histograms of the spectral values of $\widehat{\Sigma}$ when $X^{(1)},\ldots,X^{(n)}$ are i.i.d. with standard Gaussian $\mathcal{N}(0,I_p)$ distribution, with $n = 1000$ and $p = n/2$, $p = n$, and $p = 2n$, respectively.

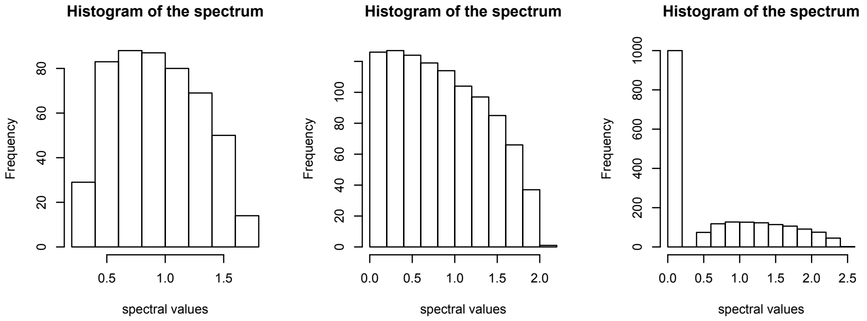


Figure 1.8 *Histogram of the spectral values of the empirical covariance matrix $\widehat{\Sigma}$, with $n =$ 1000 and $p = n/2$ (left), $p = n$ (center), $p = 2n$ (right).*

We observe in the three cases that the spectrum of the empirical covariance matrix $\widehat{\Sigma}$ is very different from the spectrum of the identity, so the empirical covariance $\widehat{\Sigma}$ is a very poor approximation of the covariance I_p in this setting. From the theory of random matrices, we know the limit distribution of the spectral values of $\widehat{\Sigma}$ when n goes to infinity and $p \sim \alpha n$ with $\alpha > 0$ (see Section 1.5 for references). The support of this limit distribution (known as the Marchenko-Pastur distribution) is actually equal to $[(1-\sqrt{\alpha})^2, (1+\sqrt{\alpha})^2]$ up to a singleton at 0 when $\alpha > 1$. This means that we cannot trust the empirical covariance matrix $\widehat{\Sigma}$ when n and p have a similar size.

1.2.4 Computational Complexity

Another burden arises in high-dimensional settings: numerical computations can become very intensive and largely exceed the available computational (and memory) resources. For example, basic operations with $p \times p$ matrices (like multiplication, inversion, etc.) require at least p^α operations with $\alpha > 2$. When p scales in thousands, iterating such operations can become quite intensive.

The computational complexity appears to be really problematic in more involved problems. For example, we have seen above that the mean square error $\|\widehat{\beta} - \beta^*\|^2$ in the linear regression model

$$y = \sum_{j=1}^{p} \beta_j^* x_j + \varepsilon$$

typically scales linearly with p. Yet, it is unlikely that all the covariates x_j are influential on the response y. So we may wish to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^* x_j + \varepsilon \quad \text{for each } m \subset \{1, \ldots, p\}. \tag{1.4}$$

Unfortunately, the cardinality of $\{m : m \subset \{1, \ldots, p\}\}$ is 2^p, which grows exponentially with p. So, when p is larger than a few tens, it is impossible to compute the 2^p estimators $\widehat{\beta}_m$ associated to the model (1.4). This issue is detailed in Chapters 2 and 4.

1.3 High-Dimensional Statistics

1.3.1 Circumventing the Curse of Dimensionality

As explained in the previous section, the high dimensionality of the data, which seems at first to be a blessing, is actually a major issue for the statistical analyses. In light of the few examples described above, the situation may appear hopeless. Fortunately, high-dimensional data are often much more low dimensional than they seem to be. Usually, they are not "uniformly" spread in $\mathbb{R}^p$, but rather concentrated around some low-dimensional structures. These structures are due to the relatively small complexity of the systems producing the data. For example,

- pixel intensities in an image are not purely-random since there exist many geometrical structures in an image;
- biological data are the outcome of a "biological system", which is strongly regulated and whose regulation network has a relatively small complexity;

PCA in action

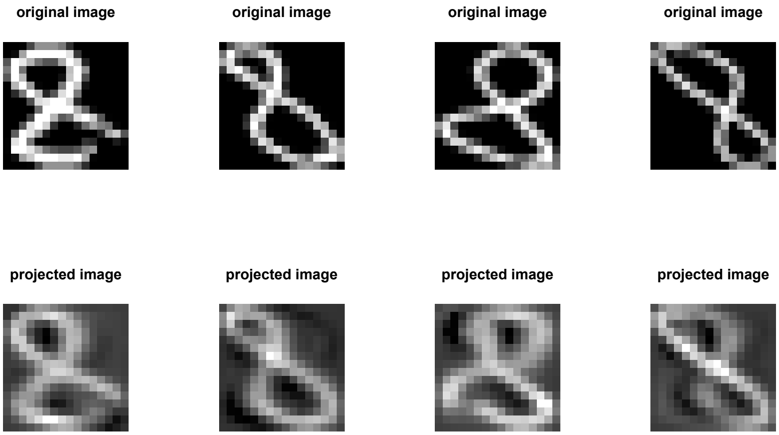


Figure 1.9 *Projection on V_{10} of some scans of the digit 8.*

The Mixed National Institute of Standards and Technology (MNIST) data set [83] gathers 1100 scans of each digit. Each scan is a 16×16 image that is encoded by a vector in $\mathbb{R}^{256}$. The above pictures represent the projections of four images onto the space V_{10}, computed according to (1.5) from the 1100 scans of the digit 8 in the MNIST database. The original images are displayed in the first row, with their projection onto V_{10} in the second row.

- marketing data strongly reflects some social structures in a population, and these structures are relatively simple; and
- technical data are the outcome of human technologies, whose complexity remains limited.

So in many cases, the data have an intrinsic low complexity, and we can hope to extract useful information from them. Actually, when the low-dimensional structures are known, we are back to some more classical "low-dimensional statistics." The major issue with high-dimensional data is that these structures are usually unknown. The main task will then be to identify at least approximately theses structures. This issue can be seen as the central issue of high-dimensional statistics.

A first approach is to try to find directly the low-dimensional structures in the data. The simplest and the most widely used technique for this purpose is the Principal Component Analysis (PCA). For any data points $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ and a given dimension $d \leq p$, the PCA computes the linear span V_d fulfilling

$$V_d \in \underset{\dim(V)\leq d}{\operatorname{argmin}} \sum_{i=1}^{n} \|X^{(i)} - \mathrm{Proj}_V(X^{(i)})\|^2, \tag{1.5}$$

where the minimum is taken over all the subspaces V of $\mathbb{R}^p$, with dimension not larger than d, and $\mathrm{Proj}_V : \mathbb{R}^p \to \mathbb{R}^p$ is the orthogonal projector onto V. We refer to Figure 1.9 for an illustration of the PCA in action and to Exercise 1.6.4 for the mathematical details.

Another approach, developed all along this book, is to perform an "estimation-oriented" search of these low-dimensional structures. With this approach, we only seek for the low-dimensional structures that are useful for our estimation problem. In principle, this approach allows for more precise results. We illustrate this feature in Figure 1.10, which is based on a data set of Sabarly *et al.* [107] gathering the measurement of 55 chemical compounds for 162 strains of the bacteria *E. coli*. Some of these bacteria are pathogens for humans, some others are commensal. Our goal is to find a classification rule that enables us to separate from the chemical measurements the strains that are pathogens from those that are commensal. In Figure 1.10, the pathogen bacteria are denoted by "IPE" (internal pathogens) and "EPE" (external pathogens) and the commensal bacteria are denoted by "Com." The left-hand-side figure displays the data projected on the two-dimensional space V_2 given by a PCA. The right-hand-side figure displays the data projected on a two-dimension space S_2,

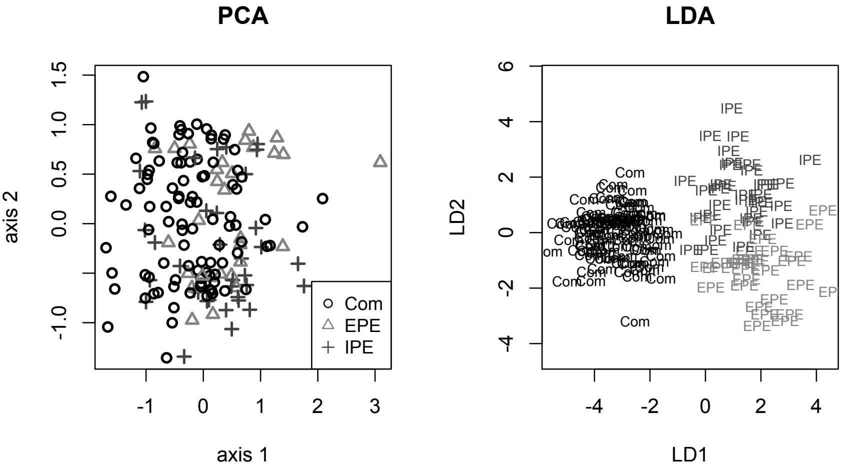


Figure 1.10 *Dimension reduction for a data set gathering 55 chemical measurements of 162 strains of E. coli. Commensal strains are labeled by "Com," and pathogen strains are labeled by "EPE" (external pathogens) and "IPE" (internal pathogens). Left: The data is projected on the plane given by a PCA. Right: The data is projected on the plane given by an LDA.*

which aims to separate at best the different classes of bacteria (more precisely, S_2 is spanned by the directions orthogonal to the separating hyperplanes of a Linear Discriminant Analysis (LDA) ; see Exercise 9.5.1 for details on the LDA). Clearly, the plane S_2 is more useful than V_2 for classifying strains of bacteria according to their pathogenicity. This better result is simply due to the fact that V_2 has been computed independently of the classification purpose, whereas S_2 has been computed in order to solve at best this classification problem.

1.3.2 A Paradigm Shift

Classical statistics provide a very rich theory for analyzing data with the following characteristics:

- a small number p of parameters
- a large number n of observations

This setting is typically illustrated by the linear regression $y = ax + b + \varepsilon$ plotted in Figure 1.11, where we have to estimate 2 parameters a and b, with $n = 100$ observations $(X_i, Y_i)_{i=1,\dots,n}$. Classical results carefully describe the asymptotic behavior of estimators when n goes to infinity (with p fixed), which makes sense in such a setting.

As explained in Section 1.1, in many fields, current data have very different characteristics:

- a huge number p of parameters
- a sample size n, which is either roughly of the same size as p, or sometimes much smaller than p

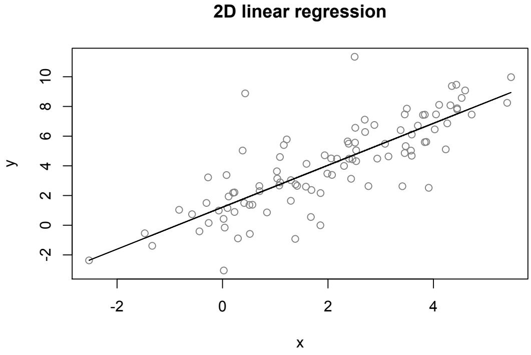


Figure 1.11 *Iconic example of classical statistics: $n = 100$ observations (gray dots) for estimating the $p = 2$ parameters of the regression line (in black).*

The asymptotic analysis with p fixed and n goes to the infinity does not make sense anymore. Worse, it can lead to very misleading conclusions. We must change our point of view on statistics!

In order to provide a theory adapted to the 21st century data, two different points of view can be adopted. A first point of view is to investigate the properties of estimators in a setting where both n and p go to infinity, with $p \sim f(n)$ for some function f; for example, $f(n) = \alpha n$, or $f(n) = n^2$, or $f(n) = e^{\alpha n}$, etc. Such a point of view is definitely more suited to modern data than the classical point of view. Yet, it is sensitive to the choice of the function f. For example, asymptotic results for $f(n) = n^2$ and $f(n) = e^{\alpha n}$ can be very different. If $p = 1000$ and $n = 33$, are you in the setting $f(n) = n^2$ or $f(n) = e^{n/5}$?

An alternative point of view is to treat n and p as they are and provide a non asymptotic analysis of the estimators, which is valid for any value of n and p. Such an analysis avoids the above caveat of the asymptotic analysis. The main drawback is that non asymptotic analyses are much more involved than asymptotic analyses. They usually require much more elaborate arguments in order to provide precise enough results.

1.3.3 Mathematics of High-Dimensional Statistics

In order to quantify non asymptotically the performances of an estimator, we need some tools to replace the classical convergence theorems used in classical statistics. A typical example of convergence theorem is the central limit theorem, which describes the asymptotic convergence of an empirical mean toward its expected value: for $f : \mathbb{R} \to \mathbb{R}$ and $X_1, \ldots, X_n$ i.i.d. such that $\text{var}(f(X_1)) < +\infty$, we have when $n \to +\infty$

$$\sqrt{\frac{n}{\text{var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X_1)\right] \right) \xrightarrow{\text{d}} Z, \quad \text{with } Z \sim \mathcal{N}(0,1).$$

Loosely speaking, the difference between the empirical mean and the statistical mean $\mathbb{E}\left[f(X_1)\right]$ behaves roughly as $\sqrt{n^{-1}\text{var}(f(X_1))}\, Z$ when n is large. Let us assume that f is L-Lipschitz, and let X_1, X_2 be i.i.d. with finite variance σ^2. We have

$$\text{var}(f(X_1)) = \frac{1}{2}\mathbb{E}\left[(f(X_1) - f(X_2))^2\right] \leq \frac{L^2}{2}\mathbb{E}\left[(X_1 - X_2)^2\right] = L^2\sigma^2,$$

so, the central limit theorem ensures that for $X_1, \ldots, X_n$ i.i.d. with finite variance σ^2 and a L-Lipschitz function f,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X_1)\right] \geq \frac{L\sigma}{\sqrt{n}} x \right) \leq \mathbb{P}(Z \geq x) \leq e^{-x^2/2}, \quad \text{for } x > 0 \quad (1.6)$$

(the last inequality follows from Lemma B.3, page 218 in Appendix B).

Concentration inequalities provide some non asymptotic versions of such results.

Assume, for example, that $X_1,\dots,X_n$ are i.i.d., with $\mathcal{N}(0,\sigma^2)$ Gaussian distribution. The Gaussian concentration inequality (Theorem B.6, page 221 in Appendix B) claims that for any L-Lipschitz function $F:\mathbb{R}^n\to\mathbb{R}$ we have

$$F(X_1,\dots,X_n)-\mathbb{E}\left[F(X_1,\dots,X_n)\right]\leq L\sigma\sqrt{2\xi},\quad \text{where } \mathbb{P}(\xi\geq t)\leq e^{-t} \text{ for } t>0.$$

When $f:\mathbb{R}\to\mathbb{R}$ is L-Lipschitz, the Cauchy–Schwartz inequality gives

$$\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\frac{1}{n}\sum_{i=1}^{n}f(Y_i)\right|\leq\frac{L}{n}\sum_{i=1}^{n}|X_i-Y_i|\leq\frac{L}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(X_i-Y_i)^2}\ ,$$

so the function $F(X_1,\dots,X_n)=n^{-1}\sum_{i=1}^{n}f(X_i)$ is $(n^{-1/2}L)$-Lipschitz. According to the Gaussian concentration inequality, we then have for $x>0$ and $n\geq 1$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}\left[f(X_1)\right]\geq\frac{L\sigma}{\sqrt{n}}x\right)\leq\mathbb{P}\left(\sqrt{2\xi}\geq x\right)=e^{-x^2/2},$$

which can be viewed as a non asymptotic version of (1.6). Concentration inequalities are central tools for the non asymptotic analysis of estimators, and we will meet them in every major proof of this book. Appendix B gathers a few classical concentration inequalities (with proofs); see also the Exercise 1.6.6 at the end of this chapter.

1.4 About This Book

1.4.1 Statistics and Data Analysis

Data science is an ever-expanding field. The world is awash in data, and there is a sustained effort by the data-analyst community for developing statistical procedures and algorithms able to process these data. Most of this effort is carried out according to the following track:

1. identification of an issue (new or not) related to some kind of data;
2. proposition of a statistical procedure based on some heuristics; and
3. implementation of this procedure on a couple of data sets (real or simulated), with comparison to some existing procedures, when available.

This line of research feeds hundreds of scientific conferences every year covering a wide range of topics, ranging from biology to economy via humanities, communication systems and astrophysics. It has led to some dazzling success, at least in the technological area.

In the above process, mathematical formalism is used all along the way, but there are (almost) no mathematics there, in the sense that there is no mathematical analysis of the implemented procedures. In particular, even if we restrict to the academic community, the statistician community, as part of the mathematical community, is a minority. Let us call “mathematical statistics” the field of research that

1. formalizes precisely a statistical issue (new or not),

2. formalizes precisely a statistical procedure for handling this issue (new or not),
3. provides a mathematical analysis (theorem) of the performance of this statistical procedure, with a special attention to its optimality and its limitations.

Since the mathematical analysis is the limiting step (due to its difficulty), the models and statistical procedures investigated in mathematical statistics are usually quite simpler than those implemented by "mainstream" data analysts. Some may directly conclude that mathematical statistics are useless and have no interest. We want to argue that mathematical statistics are important and actually have their own interest:

- Mathematical statistics provide some strong guarantees on statistical procedures and they identify to what extent we can trust these procedures.
- Mathematical statistics enable us to identify precisely the frontier between the problems where estimation is possible and those where it is hopeless.
- Mathematical statistics provide mathematical foundations to data analysis.
- Mathematical statistics can be one of the remedies against the lack of reproducibility observed in many data-based scientific fields.

Again, most of the works in mathematical statistics concern some simple models, but they provide some useful intuitions for the more involved settings met in practice.

1.4.2 Purpose of This Book

This book focuses on the mathematical foundations of high-dimensional statistics, which is clearly in the "mathematical statistics" stream. Its goal is to present the main concepts of high-dimensional statistics and to delineate in some fundamental cases the frontier between what is achievable and what is impossible.

These lecture notes concentrate on state-of-the-art techniques for handling high-dimensional data. They do not strive for gathering a comprehensive catalog of statistical methods; they try instead to explain for some selected topics the key fundamental concepts and ideas. These concepts and ideas are exposed in simple settings, which allow concentration on the main arguments by avoiding unessential technicalities. The proofs issued from the recent research literature have been intensively streamlined in order to enlighten the main arguments. The reader is invited to adapt these ideas to more complex settings in the detailed exercises at the end of each chapter. He is also welcome to share his solutions to the exercises on the dedicated wiki-website `http://high-dimensional-statistics.wikidot.com`.

1.4.3 Overview

As explained in Section 1.3, for analyzing high-dimensional data, we must circumvent two major issues:

- the intrinsic statistical difficulty related to the curse of dimensionality, and
- the computational difficulty: procedures must have a low computational complexity in order to fit the computational resources.

To bypass the statistical curse of dimensionality, the statistician must build on the low-dimensional structures hidden in the data. Model selection, mainly developed in the late '90s, is a powerful theory for tackling this problem. It provides very clear insights on the frontier between the problems where estimation is possible and those where it is hopeless. A smooth version of this theory is presented in Chapter 2. An alternative to model selection is model aggregation introduced in Chapter 3. While the underlying principles of model selection and model aggregation are essentially the same, they lead to different estimation schemes in practice. Model selection and model aggregation provide powerful tools able to break the statistical curse of dimensionality, yet they both suffer from a very high computational complexity, which is prohibitive in practice.

A powerful strategy to bypass this second curse of dimensionality is to "convexify" in some way the model selection schemes. This strategy has been intensively developed in the last decade with bright success. Chapter 4 is a concise introduction to this topic. It is mainly illustrated with the celebrated Lasso estimator, for which a thorough mathematical analysis is provided. The statistical procedures presented in Chapter 4 succeed to circumvent both the statistical and computational curses of high dimensionality. Yet, they suffer from two caveats. First, for each specific class of low-dimensional structures (which is often unknown) corresponds a specific estimator. Second, all these estimators depend on a "tuning parameter" that needs to be chosen according to the variance of the noise, usually unknown. Some estimator selection is then required in order to handle these two issues. A selection of such procedures are sketched in Chapter 5.

Chapters 6 and 7 go one step beyond in the statistical complexity. They present recent theories that have been developed in order to take advantage of handling simultaneously several statistical problems. Chapter 6 explores how we can exploit via some rank constraints the correlations between these statistical problems, and improve thereby the accuracy of the statistical procedures. Chapter 7 focuses on the simultaneous estimation of the conditional dependencies among a large set of variables via the theory of graphical models.

The last two chapters deal with issues arising in high-dimensional data classification. Chapter 8 details the mathematical foundations of multiple testing, with a special focus on False Discovering Rate (FDR) control. Chapter 9 gives a concise presentation of Vapnik's theory for supervised classification.

The theories presented in this book heavily rely on some more or less classical mathematical tools, including some probabilistic inequalities (concentration inequalities, contraction principle, Birgé's Lemma, etc.), some matrix analysis (singular value decomposition, Ky–Fan norms, etc.), and some classical analysis (subdifferential of convex functions, reproducing Hilbert spaces, etc.). The five appendices provide self-contained and compact introductions to these mathematical tools. The notations used throughout the book are gathered all together at the end of the volume.

1.5 Discussion and References

1.5.1 Take-Home Message

Being able to sense simultaneously thousands of parameters sounds like a blessing, since we collect huge amounts of data. Yet, the information is awash in the noise in high-dimensional settings. Fortunately, the useful information usually concentrates around low-dimensional structures, and building on this feature allows us to circumvent this curse of the dimensionality.

This book is an introduction to the main concepts and ideas involved in the analysis of the high-dimensional data. Its focus is on the mathematical side, with the choice to concentrate on simple settings in order to avoid unessential technical details that could blur the main arguments.

1.5.2 References

The book by Hastie, Tibshirani, and Friedman [73] is an authoritative reference for the data scientist looking for a pedagogical and (almost) comprehensive catalog of statistical procedures. The associated website
`http://statweb.stanford.edu/~tibs/ElemStatLearn/`
provides many interesting materials both for learners and academics.

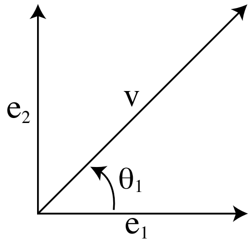
We refer to the books by Ledoux [84] and by Boucheron, Lugosi, and Massart [33] for thorough developments on concentration inequalities. The asymptotic distribution of the spectral values of a random matrix has been described first by Marchenko and Pastur [92]. We refer to Vershynin [123] for a non asymptotic analysis.

The lecture notes [53] by Donoho give an interesting point of view on high-dimensional statistics and their mathematical challenges. Finally, the papers by Jin [76], Verzelen [124], Donoho, Johnstone, and Montanari [54] or Berthet and Rigollet [25] are examples of papers providing insightful information on the frontier of successful estimation in high-dimensional settings.

1.6 Exercises

1.6.1 Strange Geometry of High-Dimensional Spaces

Figures 1.4 and 1.5 give examples of counter-intuitive results in high-dimensional spaces. Let us give another one. Let $e_1, \ldots, e_p$ be the canonical basis in $\mathbb{R}^p$, and let us denote by v the diagonal of the hypercube $[0,1]^p$. Check that the angle θ_i between the vector e_i and v fulfills



$$\cos(\theta_i) = \frac{\langle e_i, v\rangle}{\|e_i\|\,\|v\|} = \frac{1}{\sqrt{p}} \overset{p\to\infty}{\to} 0.$$

So the diagonal of a hypercube tends to be orthogonal to all the edges of the hypercube in high-dimensional spaces!

1.6.2 Volume of a p-Dimensional Ball

We will prove the Formula (1.1) for the volume $V_p(r)$ of a p-dimensional ball of radius $r > 0$.

1. Prove that the gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$, $\alpha > 0$ fulfills the equalities $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$, and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for any $\alpha > 0$. Deduce the two formulas
$$\Gamma(p+1) = p! \quad \text{and} \quad \Gamma(p+3/2) = \frac{(2p+1)(2p-1)\dots 1}{2^{p+1}}\sqrt{\pi} \quad \text{for } p \in \mathbb{N}.$$
2. Prove that $V_p(r) = r^p V_p(1)$ for any $p \geq 1$ and check that $V_1(1) = 2$ and $V_2(1) = \pi$.
3. For $p \geq 3$, prove that
$$\begin{aligned} V_p(1) &= \int_{x_1^2+x_2^2\leq 1} V_{p-2}\left(\sqrt{1-x_1^2-x_2^2}\right) dx_1\,dx_2 \\ &= V_{p-2}(1)\int_{r=0}^{1}\int_{\theta=0}^{2\pi}(1-r^2)^{p/2-1}r\,dr\,d\theta = \frac{2\pi}{p}V_{p-2}(1). \end{aligned}$$
4. Conclude that
$$V_{2p}(1) = \frac{\pi^p}{p!} \quad \text{and} \quad V_{2p+1}(1) = \frac{2^{p+1}\pi^p}{(2p+1)(2p-1)\dots 3}.$$
5. With the Stirling expansion $\Gamma(\alpha) = \alpha^{\alpha-1/2}e^{-\alpha}\sqrt{2\pi}\left(1+O(\alpha^{-1})\right)$ for $\alpha \to +\infty$, prove (1.1).

1.6.3 Tails of a Standard Gaussian Distribution

Let Z be a $\mathcal{N}(0,1)$ standard Gaussian random variable.

1. For $z > 0$, prove (with an integration by parts) that
$$\begin{aligned} \mathbb{P}(|Z| \geq z) &= \sqrt{\frac{2}{\pi}}\,\frac{e^{-z^2/2}}{z} - \sqrt{\frac{2}{\pi}}\int_z^\infty x^{-2}e^{-x^2/2}\,dx \\ &= \sqrt{\frac{2}{\pi}}\,\frac{e^{-z^2/2}}{z}\left(1+O\left(\frac{1}{z^2}\right)\right). \end{aligned}$$
2. For $Z_1,\dots,Z_p$ i.i.d. with $\mathcal{N}(0,1)$ standard Gaussian distribution and $\alpha > 0$, prove that when $p \to \infty$
$$\begin{aligned} &\mathbb{P}\left(\max_{j=1,\dots,p}|Z_j| \geq \sqrt{\alpha\log(p)}\right) \\ &= 1-\left(1-\mathbb{P}\left(|Z_1| \geq \sqrt{\alpha\log(p)}\right)\right)^p \\ &= 1-\exp\left(-\sqrt{\frac{2}{\alpha\pi}}\,\frac{p^{1-\alpha/2}}{(\log p)^{1/2}} + O\left(\frac{p^{1-\alpha/2}}{(\log p)^{3/2}}\right)\right). \end{aligned}$$

1.6.4 Principal Component Analysis

The Principal Component Analysis (PCA) is tightly linked to the Singular Value Decomposition (SVD). We refer to Appendix C for a reminder on the SVD.

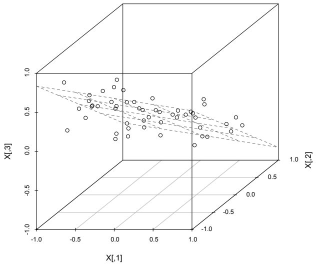
For any data points $X^{(1)},\dots,X^{(n)} \in \mathbb{R}^p$ and any dimension $d \leq p$, the PCA computes the linear span in $\mathbb{R}^p$

$$V_d \in \underset{\dim(V)\leq d}{\operatorname{argmin}} \sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|^2,$$

where Proj_V is the orthogonal projection matrix onto V. Let us denote by $\mathbf{X} = \sum_{k=1}^{r} \sigma_k u_k v_k^T$ a SVD of the $n \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} (X^{(1)})^T \\ \vdots \\ (X^{(n)})^T \end{pmatrix}$$

with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.



V_2 in dimension $p = 3$.

1. With the Theorem C.5 in Appendix C, page 233, prove that for any $d \leq r$

$$\sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|^2 = \|\mathbf{X} - \mathbf{X}\text{Proj}_V\|_F^2 \geq \sum_{k=d+1}^{r} \sigma_k^2,$$

   where $\|\cdot\|_F$ denotes the Frobenius norm $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$.
2. Write V_d for the linear space spanned by $\{v_1,\dots,v_d\}$. Prove that for any $d \leq r$ we have

$$\|\mathbf{X} - \mathbf{X}\text{Proj}_{V_d}\|^2 = \sum_{k=d+1}^{r} \sigma_k^2.$$

3. Conclude that V_d minimizes (1.5).
4. Prove that the coordinates of $\text{Proj}_{V_d} X^{(i)}$ in the orthonormal basis $(v_1,\dots,v_d)$ of V_d are given by $(\sigma_1\langle e_i, u_1\rangle,\dots,\sigma_d\langle e_i, u_d\rangle)$.

Terminology: The right-singular vectors $v_1,\dots,v_r$ are called the principal axes. The vectors $c_k = \mathbf{X}v_k = \sigma_k u_k$ for $k = 1,\dots,r$ are called the principal components. The principal component c_k gathers the coordinates of $X^{(1)},\dots,X^{(n)}$ on v_k.

Remark: Since V_d is a linear span and not an affine span, it is highly recommended to first center the data points

$$\widetilde{X}^{(i)} = X^{(i)} - \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$$

and then proceed with a PCA on the $\widetilde{X}^{(1)},\dots,\widetilde{X}^{(n)}$.

1.6.5 Basics of Linear Regression

We consider the linear regression model $Y = \mathbf{X}\beta^* + \varepsilon$, with $Y, \varepsilon \in \mathbb{R}^n$, $\beta^* \in \mathbb{R}^p$ and $\mathbf{X}$ an $n \times p$-matrix with rank p. We assume that the components $\varepsilon_1, \ldots, \varepsilon_n$ of ε are i.i.d. centered, with variance σ^2. We set $F(\beta) = \|Y - \mathbf{X}\beta\|^2$.

1. Prove that F is convex and has a gradient $\nabla F(\beta) = 2\mathbf{X}^T(\mathbf{X}\beta - Y)$.
2. Prove that the least-square estimator $\widehat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} F(\beta)$ solves $\mathbf{X}^T\mathbf{X}\widehat{\beta} = \mathbf{X}^T Y$, so $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$ when the rank of $\mathbf{X}$ is p.
3. Let A be a $p \times n$ matrix. Prove that $\mathbb{E}\left[\|A\varepsilon\|^2\right] = \sigma^2 \mathrm{Tr}(A^T A)$.
4. Conclude that mean square error of the least-square estimator $\widehat{\beta}$ is

$$\mathbb{E}\left[\|\widehat{\beta} - \beta^*\|^2\right] = \mathrm{Tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\sigma^2.$$

1.6.6 Concentration of Square Norm of Gaussian Random Variable

In the regression setting, we often need a bound on the square norm of a $\mathcal{N}(0, I_n)$ Gaussian random variable ε; see, for example, the proof of Theorem 2.2 in Chapter 2. The expectation of $\|\varepsilon\|^2$ is $\mathbb{E}\left[\|\varepsilon\|^2\right] = n$, so all we need is to get a probabilistic bound on the deviations of $\|\varepsilon\|^2$ above its expectation. Since the map $\varepsilon \to \|\varepsilon\|$ is 1-Lipschitz, the Gaussian concentration inequality (Theorem B.6, page 221, in Appendix B) ensures that

$$\mathbb{P}\left(\|\varepsilon\| \le \mathbb{E}[\|\varepsilon\|] + \sqrt{2x}\right) \ge 1 - e^{-x}, \quad \text{for any } x > 0.$$

From Jensen inequality, we have $\mathbb{E}[\|\varepsilon\|]^2 \le \mathbb{E}\left[\|\varepsilon\|^2\right] = n$, so we have the concentration inequality

$$\mathbb{P}\left(\|\varepsilon\|^2 \le n + 2\sqrt{2nx} + 2x\right) \ge 1 - e^{-x}, \quad \text{for any } x > 0. \tag{1.7}$$

In the following, we give a direct proof of (1.7) based on the simple Markov inequality (Lemma B.1, page 217, in Appendix B).

1. Check that for $0 < s < 1/2$, we have $\mathbb{E}\left[\exp\left(s\|\varepsilon\|^2\right)\right] = (1-2s)^{-n/2}$.
2. With the Markov inequality, prove that $\mathbb{P}\left(\|\varepsilon\|^2 > n + t)\right) \le e^{-s(n+t)}(1-2s)^{-n/2}$ for any $t > 0$ and $0 < s < 1/2$.
3. Check that the above bound is minimal for $s = t/(2(n+t))$, and hence

$$\mathbb{P}\left(\|\varepsilon\|^2 > n + t)\right) \le e^{-t/2}(1 + t/n)^{n/2}, \quad \text{for any } t > 0.$$

4. Check that $\log(1+u) \le u - u^2/(2+2u)$ for any $u \ge 0$, and hence

$$\mathbb{P}\left(\|\varepsilon\|^2 > n + t)\right) \le \exp\left(-\frac{t^2}{4(n+t)}\right), \quad \text{for any } t > 0.$$

5. Prove Bound (1.7) for $0 < x \le n$.

6. From the bound $\log(1+u) \le \sqrt{u}$ for $u \ge 0$, check that

$$\frac{n}{2}\log\left(1+\frac{2\sqrt{2nx}+2x}{n}\right) \le \sqrt{2nx}, \quad \text{for } x \ge n.$$

7. Prove (1.7) for $x \ge n$.

Chapter 2

Model Selection

Model selection is a key conceptual tool for performing dimension reduction and exploiting hidden structures in the data. The general idea is to compare different statistical models corresponding to different possible hidden structures and then select among them the one that is more suited for estimation. Model selection is a very powerful theory, but it suffers in many cases from a very high computational complexity that can be prohibitive. When model selection cannot be implemented due to its computational cost, it remains a good guideline for developing computationally tractable procedures, as we will see in Chapter 4.

In this chapter, we present the theory of model selection in a simple — yet useful — setting: the Gaussian regression model. References for a broader theory are given in Section 2.8.

2.1 Statistical Setting

We consider in this chapter the regression model

$$y_i = f(x^{(i)}) + \varepsilon_i, \quad i = 1, \dots, n,$$

which links a quantity of interest $y \in \mathbb{R}$ to p variables whose values are stored in a p-dimensional vector $x \in \mathbb{R}^p$. We give several examples below corresponding to the regression model.

Example 1: Sparse piecewise constant regression (variation sparsity)

> It corresponds to the case where $x \in \mathbb{R}$ and f is piecewise constant with a small number of jumps. This situation appears, for example, in biology in the analysis of the copy number variations along a DNA chain.
>
> Assume that the jumps of f are located in the set $\{z_j : j \in \mathscr{J}\}$. Then a right-continuous piecewise constant function f can be written as
>
> $$f(x) = \sum_{j \in \mathscr{J}} c_j \mathbf{1}_{x \geq z_j}.$$
>
> When the function f only has a few jumps, then only a small fraction of the $\{c_j : j \in \mathscr{J}\}$ are nonzero.

Example 2: Sparse basis/frame expansion

> It corresponds to the case where $f : \mathbb{R} \to \mathbb{R}$, and we estimate f by expanding it on a basis or frame $\{\varphi_j\}_{j\in\mathscr{J}}$
>
> $$f(x) = \sum_{j\in\mathscr{J}} c_j \varphi_j(x),$$
>
> with a small number of nonzero c_j. This situation arises, for example, for denoising, for representing cortex signals, etc. Typical examples of basis are Fourier basis, splines, wavelets, etc. The most simple example is the piecewise linear decomposition
>
> $$f(x) = \sum_{j\in\mathscr{J}} c_j (x - z_j)_+, \tag{2.1}$$
>
> where $z_1 < z_2 < \ldots$ and $(x)_+ = \max(x, 0)$.

In these first two examples, x is low dimensional, but the function to be estimated can be complex and requires to estimate a large number of coefficients. At the opposite, we can have x in a high-dimensional space $\mathbb{R}^p$ and f very simple. The most popular case is the linear regression.

Example 3: Sparse linear regression

> It corresponds to the case where f is linear: $f(x) = \langle \beta, x \rangle$ with $\beta \in \mathbb{R}^p$. We say that the linear regression is sparse when only a few coordinates of β are nonzero.

This model can be too rough to model the data. Assume, for example, that we want to model the relationship between some phenotypes and some gene expression levels. We expect from biology that only a small number of genes influence a given phenotype, but the relationship between these genes and the phenotype is unlikely to be linear. We may consider in this case a slightly more complex model.

Example 4: Sparse additive model and group-sparse regression

> In the sparse additive model, we expect that $f(x) = \sum_k f_k(x_k)$ with most of the f_k equal to 0.
>
> If we expand each function f_k on a frame or basis $\{\varphi_j\}_{j\in\mathscr{J}_k}$ we obtain the decomposition
>
> $$f(x) = \sum_{k=1}^{p} \sum_{j\in\mathscr{J}_k} c_{j,k} \varphi_j(x_k),$$
>
> where most of the vectors $\{c_{j,k}\}_{j\in J_k}$ are zero.

Such a model can be hard to fit from a small sample, since it requires to estimate a relatively large number of nonzero $c_{j,k}$. Nevertheless, in some cases, the basis expansion of f_k can be sparse itself, as in Example 2, producing a more complex pattern of sparsity. An interesting example is the following model.

Example 5: Sparse additive piecewise linear regression

The sparse additive piecewise linear model is a sparse additive model $f(x) = \sum_k f_k(x_k)$, with sparse piecewise linear functions f_k. We then have two levels of sparsity:

1. Most of the f_k are equal to 0, and
2. the nonzero f_k have a sparse expansion in the following representation

$$f_k(x_k) = \sum_{j \in \mathscr{J}_k} c_{j,k} \, (x_k - z_{j,k})_+$$

In other words, the matrix $c = [c_{j,k}]$ of the sparse additive model has only a few nonzero columns, and these nonzero columns are sparse.

It turns out that all the above models correspond to a sparse linear model, in the sense that we have a representation of $f^* = \left[f(x^{(i)})\right]_{i=1,\dots,n}$ of the form $f_i^* = \langle \alpha, \psi_i \rangle$ for $i = 1, \dots, n$, with α sparse in some sense. Let us identify this representation in the five above examples.

Examples 1, 2, 3, 4, 5 (continued): Representation $f_i^* = \langle \alpha, \psi_i \rangle$

- Sparse piecewise constant regression: $\psi_i = e_i$ with $\{e_1, \dots, e_n\}$ the canonical basis of $\mathbb{R}^n$ and $\alpha = f^*$ is piecewise constant.
- Sparse basis expansion: $\psi_i = [\varphi_j(x^{(i)})]_{j \in \mathscr{J}}$ and $\alpha = c$.
- Sparse linear regression: $\psi_i = x^{(i)}$ and $\alpha = \beta$.
- Sparse additive models: $\psi_i = [\varphi_j([x_k^{(i)}])]_{\substack{k=1,\dots,p \\ j \in \mathscr{J}_k}}$ and $\alpha = [c_{j,k}]_{\substack{k=1,\dots,p \\ j \in \mathscr{J}_k}}$.

Since the five above examples can be recast in a linear regression framework, we will consider the linear regression model as the canonical example and use the following notations.

Linear model

We have $f_i^* = \langle \beta^*, x^{(i)} \rangle$. We define the $n \times p$ matrix $\mathbf{X} = [x_j^{(i)}]_{i=1,\dots,n,\; j=1,\dots,p}$ and the n-dimensional vector $\varepsilon = [\varepsilon_i]_{i=1,\dots,n}$. With these notations, we have the synthetic formula

$$Y = f^* + \varepsilon = \mathbf{X}\beta^* + \varepsilon. \tag{2.2}$$

The above examples correspond to different sparsity patterns for β^*. We use all along the book the following terminology.

Sparsity patterns

- **Coordinate sparsity:** Only a few coordinates of β^* are nonzero. This situation arises in Examples 2 and 3.
- **Group sparsity:** The coordinates of β^* are clustered into groups, and only a few groups are nonzero. More precisely, we have a partition $\{1,\ldots,p\} = \bigcup_{k=1}^{M} G_k$, and only a few vectors $\beta^*_{G_k} = (\beta^*_j)_{j\in G_k}$ are nonzero. This situation arises in Example 4.
- **Sparse-group sparsity:** In the same notation as the group sparse setting, only a few vectors $\beta^*_{G_k}$ are nonzero and in addition they are sparse. This situation arises in Example 5.
- **Variation sparsity (or fused sparsity):** The vector $\Delta\beta^* = [\beta^*_j - \beta^*_{j-1}]_{j=2,\ldots,p}$ is coordinate-sparse. This situation arises in Example 1.

In the remainder of this chapter, we will focus on the Gaussian regression setting

$$Y_i = f^*_i + \varepsilon_i,\ i = 1,\ldots,n, \quad \text{with the } \varepsilon_i \text{ i.i.d. with } \mathcal{N}(0,\sigma^2) \text{ distribution.} \qquad (2.3)$$

2.2 To Select among a Collection of Models

2.2.1 Models and Oracle

Let us consider the problem of linear regression in the coordinate-sparse setting. If we knew that only the coordinates β^*_j with $j \in m^*$ are nonzero, then we would remove all the other variables $\{x_j :\ j \notin m^*\}$ and consider the simpler linear regression model

$$y_i = \langle \beta^*_{m^*}, x^{(i)}_{m^*} \rangle + \varepsilon_i = \sum_{j\in m^*} \beta^*_j x^{(i)}_j + \varepsilon_i.$$

In the other sparse settings described above, we would work similarly if we knew exactly the sparsity pattern (which groups are nonzero for group sparsity, where the jumps are located for variation sparsity, etc.).

More generally, if we knew that f^* belongs to some linear subspace S of $\mathbb{R}^n$, instead of estimating f^* by simply maximizing the likelihood, we would rather estimate f^* by maximizing the likelihood under the constraint that the estimator belongs to S. For example, in the coordinate-sparse setting where we know that the nonzero coordinates are the $\{\beta^*_j : j \in m^*\}$, we would take $S = \text{span}\{x_j,\ j \in m^*\}$. In our Gaussian setting (2.3), the log-likelihood is given by

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Y - f\|^2,$$

so the estimator maximizing the likelihood under the constraint that it belongs to S

is simply $\widehat{f} = \mathrm{Proj}_S Y$, where $\mathrm{Proj}_S : \mathbb{R}^n \to \mathbb{R}^n$ is the orthogonal projection operator onto S.

If we do not know *a priori* that f^* belongs to a *known* linear subspace S of $\mathbb{R}^n$, then we may wish to

1. consider a collection $\{S_m,\ m \in \mathcal{M}\}$ of linear subspaces of $\mathbb{R}^n$, called *models*;
2. associate to each subspace S_m the constrained maximum likelihood estimators $\widehat{f}_m = \mathrm{Proj}_{S_m} Y$; and
3. finally estimate f by the *best* estimator among the collection $\left\{\widehat{f}_m,\ m \in \mathcal{M}\right\}$.

To give a meaning to *best*, we need a criterion to quantify the quality of an estimator. In the following, we will measure the quality of an estimator $\widehat{f}$ of f^* by its ℓ^2-risk

$$R(\widehat{f}) = \mathbb{E}\left[\|\widehat{f} - f^*\|^2\right]. \tag{2.4}$$

Each estimator $\widehat{f}_m$ has a ℓ^2-risk $r_m = R(\widehat{f}_m)$, and the best estimator in terms of the ℓ^2-risk is the so-called *oracle estimator*

$$\widehat{f}_{m_o} \quad \text{with} \quad m_o \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \{r_m\}. \tag{2.5}$$

We emphasize that the signal f^* may not belong to the oracle model S_{m_o}. It may even not belong to any of the models $\{S_m,\ m \in \mathcal{M}\}$.

The oracle estimator is the best estimator in terms of the ℓ^2-risk, so we would like to use this estimator to estimate f^*. Unfortunately, we cannot use it in practice, since it cannot be computed from the data only. Actually, the index m_o depends on the collection of risks $\{r_m,\ m \in \mathcal{M}\}$, which is unknown to the statisticians since it depends on the unknown signal f^*.

A natural idea to circumvent this issue is to replace the risks r_m in (2.5) by some estimators $\widehat{r}_m$ of the risk and therefore estimate f^* by

$$\widehat{f}_{\widehat{m}} \quad \text{with} \quad \widehat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \{\widehat{r}_m\}. \tag{2.6}$$

The estimator $\widehat{f}_{\widehat{m}}$ can be computed from the data only, but we have no guarantee that it performs well. The main object of this chapter is to provide some suitable $\widehat{r}_m$ for which we can guarantee that the selected estimator $\widehat{f}_{\widehat{m}}$ performs almost as well as the oracle $\widehat{f}_{m_o}$.

Examples of collections of models

Let us describe the collections of models suited to the different sparsity patterns described in Section 2.1. In the following, we write $\mathbf{X}_j$ for the j-th column of $\mathbf{X}$.

Coordinate sparsity

Let us define $\mathscr{M} = \mathscr{P}(\{1,\ldots,p\})$, where $\mathscr{P}(\{1,\ldots,p\})$ denotes the set of all the subsets of $\{1,\ldots,p\}$. In this setting, we will consider the collection of models $\{S_m,\ m \in \mathscr{M}\}$, with the models S_m defined by $S_m = \text{span}\left\{\mathbf{X}_j\,,\ j \in m\right\}$ for $m \in \mathscr{M}$. The model S_m then gathers all the vectors $\mathbf{X}\beta$, with β fulfilling $\beta_j = 0$ if $j \notin m$.

Group sparsity

In this setting, we will consider a collection of models $\{S_m,\ m \in \mathscr{M}\}$ indexed by the set $\mathscr{M} = \mathscr{P}(\{1,\ldots,M\})$ of all the subsets of $\{1,\ldots,M\}$. For $m \in \mathscr{M}$, the model S_m is defined by $S_m = \text{span}\left\{\mathbf{X}_j\ :\ j \in \bigcup_{k\in m} G_k\right\}$. The model S_m then gathers all the vectors $\mathbf{X}\beta$, with β fulfilling $\beta_{G_k} = 0$ for $k \notin m$.

Sparse-group sparsity

The description of the collection of models $\{S_m,\ m \in \mathscr{M}\}$ in this setting is slightly more involved. To a subset $g \subset \{1,\ldots,M\})$, we associate the set of indices

$$\mathscr{M}_g = \left\{\{J_k\}_{k\in g}\ :\ \text{where } J_k \subset G_k \text{ for all } k \in g\right\}.$$

We defined the set of indices by

$$\mathscr{M} = \bigcup_{g\subset\{1,\ldots,M\}}\ \bigcup_{\{J_k\}_{k\in g}\in\mathscr{M}_g} \left\{(g,\{J_k\}_{k\in g})\right\},$$

and to an index $m = (g,\{J_k\}_{k\in g})$ in $\mathscr{M}$, we associate the model $S_m = \text{span}\left\{\mathbf{X}_j\ :\ j \in \bigcup_{k\in g} J_k\right\}$. The model $S_{(g,\{J_k\}_{k\in g})}$ then gathers all the vectors $\mathbf{X}\beta$, with β fulfilling $\beta_j = 0$ if $j \notin \bigcup_{k\in g} J_k$.

Variation sparsity

Writing $\Delta\beta = [\beta_k - \beta_{k-1}]_{k=1,\ldots,p}$ (with the convention $\beta_0 = 0$), we have $\beta_j = \sum_{k\leq j}[\Delta\beta]_k$ and

$$\mathbf{X}\beta = \sum_{j=1}^{p}\sum_{k\leq j}[\Delta\beta]_k\mathbf{X}_j = \sum_{k=1}^{p}[\Delta\beta]_k\sum_{j\geq k}\mathbf{X}_j\,.$$

In view of this formula, we will consider for the variation sparse setting the collection of models $\{S_m,\ m \in \mathscr{M}\}$ indexed by the set $\mathscr{M} = \mathscr{P}(\{1,\ldots,p\})$ of all the subsets of $\{1,\ldots,p\}$, with models S_m defined by $S_m = \text{span}\left\{\sum_{j\geq k}\mathbf{X}_j\,,\ k \in m\right\}$. The model S_m then gathers all the vectors $\mathbf{X}\beta$, with β fulfilling $\beta_j = \beta_{j-1}$ if $j \notin m$.

Models are (usually) wrong

We emphasize that it is unlikely that our data exactly corresponds to one of the settings described in the five examples of Section 2.1. For example, it is unlikely that a signal $f(x)$ is exactly sparse linear, yet it may exist a good sparse linear approximation of it. In particular, we do not have in mind that one of the models is exact and we seek it. We rather try to select the best model in the collection in order to estimate f^*, keeping in mind that all these models can be wrong. The best model for estimation corresponds to the oracle model S_{m_o}, which is our benchmark for comparison.

2.2.2 Model Selection Procedures

Risk r_m of $\widehat{f}_m$

Let us compute the risk $r_m = R(\widehat{f}_m)$ of the estimator $\widehat{f}_m$. Starting from $Y = f^* + \varepsilon$, we obtain the decomposition $f^* - \widehat{f}_m = (I - \mathrm{Proj}_{S_m})f^* - \mathrm{Proj}_{S_m}\varepsilon$. Then, the Pythagorean formula and Lemma A.3, page 214 in Appendix A, give

$$\begin{aligned} r_m = \mathbb{E}\left[\|f^* - \widehat{f}_m\|^2\right] &= \mathbb{E}\left[\|(I - \mathrm{Proj}_{S_m})f^*\|^2 + \|\mathrm{Proj}_{S_m}\varepsilon\|^2\right] \\ &= \|(I - \mathrm{Proj}_{S_m})f^*\|^2 + d_m\sigma^2, \end{aligned}$$

where $d_m = \dim(S_m)$. The risk r_m involves two terms. The first term $\|(I - \mathrm{Proj}_{S_m})f^*\|^2$ is a bias term that reflects the quality of S_m for approximating f^*. The second term $d_m\sigma^2$ is a variance term that increases linearly with the dimension of S_m. In particular, we notice that enlarging S_m reduces the first term but increases the second term. The oracle model S_{m_o} is then the model in the collection $\{S_m,\ m \in \mathcal{M}\}$, which achieves the best trade-off between the bias and the variance.

Unbiased estimator of the risk

A natural idea is to use an unbiased estimator $\widehat{r}_m$ of the risk r_m in Criterion (2.6). It follows from the decomposition $Y - \widehat{f}_m = (I - \mathrm{Proj}_{S_m})(f^* + \varepsilon)$ that

$$\begin{aligned} \mathbb{E}\left[\|Y - \widehat{f}_m\|^2\right] &= \mathbb{E}\left[\|(I - \mathrm{Proj}_{S_m})f^*\|^2 + 2\langle (I - \mathrm{Proj}_{S_m})f^*, \varepsilon\rangle + \|(I - \mathrm{Proj}_{S_m})\varepsilon\|^2\right] \\ &= \|(I - \mathrm{Proj}_{S_m})f^*\|^2 + (n - d_m)\sigma^2 \\ &= r_m + (n - 2d_m)\sigma^2. \end{aligned}$$

As a consequence,

$$\widehat{r}_m = \|Y - \widehat{f}_m\|^2 + (2d_m - n)\sigma^2 \tag{2.7}$$

is an unbiased estimator of the risk. Note that dropping the $-n\sigma^2$ term in $\widehat{r}_m$ does not change the choice of $\widehat{m}$ in (2.6). This choice gives the Akaike Information Criterion (AIC)

$$\widehat{m}_{\mathrm{AIC}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \|Y - \widehat{f}_m\|^2 + 2d_m\sigma^2 \right\}. \tag{2.8}$$

This criterion is very natural and popular. Nevertheless it can produce very poor results in some cases because it does not take into account the variability of the estimated risks $\widehat{r}_m$ around their mean r_m. This is, for example, the case when the number of models S_m with dimension d grows exponentially with d; see Exercise 2.9.1. It is

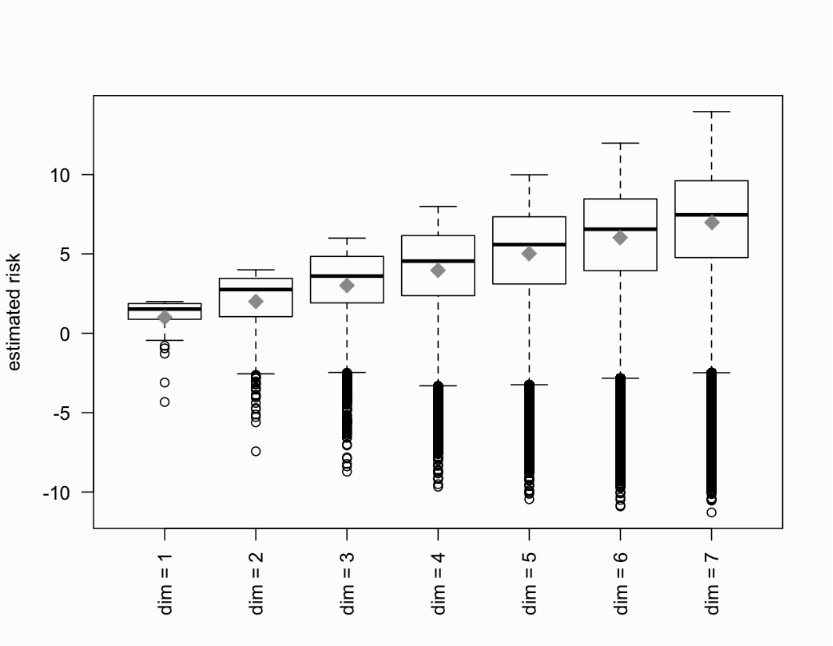


Figure 2.1 *We consider the pure noise case ($f^* = 0$) in the coordinate-sparse setting with orthogonal design, as in Exercise 2.9.1, with $p = 40$. For each dimension $d = 1,\ldots,7$, the boxplot of $\{\widehat{r}_m : d_m = d\}$ is represented, with $\widehat{r}_m$ defined by (2.7). The true risks $r_m = d_m\sigma^2$ are represented by gray diamonds. We observe that when dimension d grows there are more and more outliers, and* $\min\{\widehat{r}_m : d_m = d\}$ *tends to decrease with d. The AIC model $S_{\widehat{m}_{AIC}}$ then has a large dimension.*

due to the fact that in this setting, for large dimensions d, we have a huge number of models S_m with dimension d. Therefore, we have a huge number of estimators $\widehat{r}_m$, and due to the randomness, some of them deviate seriously from their expected value r_m. In particular, some $\widehat{r}_m$ are very small, much smaller than $\widehat{r}_{m_o}$ associated to the oracle m_o. This leads the AIC criterion to select a model $S_{\widehat{m}}$ much bigger than S_{m_o} with very high probability. We illustrate this issue in Figure 2.1.

Penalized estimator of the risk

How can we avoid the undesirable phenomenon described above? The AIC criterion (2.8) involves two terms. The first term $\|Y - \widehat{f}_m\|^2 = \|Y - \text{Proj}_{S_m}Y\|^2$ is decreasing with the size of S_m, whereas the second term $2d_m\sigma^2$ is increasing with the dimension of S_m. In order to avoid the selection of a model $S_{\widehat{m}}$ with an overly large dimension $d_{\widehat{m}}$ (as in Exercise 2.9.1), we shall replace the second term $2d_m\sigma^2$ by a term taking into account the number of models per dimension.

Following this idea, we focus henceforth on a selection criterion of the form

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \|Y - \widehat{f}_m\|^2 + \sigma^2 \mathrm{pen}(m) \right\},$$

where the function pen : $\mathcal{M} \to \mathbb{R}^+$ is called the *penalty* function. A proper strategy to build a good penalty function is to perform a non asymptotic analysis of the risk $R(\widehat{f}_{\widehat{m}})$ and then choose the penalty $\mathrm{pen}(m)$ in order to have a risk $R(\widehat{f}_{\widehat{m}})$ as close as possible to the oracle risk $R(\widehat{f}_{m_o})$. Such an analysis, leading to the choice of the selection criterion presented below, is detailed in the proof of forthcoming Theorem 2.2.

To start with, we associate to the collection of models $\{S_m,\ m \in \mathcal{M}\}$ a probability distribution $\pi = \{\pi_m,\ m \in \mathcal{M}\}$ on $\mathcal{M}$. For a given probability π and a given $K > 1$, we select $\widehat{m}$ according to the criterion

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \|Y - \widehat{f}_m\|^2 + \sigma^2 \mathrm{pen}(m) \right\},$$
$$\text{with} \quad \mathrm{pen}(m) = K \left(\sqrt{d_m} + \sqrt{2 \log(1/\pi_m)} \right)^2. \tag{2.9}$$

The resulting estimator is $\widehat{f} = \widehat{f}_{\widehat{m}}$.

At first sight, it is not obvious why we should use a penalty $\mathrm{pen}(m)$ with such a shape. We refer to Remark 2, page 40, for explanations on the shape of Criterion (2.9) in the light of the proof of Theorem 2.2.

Criterion (2.9) depends heavily on the probability π. It is then crucial to choose this probability π properly. This probability distribution can reflect our knowledge on the likelihood of being in one model rather than in another one, but most of the time it is chosen in a completely *ad-hoc* way. Actually, as we will see in Theorem 2.2, the ℓ^2-risk of the estimator $\widehat{f} = \widehat{f}_{\widehat{m}}$, with $\widehat{m}$ selected by (2.9), roughly behaves like

$$\min_{m \in \mathcal{M}} \left(r_m + \sigma^2 \log \frac{1}{\pi_m} \right). \tag{2.10}$$

Therefore, the choice of π_m will be driven by an attempt to make this term as close as possible to the oracle risk r_{m_o}. It will be the case, for example, if $\log(\pi_m^{-1})\sigma^2$ behaves like the variance $d_m\sigma^2$ of $\widehat{f}_m$: since $r_m \geq d_m\sigma^2$, the minimum (2.10) is upper bounded by $2r_{m_o}$ when $\log(\pi_m^{-1}) \leq d_m$. Unfortunately, we cannot choose $\log(\pi_m^{-1}) \leq \alpha d_m$ with α a numerical constant for any collection of models. Actually, increasing the size of $\mathcal{M}$ requires to decrease the size of π_m (since π is a probability distribution on $\mathcal{M}$), and when $\mathcal{M}$ is very large, as in the coordinate-sparse setting, the sum $\sum_{m \in \mathcal{M}} e^{-\alpha d_m}$ cannot be bounded independently of p for a fixed $\alpha > 0$.

Below, we give examples of probabilities π that produce good results. They are built by slicing the collection of models $\{S_m,\ m \in \mathcal{M}\}$ according to some notion of complexity, most of the time connected to the variance $d_m\sigma^2$ of $\widehat{f}_m$.

Examples of probability π_m

The risk r_m equals $\|(I-\text{Proj}_{S_m})f^*\|^2 + d_m\sigma^2$, where $d_m = \dim(S_m)$. Since we want the bound $\min_{m\in\mathcal{M}}\left\{r_m + \sigma^2\log\left(1/\pi_m\right)\right\}$ to be as close as possible to r_{m_o}, we build below probabilities π, such that $\log\left(1/\pi_m\right)$ remains of a size comparable to d_m. We refer to Section 2.2.1 for the definition of the collections of models $\{S_m, m\in\mathcal{M}\}$ in the different sparsity settings.

We will write henceforth C_p^d for $p!/(d!(n-d)!)$. It fulfills the following upper bound.

Lemma 2.1

For $0\le d\le p$, *we have the upper bound* $\log\left(C_p^d\right)\le d(1+\log(p/d))$, *with the convention* $0\log 0 = 0$.

Proof. The result is obvious for $d=0$ and $d=1$. Assume that the bound holds true for C_p^{d-1}. Since $(1+1/k)^k\le e$ for any integer k, we have

$$C_p^d = C_p^{d-1}\,\frac{p-d}{d} \le \left(\frac{ep}{d-1}\right)^{d-1}\frac{p}{d} \le \left(\frac{ep}{d}\right)^{d-1}\left(1+\frac{1}{d-1}\right)^{d-1}\frac{p}{d} \le \left(\frac{ep}{d}\right)^{d}.$$

We conclude by induction on d. □

Coordinate sparsity

In this setting, we slice the collection of models according to the cardinality of m, and we give the same probability to the C_p^d models with m of cardinality d. We consider two choices of probability π on $\mathcal{M}=\mathcal{P}(\{1,\ldots,p\})$.

1. A simple choice is $\pi_m = (1+1/p)^{-p}p^{-|m|}$, for which we have the upper bound $\log(1/\pi_m)\le 1+|m|\log(p)$ since $(1+1/p)^p\le e$.
2. Another choice is $\pi_m = \left(C_p^{|m|}\right)^{-1}e^{-|m|}(e-1)/(e-e^{-p})$. According to Lemma 2.1, we have

$$\log\left(\frac{1}{\pi_m}\right)\le \log(e/(e-1)) + |m|(2+\log(p/|m|)).$$

We notice that for the two above choices, $\log(\pi_m^{-1})$ roughly behaves like $2|m|\log(p)$ for $|m|$ small compared to p. There is then an additional $\log(p)$ factor compared to the AIC penalty $\text{pen}_{\text{AIC}}(m) = 2|m|\sigma^2$. This $\log(p)$ factor is due to the fact that the number of models with dimension d roughly grows with d, like $d\log(p)$ in the logarithmic scale, so this term shall appear in the definition of π_m. Exercise 2.9.1 and Section 2.4.1 show that this $\log(p)$ factor is actually unavoidable. Since the upper Bound (2.10) involves $\log(\pi_m^{-1})$, we then lose a $\log(p)$ factor between the upper Bound (2.10) and the oracle risk r_{m_o}.

Group sparsity

Here, we slice again the collection of models according to the cardinality of m, which corresponds to the number of groups G_k for which $\beta_{G_k} \neq 0$. As before, we consider two choices.

1. A simple choice $\pi_m = (1+1/M)^{-M} M^{-|m|}$, for which we have the upper bound $\log(1/\pi_m) \leq 1 + |m|\log(M)$.
2. Another choice is $\pi_m = \left(C_M^{|m|}\right)^{-1} e^{-|m|}(e-1)/(e-e^{-M})$, for which we have the upper bound $\log(1/\pi_m) \leq \log(e/(e-1)) + |m|(2+\log(M/|m|))$.

Sparse-group sparsity

In this case, we slice $\mathscr{M}$ according to both the number of groups and the cardinality of the nonzero elements in each groups. More precisely, for $m = (g, \{J_k\}_{k\in g})$, we set

$$\pi_m = \frac{1}{Z} \times \frac{e^{-|g|}}{C_M^{|g|}} \prod_{k\in g} \frac{e^{-|J_k|}}{C_{|G_k|}^{|J_k|}} \tag{2.11}$$

with Z such that $\sum_m \pi_m = 1$. In the case where all the G_k have the same size p/M, we have

$$Z = \frac{\alpha - \alpha^{M+1}}{1-\alpha} \quad \text{where } \alpha = \frac{1-e^{-p/M}}{e(e-1)}.$$

We have the upper bound

$$\log(1/\pi_m) \leq |g|(2+\log(M/|g|)) + \sum_{k\in g} |J_k|(2+\log(|G_k|/|J_k|)).$$

Variation sparsity

We can choose the same π_m as in the coordinate-sparse setting.

2.3 Risk Bound for Model Selection

2.3.1 Oracle Risk Bound

We consider a collection of models $\{S_m,\ m \in \mathscr{M}\}$, a probability distribution $\pi = \{\pi_m,\ m\in\mathscr{M}\}$ on $\mathscr{M}$, a constant $K > 1$, and the estimator $\widehat{f} = \widehat{f}_{\widehat{m}}$ given by (2.9). We have the following risk bound on $R(\widehat{f})$.

Theorem 2.2 Oracle risk bound for model selection

For the above choice of $\widehat{f}$, there exists a constant $C_K > 1$ depending only on $K > 1$, such that

$$\mathbb{E}\left[\|\widehat{f} - f^*\|^2\right] \leq C_K \min_{m\in\mathscr{M}} \left\{ \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] + \sigma^2 \log\left(\frac{1}{\pi_m}\right) + \sigma^2 \right\}. \tag{2.12}$$

Proof. The analysis of the risk of $\widehat{f}$ relies strongly on the Gaussian concentration Inequality (B.2) (see Appendix B, page 221). The strategy for deriving the risk Bound (2.12) is to start from Definition (2.9) of $\widehat{m}$, which ensures that

$$\|Y-\widehat{f}_{\widehat{m}}\|^2+\sigma^2\mathrm{pen}(\widehat{m})\leq\|Y-\widehat{f}_m\|^2+\sigma^2\mathrm{pen}(m),\quad\text{for all } m\in\mathcal{M},\tag{2.13}$$

and then to control the fluctuations of these quantities around their mean with the Gaussian concentration Inequality (B.2). The first step is to bound $\|\widehat{f}-f^*\|^2$ in terms of $\|\widehat{f}_m-f^*\|^2$ up to an additional random term. The second step is to upper bound the expectation of this additional random term.

1- Basic inequalities

Let us fix some $m\in\mathcal{M}$. Since $Y=f^*+\varepsilon$, expanding the square $\|\varepsilon+(f^*-\widehat{f})\|^2$ in the Inequality (2.13) gives

$$\|f^*-\widehat{f}\|^2\leq\|f^*-\widehat{f}_m\|^2+2\langle\varepsilon,f^*-\widehat{f}_m\rangle+\mathrm{pen}(m)\sigma^2+2\langle\varepsilon,\widehat{f}-f^*\rangle-\mathrm{pen}(\widehat{m})\sigma^2.\tag{2.14}$$

From Lemma A.3 in Appendix A, page 214, the random variable $\|\mathrm{Proj}_{S_m}\varepsilon\|^2$ follows a chi-square distribution of d_m degrees of freedom, so

$$\mathbb{E}\left[\langle\varepsilon,f^*-\widehat{f}_m\rangle\right]=\mathbb{E}\left[\langle\varepsilon,f^*-\mathrm{Proj}_{S_m}f^*\rangle\right]-\mathbb{E}\left[\|\mathrm{Proj}_{S_m}\varepsilon\|^2\right]=0-d_m\sigma^2,\tag{2.15}$$

which is nonpositive. Since

$$\begin{aligned}\mathrm{pen}(m)\sigma^2=K\left(\sqrt{d_m}+\sqrt{2\log(1/\pi_m)}\right)^2\sigma^2\ &\leq\ 2K\left(d_m+2\log(1/\pi_m)\right)\sigma^2\\ &\leq 2K\mathbb{E}\left[\|\widehat{f}_m-f^*\|^2\right]+4K\log(1/\pi_m)\,\sigma^2,\end{aligned}\tag{2.16}$$

in order to prove Bound (2.12), we only need to prove that for some constants $a>1$ and $c\geq 0$ there exists a random variable Z, such that

$$2\langle\varepsilon,\widehat{f}-f^*\rangle-\mathrm{pen}(\widehat{m})\leq a^{-1}\|\widehat{f}-f^*\|^2+Z\tag{2.17}$$

and $\mathbb{E}[Z]\leq c\sigma^2$. Actually, combining (2.14), (2.15), (2.16), and (2.17) we have

$$\frac{a-1}{a}\,\mathbb{E}\left[\|f^*-\widehat{f}\|^2\right]\ \leq\ (1+2K)\,\mathbb{E}\left[\|f^*-\widehat{f}_m\|^2\right]+4K\log(1/\pi_m)\,\sigma^2+c\sigma^2,$$

for any $m\in\mathcal{M}$, and Bound (2.12) follows.

Let us prove (2.17). We denote by $<f^*>$ the line spanned by f^*, by $\bar{S}_m$ the space $\bar{S}_m=S_m+<f^*>$, and by $\tilde{S}_m$ the orthogonal of $<f^*>$ in $\bar{S}_m$. In particular, $\bar{S}_m$ is the orthogonal sum of $<f^*>$ and $\tilde{S}_m$, which is denoted by $\bar{S}_m=<f^*>\bigoplus\tilde{S}_m$. Applying the inequality $2\langle x,y\rangle\leq a\|x\|^2+\|y\|^2/a$ for $a>1$, we obtain

$$\begin{aligned}2\langle\varepsilon,\widehat{f}-f^*\rangle\ &=\ 2\langle\mathrm{Proj}_{\bar{S}_{\widehat{m}}}\varepsilon,\widehat{f}-f^*\rangle\\ &\leq\ a\|\mathrm{Proj}_{\bar{S}_{\widehat{m}}}\varepsilon\|^2+a^{-1}\|\widehat{f}-f^*\|^2\\ &\leq\ aN^2\sigma^2+aU_{\widehat{m}}\sigma^2+a^{-1}\|\widehat{f}-f^*\|^2,\end{aligned}$$

where $N^2 = \|\mathrm{Proj}_{<f^*>}\varepsilon\|^2/\sigma^2$ and $U_m = \|\mathrm{Proj}_{\tilde{S}_m}\varepsilon\|^2/\sigma^2$. According to Lemma A.3, N^2 follows a χ^2 distribution of dimension 1, and U_m follows a χ^2 distribution of dimension d_m if $f^* \notin S_m$ and $d_m - 1$ if $f^* \in S_m$. We then have (2.17), with $Z = aN^2\sigma^2 + aU_{\widehat{m}}\sigma^2 - \mathrm{pen}(\widehat{m})\sigma^2$. All we need is to prove that there exists a constant $c \geq 0$, such that

$$\mathbb{E}[Z] = a\sigma^2 + \sigma^2\mathbb{E}\left[aU_{\widehat{m}} - \mathrm{pen}(\widehat{m})\right] \leq c\sigma^2. \tag{2.18}$$

We emphasize that the index $\widehat{m}$ depends on ε, so it is not independent of the sequence $\{U_m,\ m \in \mathcal{M}\}$. As a consequence, even if each U_m follows a χ^2 distribution, the variable $U_{\widehat{m}}$ does not follow a χ^2 distribution.

Stochastic control of $\mathbb{E}\left[aU_{\widehat{m}} - \mathrm{pen}(\widehat{m})\right]$

We remind the reader that $\mathrm{pen}(m) = K\left(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)}\right)^2$. In the following, we choose $a = (K+1)/2 > 1$, and we start from the basic inequality

$$\begin{aligned}
&\mathbb{E}\left[\frac{K+1}{2}U_{\widehat{m}} - \mathrm{pen}(\widehat{m})\right] \\
&\leq \frac{K+1}{2}\mathbb{E}\left[\max_{m\in\mathcal{M}}\left(U_m - \frac{2}{K+1}\mathrm{pen}(m)\right)\right] \\
&\leq \frac{K+1}{2}\sum_{m\in\mathcal{M}}\mathbb{E}\left[\left(U_m - \frac{2K}{K+1}\left(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)}\right)^2\right)_+\right].
\end{aligned} \tag{2.19}$$

The map $\varepsilon \to \|\mathrm{Proj}_{\tilde{S}_m}\varepsilon\|$ is 1-Lipschitz, so the Gaussian concentration Inequality (B.2), page 221, ensures that for each m there exists ξ_m with exponential distribution, such that

$$\|\mathrm{Proj}_{\tilde{S}_m}\varepsilon\| \leq \mathbb{E}\left[\|\mathrm{Proj}_{\tilde{S}_m}\varepsilon\|\right] + \sigma\sqrt{2\xi_m}.$$

Since $U_m = \|\mathrm{Proj}_{\tilde{S}_m}\varepsilon\|^2/\sigma^2$ follows a χ^2 distribution with d_m or $d_m - 1$ degrees of freedom, we have $\mathbb{E}\left[\|\mathrm{Proj}_{\tilde{S}_m}\varepsilon\|\right] \leq \mathbb{E}\left[\|\mathrm{Proj}_{\tilde{S}_m}\varepsilon\|^2\right]^{1/2} \leq \sigma\sqrt{d_m}$. As a consequence, we have the upper bound

$$\begin{aligned}
U_m &\leq \left(\sqrt{d_m} + \sqrt{2\xi_m}\right)^2 \\
&\leq \left(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)} + \sqrt{2(\xi_m - \log(1/\pi_m))_+}\right)^2 \\
&\leq \frac{2K}{K+1}\left(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)}\right)^2 + \frac{4K}{K-1}(\xi_m - \log(1/\pi_m))_+,
\end{aligned}$$

where we used for the last line the inequality $(x+y)^2 \leq (1+\alpha)x^2 + (1+\alpha^{-1})y^2$, with $\alpha = (K-1)/(K+1)$. Since $\mathbb{E}\left[(\xi_m - \log(1/\pi_m))_+\right] = \exp(-\log(1/\pi_m)) = \pi_m$, we have

$$\mathbb{E}\left[\left(U_m - \frac{2K}{K+1}\left(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)}\right)^2\right)_+\right] \leq \frac{4K}{K-1}\,\pi_m\,.$$

Combining this bound with (2.19), we finally end with

$$\mathbb{E}\left[\frac{K+1}{2}U_{\widehat{m}} - \text{pen}(\widehat{m})\right] \leq \frac{2K(K+1)}{K-1}\sigma^2 \sum_{m\in\mathscr{M}} \pi_m = \frac{2K(K+1)}{K-1}\sigma^2, \qquad (2.20)$$

since $\sum_{m\in\mathscr{M}} \pi_m = 1$. This proves (2.18) and thus (2.17). The proof of (2.16) is complete. □

Remark 1. We observe from the last inequality in the above proof that Theorem 2.2 remains valid when $\{\pi_m : m \in \mathscr{M}\}$ do not sum to one, as long as the sum $\sum_{m\in\mathscr{M}} \pi_m$ remains small (since the constant C_K is proportional to $\sum_{m\in\mathscr{M}} \pi_m$). Asking for $\{\pi_m : m \in \mathscr{M}\}$ to sum to one is therefore not mandatory, it is simply a convenient convention.

Remark 2. Let us explain the shape of Penalty (2.9) in light of the above proof. From Inequality (2.17) we observe that we need a penalty, such that there exist some $a > 1$, $c \geq 0$, and some random variables Z_m fulfilling

$$2\langle \varepsilon, \widehat{f}_m - f^* \rangle - \text{pen}(m) \leq a^{-1}\|\widehat{f}_m - f^*\|^2 + Z_m, \quad \text{for all } m \in \mathscr{M},$$

and $\mathbb{E}\left[\sup_{m\in\mathscr{M}} Z_m\right] \leq c\sigma^2$. Since

$$\begin{aligned} 2\langle \varepsilon, \widehat{f}_m - f^* \rangle &\leq a\left\langle \varepsilon, \frac{\widehat{f}_m - f^*}{\|\widehat{f}_m - f^*\|}\right\rangle^2 + a^{-1}\|\widehat{f}_m - f^*\|^2 \\ &\leq a \sup_{f\in\bar{S}_m}\left\langle \varepsilon, \frac{f}{\|f\|}\right\rangle^2 + a^{-1}\|\widehat{f}_m - f^*\|^2 \\ &\leq a\|\text{Proj}_{\bar{S}_m}\varepsilon\|^2 + a^{-1}\|\widehat{f}_m - f^*\|^2, \end{aligned}$$

we need a penalty fulfilling $\mathbb{E}\left[\sup_{m\in\mathscr{M}}\left(a\|\text{Proj}_{\bar{S}_m}\varepsilon\|^2 - \text{pen}(m)\sigma^2\right)\right] \leq c\sigma^2$. As explained in the second part of the proof of Theorem 2.2, Penalty (2.9) fulfills such a bound; see Bound (2.20).

Remark 3. We mention that there are some alternatives to Criterion (2.9) for selecting a model, for example the Goldenshluger–Lepski method. We refer to Exercise 2.9.5 for a simple version of this method and to Section 2.8 for references.

Discussion of Theorem 2.2

The risk Bound (2.12) shows that the estimator (2.9) almost performs the best trade-off between the risk r_m and the complexity term $\log(\pi_m^{-1})\sigma^2$. For collections of sufficiently small complexity, we can choose a probability π, such that $\log(\pi_m^{-1})$ is of the same order as d_m. In this case, r_m is larger (up to a constant) than the complexity term $\log(\pi_m^{-1})\sigma^2$, and the risk of the estimator $\widehat{f}$ is bounded by a constant times the risk of the oracle $\widehat{f}_{m_o}$; see Exercise 2.9.3 for an example. In particular, the AIC criterion can produce some good results in this case. For some very large collections, such as the one for coordinate sparsity, we cannot choose $\log(\pi_m^{-1})$ of the same order as d_m, and Exercise 2.9.1 shows that the AIC criterion fails. In this case, the Bound (2.12) cannot be directly compared to the oracle risk $r_{m_o} = R(\widehat{f}_{m_o})$. We refer to Exercise 2.9.2 for a discussion of Bound (2.12) in the coordinate-sparse, the group-sparse, and the sparse group-sparse settings. In these cases where (2.12) cannot be directly compared to the oracle risk r_{m_o}, two natural questions arise:

1. Can an estimator perform significantly better than the estimator (2.9)?
2. Can we choose a smaller penalty in (2.9)?

We address these two issues in the next section.

2.4 Optimality

2.4.1 Minimax Optimality

We investigate in this section the optimality of the estimator (2.9). For simplicity, we focus on the example of coordinate sparsity. Similar results can be proved in the other sparse settings. For $D \in \{1, \dots, p\}$, we define

$$V_D(\mathbf{X}) = \{\mathbf{X}\beta : \beta \in \mathbb{R}^p, |\beta|_0 = D\} \quad \text{where } |\beta|_0 = \text{Card}\left\{j : \ \beta_j \neq 0\right\}.$$

We address in this section the two following issues:

1. What is the best risk that an estimator $\widehat{f}$ can achieve uniformly on $f^* \in V_D(X)$?
2. Is the risk of $\widehat{f} = \widehat{f}_{\widehat{m}}$ given by (2.9) of the same size as this optimal risk?

Let us first upper bound the risk of the estimator (2.9) when $f^* \in V_D(\mathbf{X})$. We write in the following $\mathbb{P}_f$ for the Gaussian distribution $\mathcal{N}(f, \sigma^2 I_n)$ and the expectation $\mathbb{E}_f$ will refer to the expectation when the vector of observations Y is distributed according to $\mathbb{P}_f$. We choose the collection of models as in Section 2.2.1 for the coordinate-sparse setting and the probability

$$\pi_m = \left(C_p^{|m|}\right)^{-1} e^{-|m|} \, \frac{e-1}{e-e^{-p}}.$$

Then, the risk Bound (2.12) ensures that there exists a constant $C'_K > 1$, such that $\mathbb{E}_{f^*}\left[\|\widehat{f} - f^*\|^2\right] \leq C'_K D \log(p/D)\, \sigma^2$ uniformly on all $f^* \in V_D(\mathbf{X})$, all matrices $\mathbf{X}$, all $n, p \in \mathbb{N}$, and all $D \leq p/2$. In particular, for all $n, p \in \mathbb{N}$ and $D \leq p/2$, we have

$$\sup_{\mathbf{X}} \sup_{f^* \in V_D(\mathbf{X})} \mathbb{E}_{f^*}\left[\|\widehat{f} - f^*\|^2\right] \leq C'_K D \log\left(\frac{p}{D}\right) \sigma^2. \tag{2.21}$$

Let us now investigate whether there exists an estimator with a significantly better risk on $V_D(X)$. We define the *minimax risk* on $V_D(\mathbf{X})$ as

$$\mathbf{R}[\mathbf{X}, D] = \inf_{\widehat{f}} \sup_{f^* \in V_D(\mathbf{X})} \mathbb{E}_{f^*}\left[\|\widehat{f} - f^*\|^2\right],$$

where the infimum is taken over all the estimators. It corresponds to the best risk that an estimator $\widehat{f}$ can achieve uniformly on $f^* \in V_D(X)$.

For any integer $D_{\max}$ not larger than $p/2$, we introduce the restricted isometry constants

$$\underline{c}_{\mathbf{X}} := \inf_{\beta : |\beta|_0 \leq 2D_{\max}} \frac{\|\mathbf{X}\beta\|}{\|\beta\|} \leq \sup_{\beta : |\beta|_0 \leq 2D_{\max}} \frac{\|\mathbf{X}\beta\|}{\|\beta\|} =: \bar{c}_{\mathbf{X}}. \qquad (2.22)$$

We have the following lower bound on $\mathbf{R}[\mathbf{X}, D]$.

Theorem 2.3 Minimax risk for coordinate-sparse regression

Let us fix some $D_{\max} \leq p/5$. For any $D \leq D_{\max}$, we have the lower bound

$$\mathbf{R}[\mathbf{X}, D] \geq \frac{e}{8(2e+1)^2} \left(\frac{\underline{c}_{\mathbf{X}}}{\bar{c}_{\mathbf{X}}} \right)^2 D \log\left(\frac{p}{5D} \right) \sigma^2, \qquad (2.23)$$

with $\underline{c}_{\mathbf{X}}$ and $\bar{c}_{\mathbf{X}}$ defined by (2.22).

Before proving this result, let us compare this bound with (2.21). We observe that the lower bound in Theorem 2.3 and the upper bound in (2.21) are similar, except that the lower bound involves the ratio $(\underline{c}_{\mathbf{X}}/\bar{c}_{\mathbf{X}})^2$. We emphasize that this ratio can be equal to 0 for $D_{\max}$ large, for example, when $2D_{\max} \geq 1 + \text{rank}(\mathbf{X})$. Yet, there exists some designs $\mathbf{X}$ for which this ratio is non vanishing for $D_{\max}$ of the order of $n/\log(p)$. For example, if the matrix $\mathbf{X} = [X_{i,j}]_{i,j}$ is obtained by sampling each entry $X_{i,j}$ independently according to a standard Gaussian distribution, then Lemma 7.4 in Chapter 7 ensures that with large probability, for $D_{\max} \leq n/(32\log(p))$, we have $\underline{c}_{\mathbf{X}}/\bar{c}_{\mathbf{X}} \geq 4$. As a consequence, there exists a numerical constant $C > 0$, such that for all $\sigma^2 > 0$, for all $n, p \in \mathbb{N}$, and for all D smaller than $p/5$ and $n/(32\log(p))$, we have

$$\sup_{\mathbf{X}} \mathbf{R}[\mathbf{X}, D] \geq CD \log\left(\frac{p}{5D} \right) \sigma^2. \qquad (2.24)$$

Combining (2.24) with (2.21), we obtain for any D smaller than $p/5$ and $n/(32\log(p))$

$$CD\log\left(\frac{p}{5D}\right)\sigma^2 \leq \sup_{\mathbf{X}} \mathbf{R}[\mathbf{X}, D] \leq \sup_{\mathbf{X}} \sup_{f^* \in V_D(\mathbf{X})} \mathbb{E}_{f^*}\left[\|\widehat{f} - f^*\|^2\right] \leq C'_K D \log\left(\frac{p}{D}\right)\sigma^2.$$

Up to the size of the constants C and C'_K, the lower and upper bounds have the same form. In this sense, the estimation procedure (2.9) is optimal.

Let us now prove Theorem 2.3. The path for deriving minimax lower bounds is quite standard. It mainly relies on the following lemma (or some version of it), which enlightens that the minimax risk on a set V depends on the ability of set V to include a subset $\mathcal{V} \subset V$ of well-separated points with limited diameter. This gives an insight into the link between the minimax risk on V and the (algebraic or metric) dimension of V. Below, we use the notation $x \vee y = \max(x, y)$.

Lemma 2.4 A general lower bound

For any finite set $\mathscr{V} \subset \mathbb{R}^n$, we have the lower bound

$$\inf_{\widehat{f}} \max_{f \in \mathscr{V}} \mathbb{E}_f \left[\|\widehat{f} - f\|^2 \right] \geq$$

$$\frac{1}{4}\left(1 - \left(\frac{2e}{2e+1} \bigvee \max_{f \neq f' \in \mathscr{V}} \frac{\|f - f'\|^2}{2\sigma^2 \log(|\mathscr{V}|)}\right)\right) \times \min_{f \neq f' \in \mathscr{V}} \|f - f'\|^2. \quad (2.25)$$

Proof of the Lemma 2.4

Let $\widehat{f}$ be any estimator. We denote by $\widetilde{f} \in \operatorname{argmin}_{f \in \mathscr{V}} \|\widehat{f} - f\|$ a measurable projection of $\widehat{f}$ on $\mathscr{V}$. For any $f \in \mathscr{V}$, we have

$$\|f - \widetilde{f}\| \leq \|f - \widehat{f}\| + \|\widetilde{f} - \widehat{f}\| \leq 2\|f - \widehat{f}\|$$

and

$$\mathbb{E}_f \left[\|\widetilde{f} - f\|^2 \right] \geq \mathbb{P}_f \left[\widetilde{f} \neq f \right] \times \min_{f \neq f' \in \mathscr{V}} \|f' - f\|^2.$$

Combining the two above inequalities, we obtain

$$\max_{f \in \mathscr{V}} \mathbb{E}_f \left[\|\widehat{f} - f\|^2 \right] \geq \frac{1}{4} \max_{f \in \mathscr{V}} \mathbb{P}_f \left[\widetilde{f} \neq f \right] \times \min_{f \neq f' \in \mathscr{V}} \|f' - f\|^2. \quad (2.26)$$

In order to obtain a lower bound independent of $\widehat{f}$, we need to bound from below the probability

$$\max_{f \in \mathscr{V}} \mathbb{P}_f \left[\widetilde{f} \neq f \right] = 1 - \min_{f \in \mathscr{V}} \mathbb{P}_f \left[\widetilde{f} = f \right]$$

uniformly on $\widetilde{f}$. Such a bound can be obtained thanks to Birgé's lemma (stated and proved in Theorem B.10, Appendix B, page 227), which ensures that for any family of disjointed events $\{A_f : f \in \mathscr{V}\}$, we have

$$\min_{f \in \mathscr{V}} \mathbb{P}_f \left[A_f \right] \leq \frac{2e}{2e+1} \bigvee \max_{f \neq f' \in \mathscr{V}} \frac{\mathscr{K}(f, f')}{\log(|\mathscr{V}|)},$$

where $\mathscr{K}(f, f') = \int \log(d\mathbb{P}_f / d\mathbb{P}_{f'})\, d\mathbb{P}_f$ is the Kullback–Leibler divergence between $\mathbb{P}_f$ and $\mathbb{P}_{f'}$. Computing the Kullback–Leibler divergence $\mathscr{K}(f, f')$, we obtain

$$\mathscr{K}(f, f') = \mathbb{E}_f \left[\log\left(\frac{\exp\left(-\|Y - f\|^2/(2\sigma^2)\right)}{\exp\left(-\|Y - f'\|^2/(2\sigma^2)\right)} \right) \right]$$

$$= \frac{1}{2\sigma^2} \mathbb{E}_f \left[\|Y - f'\|^2 - \|Y - f\|^2 \right] = \frac{\|f - f'\|^2}{2\sigma^2}.$$

The events $A_f = \left\{ \widetilde{f} = f \right\}$ are disjointed, so Birgé's lemma ensures the lower bound

$$\max_{f \in \mathscr{V}} \mathbb{P}_f \left[\widetilde{f} \neq f \right] \geq 1 - \left(\frac{2e}{2e+1} \bigvee \max_{f \neq f' \in \mathscr{V}} \frac{\|f - f'\|^2}{2\sigma^2 \log(|\mathscr{V}|)} \right).$$

Combining this inequality with (2.26), we obtain (2.25). □

The lower Bound (2.25) involves two opposite terms. In order to be bounded away from zero, the first term requires that the square of the diameter of $\mathscr{V}$ divided by the logarithm of its cardinality is bounded away from $2\sigma^2$. On the contrary, in order to be large, the second term requires that the elements of $\mathscr{V}$ are well-scattered. For $\mathscr{V} \subset V$, the minimax risk on V is larger than the minimax risk on $\mathscr{V}$. Then, as mentioned above, for any set V in $\mathbb{R}^n$, the magnitude of the minimax risk on V depends on the ability of set V to include a set $\mathscr{V}$ of well-separated points with limited diameter. The remaining work for deriving a lower bound on the minimax risk on V is to exhibit a subset $\mathscr{V} \subset V$ of well-separated points with diameter bounded away from $\sigma\sqrt{2\log(|\mathscr{V}|)}$. We build below such a subset $\mathscr{V}$ for $V = V_D(\mathbf{X})$.

Proof of Theorem 2.3

In light of Lemma 2.4, we will build a subset $\mathscr{V} \subset V_D(\mathbf{X})$ of well-separated points fulfilling

$$\max_{f \neq f' \in \mathscr{V}} \frac{\|f - f'\|^2}{2\sigma^2} \leq \frac{2e}{2e+1} \log(|\mathscr{V}|). \tag{2.27}$$

We rely on the following combinatorial lemma.

Lemma 2.5 Spreading points in a sparse hypercube

For any positive integer D less than $p/5$, there exists a set $\mathscr{C}$ in $\{0,1\}^p_D := \{x \in \{0,1\}^p : |x|_0 = D\}$, fulfilling

$$\begin{aligned} & |\beta - \beta'|_0 > D, \quad \textit{for all } \beta \neq \beta' \in \mathscr{C} \\ \textit{and} \quad & \log|\mathscr{C}| \geq \frac{D}{2}\log\left(\frac{p}{5D}\right). \end{aligned}$$

We refer to Exercise 2.9.6, page 58, for a proof of this lemma.

The set $\mathscr{C}$ of Lemma 2.5 gathers coordinate-sparse vectors, which are well-separated. The set $\mathscr{V} = \mathbf{X}\mathscr{C}$ is included in $V_D(\mathbf{X})$, yet it may not fulfill (2.27). In order to fulfill this condition, we rescale this set by an appropriate factor $r > 0$.

We set

$$r^2 = \frac{e}{2(2e+1)} \times \frac{\sigma^2}{\bar{c}_{\mathbf{X}}^2} \times \log\left(\frac{p}{5D}\right),$$

with $\bar{c}_{\mathbf{X}}$ defined by (2.22) and $\mathscr{V} = \{r\mathbf{X}\beta : \beta \in \mathscr{C}\}$. For any $\beta, \beta' \in \mathscr{C}$, we have $|\beta - \beta'|_0 \leq 2D$ and $\|\beta - \beta'\|^2 \leq 2\|\beta\|^2 + 2\|\beta'\|^2 \leq 4D$, so

$$\begin{aligned} \max_{f \neq f' \in \mathscr{V}} \frac{\|f - f'\|^2}{2\sigma^2} &= r^2 \max_{\beta \neq \beta' \in \mathscr{C}} \frac{\|\mathbf{X}(\beta - \beta')\|^2}{2\sigma^2} \\ &\leq \bar{c}_{\mathbf{X}}^2 r^2 \max_{\beta \neq \beta' \in \mathscr{C}} \frac{\|\beta - \beta'\|^2}{2\sigma^2} \\ &\leq \bar{c}_{\mathbf{X}}^2 r^2 \frac{4D}{2\sigma^2} \leq \frac{2e}{2e+1} \times \frac{D}{2}\log\left(\frac{p}{5D}\right) \leq \frac{2e}{2e+1}\log(|\mathscr{C}|). \end{aligned}$$

When $\underline{c}_{\mathbf{X}} = 0$, Theorem 2.3 is obvious, so we focus in the following on the case where $\underline{c}_{\mathbf{X}} > 0$. Then $|\mathscr{V}| = |\mathscr{C}|$ and combining Lemma 2.4 with the above inequality gives

$$\begin{aligned}\inf_{\widehat{f}} \max_{f\in\mathscr{V}} \mathbb{E}_f\left[\|\widehat{f}-f\|^2\right] &\geq \frac{1}{4(2e+1)} \min_{f\neq f'\in\mathscr{V}} \|f-f'\|^2 \\ &\geq \frac{\underline{c}_{\mathbf{X}}^2 r^2}{4(2e+1)} \min_{\beta\neq\beta'\in\mathscr{C}} \|\beta-\beta'\|^2.\end{aligned}$$

Since $\mathscr{C} \subset \{0,1\}^D$ and $|\beta-\beta'|_0 > D$ for all $\beta \neq \beta' \in \mathscr{C}$, we have $\|\beta-\beta'\|^2 \geq D$ for all $\beta \neq \beta' \in \mathscr{C}$, and then

$$\inf_{\widehat{f}} \max_{f\in\mathscr{V}} \mathbb{E}_f\left[\|\widehat{f}-f\|^2\right] \geq \frac{\underline{c}_{\mathbf{X}}^2 r^2 D}{4(2e+1)} \geq \frac{e}{8(2e+1)^2}\left(\frac{\underline{c}_{\mathbf{X}}}{\bar{c}_{\mathbf{X}}}\right)^2 D\log\left(\frac{p}{5D}\right)\sigma^2.$$

The proof of Theorem 2.3 is complete. □

2.4.2 Frontier of Estimation in High Dimensions

The minimax lower Bound (2.23) provides insightful information on the frontier between the statistical problems that can be successfully solved and those that are hopeless. Again, we only discuss the sparse coordinate case.

The prediction risk $\mathbb{E}\left[\|\widehat{f}-f^*\|^2\right]$ considered in this chapter involves the expectation of the square of the Euclidean norm of an n-dimensional vector. Since the square of the Euclidean norm of an n-dimensional vector with entries of order $\varepsilon > 0$ grows like $n\varepsilon^2$ with n, it is meaningful to discuss the accuracy of an estimator in terms of the size of the scaled risk $\mathbb{E}\left[n^{-1}\|\widehat{f}-f^*\|^2\right]$.

From Bounds (2.21) and (2.23), we know that the minimax scaled-risk

$$\inf_{\widehat{f}} \sup_{f^*\in V_D(\mathbf{X})} \mathbb{E}_{f^*}\left[n^{-1}\|\widehat{f}-f^*\|^2\right] \quad \text{is of order} \quad \frac{D}{n}\log\left(\frac{p}{D}\right)\sigma^2,$$

as long as $\underline{c}_{\mathbf{X}}/\bar{c}_{\mathbf{X}} \approx 1$ and $D \leq p/5$. In practice, it means that when $f^* \in V_D(\mathbf{X})$ with $D\sigma^2\log(p/D)$ small compared to n, a procedure like (2.9) will produce an accurate estimation of f^*. On the contrary, when $D\sigma^2\log(p/D)$ is large compared to n, no estimator can provide a reliable estimation uniformly on $f^* \in V_D(\mathbf{X})$. In particular, if the dimension p is larger than e^n, accurate estimation is hopeless.

2.4.3 Minimal Penalties

We consider again the example of coordinate sparsity. Can we choose a smaller penalty than $\text{pen}(m) = K\left(\sqrt{d_m}+\sqrt{2\log(1/\pi_m)}\right)^2$, with $K > 1$ and π_m described in Section 2.2.2 ?

In the coordinate-sparse and group-sparse settings the answer is no, in the sense that we can prove that when $f^* = 0$, the selected $\widehat{m}$ can have a very large size when $K < 1$; see Exercise 2.9.1, parts B and C. In particular, the AIC criterion can lead to a strong overfitting.

2.5 Computational Issues

We have seen in the previous sections that the estimator (2.9) has some very nice statistical properties. Unfortunately, in many cases, it cannot be implemented in practice due to its prohibitive computational complexity.

Actually, computing (2.9) requires to screen the whole family $\mathscr{M}$ except in a few cases listed below. The size of the families $\mathscr{M}$ introduced in Section 2.2.1 becomes huge for moderate or large p. For example, the cardinality of the family $\mathscr{M}$ corresponding to the coordinate-sparse setting is 2^p, for the group-sparse setting it is 2^M, etc. As a consequence, the complexity of the selection procedure (2.9) becomes prohibitive for p or M larger than a few tens; see the table below.

number of variables (or groups)	p	10	20	40	80	270
cardinality of $\mathscr{M}$	2^p	1 024	$1.1\,10^6$	$1.1\,10^{12}$	$1.2\,10^{24}$	number of particles in the universe

Model selection is therefore a very powerful conceptual tool for understanding the nature of high-dimensional statistical problems, but it cannot be directly implemented except in the few cases listed below. Yet, it is a useful baseline for deriving computationally feasible procedures, as we will see in Chapter 4 and 5.

Cases where model selection can be implemented

There are two cases where we can avoid to screen the whole family $\mathscr{M}$ for minimizing (2.9).

1. The first case is when the columns of $\mathbf{X}$ are orthogonal. Minimizing (2.9) then essentially amounts to threshold the values of $\mathbf{X}_j^T Y$, for $j = 1, \ldots, p$, which has a very low complexity; see Exercise 2.9.1, part A, for details.
2. The second case is for the variation sparse setting, where minimizing (2.9) can be performed via dynamic programming with an overall complexity of the order of n^3. We refer to Exercise 2.9.4 for the details.

Finally, we emphasize that for small families $\{S_m,\ m \in \mathscr{M}\}$ of models (for example, given by a PCA of $\mathbf{X}$), the algorithmic complexity remains small and the estimator (2.9) can be implemented; see Exercise 2.9.3 for examples.

Approximate computation

As mentioned above, some computationally efficient procedures can be derived from the model selection Criterion (2.9). We refer to Chapters 4 and 5 for details. An alternative point of view is to perform an approximate minimization of Criterion (2.9). Many algorithms, of a deterministic or stochastic nature, have been proposed. We briefly described one of the most popular ones: the forward–backward minimization.

For simplicity, we restrict to the coordinate-sparse setting and only describe one of the multiple versions of the forward–backward algorithm. The principe of forward–backward minimization is to approximately minimize Criterion (2.9) by alternatively trying to add and remove variables from the model. More precisely, it starts from the null model and then builds a sequence of models $(m_t)_{t\in\mathbb{N}}$, where two consecutive models differ by only one variable. At each step t, the algorithm alternatively tries to add or remove a variable to m_t in order to decrease Criterion (2.9). When the criterion cannot be improved by the addition or deletion of one variable, the algorithm stops and returns the current value. Below $\text{crit}(m)$ refers to Criterion (2.9).

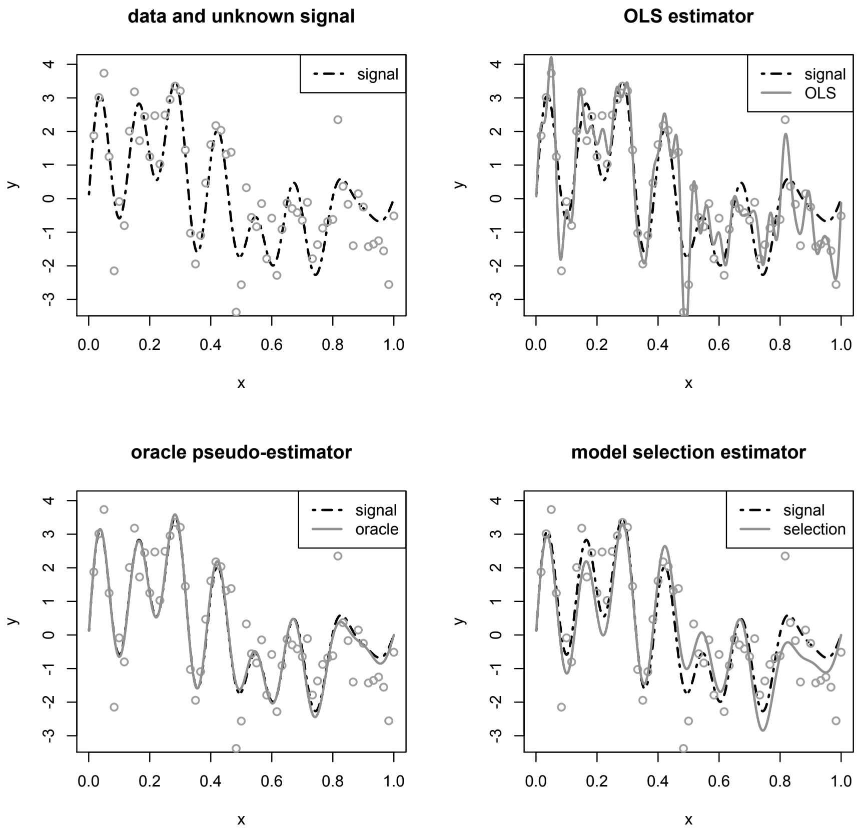
Forward–backward algorithm

```
Initialization: start from $m_0 = \emptyset$ and $t = 0$.
Iterate: until convergence
  • forward step:
      – search $j_{t+1} \in \text{argmin}_{j\notin m_t} \text{crit}(m_t \cup \{j\})$
      – if $\text{crit}(m_t \cup \{j_{t+1}\}) \leq \text{crit}(m_t)$ then $m_{t+1} = m_t \cup \{j_{t+1}\}$ else $m_{t+1} = m_t$
      – increase $t$ of one unit
  • backward step:
      – search $j_{t+1} \in \text{argmin}_{j\in m_t} \text{crit}(m_t \setminus \{j\})$
      – if $\text{crit}(m_t \setminus \{j_{t+1}\}) < \text{crit}(m_t)$ then $m_{t+1} = m_t \setminus \{j_{t+1}\}$ else $m_{t+1} = m_t$
      – increase $t$ of one unit
Output: $\widehat{f}_{m_t}$
```

In practice, the forward–backward algorithm usually converges quickly. So the final estimator $\widehat{f}_{m_t}$ can be computed efficiently. Yet, we emphasize that there is no guarantee at all, that the final estimator $\widehat{f}_{m_t}$ has some good statistical properties. Some strong results are nevertheless given by Zhang [130] for a variant of this algorithm.

2.6 Illustration

We briefly illustrate the theory of model selection on a simulated example. The graphics are reported in Figure 2.2. We have $n = 60$ noisy observation $Y_i = f(i/n) + \varepsilon_i$ (gray dots in Figure 2.2) of an unknown undulatory signal $f : [0,1] \to \mathbb{R}$ (dotted


data and unknown signal

OLS estimator

oracle pseudo-estimator

model selection estimator

Figure 2.2 *Top-left: The observed data (gray points) and the unknown signal f (dotted line) that we want to estimate. Top-right: OLS estimator (in gray). Bottom-left: Oracle pseudo-estimator (gray). Bottom-right: The estimator (2.9) in gray.*

curve) at times $i/60$ for $i = 1,\dots,60$, with $\varepsilon_1,\dots,\varepsilon_{60}$ i.i.d. with $\mathcal{N}(0,1)$ Gaussian distribution. We want to estimate this signal by expanding the observations on the Fourier basis. More precisely, we consider the setting of Example 2, page 28, with the φ_j given by

$$\begin{cases} \varphi_1(x) = 1 & \\ \varphi_{2k}(x) = \cos(2\pi kx) & \text{for } k = 1,\dots,25 \\ \varphi_{2k+1}(x) = \sin(2\pi kx) & \text{for } k = 1,\dots,25. \end{cases}$$

The upper-right plot in Figure 2.2 represents the classical estimator $\widehat{f}^{\mathrm{OLS}}$ (in gray) obtained by simple likelihood maximization

$$\widehat{f}^{\mathrm{OLS}}(x) = \sum_{j=1}^{p} \widehat{\beta}_j^{\mathrm{OLS}} \varphi_j(x) \quad \text{with} \quad \widehat{\beta}^{\mathrm{OLS}} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{p} \beta_j \varphi_j(i/n) \Big)^2 .$$

We observe that the estimator $\widehat{f}^{\mathrm{OLS}}$ is overly wavy and provides a poor estimation of f. If we knew the oracle model m_o, then we could estimate f with $\widehat{f}_{m_o}$. This pseudo-estimator is plotted in gray at the bottom left of Figure 2.2. We observe that it is very close to f. In practice, we do not know the oracle model m_o, but we can select a model $\widehat{m}$ according to the model selection procedure (2.9), with a collection of models and a probability distribution π corresponding to the coordinate-sparse setting. The constant K in the penalty term is chosen to be equal to 1.1. The resulting estimator $\widehat{f}_{\widehat{m}}$ is plotted in gray at the bottom right of Figure 2.2. We observe that this estimator is much better than $\widehat{f}^{\mathrm{OLS}}$, and it is almost as good as $\widehat{f}_{m_o}$.

2.7 An Alternative Point of View on Model Selection

Before concluding this chapter, we point out an alternative strategy for deriving a selection criterion based on the Bayesian paradigm. In this perspective, the mean f^* of Y is assumed to be the outcome of the following sampling scheme: A model S_{m^*} is sampled according to a distribution $(\pi_m)_{m\in\mathcal{M}}$, and then f^* is sampled according to a distribution $d\Pi(f|m^*)$ on S_{m^*}. The mean f^* of Y is then sampled according to the distribution

$$d\Pi(f) = \sum_{m\in\mathcal{M}} \pi_m \, d\Pi(f|m),$$

with $(\pi_m)_{m\in\mathcal{M}}$ a distribution on $\mathcal{M}$ and $d\Pi(f|m)$ a distribution on S_m. In such a probabilistic setting, where both m^*, f^*, and Y are random variables, it makes sense to consider the probability $\Pi(m|Y)$ of a model S_m given the observations Y. Computing this conditional probability, we obtain

$$\Pi(m|Y) = \frac{\pi_m d\pi(Y|m)}{d\pi(Y)} = \frac{\pi_m \int_{f\in S_m} e^{-\|Y-f\|^2/(2\sigma^2)}\, d\Pi(f|m)}{\sum_{m'} \pi_{m'} \int_{f\in S_{m'}} e^{-\|Y-f\|^2/(2\sigma^2)}\, d\Pi(f|m')} .$$

Under some technical conditions on the distribution $d\Pi(f|m)$, an asymptotic expansion when $n \to \infty$ gives for $m, m' \in \mathcal{M}$

$$\log\left(\frac{\Pi(m|Y)}{\Pi(m'|Y)}\right) \overset{n\to\infty}{\approx} \frac{\|Y-\widehat{f}_{m'}\|^2 - \|Y-\widehat{f}_m\|^2}{2\sigma^2} + \frac{d_{m'}-d_m}{2}\log(n) + \log\left(\frac{\pi_m}{\pi_{m'}}\right) + O(1); \quad (2.28)$$

see [119]. This asymptotic expansion suggests to choose m by minimizing the criterion

$$\operatorname{crit}(m) = \|Y-\widehat{f}_m\|^2 + \sigma^2 d_m \log(n) + 2\sigma^2 \log\left(\pi_m^{-1}\right). \quad (2.29)$$

Assuming a uniform distribution π_m on $\mathcal{M}$, we obtain the popular Bayesian Information Criterion (BIC)

$$\widehat{m}_{\mathrm{BIC}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \|Y - \widehat{f}_m\|^2 + \sigma^2 d_m \log(n) \right\}. \tag{2.30}$$

The term $\sigma^2 d_m \log(n)$ appearing in the BIC criterion increases more rapidly than the term $2d_m\sigma^2$ appearing in the AIC criterion. Yet, the model $\widehat{m}_{\mathrm{BIC}}$ can still be overly large when there is an exponential number of models per dimension; see Exercise 2.9.1, part B. This failure has two origins. First, the asymptotic expansion (2.28) fails to be valid in a high-dimensional setting. Second, the choice of a uniform distribution π on $\mathcal{M}$ does not enable the BIC estimator $\widehat{f}_{\widehat{m}_{\mathrm{BIC}}}$ to adapt to the complexity of $\mathcal{M}$ by taking into account the stochastic variability in (2.28). Actually, Criterion (2.29) is quite similar to Criterion (2.9), and choosing π_m as in Section 2.2.2 would avoid the overfitting described in Exercise 2.9.1.

2.8 Discussion and References

2.8.1 Take-Home Message

Model selection is a powerful theory for conceptualizing estimation in high-dimensional settings. Except in the orthogonal setting of Exercise 2.9.1, the variation sparse setting of Exercise 2.9.4 and the nested setting of Exercise 2.9.3, its computational complexity is prohibitive for implementation in high-dimensional settings. Yet, it remains a good baseline for designing computationally efficient procedures and for building criteria for selecting among different estimators. These two issues are developed in Chapters 4 and 5.

2.8.2 References

Model selection ideas based on penalized criteria date back to Mallows [91] and Akaike [1] who have introduced the AIC criterion. The BIC criterion has been later derived by Schwartz [110]. The non asymptotic analysis of model selection can be found in Barron, Birgé, and Massart [18] and Birgé and Massart [29]. In particular, Theorem 2.2 is adapted from Birgé and Massart [29].

The concept of model selection can be applied in many different settings, including density estimation, classification, nonparametric regression, etc. In particular, model selection can handle models more general than linear spans. We refer to Massart's St-Flour lecture notes [93] for a comprehensive overview on model selection based on penalized empirical risk criteria. An alternative to these penalized empirical risk criteria is the Goldenshluger–Lepski method; see Goldenschluger and Lepski [69], Chagny [44], and the references therein. A simplified version of this method is presented in Exercise 2.9.5.

The results presented above can be extended to the setting where the variance σ^2 of the noise is unknown; see Baraud *et al.* [16]. The rough idea is to plug for each

m an estimator $\widehat{\sigma}_m^2$ of σ^2 in Criterion (2.9) and to adapt the shape of the penalty $\text{pen}(m)$ accordingly. Another strategy is the slope heuristic developed by Birgé and Massart [30] (see also Arlot and Massart [8] and Baudris *et al.* [20] for a review). The slope heuristic is based on the observation that the dimension $d_{\widehat{m}}$ is small for $K > 1$ and large for $K < 1$. The rough idea is then to replace $K\sigma^2$ by a constant $\kappa > 0$ in (2.9) and search for a constant $\widehat{\kappa}_c$, such that $d_{\widehat{m}}$ is large for $\kappa < \widehat{\kappa}_c$ and small for $\kappa > \widehat{\kappa}_c$. The Criterion (2.9) is finally applied with $K\sigma^2$ replaced by $2\widehat{\kappa}_c$.

The path for deriving minimax lower bounds as in Section 2.4.1 or Exercise 2.9.6 is quite standard. Again, we refer to Massart's St-Flour lecture notes [93] for a broader theory. We also refer to Verzelen [124] and Jin [76] for a discussion of the frontier of successful estimation in high-dimensional settings.

2.9 Exercises

2.9.1 Orthogonal Design

We consider the linear model (2.2) in the coordinate-sparse setting. We assume in this exercise that the columns of $\mathbf{X}$ are orthogonal. We consider the family $\mathscr{M}$ and the models S_m of the coordinate-sparse setting and take in Criterion (2.9) the penality $\text{pen}(m) = \lambda|m|$ for some $\lambda > 0$. Note that with $\lambda = K\big(1 + \sqrt{2\log(p)}\big)^2$, this penalty is almost equal to the penalty $\text{pen}(m) = K\big(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)}\big)^2$, with the probability $\pi_m = (1+1/p)^{-p}p^{-|m|}$.

For $\lambda > 0$, we define $\widehat{m}_\lambda$ as a minimizer of (2.9) with $\text{pen}(m) = \lambda|m|$.

A) Hard thresholding

1. Check that, in this setting, Criterion (2.9) is equal to

$$\|Y - \text{Proj}_{S_m} Y\|^2 + \text{pen}(m)\sigma^2 = \|Y\|^2 + \sum_{j\in m}\left(\lambda\sigma^2 - \left(\frac{\mathbf{X}_j^T Y}{\|\mathbf{X}_j\|}\right)^2\right),$$

   where $\mathbf{X}_j$ is the j-th column of $\mathbf{X}$.

2. Prove that $\widehat{m}_\lambda$ is given by $\widehat{m}_\lambda = \Big\{j:\ (\mathbf{X}_j^T Y)^2 > \lambda\|\mathbf{X}_j\|^2\sigma^2\Big\}$.
3. What is the consequence in terms of computational complexity?

B) Minimal penalties

We next check that the penalty $\text{pen}(m) = 2|m|\log(p)$ is minimal in some sense. We assume in this part that $f^* = 0$ so that the oracle model m_o is the void set. An accurate value of λ must then be such that $\mathbb{E}[|\widehat{m}_\lambda|]$ is small.

1. For $j = 1,\ldots,p$, we set $Z_j = \mathbf{X}_j^T\varepsilon/(\|\mathbf{X}_j\|\sigma)$. Prove that the Z_j are i.i.d. with $\mathscr{N}(0,1)$ distribution.
2. Prove that $|\widehat{m}_\lambda|$ follows a binomial distribution with parameters p and $\eta_\lambda = \mathbb{P}\big(Z^2 > \lambda\big)$, with Z a standard Gaussian random variable.

3. We recall that the AIC criterion is defined by (2.8). Check that $\mathbb{E}\left[|\widehat{m}_{\mathrm{AIC}}|\right] \approx 0.16\,p$. In particular, the AIC criterion is not suited for this context.

4. We recall that for a standard Gaussian random variable Z, we have
$$\mathbb{P}\left(Z^2 > x\right) \overset{x\to\infty}{\sim} \sqrt{\frac{2}{\pi x}}\, e^{-x/2}.$$
Check that for $K > 0$ we have
$$\mathbb{E}\left[|\widehat{m}_{2K\log(p)}|\right] \overset{p\to\infty}{\sim} \frac{p^{1-K}}{\sqrt{\pi K \log(p)}}.$$
Conclude that for $K < 1$ the mean size of the selected model $\widehat{m}_{2K\log(p)}$ grows like a fractional power of p. Choosing $\mathrm{pen}(m) = 2K|m|\log(p)$ with $K < 1$ can then produce poor results in the coordinate-sparse setting. In this sense the penalty $\mathrm{pen}(m) = 2|m|\log(p)$ is minimal in this setting.

5. We recall that the BIC criterion is defined by (2.30). Check that for $p \sim n$ we have
$$\mathbb{E}\left[|\widehat{m}_{\mathrm{BIC}}|\right] \overset{p\to\infty}{\sim} \sqrt{\frac{2p}{\pi \log(p)}}.$$
In particular, the BIC criterion is not suited for this context.

C) Overfitting with $K < 1$

We do not assume anymore that $f^* = 0$. We consider again the penalty $\mathrm{pen}(m) = \lambda|m|$. We prove here that the choice $\lambda = 2K\log(p)$ with $K < 1$ produces an estimator with a risk much larger than the minimax risk.

1. As before, we set $Z_j = \mathbf{X}_j^T \varepsilon/(\|\mathbf{X}_j\|\sigma)$. We assume henceforth that $f^* \in S_{m^*}$ with $|m^*| = D^*$. Writing P_j for the projection on the line spanned by $\mathbf{X}_j$, prove that
$$\begin{aligned}\|\widehat{f}_{\widehat{m}_\lambda} - f^*\|^2 &= \|f^* - \sum_{j\in\widehat{m}_\lambda} P_j f^*\|^2 + \sum_{j\in\widehat{m}_\lambda} Z_j^2\sigma^2 \\ &\geq \sum_{j\in\widehat{m}_\lambda\setminus m^*} Z_j^2\sigma^2 \;\geq\; (|\widehat{m}_\lambda| - D^*)\lambda\sigma^2.\end{aligned}$$

2. Prove that for $a, x \in \mathbb{R}^+$, we have
$$\int_{x-a}^{x} e^{-z^2/2}\,dz \geq \int_{x}^{x+a} e^{-z^2/2}\,dz.$$
As a consequence, for a standard Gaussian random variable Z and for any $a, x \in \mathbb{R}$, prove that $\mathbb{P}\left((Z+a)^2 > x^2\right) \geq \mathbb{P}\left(Z^2 > x^2\right)$.

3. Check that $\mathbb{E}\left[|\widehat{m}_\lambda|\right] \geq p\,\mathbb{P}\left(Z^2 > \lambda\right)$, with Z a standard Gaussian random variable.

4. For $K < 1$ and $D^* \ll p^{1-K}(\log(p))^{-1/2}$, prove that
$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}_{2K\log(p)}} - f^*\|^2\right] \geq$$
$$\left(\mathbb{E}\left[|\widehat{m}_{2K\log(p)}|\right] - D^*\right) 2K\log(p)\sigma^2 \overset{p\to\infty}{\sim} p^{1-K}\sigma^2\sqrt{\frac{4K\log(p)}{\pi}}.$$

5. We use the notations $V_D(\mathbf{X})$ and $R[X,D]$ defined in Section 2.4.1. Conclude that for $K<1$ and $D^* \ll p^{1-K}(\log(p))^{-1/2}$, we have for any $f^* \in V_{D^*}(\mathbf{X})$

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}_{2K\log(p)}} - f^*\|^2\right] \gg R[X,D^*], \quad \text{as } p \to +\infty.$$

In particular, the choice $\lambda = 2K\log(p)$ with $K<1$ produces an estimator with a risk overly large on $V_{D^*}(\mathbf{X})$ for $D^* \ll p^{1-K}(\log(p))^{-1/2}$.

This last part emphasizes that the AIC estimator overfits in this setting, as well as the BIC estimator when $p \sim n^{\alpha}$ with $\alpha > 1/2$. They should then be avoided.

2.9.2 Risk Bounds for the Different Sparsity Settings

We consider the linear regression model (2.2).

A) Coordinate-sparse setting

We focus here on the coordinate-sparse setting. We use the models defined in Section 2.2.1 for this setting and the weights $\pi_m = \left(C_p^{|m|}\right)^{-1} e^{-|m|}(e-1)/(e-e^{-p})$.

1. We define $\text{supp}(\beta) = \left\{j : \beta_j \neq 0\right\}$ and $|\beta|_0 = \text{card}(\text{supp}(\beta))$. Prove that for any $m \in \mathcal{M}$ we have

$$\mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] = \inf_{\beta:\text{supp}(\beta)=m} \left\{\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2 + |\beta|_0\sigma^2\right\}.$$

2. With Theorem 2.2, prove that there exists a constant $C_K > 1$ depending only on $K>1$, such that the estimator $\widehat{f} = \mathbf{X}\widehat{\beta}$ defined by (2.9) fulfills

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta^*\|^2\right] \\
&\le C_K \min_{m \in \mathcal{M}\setminus\emptyset} \inf_{\beta:\text{supp}(\beta)=m} \left\{\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2 + |\beta|_0\left[1+\log\left(\frac{p}{|\beta|_0}\right)\right]\sigma^2\right\} \\
&\le C_K \inf_{\beta\neq 0} \left\{\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2 + |\beta|_0\left[1+\log\left(\frac{p}{|\beta|_0}\right)\right]\sigma^2\right\}. \qquad (2.31)
\end{aligned}$$

B) Group-sparse setting

We focus here on the group-sparse setting, and we assume that the M groups have the same cardinality p/M. We use the models defined in Section 2.2.1 for this setting and the weights $\pi_m = \left(C_M^{|m|}\right)^{-1} e^{-|m|}(e-1)/(e-e^{-M})$.

1. We write $\mathcal{K}(\beta) = \left\{k : \beta_{G_k} \neq 0\right\}$. Prove that

$$\mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] = \inf_{\beta:\mathcal{K}(\beta)=m} \left\{\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2 + \frac{|\mathcal{K}(\beta)|p}{M}\,\sigma^2\right\}.$$

2. With Theorem 2.2, prove that there exists a constant $C_K > 1$ depending only on $K > 1$, such that the estimator $\widehat{f} = \mathbf{X}\widehat{\beta}$ defined by (2.9) fulfills

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta^*\|^2\right]\\ &\leq C_K \min_{m\in\mathscr{M}\setminus\emptyset}\left\{\mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] + |m|\left(1+\log(M/|m|)\right)\sigma^2\right\}\\ &\leq C_K \inf_{\beta\neq 0}\left\{\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2 + |\mathscr{K}(\beta)|\left[1+\frac{p}{M}+\log\left(\frac{M}{|\mathscr{K}(\beta)|}\right)\right]\sigma^2\right\}.\end{aligned}$$

3. Observe that for some β we have $|\mathscr{K}(\beta)| = |\beta|_0 M/p$ and

$$|\mathscr{K}(\beta)|\left[1+\frac{p}{M}+\log\left(\frac{M}{|\mathscr{K}(\beta)|}\right)\right] = |\beta|_0 + \frac{M}{p}|\beta|_0\left[1+\log\left(\frac{p}{|\beta|_0}\right)\right].$$

This inequality enlightens the gain of using the group-sparse models instead of the coordinate-sparse model when β^* is group-sparse and $M \ll p$.

C) Sparse group-sparse setting

We focus here on the sparse group-sparse setting, and we assume again that the M groups have the same cardinality p/M. We use the models and the probability π defined in Section 2.2.1 and 2.2.2 for this setting.

1. We write $\mathscr{J}_k(\beta) = \left\{j \in G_k : \beta_j \neq 0\right\}$. Prove that

$$\mathbb{E}\left[\|\widehat{f}_{(g,\{J_k\}_{k\in g})} - f^*\|^2\right] = \inf_{\beta:\mathscr{K}(\beta)=g,\,\mathscr{J}_k(\beta)=J_k}\left\{\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2 + \sum_{k\in g}|\mathscr{J}_k(\beta)|\,\sigma^2\right\}.$$

2. With Theorem 2.2, prove that there exists a constant $C_K > 1$ depending only on $K > 1$, such that the estimator $\widehat{f} = \mathbf{X}\widehat{\beta}$ defined by (2.9) fulfills

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta^*\|^2\right] \leq C_K \inf_{\beta\neq 0}\Bigg\{&\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2 + |\mathscr{K}(\beta)|\left[1+\log\left(\frac{M}{|\mathscr{K}(\beta)|}\right)\right]\sigma^2\\ &+\sum_{k\in g}|\mathscr{J}_k(\beta)|\left[1+\log\left(\frac{p}{M|\mathscr{J}_k(\beta)|}\right)\right]\sigma^2\Bigg\}.\end{aligned}$$

Compared with the group-sparse setting, we observe that the term $|\mathscr{K}(\beta)|p/M$ is replaced by

$$\sum_{k\in g}|\mathscr{J}_k(\beta)|\left[1+\log\left(\frac{p}{M|\mathscr{J}_k(\beta)|}\right)\right],$$

which can be much smaller when $|\beta|_0 \ll |\mathscr{K}(\beta)|p/M$.

2.9.3 Collections of Nested Models

We consider a collection of models $\{S_m,\ m \in \mathscr{M}\}$ indexed by $\mathscr{M} = \{1,\ldots,M\}$ and fulfilling $\dim(S_m) = m$. Such a collection of models arises naturally, when expanding a signal (as in Example 2) on the m first coefficients of the Fourier basis $\{\varphi_j : j \in \mathbb{N}\}$. In this case we have $S_m = \text{span}\{\varphi_j : j = 1,\dots,m\}$. Another example is when S_m is spanned by the m first principal axes given by a Principal Component Analysis of $\mathbf{X}$.

For any $\alpha > 0$, we choose the probability distribution on $\mathscr{M}$ defined by $\pi_m = e^{-\alpha m}(e^{\alpha} - 1)/(1 - e^{-\alpha M})$ for $m \in \mathscr{M}$. In this case, we can compare directly the risk of the estimator (2.9) to the risk of the oracle $\widehat{f}_{m_o}$.

1. Check that the estimator defined by (2.9) fulfills the oracle inequality

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}} - f^*\|^2\right] \le C_{K,\alpha} \inf_{m=1,\dots,M} \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right],$$

   for some constant $C_{K,\alpha} > 1$ depending only on K and α. In particular, the performance of the estimator (2.9) is almost as good as the performance of the oracle $\widehat{f}_{m_o}$ up to a universal (multiplicative) constant $C_{K,\alpha}$.

2. For $K \in]1,2[$, we set $\alpha = (\sqrt{2/K} - 1)^2$ and $\pi_m = \exp(-\alpha m)$.
   (a) Compute the penalty $\text{pen}(m) = K\left(\sqrt{m} + \sqrt{\log(1/\pi_m)}\right)^2$ for this choice of π_m. What do you recognize?
   (b) Compute the sum $\sum_{m \in \mathscr{M}} \pi_m$. This sum is not equal to one, but we know from the remark below the proof of Theorem 2.2 that the risk Bound (2.12) remains valid when the sum $\sum_{m \in \mathscr{M}} \pi_m$ is not equal to one, as long as the sum remains of a moderate size. What it the size of this sum for $K = 1.1$?

The conclusion is that the AIC criterion fulfills an oracle inequality like (2.12) when the collection of models is nested. Using the AIC criterion in this setting thus perfectly makes sense.

2.9.4 Segmentation with Dynamic Programming

We consider here the setting where the coordinates f_i^* are piecewise constant as described in Example 1, page 27. It corresponds to the variation sparse setting with $p = n$ and with $\mathbf{X}_1,\dots,\mathbf{X}_n$ the canonical basis of $\mathbb{R}^n$.

We want to estimate f^* by model selection. We then consider the family of models introduced in Section 2.2.2 for the variation sparse setting. This family is indexed by the 2^n subsets of $\{1,\dots,n\}$. A naive minimization of Criterion (2.9) would then require 2^n evaluations of the criterion, which is prohibitive in practice. Yet, in this setting, Criterion (2.9) can be minimized much more efficiently by dynamic programing. This exercise is adapted from Lebarbier [82].

We write $N^2(k) = Y_1^2 + \dots + Y_{k-1}^2$ for $2 \le k \le n$ and $N^2(1) = 0$. For $1 \le k < j \le n+1$, we set

$$\bar{y}_{(k,j)} = \frac{1}{j-k}\sum_{i=k}^{j-1} Y_i \quad \text{and} \quad R^2(k,j) = \sum_{i=k}^{j-1}\left(Y_i - \bar{y}_{(k,j)}\right)^2.$$

1. For $m = \{i_1, \ldots, i_D\} \subset \{1, \ldots, n\}$, check that $\widehat{f}_m$ is given by

$$\widehat{f}_m = \sum_{q=1}^{D} \bar{y}_{i_q, i_{q+1}} \mathbf{1}_{i_q}^{i_{q+1}},$$

where $i_{D+1} = n+1$ and where $\mathbf{1}_k^j = \mathbf{X}_k + \ldots + \mathbf{X}_{j-1}$ is the vector with i^{th} coordinate equal to 1 if $k \leq i < j$ and equal to 0 else.

2. We assume in the following that the probability π_m depends on m only through its cardinality $|m|$, as in the examples given in Section 2.2.2 for the variation sparse setting. We will write $\text{pen}(|m|)$ instead of $\text{pen}(m)$ in order to emphasize this dependence. We also define for $0 \leq D \leq n$

$$\widehat{m}_D = \operatorname*{argmin}_{m\,:\,|m|=D} \|Y - \widehat{f}_m\|^2 \quad \text{and} \quad C_n(D) = \|Y - \widehat{f}_{\widehat{m}_D}\|^2.$$

Prove that the minimizer $\widehat{m}$ of (2.9) is given by $\widehat{m} = \widehat{m}_{\widehat{D}}$, where

$$\widehat{D} = \operatorname*{argmin}_{D=0,\ldots,n} \left\{ C_n(D) + \sigma^2 \text{pen}(D) \right\}.$$

3. For $1 \leq D \leq n$, prove that $C_n(D)$ and $\widehat{m}_D$ are solutions of

$$C_n(D) = \min_{1 \leq i_1 < \ldots < i_D \leq n} \left\{ N^2(i_1) + \sum_{q=1}^{D} R^2(i_q, i_{q+1}) \right\}$$

$$\text{and} \quad \widehat{m}_D = \operatorname*{argmin}_{1 \leq i_1 < \ldots < i_D \leq n} \left\{ N^2(i_1) + \sum_{q=1}^{D} R^2(i_q, i_{q+1}) \right\},$$

with the convention $i_{D+1} = n+1$.

4. Check that for $2 \leq D \leq n$, we have the recursive formula

$$C_n(D) = \min_{i=D,\ldots,n} \left\{ C_{i-1}(D-1) + R^2(i, n+1) \right\}. \tag{2.32}$$

5. Check that computing all the $N^2(k)$ and $R^2(k,j)$ for $1 \leq k < j \leq n+1$ requires $O(n^3)$ operations. Building on the recursive Formula (2.32), propose an algorithm that computes $C_n(D)$ and $\widehat{m}_D$ for $D = 1, \ldots, n$ with $O(n^3)$ operations. Conclude that $\widehat{m}$ can be computed with only $O(n^3)$ operations.

2.9.5 Goldenshluger–Lepski Method

Goldenshluger–Lepski method [86, 69] is an alternative to the penalized empirical risk Criterion (2.9). It is initially designed for selecting the bandwidth of kernel estimators, but the method has been recently extended to model selection by Chagny [44]. The following exercise is an adaptation of this work.

We consider a set $\mathscr{M} = \{1, \ldots, M\}$ and a collection of models $\{S_m,\ m \in \mathscr{M}\}$, such that $S_m \subset S_{m'}$ for $m \leq m'$. The main idea underlying Goldenshluger–Lepski method is to estimate directly the bias term $\|f^* - \mathrm{Proj}_{S_m} f^*\|^2$ by a term $B(m)$ involving $\|\widehat{f}_{m'} - \widehat{f}_{m \wedge m'}\|^2$ with $m' \neq m$. More precisely, let us consider a probability distribution $\pi = \{\pi_m,\ m \in \mathscr{M}\}$ on $\mathscr{M}$ and define $\mathrm{pen}(m)$ as in (2.9). In the following, we consider the model selection procedure

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathscr{M}} \left\{B(m) + \mathrm{pen}(m)\sigma^2\right\},$$
$$\text{where} \quad B(m) = \max_{m' \in \mathscr{M}} \left[\|\widehat{f}_{m'} - \widehat{f}_{m' \wedge m}\|^2 - \mathrm{pen}(m')\sigma^2\right]_+ . \qquad (2.33)$$

1. Considering apart the cases $m \leq \widehat{m}$ and $m > \widehat{m}$, prove that

$$\begin{aligned} \|\widehat{f}_{\widehat{m}} - f^*\|^2 &\leq 2\max\left(\|\widehat{f}_{\widehat{m}} - \widehat{f}_{\widehat{m} \wedge m}\|^2, \|\widehat{f}_m - \widehat{f}_{\widehat{m} \wedge m}\|^2\right) + 2\|\widehat{f}_m - f^*\|^2 \\ &\leq 2(B(\widehat{m}) + \mathrm{pen}(m)\sigma^2 + B(m) + \mathrm{pen}(\widehat{m})\sigma^2) + 2\|\widehat{f}_m - f^*\|^2 \\ &\leq 4(B(m) + \mathrm{pen}(m)\sigma^2) + 2\|\widehat{f}_m - f^*\|^2. \end{aligned}$$

2. Prove that for any $m \in \mathscr{M}$ and $\eta > 0$ we have

$$\begin{aligned} B(m) \leq& \frac{\eta+1}{\eta} \max_{m' \geq m} \|\mathrm{Proj}_{S_{m'}} f^* - \mathrm{Proj}_{S_m} f^*\|^2 \\ &+ \sum_{m' \geq m} \left[(1+\eta)\|\mathrm{Proj}_{S_{m'}} \varepsilon - \mathrm{Proj}_{S_m} \varepsilon\|^2 - \mathrm{pen}(m')\sigma^2\right]_+ \\ \leq& \frac{\eta+1}{\eta} \|f^* - \mathrm{Proj}_{S_m} f^*\|^2 + \sum_{m' \in \mathscr{M}} \left[(1+\eta)\|\mathrm{Proj}_{S_{m'}} \varepsilon\|^2 - \mathrm{pen}(m')\sigma^2\right]_+ . \end{aligned}$$

We set in the following $\eta = (K-1)/2$, where $K > 1$ is the constant involved in the definition of $\mathrm{pen}(m)$. We lift from the proof of Theorem 2.2 that the expectation of the above sum is upper bounded by $2K(K+1)\sigma^2/(K-1)$.

3. Prove the following risk bound for the estimator $\widehat{f}_{\widehat{m}}$ with $\widehat{m}$ selected according to (2.33)

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}} - f^*\|^2\right] \leq \inf_{m \in \mathscr{M}} \left\{\frac{6K+2}{K-1} \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] + 4\mathrm{pen}(m)\sigma^2\right\} + \frac{8K(K+1)}{K-1}\sigma^2.$$

   Compare this bound with the bound in Theorem 2.2.

4. Let us assume that $d_m = m$, for $m = 1, \ldots, M$. Check that with the choice $\pi_m = e^{-m}(e-1)/(1-e^{-M})$, we have

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}} - f^*\|^2\right] \leq C_K \inf_{m=1,\ldots,M} \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right]$$

   for some constant C_K depending only on K.

Remark: Note that the positive part in the definition of $B(m)$ did not play a role in the above analysis. Yet, this positive part is natural, since $B(m)$ is an estimator of the bias term $\|f^* - \mathrm{Proj}_{S_m} f^*\|^2$, which is always non-negative. Similarly, we can consider maximizing only over the $m' \geq m$ in the definition of $B(m)$. As an exercise, you can check that if we remove the positive part in the definition of $B(m)$ and maximize only on $m' \geq m$, then the selection criteria (2.9) and (2.33) are completely equivalent.

2.9.6 Minimax Lower Bounds

A) Spreading points in a sparse hypercube

We prove in this part Lemma 2.5, page 44. Let D be any positive integer less than $p/5$ and consider a maximal set $\mathscr{C}$ in $\{0,1\}^p_D := \{x \in \{0,1\}^p : |x|_0 = D\}$, fulfilling

$$|\beta - \beta'|_0 > D, \quad \text{for all } \beta \neq \beta' \in \mathscr{C}.$$

In the next questions, we will prove that

$$\log|\mathscr{C}| \geq \frac{D}{2} \log\left(\frac{p}{5D}\right). \tag{2.34}$$

1. Check that
$$\{0,1\}^p_D = \bigcup_{\beta \in \mathscr{C}} \left\{x \in \{0,1\}^p_D : |x - \beta|_0 \leq D\right\}.$$
Deduce from this covering the upper bound $C_p^D \leq |\mathscr{C}| \max_{\beta \in \mathscr{C}} |B(\beta, D)|$, where $B(\beta, D) = \left\{x \in \{0,1\}^p_D : |x - \beta|_0 \leq D\right\}$.
2. Writing d for the integer part of $D/2$, check the sequence of inequalities
$$\frac{|B(\beta,D)|}{C_p^D} = \sum_{k=0}^{d} \frac{C_D^k C_{p-D}^k}{C_p^D} \leq 2^D \frac{C_p^d}{C_p^D} \leq 2^D \left(\frac{D}{p-D+1}\right)^{D-d} \leq \left(\frac{5D}{p}\right)^{D/2}.$$
3. Conclude the proof of (2.34).

B) Some other minimax lower bounds

Let d be a distance on $\mathbb{R}^p$ and fix $q \geq 1$. As in Section 2.4.1, we write $\mathbb{P}_f$ for the Gaussian distribution $\mathcal{N}(f, \sigma^2 I_n)$ and the expectation $\mathbb{E}_f$ will refer to the expectation when the vector of observations Y is distributed according to $\mathbb{P}_f$.

1. Let $\mathscr{C}$ be any finite subset of $\mathbb{R}^p$. Adapting the proof of Lemma 2.4, prove that
$$\inf_{\widehat{\beta}} \max_{\beta \in \mathscr{C}} \mathbb{E}_{\mathbf{X}\beta}\left[d(\widehat{\beta}, \beta)^q\right] \geq \frac{1}{2^q}\left(1 - \frac{2e}{2e+1} \bigvee \max_{\beta \neq \beta' \in \mathscr{C}} \frac{\|\mathbf{X}(\beta - \beta')\|^2}{2\sigma^2 \log(|\mathscr{C}|)}\right) \times \min_{\beta \neq \beta' \in \mathscr{C}} d(\beta, \beta')^q.$$

We fix $D_{\max} \leq p/5$. For $D \leq D_{\max}$, we set

$$r^2 = \frac{e}{2(2e+1)} \times \frac{\sigma^2}{\bar{c}_{\mathbf{X}}^2} \times \log\left(\frac{p}{5D}\right),$$

with $\bar{c}_{\mathbf{X}}$ defined by (2.22). We consider below the set $\mathscr{C}_r = \{r\beta : \beta \in \mathscr{C}\}$, with $\mathscr{C}$ from the previous part and the distance d induced by the norm $|x|_q = \left(\sum_j |x_j|^q\right)^{1/q}$.

2. Prove that
$$\inf_{\widehat{\beta}} \max_{\beta \in \mathscr{C}_r} \mathbb{E}_{\mathbf{X}\beta}\left[|\widehat{\beta} - \beta|_q^q\right] \geq \frac{r^q D}{2^q(2e+1)}.$$

3. Conclude that for any $q \geq 1$, we have the lower bound

$$\inf_{\widehat{\beta}} \sup_{\beta:|\beta|_0=D} \mathbb{E}_{\mathbf{X}\beta}\left[|\widehat{\beta}-\beta|_q^q\right] \geq \frac{e^{q/2}}{2^{q+1}(2e+1)^{1+q/2}} \left(\frac{\sigma}{\bar{c}_{\mathbf{X}}}\right)^q D\left(\log\left(\frac{p}{5D}\right)\right)^{q/2}.$$

Chapter 3

Aggregation of Estimators

Estimators aggregation is an alternative to model selection for exploiting possible unknown structures in the data. The main idea is to use a convex combination of a collection of estimators instead of selecting one among them. Estimator aggregation shares all the good properties of model selection. Unfortunately, it suffers from the same high-computational cost. Some approximate computations can be implemented; some of them are described at the end of the chapter.

In the following, we present the aggregation of estimators in the same statistical setting as in the previous chapter.

3.1 Introduction

The model $\widehat{m}$ selected in the model selection procedure (2.9) can provide some interesting information on the data. Yet, when the objective is only to predict at best f^* (as in the Examples 2, 4, and 5 in Section 2.1), we only want to have a ℓ^2-risk (2.4) as small as possible, and we do not care to select (or not) a model. In this case, instead of selecting an estimator $\widehat{f}_{\widehat{m}}$ in $\{\widehat{f}_m,\ m \in \mathcal{M}\}$, we may prefer to estimate f^* by a convex (or linear) combination $\widehat{f}$ of the $\{\widehat{f}_m,\ m \in \mathcal{M}\}$

$$\widehat{f} = \sum_{m\in\mathcal{M}} w_m \widehat{f}_m\,, \quad \text{with} \quad w_m \geq 0 \quad \text{and} \quad \sum_{m\in\mathcal{M}} w_m = 1\,. \tag{3.1}$$

Of course, selecting a model is a special case of convex combination of the $\{\widehat{f}_m,\ m \in \mathcal{M}\}$, with the weights $w_{\widehat{m}} = 1$ and $w_m = 0$ for $m \neq \widehat{m}$. Removing the requirement that all the weights w_m are equal to 0 except one of them allows for more flexibility and can possibly provide more stable estimators (in the sense that a small perturbation of the data Y only induces a small change in the estimator $\widehat{f}$).

3.2 Gibbs Mixing of Estimators

As in Chapter 2, page 31, we consider a collection of models $\{S_m,\ m \in \mathcal{M}\}$. We remind the reader that an unbiased estimator of the risk $r_m = \mathbb{E}\left[\|f^* - \widehat{f}_m\|^2\right]$ of the estimator $\widehat{f}_m = \mathrm{Proj}_{S_m} Y$ is

$$\widehat{r}_m = \|Y - \widehat{f}_m\|^2 + 2d_m\sigma^2 - n\sigma^2, \quad \text{with } d_m = \dim(S_m);$$

see (2.7) in Chapter 2, page 33. We associate to $\beta > 0$ and a probability distribution π on $\mathscr{M}$ the Gibbs mixing $\widehat{f}$ of the collection of estimators $\{\widehat{f}_m,\ m \in \mathscr{M}\}$

$$\widehat{f} = \sum_{m\in\mathscr{M}} w_m \widehat{f}_m, \quad \text{with } w_m = \frac{\pi_m e^{-\beta\widehat{r}_m/\sigma^2}}{\mathscr{Z}}, \quad \text{where } \mathscr{Z} = \sum_{m\in\mathscr{M}} \pi_m e^{-\beta\widehat{r}_m/\sigma^2}. \tag{3.2}$$

The Gibbs distribution w corresponds to the distribution minimizing over the set of probability q on $\mathscr{M}$ the functional

$$\mathscr{G}(q) = \sum_m q_m \widehat{r}_m + \frac{\sigma^2}{\beta}\mathscr{K}(q,\pi), \tag{3.3}$$

where $\mathscr{K}(q,\pi) = \sum_{m\in\mathscr{M}} q_m \log(q_m/\pi_m) \geq 0$ is the Kullback–Leibler divergence between the probabilities q and π; see Exercise 3.7.1. We will use this property in the next section for analyzing the risk of $\widehat{f}$.

Remark. The estimator $\widehat{f}_{\widetilde{m}}$ with the largest weight $w_{\widetilde{m}}$ corresponds to the model $\widetilde{m}$ minimizing the criterion

$$\widehat{r}_m + \frac{\sigma^2}{\beta}\log\left(\frac{1}{\pi_m}\right) = \|Y - \widehat{f}_m\|^2 + 2d_m\sigma^2 + \frac{\sigma^2}{\beta}\log\left(\frac{1}{\pi_m}\right) - n\sigma^2, \tag{3.4}$$

which is quite similar to the model selection Criterion (2.9), since $n\sigma^2$ plays no role in the minimization. Since the weights w_m are inversely proportional to the exponential of (3.4), when one of the model has a Criterion (3.4) much smaller than the others, the estimator $\widehat{f}$ is very close to the estimator $\widehat{f}_{\widetilde{m}}$, with $\widetilde{m}$ minimizing (3.4). Estimator aggregation then significantly differs from estimator selection only when there are several models approximately minimizing Criterion (3.4). We refer to Exercise 3.7.4 for a brief analysis of the virtues of estimator aggregation in this context.

3.3 Oracle Risk Bound

The analysis of the risk of the estimator $\widehat{f}$ relies on the famous Stein's formula.

Proposition 3.1 Stein's formula

Let Y be an n-dimensional Gaussian vector with mean μ and covariance matrix $\sigma^2 I_n$. For any function $F : \mathbb{R}^n \to \mathbb{R}^n$, $F(x) = (F_1(x),\ldots,F_n(x))$ continuous, with piecewise continuous partial derivatives, fulfilling for all $i \in \{1,\ldots,n\}$

(a) $\lim_{|y_i|\to\infty} F_i(y_1,\ldots,y_n)e^{-(y_i-\mu_i)^2/2\sigma^2} = 0$ *for all* $(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n) \in \mathbb{R}^{n-1}$

(b) $\mathbb{E}\left[|\partial_i F_i(Y)|\right] < \infty$,

we have

$$\mathbb{E}\left[\|F(Y) - \mu\|^2\right] = \mathbb{E}\left[\|F(Y) - Y\|^2 - n\sigma^2 + 2\sigma^2 \mathrm{div}(F)(Y)\right], \tag{3.5}$$

where $\mathrm{div}(F) = \sum_i \partial_i F_i$.

Proof.
An integration by parts gives

$$\sigma^2 \int_{y_i \in \mathbb{R}} \partial_i F_i(y_1,\ldots,y_n) e^{-(y_i-\mu_i)^2/2\sigma^2}\, dy_i \\ = \int_{y_i \in \mathbb{R}} (y_i - \mu_i) F_i(y_1,\ldots,y_n) e^{-(y_i-\mu_i)^2/2\sigma^2}\, dy_i\,,$$

for all $(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n) \in \mathbb{R}^{n-1}$. Taking the expectation with respect to $Y_1,\ldots,Y_{i-1},Y_{i+1},\ldots,Y_n$, it follows that $\mathbb{E}\left[(Y_i-\mu_i)F_i(Y)\right] = \sigma^2 \mathbb{E}\left[\partial_i F_i(Y)\right]$, and finally $\mathbb{E}\left[\langle Y-\mu, F(Y)\rangle\right] = \sigma^2 \mathbb{E}\left[\mathrm{div}(F)(Y)\right]$.

We conclude by expanding

$$\begin{aligned}\mathbb{E}\left[\|F(Y)-\mu\|^2\right] &= \mathbb{E}\left[\|F(Y)-Y\|^2\right] + \mathbb{E}\left[\|Y-\mu\|^2\right] + 2\mathbb{E}\left[\langle F(Y)-Y, Y-\mu\rangle\right] \\ &= \mathbb{E}\left[\|F(Y)-Y\|^2\right] + n\sigma^2 + 2\sigma^2 \mathbb{E}\left[\mathrm{div}(F)(Y)\right] - 2n\sigma^2,\end{aligned}$$

where the last equality comes from $\mathbb{E}\left[\langle Y, Y-\mu\rangle\right] = \mathbb{E}\left[\|Y-\mu\|^2\right] = n\sigma^2$, since Y is a Gaussian vector with mean μ and covariance matrix $\sigma^2 I_n$. The proof of Stein's formula is complete. □

We consider a collection of models $\{S_m,\ m \in \mathcal{M}\}$, a probability distribution $\{\pi_m,\ m \in \mathcal{M}\}$ on $\mathcal{M}$, a constant $\beta > 0$, and the estimator $\widehat{f}$ given by (3.2). We have the following risk bound on $R(\widehat{f})$.

Theorem 3.2 Oracle risk bound

For $\beta \leq 1/4$ we have

$$\mathbb{E}\left[\|\widehat{f} - f^*\|^2\right] \leq \min_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] + \frac{\sigma^2}{\beta} \log\left(\frac{1}{\pi_m}\right) \right\}. \tag{3.6}$$

Before proving this result, let us comment on it. The risk Bound (3.6) is very similar to the risk Bound (2.12), page 37, for model selection. The main difference lies in the constant in front of $\mathbb{E}\left[\|\widehat{f}_m - f\|^2\right]$, which is exactly one in (3.6). For this reason, Bound (3.6) is better than Bound (2.12). In particular, the aggregated estimator $\widehat{f}$ given by (3.2) fulfills all the good properties of the estimator (2.9) obtained by model selection.

Proof.
The Stein formula applied to $F(Y) = \widehat{f}$ and the linearity of the divergence operator ensure that

$$\widehat{r} = \|\widehat{f} - Y\|^2 - n\sigma^2 + 2\sigma^2 \sum_{m \in \mathcal{M}} \mathrm{div}(w_m \widehat{f}_m) \quad \text{fulfills} \quad \mathbb{E}\left[\widehat{r}\right] = \mathbb{E}\left[\|\widehat{f} - f^*\|^2\right].$$

After a simplification of the expression of $\widehat{r}$, the result will follow from the fact that w minimizes (3.3).

Upper bound on $\widehat{r}$. We first expand

$$\begin{aligned}
\|Y-\widehat{f}\|^2 &= \Big\langle \sum_{m\in\mathscr{M}} w_m(Y-\widehat{f}_m)\ ,\ Y-\widehat{f}\Big\rangle \\
&= \sum_{m\in\mathscr{M}} w_m\langle Y-\widehat{f}_m, Y-\widehat{f}_m\rangle + \sum_{m\in\mathscr{M}} w_m\langle Y-\widehat{f}_m, \widehat{f}_m-\widehat{f}\rangle \\
&= \sum_{m\in\mathscr{M}} w_m\|Y-\widehat{f}_m\|^2 + \underbrace{\sum_{m\in\mathscr{M}} w_m\langle Y-\widehat{f}, \widehat{f}_m-\widehat{f}\rangle}_{=0} - \sum_{m\in\mathscr{M}} w_m\|\widehat{f}-\widehat{f}_m\|^2,
\end{aligned}$$

and notice that the divergence of $w_m\widehat{f}_m$ is given by

$$\operatorname{div}(w_m\widehat{f}_m) = w_m\operatorname{div}(\widehat{f}_m) + \langle \widehat{f}_m, \nabla w_m\rangle.$$

Plugging these two formulas in the definition of $\widehat{r}$, we obtain

$$\begin{aligned}
\widehat{r} = \sum_{m\in\mathscr{M}} w_m\left(\|Y-\widehat{f}_m\|^2 - n\sigma^2 + 2\sigma^2\operatorname{div}(\widehat{f}_m)\right) \\
+ \sum_{m\in\mathscr{M}} w_m\left(2\sigma^2\Big\langle \widehat{f}_m, \frac{\nabla w_m}{w_m}\Big\rangle - \|\widehat{f}-\widehat{f}_m\|^2\right).
\end{aligned}$$

For a linear function $G(Y) = AY$, we have $\partial_i G_i(Y) = A_{ii}$, so $\operatorname{div}(G)(Y) = \operatorname{Tr}(A)$. Since $\widehat{f}_m = \operatorname{Proj}_{S_m}(Y)$, the divergence of $\widehat{f}_m$ is the trace of $\operatorname{Proj}_{S_m}$, which equals the dimension d_m of the space S_m. Then, we have

$$\|Y-\widehat{f}_m\|^2 - n\sigma^2 + 2\sigma^2\operatorname{div}(\widehat{f}_m) \ = \ \|Y-\widehat{f}_m\|^2 + (2d_m-n)\sigma^2 \ = \ \widehat{r}_m.$$

In addition, we have

$$\begin{aligned}
\frac{\nabla w_m}{w_m} &= -2\frac{\beta}{\sigma^2}(Y-\widehat{f}_m) - \frac{\nabla\mathscr{Z}}{\mathscr{Z}} = -2\frac{\beta}{\sigma^2}\Big((Y-\widehat{f}_m) - \sum_{m'\in\mathscr{M}} w_{m'}(Y-\widehat{f}_{m'})\Big) \\
&= -2\frac{\beta}{\sigma^2}\left(\widehat{f}-\widehat{f}_m\right),
\end{aligned}$$

so

$$\widehat{r} = \sum_{m\in\mathscr{M}} w_m\widehat{r}_m - 4\beta\sum_{m\in\mathscr{M}} w_m\langle \widehat{f}_m, \widehat{f}-\widehat{f}_m\rangle - \sum_{m\in\mathscr{M}} w_m\|\widehat{f}-\widehat{f}_m\|^2.$$

We notice that $\sum_{m\in\mathscr{M}} w_m\langle \widehat{f}, \widehat{f}-\widehat{f}_m\rangle = 0$ so

$$\begin{aligned}
\widehat{r} &= \sum_{m\in\mathscr{M}} w_m\widehat{r}_m + (4\beta-1)\sum_{m\in\mathscr{M}} w_m\|\widehat{f}-\widehat{f}_m\|^2 \\
&\leq \sum_{m\in\mathscr{M}} w_m\widehat{r}_m\,, \qquad (3.7)
\end{aligned}$$

where we used for the last line the condition $\beta \leq 1/4$.

Optimality condition. Since $\mathscr{K}(w,\pi) \geq 0$ and w minimizes $\mathscr{G}$ over the set of probability on $\mathscr{M}$ (see Exercise 3.7.1, page 69), we have

$$\widehat{r} \leq \mathscr{G}(w) \leq \mathscr{G}(q) = \sum_{m\in\mathscr{M}} q_m \widehat{r}_m + \frac{\sigma^2}{\beta}\mathscr{K}(q,\pi), \quad \text{for any probability } q \text{ on } \mathscr{M}.$$

Taking the expectation, we obtain for any probability q on $\mathscr{M}$

$$\mathbb{E}\left[\|\widehat{f}-f^*\|^2\right] = \mathbb{E}[\widehat{r}] \ \leq \ \mathbb{E}[\mathscr{G}(q)] = \sum_{m\in\mathscr{M}} q_m r_m + \frac{\sigma^2}{\beta}\mathscr{K}(q,\pi), \tag{3.8}$$

with $r_m = \mathbb{E}[\widehat{r}_m] = \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right]$. The right-hand side of the above bound is minimum for $q_m = \pi_m e^{-\beta r_m/\sigma^2}/\mathscr{Z}'$, with $\mathscr{Z}' = \sum_m \pi_m e^{-\beta r_m/\sigma^2}$. Since $\mathscr{Z}' \geq \pi_m e^{-\beta r_m/\sigma^2}$ for all $m \in \mathscr{M}$, we observe that

$$\sum_{m\in\mathscr{M}} q_m r_m + \frac{\sigma^2}{\beta}\mathscr{K}(q,\pi) = -\frac{\sigma^2}{\beta}\log(\mathscr{Z}') \leq \min_{m\in\mathscr{M}}\left\{r_m - \frac{\sigma^2}{\beta}\log(\pi_m)\right\}. \tag{3.9}$$

Combining (3.8) and (3.9), we obtain (3.6). □

3.4 Numerical Approximation by Metropolis–Hastings

Similarly to model selection, when the cardinality of $\mathscr{M}$ is large, the Gibbs mixing of estimators (3.2) is computationally intractable, since it requires at least card$(\mathscr{M})$ computations. To overcome this issue, one possibility is to compute an approximation of (3.2) by using Markov Chain Monte-Carlo (MCMC) techniques.

Let $F : \mathscr{M} \to \mathbb{R}$ be a real function on $\mathscr{M}$ and w be an arbitrary probability distribution on $\mathscr{M}$ fulfilling $w_m > 0$ for all $m \in \mathscr{M}$. The Metropolis–Hastings algorithm is a classical MCMC technique to approximate

$$\mathbb{E}_w[F] := \sum_{m\in\mathscr{M}} w_m F(m). \tag{3.10}$$

The underlying idea of the Metropolis–Hastings algorithm is to generate an ergodic[1] Markov chain $(M_t)_{t\in\mathbb{N}}$ in $\mathscr{M}$ with stationary distribution w and to approximate $\mathbb{E}_w[F]$ by the average $T^{-1}\sum_{t=1}^{T} F(M_t)$, with T large. This approximation relies on the fact that for any ergodic Markov chain $(M_t)_{t\in\mathbb{N}}$ in $\mathscr{M}$ with stationary distribution w, we have almost surely the convergence

$$\frac{1}{T}\sum_{t=1}^{T} F(M_t) \overset{T\to\infty}{\to} \mathbb{E}_w[F]. \tag{3.11}$$

[1] We refer to [99] for lecture notes on Markov chains.

We remind the reader that a Markov chain on $\mathscr{M}$ with transition probability $Q(m,m')$ is said to be *reversible according to the probability distribution* w if

$$w_m Q(m,m') = w_{m'} Q(m',m) \quad \text{for all } m,m' \in \mathscr{M}\,. \tag{3.12}$$

An aperiodic and irreducible Markov chain that is reversible according to the probability w automatically fulfills (3.11). The Metropolis–Hastings algorithm proposes a way to build a Markov chain $(M_t)_{t\in\mathbb{N}}$ fulfilling the requirements that

- $(M_t)_{t\in\mathbb{N}}$ is aperiodic, irreducible, and reversible according to the probability w,
- the numerical generation of the chain $(M_t)_{t=1,\dots,T}$ is computationally efficient.

To generate such a Markov chain, the principle of the Metropolis–Hastings algorithm is the following. Let $\Gamma(m,m')$ be the transition probability of an aperiodic irreducible random walk on $\mathscr{M}$ and define

$$\begin{aligned} Q(m,m') &= \Gamma(m,m') \wedge \frac{w_{m'}\Gamma(m',m)}{w_m} \quad \text{for all } m \neq m' \qquad (3.13)\\ \text{and } Q(m,m) &= 1 - \sum_{m':\ m'\neq m} Q(m,m') \quad \text{for all } m \in \mathscr{M}. \end{aligned}$$

With such a definition, w and Q obviously fulfill (3.12). Therefore, the Markov chain $(M_t)_{t\in\mathbb{N}}$ with transition Q fulfills for any initial condition

$$\frac{1}{T}\sum_{t=1}^{T} F(M_t) \overset{T\to\infty}{\to} \sum_{m\in\mathscr{M}} w_m F(m) \quad \text{a.s.}$$

From a practical point of view, the Markov chain $(M_t)_t$ can be generated as follows. We start from the transition $\Gamma(m,m')$, which is the transition probability of an aperiodic irreductible random walk on $\mathscr{M}$, and then implement the following algorithm.

Metropolis–Hastings algorithm

Initialization: Pick an arbitrary $M_1 \in \mathscr{M}$ and choose a burn-in time T_0.

Iterate: For $t = 1,\dots,T$

1. From the current state M_t, generate M'_{t+1} according to the distribution $\Gamma(M_t,.)$
2. Set
$$p_{t+1} = 1 \wedge \frac{w_{M'_{t+1}}\Gamma(M'_{t+1},M_t)}{w_{M_t}\Gamma(M_t,M'_{t+1})}$$
3. $$\begin{cases} \text{With probability } p_{t+1} & : \ \text{set } M_{t+1} = M'_{t+1}, \\ \text{otherwise} & : \ \text{set } M_{t+1} = M_t. \end{cases}$$

Output:
$$\frac{1}{T-T_0}\sum_{t=T_0+1}^{T} F(M_t)$$

Let us come back to the problem of approximating the Gibbs aggregation (3.2) of estimators. In this case, the function F is $F(m) = \widehat{f}_m$, and the probability distribution is

$$w_m = \frac{\pi_m e^{-\beta \widehat{r}_m/\sigma^2}}{\mathscr{Z}}, \quad \text{where } \mathscr{Z} = \sum_{m \in \mathscr{M}} \pi_m e^{-\beta \widehat{r}_m/\sigma^2}.$$

Computing a single w_m is very time consuming when $\mathscr{M}$ is very large, since computing $\mathscr{Z}$ requires to sum card$(\mathscr{M})$ terms. Fortunately, the Metropolis–Hastings algorithm does not require computing the weights w_m but only some ratios,

$$\frac{w_{m'}}{w_m} = \frac{\pi_{m'} e^{-\beta \widehat{r}_{m'}/\sigma^2}}{\pi_m e^{-\beta \widehat{r}_m/\sigma^2}},$$

that can be easily evaluated, since the term $\mathscr{Z}$ cancels.

To implement the Metropolis–Hastings algorithm for evaluating (3.2), we must choose a transition probability Γ. The choice of this transition probability Γ is delicate in practice, since it can dramatically change the speed of the convergence in (3.11). Ideally, the transition Γ must be simple enough so that we can efficiently generate M'_{t+1} from M_t, and it must lead to a “good” exploration of $\mathscr{M}$ in order to avoid getting trapped in a small neighborhood. We refer to [104, 105] for discussions on the choice of Γ and much more on the Metropolis–Hastings algorithm. Below, we give an example of implementation of the Metropolis–Hastings algorithm for evaluating (3.2) in the coordinate-sparse setting.

Example: Approximating (3.2) in the sparse regression setting

We consider the collection of estimators $\{\widehat{f}_m : m \in \mathscr{M}\}$ and the probability π introduced in sections 2.2.1 and 2.2.2 for the coordinate-sparse setting. We write $\mathscr{V}(m)$ for the set of all the subsets m' that can be obtained from m by adding or removing one integer $j \in \{1, \ldots, p\}$ to m. If we take the uniform distribution on $\mathscr{V}(m)$ as proposal distribution $\Gamma(m, .)$, then $\Gamma(m, m') = \Gamma(m', m)$ for any $m' \in \mathscr{V}(m)$. As a consequence, we only need to compute $w_{M'_{t+1}}/w_{M'_t}$ at the second step of the Metropolis–Hastings algorithm. It is crucial to compute this ratio efficiently, since it will be iterated many times. We have seen that

$$\frac{w_{m'}}{w_m} = \frac{\pi_{m'} e^{-\beta \widehat{r}_{m'}/\sigma^2}}{\pi_m e^{-\beta \widehat{r}_m/\sigma^2}},$$

so we need to compute $\pi_{m'}/\pi_m$ efficiently when $m' \in \mathscr{V}(m)$. Let us consider the choice $\pi_m \propto e^{-|m|}/C_p^{|m|}$ proposed for the coordinate-sparse setting. We have the simple formulas

$$\frac{\pi_{m'}}{\pi_m} = \frac{|m|+1}{e(p-|m|)} \text{ for } m' = m \cup \{j\} \quad \text{and} \quad \frac{\pi_{m'}}{\pi_m} = \frac{e(p-|m|+1)}{|m|} \text{ for } m' = m \setminus \{j\}.$$

Convergence rate of the approximation

The main weakness of this Metropolis–Hastings approximation of $\widehat{f}$ is that we do not know the convergence rate in (3.11). In particular, we have no guarantee that we can achieve a reasonably good approximation of (3.2) by the Metropolis–Hastings algorithm in a reasonable time. We emphasize that there is no hope to evaluate precisely all the weights $\{w_m : m \in \mathscr{M}\}$ in less than card$(\mathscr{M})$ iterations since the Markov chain $(M_t)_{t\in\mathbb{N}}$ needs to visit each $m \in \mathscr{M}$ many times in order to evaluate precisely each w_m. Yet, if only a few weights w_m are significantly positive, we only need to evaluate properly these few ones, which may happen in a short amount of time. However, we emphasize that no result guarantees that these few larger weights w_m are estimated accurately after a reasonable amount of time.

To sum up, the Metropolis–Hastings algorithm presented above can be viewed as a stochastic version of the forward–backward algorithm presented in Section 2.5, where we average the estimators $(\widehat{f}_{M_t})_{t=1,\ldots,T}$ instead of selecting the one with smallest Criterion (2.9).

3.5 Numerical Illustration

Let us illustrate the estimator aggregation on the simulated example of Section 2.6. We use again the models and probability distribution suited for the coordinate-sparse setting. The Gibbs mixing $\widehat{f}$ defined by (3.2) with $\beta = 1/4$ is plotted in Figure 3.1. Comparing Figure 3.1 with Figure 2.2, we observe that the estimators (3.2) and (2.9) are very similar in this example. We refer to Exercise 3.7.2 for a comparison of the Gibbs mixing and model selection in the orthonormal setting.

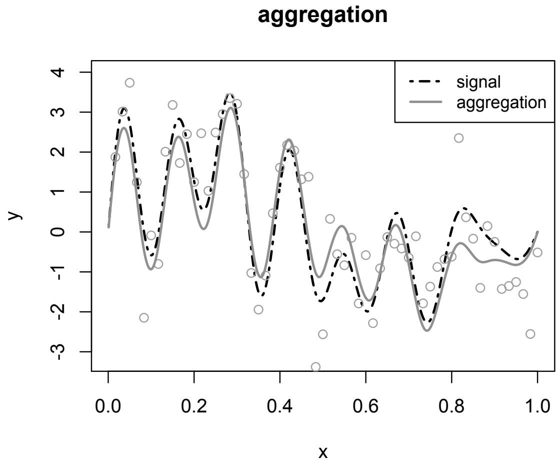


Figure 3.1 *Dotted line: The unknown signal. Gray dots: The noisy observations. Gray line: The estimator (3.2).*

3.6 Discussion and References

3.6.1 Take-Home Message

The Gibbs aggregation of estimators shares all the nice statistical properties of model selection estimators, with a slightly better risk Bound (3.6). Unfortunately, it suffers from the same computational complexity, and in many cases it is not possible to directly implement it in practice. A possible direction to overcome this issue is to approximate (3.2) by a Metropolis–Hastings algorithm, which is some kind of stochastic version of the forward–backward algorithm of Section 2.5 with averaging. No precise convergence rate is known for this algorithm, so we have no guarantee on the quality of the finite-time approximation.

3.6.2 References

Ideas on aggregation of estimators in statistics date back at least to Barron [19] and Catoni [42, 43]. Most of the material presented here comes from the seminal paper of Leung and Barron [88]. The computational aspects presented in Section 3.4 have been proposed by Tsybakov and Rigollet [102]. We refer to Tsybakov and Rigollet [103] for a recent review on the topic and more examples of the use of the Metropolis–Hastings algorithm. We also point out the paper by Sanchez-Perez [108], which provides a convergence bound on the Metropolis–Hastings approximation in a slightly different setting.

The case of unknown variance is more tricky and has attracted less attention. We refer to Giraud [64] and Gerchinovitz [62] for two points of view on this case. Finally, the Stein formula for non-Gaussian noise has been investigated in Dalalyan and Tsybakov [48].

3.7 Exercises

3.7.1 Gibbs Distribution

In this exercise, we first check that the Kullback–Leibler divergence $\mathscr{K}(q,\pi)$ is non-negative on the set $P(\mathscr{M})$ of probability distributions on $\mathscr{M}$. Then, we prove that the Gibbs distribution w defined by (3.2), page 62, minimizes the function $\mathscr{G}$ defined by (3.3).

For $\lambda \in \mathbb{R}$ and $q \in (\mathbb{R}^+)^{\mathscr{M}}$ we set,

$$\mathscr{L}^{(\lambda)}(q) = \mathscr{G}(q) + \lambda \sum_{m\in\mathscr{M}} q_m \,.$$

1. From the convexity of $x \to -\log(x)$ on $\mathbb{R}^+$, prove that the Kullback–Leibler divergence $\mathscr{K}(q,\pi) = \sum_m q_m \log(q_m/\pi_m)$ is non-negative for $q,\pi \in P(\mathscr{M})$.
2. Prove that $x \to x\log(x)$ is convex on $\mathbb{R}^+$ and $q \to \mathscr{L}^{(\lambda)}(q)$ is convex on $(\mathbb{R}^+)^{\mathscr{M}}$.
3. Prove that the minimum of $\mathscr{L}^{(\lambda)}$ on $(\mathbb{R}^+)^{\mathscr{M}}$ is achieved for

$$q_m^{(\lambda)} = \pi_m \exp(-\beta\widehat{r}_m/\sigma^2)\exp(-1-\beta\lambda/\sigma^2).$$

4. Conclude that

$$\left(\frac{\pi_m e^{-\beta \widehat{r}_m/\sigma^2}}{\sum_{m'\in\mathcal{M}} \pi_{m'} e^{-\beta \widehat{r}_{m'}/\sigma^2}} : m \in \mathcal{M}\right) \in \underset{q\in P(\mathcal{M})}{\operatorname{argmin}} \mathcal{G}(q).$$

3.7.2 Orthonormal Setting with Power Law Prior

We consider the linear regression setting (2.2), in the case where the columns of $\mathbf{X}$ are orthonormal. We choose the collection of models described in Section 2.2.1 for the sparse-coordinate setting and take the distribution π on $\mathcal{M}$ given by

$$\pi_m = \left(1+\frac{1}{p}\right)^{-p} p^{-|m|}.$$

1. Let us write $Z_j = \langle Y, \mathbf{X}_j\rangle$, for $j = 1,\dots,p$. Prove that

$$\widehat{f} = \sum_{j=1}^{p} \frac{\exp\left(\beta Z_j^2/\sigma^2\right)}{\exp\left(2\beta + \log(p)\right) + \exp\left(\beta Z_j^2/\sigma^2\right)}\, Z_j\, \mathbf{X}_j\,.$$

2. What is the consequence in terms of computational complexity?
3. Compare qualitatively this mixing procedure to the model selection procedure in this setting.

3.7.3 Group-Sparse Setting

In the spirit of the algorithm for the coordinate-sparse setting, propose a Metropolis–Hastings algorithm for computing $\widehat{f}$ in the group-sparse setting (with the collection of models and the probability π given in Chapter 2).

3.7.4 Gain of Combining

For $\beta \le 1/4$ and $\delta > 0$, we set

$$\mathcal{M}_\delta = \left\{ m \in \mathcal{M} : r_m - \frac{\sigma^2}{\beta}\log(\pi_m) \le \inf_{m\in\mathcal{M}} \left\{ r_m - \frac{\sigma^2}{\beta}\log(\pi_m)\right\} + \delta \right\}.$$

By adapting the end the proof of Theorem 3.2, prove that

$$\mathbb{E}\left[\|\widehat{f} - f^*\|^2\right] \le \min_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[\|\widehat{f}_m - f^*\|^2\right] + \frac{\sigma^2}{\beta}\log\left(\frac{1}{\pi_m}\right)\right\} + \inf_{\delta>0}\left\{\delta - \frac{\sigma^2}{\beta}\log|\mathcal{M}_\delta|\right\}.$$

In particular, when M estimators are similarly good, the risk bound is reduced by a term of order $\sigma^2 \log(M)$.

3.7.5 Online Aggregation

We consider here a slightly different statistical setting. We assume that we observe some temporal data $(X_t, Y_t)_{t=1,\ldots,T}$, with $Y_t \in \mathbb{R}$ and $X_t \in \mathbb{R}^p$. For example, Y_t can be an air pollution indicator (as the concentration of ozone), and X_t can gather some atmospheric measures, as well as some past values of Y. Our goal is to predict at each time t the value Y_t based on the observations X_t and $(X_i, Y_i)_{i=1,\ldots,t-1}$.

We assume that we have a collection $\left\{\widehat{f}_m = \left[\widehat{f}_m(t)\right]_{t=1,\ldots,T},\ m \in \mathscr{M}\right\}$ of estimators with $\widehat{f}_m(t) = F_{m,t}\left(X_t, (X_i, Y_i)_{i=1,\ldots,t-1}\right)$ for some measurable function $F_{m,t}$. We denote by

$$r_m(T) = \sum_{t=1}^{T} \left(Y_t - \widehat{f}_m(t)\right)^2$$

the cumulated prediction risk of the estimator $\widehat{f}_m$. We want to combine the estimators $\{\widehat{f}_m,\ m \in \mathscr{M}\}$ in order to have a cumulated prediction risk $r(T)$ almost as good as the best $r_m(T)$. We consider a probability distribution π on $\mathscr{M}$, and we denote by

$$\widehat{r}_m(t) = \sum_{i=1}^{t-1} \left(Y_i - \widehat{f}_m(i)\right)^2$$

the estimated prediction risk at time t, with $\widehat{r}_m(1) = 0$. At each time t, we build the Gibbs mixture as in (3.2)

$$\widehat{f}(t) = \sum_{m \in \mathscr{M}} w_m(t) \widehat{f}_m(t), \quad \text{with } w_m(t) = \frac{\pi_m e^{-\beta \widehat{r}_m(t)}}{\mathscr{Z}_t} \quad \text{and } \mathscr{Z}_t = \sum_{m \in \mathscr{M}} \pi_m e^{-\beta \widehat{r}_m(t)}.$$

We emphasize that the weights $w_m(t)$ only depend on the observations available at time t.

We assume in the following that the observations Y_t belong to a bounded interval $[-B, B]$, which is known. We also assume that the predictors $\widehat{f}_m(t)$ belong to this interval. We will prove that in this setting, for a suitable choice of β, we have an oracle inequality similar (3.6).

1. Prove that

$$\begin{aligned}
&\sum_{t=1}^{T} \log\left(\sum_{m \in \mathscr{M}} w_m(t) \exp\left(-\beta \left(Y_t - \widehat{f}_m(t)\right)^2\right)\right) \\
&= \log\left(\sum_{m \in \mathscr{M}} \pi_m \exp\left(-\beta \sum_{t=1}^{T} \left(Y_t - \widehat{f}_m(t)\right)^2\right)\right) \\
&\geq \max_{m \in \mathscr{M}} \left\{\log(\pi_m) - \beta r_m(T)\right\}.
\end{aligned}$$

2. Check that $x \to \exp(-x^2)$ is concave on the interval $[-2^{-1/2}, 2^{-1/2}]$. Prove that for $\beta \leq 1/(8B^2)$ we have

$$\sum_{m \in \mathscr{M}} w_m(t) \exp\left(-\beta \left(Y_t - \widehat{f}_m(t)\right)^2\right) \leq \exp\left(-\beta \left(Y_t - \widehat{f}(t)\right)^2\right).$$

3. Combining the above results, conclude that for $\beta \leq 1/(8B^2)$

$$\sum_{t=1}^{T} \left(Y_t - \widehat{f}(t)\right)^2 \leq \min_{m \in \mathscr{M}} \left\{ \sum_{t=1}^{T} \left(Y_t - \widehat{f}_m(t)\right)^2 + \frac{1}{\beta} \log\left(\frac{1}{\pi_m}\right) \right\}.$$

We emphasize that up to this point, the only assumption on the distribution of the data is that the variables Y_t are bounded. In order to state a result similar to (3.6), we consider in the following the setting

$$Y_t = f(X_t) + \varepsilon_t, \quad t = 1, \ldots, T,$$

where ε_t has mean zero, finite variance σ_t^2 and is independent of X_t and $(X_i, Y_i)_{i=1,\ldots,t-1}$. We still assume that $|Y_t| \leq B$.

4. Prove that for $\beta = 1/(8B^2)$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{f}(t) - f(X_t)\right)^2\right] \leq \min_{m \in \mathscr{M}} \left\{ \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{f}_m(t) - f(X_t)\right)^2\right] + \frac{8B^2}{T} \log\left(\frac{1}{\pi_m}\right) \right\}.$$

Chapter 4

Convex Criteria

We have seen in Chapter 2 that the model selection procedure (2.9) has some nice statistical properties but suffers from a prohibitive computational cost in many cases. For example, in the coordinate-sparse setting, the algorithmic complexity for computing (2.9) is exponential in p, so it cannot be implemented in moderate or high-dimensional settings. To circumvent this issue, a standard trick is to derive from the NP-hard problem (2.9) a convex criterion that can be minimized efficiently. The main point is then to check that the estimator derived from this convex criterion is almost as good as the estimator (2.9), at least for some classes of matrix $\mathbf{X}$. This chapter is devoted to this issue and to some related computational aspects.

4.1 Reminder on Convex Multivariate Functions

In this chapter, we will investigate some estimators that are obtained by minimizing some convex criteria. In order to analyze these estimators, we will need some basic results from convex analysis. This section is a brief reminder on the subdifferentials of convex functions. We refer to Appendix D for the details.

4.1.1 Subdifferentials

A function $F:\mathbb{R}^n\to\mathbb{R}$ is convex if $F(\lambda x+(1-\lambda)y)\leq\lambda F(x)+(1-\lambda)F(y)$ for all $x,y\in\mathbb{R}^n$ and $\lambda\in[0,1]$. When a function F is convex and differentiable, we have

$$F(y)\geq F(x)+\langle\nabla F(x),y-x\rangle\,,\quad\text{for all } x,y\in\mathbb{R}^n;$$

see Lemma D.1 in Appendix D, page 235. For any convex function F (possibly non differentiable), we introduce the subdifferential ∂F of F, defined by

$$\partial F(x)=\{w\in\mathbb{R}^n:\ F(y)\geq F(x)+\langle w,y-x\rangle\ \text{ for all } y\in\mathbb{R}^n\}\,.\qquad(4.1)$$

A vector $w\in\partial F(x)$ is called a subgradient of F in x. It is straightforward to check (see Lemma D.2 in Appendix D, page 235) that F is convex if and only if the set $\partial F(x)$ is non empty for all $x\in\mathbb{R}^n$. Furthermore, when F is convex and differentiable, $\partial F(x)=\{\nabla F(x)\}$; see again Lemma D.2 in Appendix D.

Examples of subdifferentials

We refer to Lemma D.5, page 237, for the derivation of the following subdifferentials.

1. The subdifferential of the ℓ^1 norm $|x|_1 = \sum_j |x_j|$ is given by

$$\partial|x|_1 = \left\{w \in \mathbb{R}^n : \ w_j = \text{sign}(x_j) \text{ for } x_j \neq 0, \ w_j \in [-1,1] \text{ for } x_j = 0\right\},$$

where $\text{sign}(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x\leq 0}$.
Equivalently, $\partial|x|_1 = \{\phi : \langle \phi, x\rangle = |x|_1 \text{ and } |\phi|_\infty \leq 1\}$.

2. The subdifferential of the ℓ^∞-norm $|x|_\infty = \max_j |x_j|$ is given by $\partial|x|_\infty = \{w \in \mathbb{R}^n : \ |w|_1 \leq 1 \text{ and } \langle w, x\rangle = |x|_\infty\}$. For $x \neq 0$, writing $J_* = \left\{j : \ |x_j| = |x|_\infty\right\}$, a vector w is a subgradient of $|x|_\infty$ if and only if it fulfills

$$w_j = 0 \text{ for } j \notin J_* \ \text{ and } \ w_j = \lambda_j \text{sign}(x_j) \text{ for } j \in J_* \ \text{ where } \ \lambda_j \geq 0 \ \text{ and } \ \sum_{j \in J_*} \lambda_j = 1\,.$$

4.1.2 Two Useful Properties

We recall two useful properties of convex functions.

1. The subdifferential of a convex function $F : \mathbb{R}^n \to \mathbb{R}$ is monotone:

$$\langle w_x - w_y, x - y\rangle \geq 0, \quad \text{for all } w_x \in \partial F(x) \text{ and } w_y \in \partial F(y).$$

Actually, by definition we have $F(y) \geq F(x) + \langle w_x, y - x\rangle$ and $F(x) \geq F(y) + \langle w_y, x - y\rangle$. Summing these two inequalities gives $\langle w_x - w_y, x - y\rangle \geq 0$.

2. The minimizers of a convex function $F : \mathbb{R}^n \to \mathbb{R}$ are characterized by

$$x_* \in \underset{x \in \mathbb{R}^n}{\text{argmin}}\, F(x) \quad \iff \quad 0 \in \partial F(x_*). \tag{4.2}$$

This immediately follows from the fact that $F(y) \geq F(x_*) + \langle 0, y - x_*\rangle$ for all $y \in \mathbb{R}^n$ in both cases.

4.2 Lasso Estimator

In this section, we explain the main ideas of the chapter in the coordinate-sparse setting described in Chapter 2, Section 2.1. The other settings are investigated in Section 4.3. In particular, we focus on the linear model $Y = \mathbf{X}\beta^* + \varepsilon$, and we will assume for simplicity that the columns of $\mathbf{X}$ have ℓ^2-norm 1.

Let us consider the family of estimators $\left\{\widehat{f}_m : m \in \mathcal{M}\right\}$ introduced in Chapter 2, Section 2.2, page 31, for the coordinate-sparse setting. We remind the reader that the model selection estimator obtained by minimizing (2.9) in the coordinate-sparse setting has some very good statistical properties, but the minimization of (2.9) has a prohibitive computational cost for moderate to large p. Our goal below is to derive from (2.9) a convex criterion that can be minimized efficiently even for large p. For deriving this convex criterion, we will start from a slight variation of the selection Criterion (2.9) in the coordinate-sparse setting with $\pi_m = (1+1/p)^{-p} p^{-|m|}$, namely

$$\widehat{m} \in \underset{m \in \mathscr{M}}{\operatorname{argmin}} \left\{ \|Y - \mathbf{X}\widehat{\beta}_m\|^2 + \lambda |m| \right\}, \quad \text{with } \lambda = \left(1 + \sqrt{2\log(p)}\right)^2 \sigma^2,$$

where $\widehat{\beta}_m$ is defined by $\widehat{f}_m = \mathbf{X}\widehat{\beta}_m$. We write $\operatorname{supp}(\beta) = \{j : \beta_j \neq 0\}$ for the support of $\beta \in \mathbb{R}^p$, and we observe that for all $m \subset \{1, \ldots, p\}$, the models S_m for the coordinate-sparse setting (defined in Section 2.2.1, page 32) can be written as $S_m = \{\mathbf{X}\beta : \operatorname{supp}(\beta) = m\}$. The estimator $\widehat{f}_m = \operatorname{Proj}_{S_m} Y$ is then equal to $\mathbf{X}\widehat{\beta}_m$, with $\widehat{\beta}_m \in \operatorname{argmin}_{\beta:\ \operatorname{supp}(\beta)=m} \|Y - \mathbf{X}\beta\|^2$, so we have

$$\widehat{m} \quad \in \quad \underset{m \in \mathscr{M}}{\operatorname{argmin}} \quad \min_{\beta:\ \operatorname{supp}(\beta)=m} \left\{ \|Y - \mathbf{X}\beta\|^2 + \lambda |\beta|_0 \right\},$$

with $|\beta|_0 = \operatorname{card}(\operatorname{supp}(\beta))$. Slicing the minimization of $\beta \to \|Y - \mathbf{X}\beta\|^2 + \lambda |\beta|_0$ according to the β with support in $m \subset \{1, \ldots, p\}$, we obtain the identity

$$\widehat{\beta}_{\widehat{m}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - \mathbf{X}\beta\|^2 + \lambda |\beta|_0 \right\}. \tag{4.3}$$

The function $\beta \to \|Y - \mathbf{X}\beta\|^2$ is smooth and convex, so it can be handled easily in the minimization of (4.3). The troubles come from $|\beta|_0$, which is non smooth and non convex. The main idea for deriving from (4.3) a criterion easily amenable to minimization is

$$\text{to replace } |\beta|_0 = \sum_{j=1}^{p} \mathbf{1}_{\beta_j \neq 0} \text{ by } |\beta|_1 = \sum_{j=1}^{p} |\beta_j|, \text{ which is convex.}$$

For $\lambda > 0$, we can then relax the minimization problem (4.3) by considering the convex surrogate

$$\widehat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \mathscr{L}(\beta), \quad \text{where } \mathscr{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda |\beta|_1. \tag{4.4}$$

The estimator $\widehat{\beta}_\lambda$ is called the Lasso estimator. The solution to the minimization problem (4.4) may not be unique, but the resulting estimator $\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda$ is always unique. We refer to Exercise 4.5.1 for a proof of this result and for a criterion ensuring the uniqueness of the solution to (4.4). Criterion (4.4) is convex, and we will describe in Section 4.2.4 some efficient procedures for minimizing it. So the Lasso estimator has the nice feature that it can be computed even for large p. Yet, does the Lasso estimator have some good statistical properties? We will see below that the support of this estimator is a subset of $\{1, \ldots, p\}$ for λ large enough, and Theorem 4.1 will provide a risk bound for the resulting estimator $\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda$.

4.2.1 Geometric Insights

Let us denote by $B_{\ell^1}(R)$ the ℓ^1-ball of radius R defined by $B_{\ell^1}(R) = \{\beta \in \mathbb{R}^p : |\beta|_1 \leq R\}$. We set $\widehat{R}_\lambda = |\widehat{\beta}_\lambda|_1$, and we notice that $\widehat{R}_\lambda$ decreases when λ increases. By Lagrangian duality, the Lasso estimator $\widehat{\beta}_\lambda$ is solution of

$$\widehat{\beta}_\lambda \in \underset{\beta \in B_{\ell^1}(\widehat{R}_\lambda)}{\operatorname{argmin}} \|Y - \mathbf{X}\beta\|^2. \tag{4.5}$$

In Figure 4.1, the level sets of the function $\beta \to \|Y - \mathbf{X}\beta\|^2$ are plotted (dashed lines), together with the constraint $|\beta|_1 \leq R$ for decreasing values of R in an example with $p = 2$. We remark that for R small enough (which corresponds to λ large enough), some coordinate $[\widehat{\beta}_\lambda]_j$ are set to 0. This illustrates the fact that the Lasso estimator selects variables for λ large enough.

Remark. We point out that the selection of variables for R small enough comes from the non smoothness of the ℓ^1-ball $B_{\ell^1}(R)$. Actually, replacing the ℓ^1-ball by a ℓ^p-ball $B_{\ell^p}(R)$ with $1 < p < +\infty$ would not lead to variable selection. We refer to Exercise 4.5.7 for the analysis of the estimator (4.4) when $|\beta|_1$ is replaced by $\|\beta\|^2$.

4.2.2 Analytic Insights

Let us better understand the variable selection observed above by analyzing the shape of the solution of (4.4). The subdifferential of the function $\mathscr{L}$ is

$$\partial\mathscr{L}(\beta) = \left\{-2\mathbf{X}^T(Y - \mathbf{X}\beta) + \lambda z \, : \, z \in \partial|\beta|_1\right\},$$

so the first-order optimality condition (4.2) ensures the existence of $\widehat{z} \in \partial|\widehat{\beta}_\lambda|_1$ fulfilling $-2\mathbf{X}^T(Y - \mathbf{X}\widehat{\beta}_\lambda) + \lambda\widehat{z} = 0$. According to the description of the subdifferential of the ℓ^1-norm given in Section 4.1, we obtain

$$\mathbf{X}^T\mathbf{X}\widehat{\beta}_\lambda = \mathbf{X}^T Y - \frac{\lambda}{2}\widehat{z}. \tag{4.6}$$

for some $\widehat{z} \in \mathbb{R}^p$, fulfilling $\widehat{z}_j = \operatorname{sign}([\widehat{\beta}_\lambda]_j)$ when $[\widehat{\beta}_\lambda]_j \neq 0$ and $\widehat{z}_j \in [-1, 1]$ when $[\widehat{\beta}_\lambda]_j = 0$. Let us investigate the selection of variables induced by this formula.

Orthonormal setting

We first consider the simple case where the columns of $\mathbf{X}$ are orthonormal[1] and thus $\mathbf{X}^T\mathbf{X} = I_p$. In this case, Equation (4.6) gives

$$[\widehat{\beta}_\lambda]_j + \frac{\lambda}{2}\operatorname{sign}([\widehat{\beta}_\lambda]_j) = \mathbf{X}_j^T Y \quad \text{for } [\widehat{\beta}_\lambda]_j \neq 0,$$

[1] It means that the columns of $\mathbf{X}$ are orthogonal with norm 1. Notice that this enforces $p \leq n$.

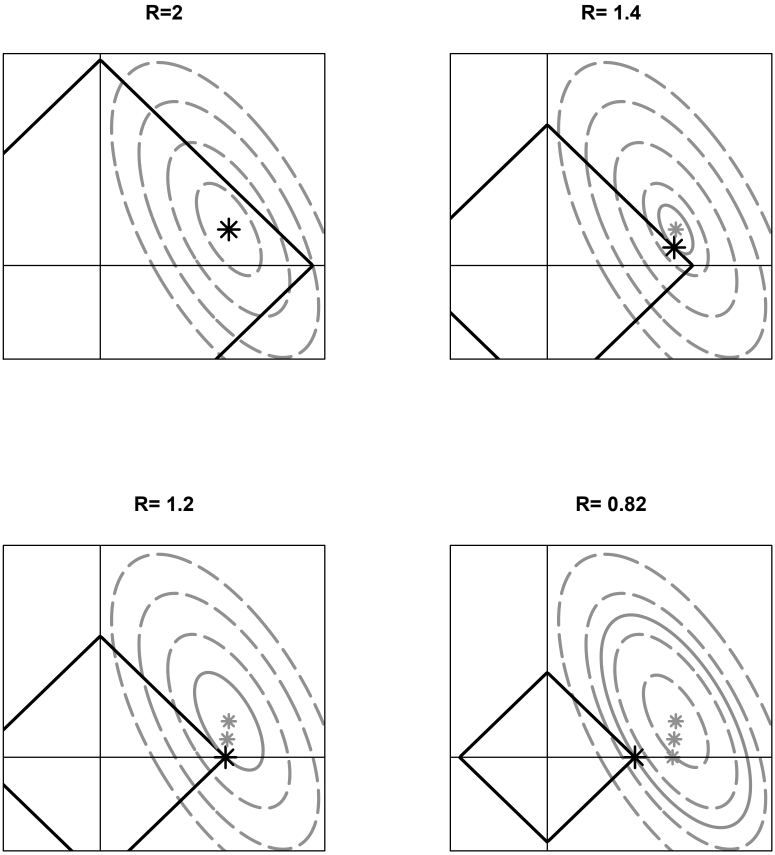


Figure 4.1 *Dashed gray lines represent the level sets of the function* $\beta \to \|Y-\mathbf{X}\beta\|^2$*. Black plain lines represent the* ℓ^1*-balls* $B_{\ell^1}(R)$ *for* $R=2$*,* $R=1.4$*,* $R=1.2$ *and* $R=0.82$*. The dark stars represent* $\widehat{\beta}_R = \mathrm{argmin}_{|\beta|_1\leq R}\|Y-\mathbf{X}\beta\|^2$ *for the current value of R. When* $R=2$*, the* ℓ^1*-norm of* $\widehat{\beta}_R$ *is smaller than R, so* $\widehat{\beta}_R$ *coincides with* $\widehat{\beta}^{ols} = \mathrm{argmin}_{\beta}\|Y-\mathbf{X}\beta\|^2$*. When R is smaller than 1.2, the second coordinate of* $\widehat{\beta}_R$ *is equal to 0.*

which enforces both $\mathrm{sign}([\widehat{\beta}_\lambda]_j) = \mathrm{sign}(\mathbf{X}_j^T Y)$ and $[\widehat{\beta}_\lambda]_j = \mathbf{X}_j^T Y - \lambda\,\mathrm{sign}(\mathbf{X}_j^T Y)/2$ when $[\widehat{\beta}_\lambda]_j \neq 0$. In particular, we notice that we cannot have $[\widehat{\beta}_\lambda]_j \neq 0$ when $|\mathbf{X}_j^T Y| \leq \lambda/2$. Therefore, we have $[\widehat{\beta}_\lambda]_j = 0$ when $|\mathbf{X}_j^T Y| \leq \lambda/2$ and $[\widehat{\beta}_\lambda]_j = \mathbf{X}_j^T Y - \lambda\,\mathrm{sign}(\mathbf{X}_j^T Y)/2$ otherwise. To sum up the above analysis, in the orthonormal setting, the Lasso estimator (4.4) is given by

$$[\widehat{\beta}_\lambda]_j = \mathbf{X}_j^T Y \left(1 - \frac{\lambda}{2|\mathbf{X}_j^T Y|}\right)_+ , \quad j = 1, \ldots, p, \quad \text{with } (x)_+ = \max(x, 0). \qquad (4.7)$$

It then selects the coordinates j such that $|\langle \mathbf{X}_j, Y\rangle| > \lambda/2$.

It is interesting to compare the variables selected by the Lasso estimator to those selected by the initial model-selection estimator (4.3). According to Exercise 2.9.1, Question A.2, the estimator (4.3) selects in the orthonormal setting the coordinates j such that $|\langle \mathbf{X}_j, Y\rangle| > \sqrt{\lambda}$. Therefore, replacing λ in (4.4) by $2\sqrt{\lambda}$ both estimators (4.3) and (4.4) select the same variables in the orthonormal setting.

Non-orthogonal setting

When the columns of $\mathbf{X}$ are not orthogonal, there is no analytic formula for $\widehat{\beta}_\lambda$, and the Lasso estimator will not select the same variables as (4.3) in general. Write $\widehat{m}_\lambda = \{j : [\widehat{\beta}_\lambda]_j \neq 0\}$ for the support of $\widehat{\beta}_\lambda$. Equation (4.6) gives

$$\begin{aligned} 0 \;\leq\; \widehat{\beta}_\lambda^T \mathbf{X}^T \mathbf{X} \widehat{\beta}_\lambda &= \langle \widehat{\beta}_\lambda, \mathbf{X}^T Y - \lambda \widehat{z}/2 \rangle \\ &= \sum_{j \in \widehat{m}_\lambda} [\widehat{\beta}_\lambda]_j \left(\mathbf{X}_j^T Y - \frac{\lambda}{2} \operatorname{sign}([\widehat{\beta}_\lambda]_j) \right), \end{aligned}$$

from which we deduce that $\widehat{\beta}_\lambda = 0$ for $\lambda \geq 2|\mathbf{X}^T Y|_\infty$. When $\lambda < 2|\mathbf{X}^T Y|_\infty$, the Lasso estimator $\widehat{\beta}_\lambda$ is nonzero, but there is no simple formula describing its support.

Finally, we can compare the estimator $\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda$ to the estimator $\widehat{f}_{\widehat{m}_\lambda}$, where $\widehat{f}_m = \operatorname{Proj}_{S_m} Y$ with S_m, as in the beginning of Section 4.2. We denote by A^+ the Moore–Penrose pseudo-inverse of a matrix A (see Appendix C.2 for a reminder on this pseudo-inverse). The matrix $(A^T)^+ A^T$ equals the projection on the range of A, so $A = (A^T)^+ A^T A$. Accordingly, for $\lambda < 2|\mathbf{X}^T Y|_\infty$, we derive from (4.6)

$$\begin{aligned} \widehat{f}_\lambda = \mathbf{X}_{\widehat{m}_\lambda} [\widehat{\beta}_\lambda]_{\widehat{m}_\lambda} &= \left(\mathbf{X}_{\widehat{m}_\lambda}^T\right)^+ \left(\mathbf{X}_{\widehat{m}_\lambda}^T Y - \frac{\lambda}{2} \operatorname{sign}([\widehat{\beta}_\lambda]_{\widehat{m}_\lambda}) \right) \\ &= \operatorname{Proj}_{\widehat{S}_\lambda} Y - \frac{\lambda}{2} \left(\mathbf{X}_{\widehat{m}_\lambda}^T\right)^+ \operatorname{sign}([\widehat{\beta}_\lambda]_{\widehat{m}_\lambda}), \end{aligned}$$

where $\widehat{S}_\lambda = \operatorname{range}(\mathbf{X}_{\widehat{m}_\lambda}) = \operatorname{span}\{\mathbf{X}_j : j \in \widehat{m}_\lambda\}$ and where $\operatorname{sign}(v)$ represents the vector with coordinates $\operatorname{sign}(v_j)$. We observe in particular that $\widehat{f}_\lambda$ differs from $\widehat{f}_{\widehat{m}_\lambda} = \operatorname{Proj}_{\widehat{S}_\lambda} Y$ by an additional term proportional to λ. As we will discuss in Section 4.2.5, this additional term induces a shrinkage of the estimator $\widehat{f}_{\widehat{m}_\lambda}$ toward 0. The intensity of this shrinkage is proportional to λ.

In the next two sections, we state a risk bound for the Lasso estimator, and we describe two numerical schemes for computing it.

4.2.3 Oracle Risk Bound

We have proved in Chapter 2 that the risk of the model selection estimator (2.9) can be nicely bounded; see Theorem 2.2 and Exercise 2.9.2, part A. We derive in this section a risk bound for the Lasso estimator $\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda$, which is similar, at least for some classes of design matrix $\mathbf{X}$.

The best risk bounds available in the literature involve the so-called compatibility constant

$$\kappa(\beta) = \min_{v\in\mathscr{C}(\beta)} \left\{ \frac{\sqrt{|m|}\,\|\mathbf{X}v\|}{|v_m|_1} \right\},$$
$$\text{where } m = \mathrm{supp}(\beta) \text{ and } \mathscr{C}(\beta) = \{v \in \mathbb{R}^p :\ 5|v_m|_1 > |v_{m^c}|_1\}. \quad (4.8)$$

This compatibility constant is a measure of the lack of orthogonality of the columns of $\mathbf{X}_m$; see Exercise 4.5.3. We emphasize that it can be very small for some matrices $\mathbf{X}$. We refer again to Exercise 4.5.3 for a simple lower bound on $\kappa(\beta)$.

A deterministic bound

We first state a deterministic bound and then derive a risk bound from it.

Theorem 4.1 A deterministic bound

For $\lambda \geq 3|\mathbf{X}^T\varepsilon|_\infty$ we have

$$\|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 \leq \inf_{\beta\in\mathbb{R}^p\setminus\{0\}} \left\{ \|\mathbf{X}(\beta-\beta^*)\|^2 + \frac{\lambda^2}{\kappa(\beta)^2}|\beta|_0 \right\}, \quad (4.9)$$

with $\kappa(\beta)$ defined by (4.8).

Proof. The proof mainly relies on the optimality condition (4.2) for (4.4) and some simple (but clever) algebra.

Optimality condition: We have $0 \in \partial\mathscr{L}(\widehat{\beta}_\lambda)$. Since any $\widehat{w} \in \partial\mathscr{L}(\widehat{\beta}_\lambda)$ can be written as $\widehat{w} = -2\mathbf{X}^T(Y - \mathbf{X}\widehat{\beta}_\lambda) + \lambda\widehat{z}$ with $\widehat{z} \in \partial|\widehat{\beta}_\lambda|_1$, using $Y = \mathbf{X}\beta^* + \varepsilon$ we obtain that there exists $\widehat{z} \in \partial|\widehat{\beta}_\lambda|_1$ such that $2\mathbf{X}^T(\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta^*) - 2\mathbf{X}^T\varepsilon + \lambda\widehat{z} = 0$. In particular, for all $\beta \in \mathbb{R}^p$

$$2\langle \mathbf{X}(\widehat{\beta}_\lambda - \beta^*), \mathbf{X}(\widehat{\beta}_\lambda - \beta)\rangle - 2\langle \mathbf{X}^T\varepsilon, \widehat{\beta}_\lambda - \beta\rangle + \lambda\langle \widehat{z}, \widehat{\beta}_\lambda - \beta\rangle = 0. \quad (4.10)$$

Convexity: Since $|.|_1$ is convex, the subgradient monotonicity ensures that $\langle \widehat{z}, \widehat{\beta}_\lambda - \beta\rangle \geq \langle z, \widehat{\beta}_\lambda - \beta\rangle$ for all $z \in \partial|\beta|_1$. Therefore, Equation (4.10) gives

$$\text{for all } \beta \in \mathbb{R}^p \text{ and for all } z \in \partial|\beta|_1 \text{ we have,}$$
$$2\langle \mathbf{X}(\widehat{\beta}_\lambda - \beta^*), \mathbf{X}(\widehat{\beta}_\lambda - \beta)\rangle \ \leq\ 2\langle \mathbf{X}^T\varepsilon, \widehat{\beta}_\lambda - \beta\rangle - \lambda\langle z, \widehat{\beta}_\lambda - \beta\rangle. \quad (4.11)$$

The next lemma provides an upper bound on the right-hand side of (4.11).

Lemma 4.2

We set m =supp(β). There exists $z \in \partial|\beta|_1$, such that for $\lambda \geq 3|\mathbf{X}^T\varepsilon|_\infty$ we have

1. *the inequality* $2\langle \mathbf{X}^T\varepsilon, \widehat{\beta}_\lambda - \beta\rangle - \lambda\langle z, \widehat{\beta}_\lambda - \beta\rangle \leq 2\lambda|(\widehat{\beta}_\lambda - \beta)_m|_1$,
2. *and* $5|(\widehat{\beta}_\lambda - \beta)_m|_1 > |(\widehat{\beta}_\lambda - \beta)_{m^c}|_1$ *when* $\langle \mathbf{X}(\widehat{\beta}_\lambda - \beta^*), \mathbf{X}(\widehat{\beta}_\lambda - \beta)\rangle > 0$.

Proof of the lemma

1. Since $\partial|z|_1 = \left\{z \in \mathbb{R}^p : z_j = \text{sign}(\beta_j) \text{ for } j \in m \text{ and } z_j \in [-1,1] \text{ for } j \in m^c\right\}$, we can choose $z \in \partial|\beta|_1$, such that $z_j = \text{sign}([\widehat{\beta}_\lambda - \beta]_j) = \text{sign}([\widehat{\beta}_\lambda]_j)$ for all $j \in m^c$. Using the duality bound $\langle x, y\rangle \leq |x|_\infty |y|_1$, we have for this choice of z
$$\begin{aligned}
&2\langle \mathbf{X}^T\varepsilon, \widehat{\beta}_\lambda - \beta\rangle - \lambda\langle z, \widehat{\beta}_\lambda - \beta\rangle \\
&= 2\langle \mathbf{X}^T\varepsilon, \widehat{\beta}_\lambda - \beta\rangle - \lambda\langle z_m, (\widehat{\beta}_\lambda - \beta)_m\rangle - \lambda\langle z_{m^c}, (\widehat{\beta}_\lambda - \beta)_{m^c}\rangle \\
&\leq 2|\mathbf{X}^T\varepsilon|_\infty|\widehat{\beta}_\lambda - \beta|_1 + \lambda|(\widehat{\beta}_\lambda - \beta)_m|_1 - \lambda|(\widehat{\beta}_\lambda - \beta)_{m^c}|_1 \\
&\leq \frac{5\lambda}{3}|(\widehat{\beta}_\lambda - \beta)_m|_1 - \frac{\lambda}{3}|(\widehat{\beta}_\lambda - \beta)_{m^c}|_1 \qquad (4.12)\\
&\leq 2\lambda|(\widehat{\beta}_\lambda - \beta)_m|_1,
\end{aligned}$$
where we used $3|\mathbf{X}^T\varepsilon|_\infty \leq \lambda$ and $|\widehat{\beta}_\lambda - \beta|_1 = |(\widehat{\beta}_\lambda - \beta)_{m^c}|_1 + |(\widehat{\beta}_\lambda - \beta)_m|_1$ for the Bound (4.12).
2. When $\langle \mathbf{X}(\widehat{\beta}_\lambda - \beta^*), \mathbf{X}(\widehat{\beta}_\lambda - \beta)\rangle > 0$, combining (4.11) with (4.12) give the inequality $5|(\widehat{\beta}_\lambda - \beta)_m|_1 > |(\widehat{\beta}_\lambda - \beta)_{m^c}|_1$. □

We now conclude the proof of Theorem 4.1. Al-Kashi formula gives

$$2\langle \mathbf{X}(\widehat{\beta}_\lambda - \beta^*), \mathbf{X}(\widehat{\beta}_\lambda - \beta)\rangle = \|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 + \|\mathbf{X}(\widehat{\beta}_\lambda - \beta)\|^2 - \|\mathbf{X}(\beta - \beta^*)\|^2.$$

When this quantity is nonpositive, we have directly (4.9). When this quantity is positive, we can combine it with (4.11) and apply successively the first part of the above lemma, the second part of the lemma with (4.8), and finally $2ab \leq a^2 + b^2$ to get that for all $\beta \in \mathbb{R}^p$

$$\begin{aligned}
\|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 + \|\mathbf{X}(\widehat{\beta}_\lambda - \beta)\|^2 &\leq \|\mathbf{X}(\beta - \beta^*)\|^2 + 2\lambda|(\widehat{\beta}_\lambda - \beta)_m|_1 \\
&\leq \|\mathbf{X}(\beta - \beta^*)\|^2 + \frac{2\lambda\sqrt{|\beta|_0}}{\kappa(\beta)}\|\mathbf{X}(\widehat{\beta}_\lambda - \beta)\| \\
&\leq \|\mathbf{X}(\beta - \beta^*)\|^2 + \frac{\lambda^2|\beta|_0}{\kappa(\beta)^2} + \|\mathbf{X}(\widehat{\beta}_\lambda - \beta)\|^2.
\end{aligned}$$

The proof of Theorem 4.1 is complete. □

If the tuning parameter λ of the Lasso estimator is such that $\lambda \geq 3|\mathbf{X}^T\varepsilon|_\infty$ with high probability, then (4.9) holds true with high probability for this choice of λ. We state in the next corollary such a risk bound in the Gaussian setting (2.3).

Corollary 4.3 Risk bound for the Lasso

Assume that all the columns of $\mathbf{X}$ *have norm 1 and that the noise* $(\varepsilon_i)_{i=1,\dots,n}$ *is i.i.d. with* $\mathcal{N}(0,\sigma^2)$ *distribution.*

Then, for any $L>0$*, the Lasso estimator with tuning parameter*

$$\lambda = 3\sigma\sqrt{2\log(p)+2L}$$

fulfills with probability at least $1-e^{-L}$ *the risk bound*

$$\|\mathbf{X}(\widehat{\beta}_\lambda-\beta^*)\|^2 \leq \inf_{\beta\neq 0}\left\{\|\mathbf{X}(\beta-\beta^*)\|^2+\frac{18\sigma^2(L+\log(p))}{\kappa(\beta)^2}\,|\beta|_0\right\}, \qquad (4.13)$$

with $\kappa(\beta)$ *defined by (4.8).*

Proof. All we need is to prove that $|\mathbf{X}^T\varepsilon|_\infty = \max_{j=1,\dots,p}|\mathbf{X}_j^T\varepsilon|$ is smaller than $\lambda/3 = \sigma\sqrt{2\log(p)+2L}$ with probability at least $1-e^{-L}$. Combining the union bound with the fact that each $\mathbf{X}_j^T\varepsilon$ is distributed according to a $\mathcal{N}(0,\sigma^2)$ Gaussian distribution, we obtain

$$\begin{aligned}\mathbb{P}\left(|\mathbf{X}^T\varepsilon|_\infty > \sigma\sqrt{2\log(p)+2L}\right) &\leq \sum_{j=1}^p \mathbb{P}\left(|\mathbf{X}_j^T\varepsilon| > \sigma\sqrt{2\log(p)+2L}\right)\\ &\leq p\,\mathbb{P}(\sigma|Z|>\sigma\sqrt{2\log(p)+2L}),\end{aligned}$$

with Z distributed according to a $\mathcal{N}(0,1)$ Gaussian distribution. From Lemma B.3, page 218, in Appendix B, we have $\mathbb{P}(|Z|\geq x)\leq e^{-x^2/2}$ for all $x\geq 0$, so the probability $\mathbb{P}\left(|\mathbf{X}^T\varepsilon|_\infty > \sigma\sqrt{2\log(p)+2L}\right)$ is upper bounded by e^{-L}, which concludes the proof of (4.13). □

Discussion of Corollary 4.3

We can compare directly the risk Bound (4.13) for the Lasso estimator to the risk Bound (2.12), page 37, for model selection in the coordinate-sparse setting. Actually, from Inequality (2.31) in Exercise 2.9.2, page 53, we know that there exists a constant $C_K>1$ depending only on $K>1$, such that the model selection estimator $\widehat{\beta}$ defined by (2.9) fulfills the inequality

$$\mathbb{E}\left[\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2\right] \leq C_K \inf_{\beta\neq 0}\left\{\|\mathbf{X}(\beta-\beta^*)\|^2+|\beta|_0\left[1+\log\left(\frac{p}{|\beta|_0}\right)\right]\sigma^2\right\}.$$

Compared to this bound, the risk Bound (4.13) has the nice feature to have a constant one in front of the term $\|\mathbf{X}(\beta-\beta^*)\|^2$, but the complexity term $|\beta|_0\log(p)\sigma^2$ is inflated by a factor $\kappa(\beta)^{-2}$, which can be huge, even infinite, when the columns of $\mathbf{X}$ are far from being orthogonal. The Lasso estimator can actually behave very poorly when $\kappa(\beta^*)$ is small; see, e.g., the second example described in Section 6.3

of Baraud *et al.* [16]. Recent results by Zhang *et al.* [131] suggest that this constant $\kappa(\beta^*)$ is unavoidable in the sense that for some matrices $\mathbf{X}$ the constant $\kappa(\beta^*)^{-2}$ necessarily appears in an upper bound of $\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta^*\|^2$ for any estimator $\widehat{\beta}$ with polynomial algorithmic complexity (see the original paper for a precise statement).

To sum up the above discussion, compared to the model selection estimator (2.9), the Lasso estimator (4.4) is not universally optimal, but it is good in many cases, and, crucially, it can be computed efficiently even for p large. Compared to the forward–backward algorithm described Section 2.5 and the Metropolis–Hastings algorithm described in Section 3.4 (stopped after T iterations), we can provide a risk bound for the Lasso estimator in a non asymptotic setting for any design matrix $\mathbf{X}$. We can also give some conditions that ensure that the support of the Lasso estimator $\widehat{\beta}_\lambda$ is equal to the support of β^*; see Exercise 4.5.2.

4.2.4 Computing the Lasso Estimator

To compute the Lasso estimator (4.4), we need to minimize the function $\beta \to \mathscr{L}(\beta)$, which is convex but non differentiable. We briefly describe below three numerical schemes for computing $\widehat{\beta}_\lambda$.

Coordinate descent

Repeatedly minimizing $\mathscr{L}(\beta_1,\dots,\beta_p)$ with respect to each coordinate β_j is a simple and efficient scheme for minimizing (4.4). This algorithm converges to the Lasso estimator thanks to the convexity of $\mathscr{L}$.

We remind the reader that the columns $\mathbf{X}_j$ of $\mathbf{X}$ are assumed to have norm 1. Setting $R_j = \mathbf{X}_j^T\left(Y - \sum_{k\neq j}\beta_k\mathbf{X}_k\right)$, the partial derivative of the function $\mathscr{L}$, with respect to the variable β_j, is

$$\partial_j\mathscr{L}(\beta) = -2\mathbf{X}_j^T(Y-\mathbf{X}\beta) + \lambda\frac{\beta_j}{|\beta_j|} = 2\beta_j - 2R_j + \lambda\frac{\beta_j}{|\beta_j|}, \quad \text{for all } \beta_j \neq 0.$$

Since $\mathscr{L}$ is convex, the minimizer of $\beta_j \to \mathscr{L}(\beta_1,\dots,\beta_{j-1},\beta_j,\beta_{j+1},\dots,\beta_p)$ is the solution in β_j of $\partial_j\mathscr{L}(\beta_1,\dots,\beta_{j-1},\beta_j,\beta_{j+1},\dots,\beta_p) = 0$ when such a solution exists, and it is $\beta_j = 0$ otherwise. Therefore, the function $\beta_j \to \mathscr{L}(\beta_1,\dots,\beta_{j-1},\beta_j,\beta_{j+1},\dots,\beta_p)$ is minimum in

$$\beta_j = R_j\left(1 - \frac{\lambda}{2|R_j|}\right)_+ \quad \text{with } R_j = \mathbf{X}_j^T\left(Y - \sum_{k\neq j}\beta_k\mathbf{X}_k\right). \tag{4.14}$$

Repeatedly computing $\beta_1,\dots,\beta_p,\beta_1,\dots,\beta_p,\dots$ according to (4.14) gives the coordinate descent algorithm summarized below.

```
Coordinate descent algorithm
Initialization: $\beta = \beta_{\text{init}}$ with $\beta_{\text{init}} \in \mathbb{R}^p$ arbitrary.
Repeat, until convergence of $\beta$, the loop:
    for $j = 1, \ldots, p$
        $\beta_j = R_j \left(1 - \frac{\lambda}{2|R_j|}\right)_+, \quad \text{with } R_j = \mathbf{X}_j^T \Big(Y - \sum_{k \neq j} \beta_k \mathbf{X}_k\Big).$
Output: $\beta$
```

When we want to compute $\{\widehat{\beta}_\lambda : \lambda \in \Lambda\}$ for a grid $\Lambda = \{\lambda_1, \ldots, \lambda_T\}$ of values ranked in decreasing order, it is advised to compute first $\widehat{\beta}_{\lambda_1}$ starting from $\beta_{\text{init}} = 0$, then $\widehat{\beta}_{\lambda_2}$ starting from $\beta_{\text{init}} = \widehat{\beta}_{\lambda_1}$, then $\widehat{\beta}_{\lambda_3}$ starting from $\beta_{\text{init}} = \widehat{\beta}_{\lambda_2}$, etc.

The coordinate descent algorithm is implemented in the `R` package `glmnet` available at `http://cran.r-project.org/web/packages/glmnet/`. For illustration, we give below the `R` code for analyzing the data set `diabetes`, which records the age, the sex, the body mass index, the average blood pressure, some serum measurements, and a quantitative measure of disease progression for $n = 442$ diabetes patients. The goal is to predict from the other variables the measure of disease progression.

```
data(diabetes, package="lars")
library(glmnet)
attach(diabetes)
fit = glmnet(x,y)
plot(fit)
coef(fit,s=1) # extract coefficients at a single value of lambda
predict(fit,newx=x[1:10,],s=1) # make predictions
```

The instruction `plot(fit)` produces a plot of the values of the coordinates of the Lasso estimator $\widehat{\beta}_\lambda$ when λ decreases: The abscissa in the plot corresponds to $|\widehat{\beta}_\lambda|_1$ and the line number j corresponds to the set of points $\left\{\left(|\widehat{\beta}_\lambda|_1, [\widehat{\beta}_\lambda]_j\right) : \lambda \geq 0\right\}$. It is displayed in Figure 4.2. The left-hand side corresponds to $\lambda = +\infty$, the right-hand side corresponds to $\lambda = 0$. We observe that only a few coefficients are nonzero for λ large (left-hand side), and that enlarging λ tends to shrink all the coefficients toward 0. We refer to Chapter 5 for the issue of choosing at best λ.

FISTA algorithm

For $\alpha \in \mathbb{R}^p$, we have noticed Section 4.2.2, page 78, that the minimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\beta - \alpha\|^2 + \lambda |\beta|_1 \right\}$$

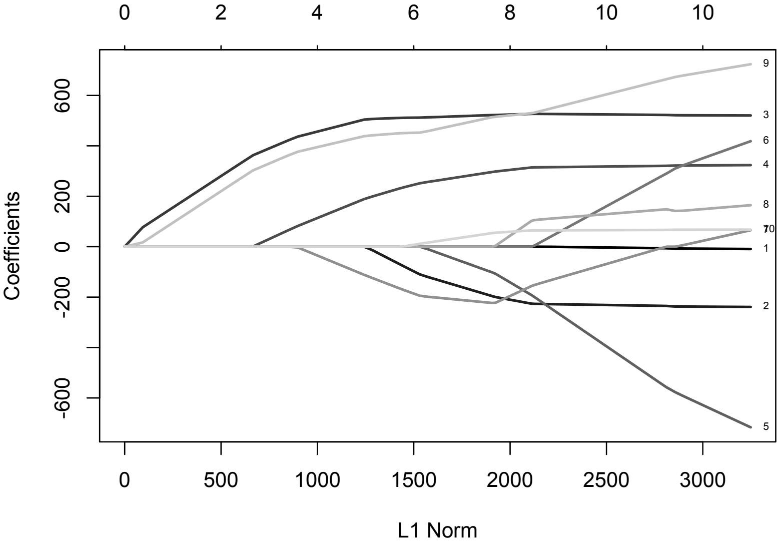


Figure 4.2 *The line j represents the value of the j^{th} coordinate $[\widehat{\beta}_\lambda]_j$ of the Lasso estimator $\widehat{\beta}_\lambda$, when λ decreases from $+\infty$ to 0 on the* `diabetes` *data set.*

has an explicit solution given by

$$S_\lambda(\alpha) = \begin{bmatrix} \alpha_1(1-\lambda/|\alpha_1|)_+ \\ \vdots \\ \alpha_p(1-\lambda/|\alpha_p|)_+ \end{bmatrix}.$$

The Fast Iterative Shrinkage Thresholding Algorithm (FISTA) builds on this formula for computing recursively an approximation of the solution to the minimization problem (4.4). Setting $F(\beta) = \|Y - \mathbf{X}\beta\|^2$, we have for any $b, \beta \in \mathbb{R}^p$

$$\mathscr{L}(b) = F(b) + \lambda|b|_1 = F(\beta) + \langle b-\beta, \nabla F(\beta)\rangle + O(\|b-\beta\|^2) + \lambda|b|_1.$$

For a small $\eta > 0$, starting from $\beta_1 = 0$, we can then iterate until convergence

$$\begin{aligned} \beta_{t+1} &= \operatorname*{argmin}_{b\in\mathbb{R}^p}\left\{F(\beta_t) + \langle b-\beta_t, \nabla F(\beta_t)\rangle + \frac{1}{2\eta}\|b-\beta_t\|^2 + \lambda|b|_1\right\} \\ &= S_{\lambda\eta}\big(\beta_t - \eta\nabla F(\beta_t)\big). \end{aligned}$$

When $\lambda = 0$, since $S_0(\alpha) = \alpha$, the above algorithm simply amounts to a minimization of F by gradient descent with step size η. This algorithm can be accelerated by using Nesterov's acceleration trick [98] leading to FISTA algorithm described below.

```
FISTA algorithm

Initialization: $\beta_1 = \alpha_1 = 0 \in \mathbb{R}^p$, $\mu_1 = 1$, $t = 1$, and $\eta = |2\mathbf{X}^T\mathbf{X}|_{\text{op}}^{-1}$.

Iterate until convergence in $\mathbb{R}^p$ of the sequence $\beta_1, \beta_2, \ldots$

        $\mu_{t+1} = 2^{-1}\left(1 + \sqrt{1 + 4\mu_t^2}\right)$   and   $\delta_t = (1 - \mu_t)\mu_{t+1}^{-1}$
        $\beta_{t+1} = S_{\lambda\eta}\left((I - 2\eta\mathbf{X}^T\mathbf{X})\alpha_t + 2\eta\mathbf{X}^T Y\right)$
        $\alpha_{t+1} = (1 - \delta_t)\beta_{t+1} + \delta_t\beta_t$
        increase $t$ of one unit

Output: $\beta_t$
```

FISTA algorithm has been proved to fulfill some very good convergence properties. We refer to Bubeck [34] for a recent review on convex optimization for machine learning, including an analysis of FISTA algorithm.

LARS algorithm

LARS algorithm is an alternative to coordinate descent and FISTA algorithms. We observe in Figure 4.2 that the map $\lambda \to \widehat{\beta}_\lambda$ is piecewise linear. This observation can be easily explained from the first-order optimality condition (4.6), which gives

$$\mathbf{X}_{\widehat{m}_\lambda}^T \mathbf{X}_{\widehat{m}_\lambda} [\widehat{\beta}_\lambda]_{\widehat{m}_\lambda} = \mathbf{X}_{\widehat{m}_\lambda}^T Y - \frac{\lambda}{2}\operatorname{sign}([\widehat{\beta}_\lambda]_{\widehat{m}_\lambda}), \quad \text{where } \widehat{m}_\lambda = \operatorname{supp}(\widehat{\beta}_\lambda). \tag{4.15}$$

For the values of λ where $\widehat{m}_\lambda$ remains constant, the above equation enforces that $\widehat{\beta}_\lambda$ depends linearly on λ. The LARS algorithm computes the sequence $\{\widehat{\beta}_{\lambda_1}, \widehat{\beta}_{\lambda_2}, \ldots\}$ of Lasso estimators, with $\lambda_1 > \lambda_2 > \ldots$ corresponding to the breakpoints of the path $\lambda \to \beta_\lambda$. At each breakpoint λ_k two situations may occur. Either one of the coordinate of $\widehat{\beta}_\lambda$ tends to 0 when λ tends to λ_k from above, in which case the support of $\widehat{\beta}_{\lambda_k}$ is obtained from the support of $\widehat{\beta}_\lambda$ with $\lambda > \lambda_k$ by removing this coordinate. Or Equation (4.15) requires to add one coordinate in $\widehat{m}_\lambda$ when λ becomes smaller than λ_k.

The LARS algorithm is implemented in the `R` package `lars` available on the CRAN `http://cran.r-project.org/web/packages/lars/`.
It has been reported to be computationally less efficient than the coordinate descent and FISTA algorithms when p is very large. We give below the `R` code for analyzing the `diabetes` data set with LARS.

```
library(lars)
data(diabetes)
attach(diabetes)
fit = lars(x,y)
plot(fit)
```

4.2.5 Removing the Bias of the Lasso Estimator

Let us come back to our simulated example of Section 2.6, page 47. In Figure 4.3, we plot the Lasso estimator $\widehat{f}_{\lambda} = \mathbf{X}\widehat{\beta}_{\lambda}$, with $\lambda = 3\sigma\sqrt{2\log(p)}$ as suggested by Corollary 4.3.

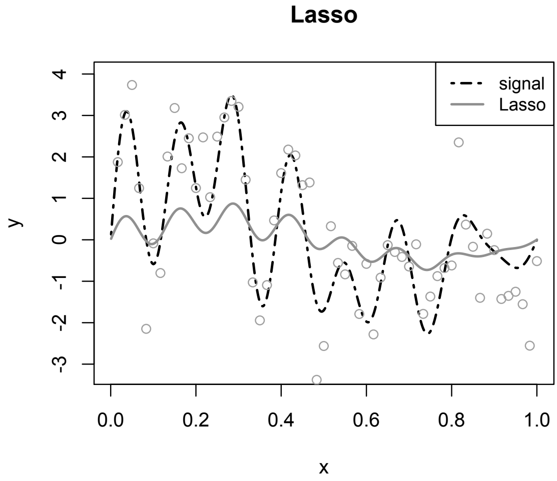


Figure 4.3 *Dotted line: the unknown signal. Gray dots: the noisy observations. Gray line: the Lasso estimator with* $\lambda = 3\sigma\sqrt{2\log(p)}$.

We observe that the Lasso estimator reproduces the oscillations of the signal f, but these oscillations are shrunk toward zero. This shrinkage is easily understood when considering the minimization problem (4.4). Actually, the ℓ^1 penalty has the nice feature to favor sparse solution, but it does also favor the β with small ℓ^1 norm, and thus it induces a shrinkage of the signal. This shrinkage can be seen explicitly in the orthonormal setting described in Section 4.2.2: The ℓ^1 penalty selects the j such that $|\mathbf{X}_j^T Y| > \lambda/2$, but it does also shrink the coordinates $\mathbf{X}_j^T Y$ by a factor $\big(1 - \lambda/(2|\mathbf{X}_j^T Y|)\big)_+$; see Equation (4.7).

A common trick to remove this shrinkage is to use as final estimator the so-called Gauss-Lasso estimator

$$\widehat{f}_{\lambda}^{\text{Gauss}} = \text{Proj}_{\widehat{S}_{\lambda}} Y, \quad \text{where} \quad \widehat{S}_{\lambda} = \text{span}\left\{\mathbf{X}_j : j \in \widehat{m}_{\lambda}\right\}.$$

In other words, with the notations of Chapter 2, the Lasso estimator (4.4) is computed in order to select the model $\widehat{m}_{\lambda} = \text{supp}(\widehat{\beta}_{\lambda})$, and the signal is estimated by $\widehat{f}_{\widehat{m}_{\lambda}} = \text{Proj}_{S_{\widehat{m}_{\lambda}}} Y$. The result for our example is displayed in Figure 4.4. We notice that the shrinkage effect is completely removed.

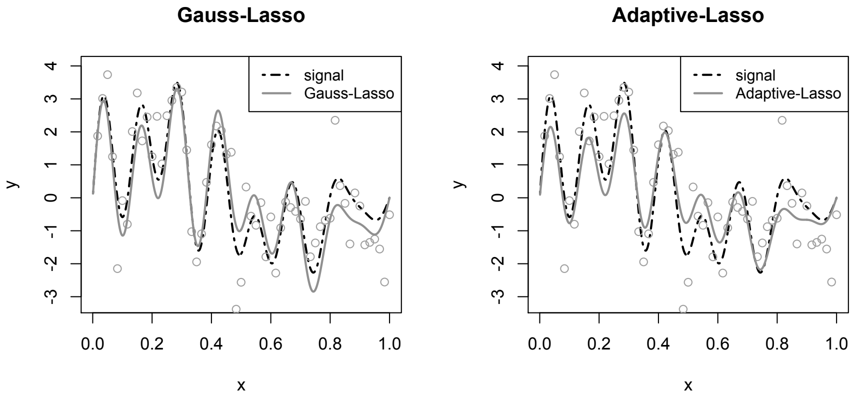


Figure 4.4 *Dotted line: the unknown signal. Gray dots: the noisy observations. Left: the Gauss-Lasso estimator (gray line). Right: the adaptive-Lasso estimator (gray line).*

Another trick for reducing the shrinkage is to compute first the Gauss-Lasso estimator $\widehat{f}_\lambda^{\text{Gauss}} = \mathbf{X}\widehat{\beta}_\lambda^{\text{Gauss}}$ and then estimate β^* with the so-called Adaptive-Lasso estimator

$$\widehat{\beta}_{\lambda,\mu}^{\text{adapt}} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{ \|Y - \mathbf{X}\beta\|^2 + \mu \sum_{j=1}^{p} \frac{|\beta_j|}{|(\widehat{\beta}_\lambda^{\text{Gauss}})_j|} \right\}.$$

The above minimization problem remains convex and it can be solved easily by a coordinate descent algorithm. Let us give a heuristic for considering this estimator. When $\beta \approx \widehat{\beta}_\lambda^{\text{Gauss}}$, we have $\sum_j |\beta_j|/|(\widehat{\beta}_\lambda^{\text{Gauss}})_j| \approx |\beta|_0$, so the above minimization problem can be viewed as an approximation of the initial minimization problem (4.3). This analogy suggests to choose $\mu = (1+\sqrt{2\log(p)})^2\sigma^2$ as in the initial problem (4.3). The Adaptive-Lasso estimator for this value of μ is displayed in Figure 4.4. In this example, the Gauss-Lasso and the Adaptive-Lasso are very similar. In practice, the Adaptive-Lasso is very popular since it tends to select more accurately the variables than the Gauss-Lasso estimator.

4.3 Convex Criteria for Various Sparsity Patterns

The Lasso estimator provides a computationally efficient estimator for the coordinate-sparse setting. We describe below some variations on the Lasso estimator, suited for the other sparsity patterns described Section 2.1 in Chapter 2.

4.3.1 Group–Lasso (Group Sparsity)

We focus in this section on the group-sparse setting described Section 2.1 in Chapter 2. We consider the collection of models described on page 32 for this setting, and we define $\beta_{G_k} = (\beta_j)_{j\in G_k}$. We start from a slight variation of the selection Criterion (2.9) in the group-sparse setting with $\pi_m = (1+1/M)^{-M} M^{-|m|}$, namely

$$\widehat{m} \in \operatorname*{argmin}_{m} \left\{ \|Y - \widehat{f}_m\|^2 + \lambda |m| \right\}, \quad \text{with } \lambda = \left(1+\sqrt{2\log(M)}\right)^2 \sigma^2.$$

We write $\mathscr{K}(\beta) = \left\{k : \beta_{G_k} \neq 0\right\}$, and as in Section 4.2, we observe that the estimator $\widehat{f}_m = \text{Proj}_{S_m} Y$ is equal to $\mathbf{X}\widehat{\beta}_m$, with $\widehat{\beta}_m \in \text{argmin}_{\beta:\ \mathscr{K}(\beta)=m} \|Y - \mathbf{X}\beta\|^2$. Therefore,

$$\widehat{m} \quad \in \quad \operatorname*{argmin}_{m} \min_{\beta:\ \mathscr{K}(\beta)=m} \left\{ \|Y - \mathbf{X}\beta\|^2 + \lambda |\mathscr{K}(\beta)| \right\},$$

and slicing the minimization of $\beta \to \|Y - \mathbf{X}\beta\|^2 + \lambda|\mathscr{K}(\beta)|$ according to the β with $\mathscr{K}(\beta) = m \subset \{1,\ldots,M\}$, we obtain the identity

$$\widehat{\beta}_{\widehat{m}} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{ \|Y - \mathbf{X}\beta\|^2 + \lambda |\mathscr{K}(\beta)| \right\}. \tag{4.16}$$

As in Section 4.2, we want to replace the non convex term $|\mathscr{K}(\beta)| = \sum_k \mathbf{1}_{\beta_{G_k}\neq 0}$ by a convex surrogate. In the coordinate-sparse setting, we have replaced the indicator function $\mathbf{1}_{\beta_j \neq 0}$ by $|\beta_j|$. Following the same idea, we can replace $\mathbf{1}_{\beta_{G_k}\neq 0}$ by $\|\beta_{G_k}\|/\sqrt{|G_k|}$. We divide the norm $\|\beta_{G_k}\|$ by the square-root of the size of the group in order to penalize similarly large and small groups.

For $\lambda = (\lambda_1,\ldots,\lambda_M) \in (\mathbb{R}^+)^M$, the group–Lasso estimator $\widehat{\beta}_\lambda$ is defined as the minimizer of the convex criterion

$$\widehat{\beta}_\lambda \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \mathscr{L}(\beta), \quad \text{where } \mathscr{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \sum_{k=1}^{M} \lambda_k \|\beta_{G_k}\|. \tag{4.17}$$

Let us understand why Criterion (4.17) promotes solutions where some groups of coordinates β_{G_k} are zero, leading to group selection.

We first observe that the geometric picture is the same as in Figure 4.1 for the Lasso, except that here the axes will represent the linear subspaces $\mathbb{R}^{G_1}$ and $\mathbb{R}^{G_2}$. When the minimum is achieved at one of the vertex, the group corresponding to the vertex is selected. From an analytical point of view, the subdifferential of $\sum_k \lambda_k \|\beta_{G_k}\|$ is given by

$$\partial \sum_k \lambda_k \|\beta_{G_k}\| = \left\{ z \in \mathbb{R}^p :\ z_{G_k} = \lambda_k \beta_{G_k}/\|\beta_{G_k}\| \text{ if } \|\beta_{G_k}\| > 0,\ \|z_{G_k}\| \leq \lambda_k \text{ else} \right\}.$$

Similarly to the Lasso, the rigidity of the subdifferential $\partial\|\beta_{G_k}\|$ for $\|\beta_{G_k}\| > 0$ will enforce some $[\widehat{\beta}_\lambda]_{G_k}$ to be zero for λ_k large enough.

Risk bound

We can state a risk bound for the group–Lasso estimator similar to Corollary 4.3 for the Lasso estimator. For simplicity, we assume that each group G_k has the same cardinality $T = p/M$, and we focus on the case where $\lambda_1 = \ldots = \lambda_M = \lambda$.

Similarly to the coordinate-sparse case we introduce the group-compatibility constant

$$\kappa_G(\beta) = \min_{v \in \mathscr{C}_G(\mathscr{K}(\beta))} \left\{ \frac{\sqrt{\mathrm{card}(\mathscr{K}(\beta))}\ \|\mathbf{X}v\|}{\sum_{k \in \mathscr{K}(\beta)} \|v_{G_k}\|} \right\}, \quad \text{where } \mathscr{K}(\beta) = \left\{k : \beta_{G_k} \neq 0\right\},$$

$$\text{and } \mathscr{C}_G(\mathscr{K}) = \left\{ v : \sum_{k \in \mathscr{K}^c} \|v_{G_k}\| < 5 \sum_{k \in \mathscr{K}} \|v_{G_k}\| \right\}. \quad (4.18)$$

We have the following risk bound for the group–Lasso estimator, which is similar to Corollary 4.3 for the Lasso estimator. We write henceforth $\mathbf{X}_{G_k}$ for the submatrix obtained by keeping only the columns of $\mathbf{X}$ with index in G_k.

Theorem 4.4 Risk bound for the group–Lasso

Assume that all the columns of $\mathbf{X}$ *have norm 1 and that the noise* $(\varepsilon_i)_{i=1,\ldots,n}$ *is i.i.d. with* $\mathscr{N}(0, \sigma^2)$ *distribution. We set* $\phi_G = \max_{k=1,\ldots,M} |\mathbf{X}_{G_k}|_{\mathrm{op}}$, *and we assume that all groups have the same cardinality* $T = p/M$.

Then, for any $L > 0$ *the group–Lasso estimator with tuning parameter*

$$\lambda = 3\sigma\left(\sqrt{T} + \phi_G\sqrt{2L + 2\log M}\right) \quad (4.19)$$

fulfills with probability at least $1 - e^{-L}$ *the risk bound*

$$\begin{aligned} &\|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 \\ &\leq \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta - \beta^*)\|^2 + \frac{18\sigma^2}{\kappa_G(\beta)^2} \mathrm{card}(\mathscr{K}(\beta)) \left(T + 2\phi_G^2 (L + \log M)\right) \right\}, \quad (4.20) \end{aligned}$$

with $\kappa_G(\beta)$ *and* $\mathscr{K}(\beta)$ *defined in (4.18).*

We refer to Exercise 4.5.4 for a proof of this result.

Let us compare this risk bound to the risk Bound (4.13) for the Lasso estimator. We first observe that when β has a pure group-sparse structure, we have $T\mathrm{card}(\mathscr{K}(\beta)) = |\beta|_0$. Since no $\log(p)$ is involved in this term, it is then smaller than $|\beta|_0 \log(p)$. There is an additional term in (4.20) proportional to $\mathrm{card}(\mathscr{K}(\beta))\log(M)$. Since $\mathrm{card}(\mathscr{K}(\beta))$ is smaller than $|\beta|_0$ and M is smaller than p, this additional term is also smaller than $|\beta|_0 \log(p)$. Inequality (4.20) then gives a tighter bound than (4.13) when the vector β has a group-sparse structure.

Computing the group–Lasso

Similarly to the Lasso, for computing the group–Lasso estimator, we can apply a block coordinate descent algorithm. The principle is simply to alternate minimization over each block of variables β_{G_k}. For $\beta_{G_k} \neq 0$, the gradient with respect to the block G_k of the function $\mathscr{L}$ defined in (4.17) is

$$\nabla_{\beta_{G_k}} \mathscr{L}(\beta) = -2\mathbf{X}_{G_k}^T R_k + 2\mathbf{X}_{G_k}^T \mathbf{X}_{G_k} \beta_{G_k} + \lambda_k \frac{\beta_{G_k}}{\|\beta_{G_k}\|}, \quad \text{with } R_k = Y - \sum_{j \notin G_k} \beta_j \mathbf{X}_j.$$

Since $\beta_{G_k} \to \mathscr{L}(\beta_{G_1}, \ldots, \beta_{G_M})$ is convex and tends to $+\infty$ when $\|\beta_{G_k}\|$ tends to $+\infty$, the minimum of $\beta_{G_k} \to \mathscr{L}(\beta_{G_1}, \ldots, \beta_{G_M})$ is either the solution of

$$\beta_{G_k} = \left(\mathbf{X}_{G_k}^T \mathbf{X}_{G_k} + \frac{\lambda_k}{2\|\beta_{G_k}\|} I \right)^{-1} \mathbf{X}_{G_k}^T R_k$$

if it exists or it is 0. For $\alpha > 0$, let us define $x_\alpha = \left(\mathbf{X}_{G_k}^T \mathbf{X}_{G_k} + \alpha I \right)^{-1} \mathbf{X}_{G_k}^T R_k$. The minimum of $\beta_{G_k} \to \mathscr{L}(\beta_{G_1}, \ldots, \beta_{G_M})$ is nonzero if and only if there exists $\alpha > 0$, such that $\alpha \|x_\alpha\| = \lambda_k/2$, and then $\beta_{G_k} = x_\alpha$. According to Exercise 4.5.4, there exists $\alpha > 0$ fulfilling $\alpha \|x_\alpha\| = \lambda_k/2$ if and only if $\|\mathbf{X}_{G_k}^T R_k\| > \lambda_k/2$. Let us summarize the resulting minimization algorithm.

Block descent algorithm

```
Initialization: $\beta = \beta_{\text{init}}$ with $\beta_{\text{init}}$ arbitrary.

Iterate until convergence
    for $k = 1, \ldots, M$
        • $R_k = Y - \sum_{j \notin G_k} \beta_j \mathbf{X}_j$
        • if $\|\mathbf{X}_{G_k}^T R_k\| \leq \lambda_k/2$ then $\beta_{G_k} = 0$
        • if $\|\mathbf{X}_{G_k}^T R_k\| > \lambda_k/2$, solve $\beta_{G_k} = \left( \mathbf{X}_{G_k}^T \mathbf{X}_{G_k} + \frac{\lambda_k}{2\|\beta_{G_k}\|} I \right)^{-1} \mathbf{X}_{G_k}^T R_k$

Output: $\beta$
```

An implementation of the group–Lasso is available in the `R` package `gglasso` at `http://cran.r-project.org/web/packages/gglasso/`.

4.3.2 Sparse–Group Lasso (Sparse–Group Sparsity)

In the case of sparse-group sparsity, as described in Section 2.1 of Chapter 2, the nonzero groups β_{G_k} are coordinate sparse. To obtain such a sparsity pattern, we can add a ℓ^1 penalty to the group–Lasso criterion, leading to the Sparse–Group Lasso

$$\widehat{\beta}_{\lambda,\mu} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \mathscr{L}(\beta), \quad \text{where } \mathscr{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \sum_{k=1}^{M} \lambda_k \|\beta_{G_k}\| + \mu |\beta|_1 . \tag{4.21}$$

A similar analysis as for the Lasso and group–Lasso shows that this estimator has a sparse-group sparsity pattern. Risk bounds similar to (4.13) and (4.20) can be proved. An implementation of the Sparse–Group Lasso is available in the `R` package `SGL` at `http://cran.r-project.org/web/packages/SGL/`.

4.3.3 Fused–Lasso (Variation Sparsity)

In the case of variation sparsity, only a few increments $\beta^*_{j+1} - \beta^*_j$ are nonzero. Therefore, we can penalize the residual sum of squares by the ℓ^1 norm of the increments of β, leading to the so-called fused-Lasso estimator

$$\widehat{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \mathscr{L}(\beta), \quad \text{where } \mathscr{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|. \tag{4.22}$$

Setting $\Delta_j = \beta_{j+1} - \beta_j$ for $j = 1, \ldots, p-1$ and $\Delta_0 = \beta_1$, we observe that

$$\mathscr{L}(\beta) = \left\| Y - \sum_{j=0}^{p-1} \Big(\sum_{k=j+1}^{p} \mathbf{X}_k \Big) \Delta_j \right\|^2 + \lambda \sum_{j=1}^{p-1} |\Delta_j|.$$

So computing the fused-Lasso estimator essentially amounts to solving a Lasso problem after a change of variables.

4.4 Discussion and References

4.4.1 Take-Home Message

A successful strategy to bypass the prohibitive computational complexity of model selection is to convexify the model selection criteria. The resulting estimators (Lasso, group–Lasso, fused-Lasso, etc.) are not universally optimal, but they are good in many cases (both in theory and in practice). They can be computed in high-dimensional settings, and they are widely used in science. Furthermore, some risk bounds have been derived for these estimators, providing guarantees on their performances.

4.4.2 References

The Lasso estimator has been introduced conjointly by Tibshirani [117] and Chen *et al.* [46]. The variants presented above and in the next exercises have been proposed by Zou [132], Tibshirani *et al.* [118], Yuan and Lin [129], Friedman *et al.* [60], Candès and Tao [40], and Zou *et al.* [133]. We refer to Bach *et al.* [12] for more complex sparsity patterns.

The theoretical analysis of the performance of the Lasso estimator has given rise to many papers, including Bickel *et al.* [26], Bunea *et al.* [38], Meinshausen and Bühlmann [95], Van de Geer [120], and Wainwright [126]. The analysis presented in this chapter is an adaptation of the results of Koltchinskii, Lounici, and Tsybakov [79] on trace regression. The analysis of the support recovery presented in Exercise 4.5.2 is adapted from Wainwright [126].

We observe that a sensible choice of the tuning parameter λ of the Lasso estimator (or its variants) depends on the variance σ^2 of the noise, which is usually unknown. We will describe in the next chapter some procedures for selecting among different estimators. In particular, these procedure will allow to select (almost) optimally the tuning parameter λ. We also point out some variants of the Lasso estimator, which do not require the knowledge of the variance for selecting their tuning parameter. The most popular variant is probably the square-root/scaled Lasso estimator [22, 3, 116] described in Section 5.4, Chapter 5; see Giraud *et al.* [67] for a review.

Finally, the numerical aspects presented Section 4.2.4 are from Efron *et al.* [57], Friedman *et al.* [58], and Beck and Teboulle [21]. We refer to Bubeck [34] for a recent review on convex optimization.

4.5 Exercises

4.5.1 When Is the Lasso Solution Unique?

The solution $\widehat{\beta}_\lambda$ of the minimization problem (4.4) is not always unique. In this exercise, we prove that the fitted value $\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda$ is always unique, and we give a criterion that enables us to check whether a solution is unique or not.

1. Let $\widehat{\beta}_\lambda^{(1)}$ and $\widehat{\beta}_\lambda^{(2)}$ be two solutions of (4.4) and set $\widehat{\beta} = \left(\widehat{\beta}_\lambda^{(1)} + \widehat{\beta}_\lambda^{(2)}\right)/2$. From the strong convexity of $x \longrightarrow \|x\|^2$ prove that if $\mathbf{X}\widehat{\beta}_\lambda^{(1)} \neq \mathbf{X}\widehat{\beta}_\lambda^{(2)}$, then we have

$$\|Y - \mathbf{X}\widehat{\beta}\|^2 + \lambda|\widehat{\beta}|_1 < \frac{1}{2}\left(\|Y - \mathbf{X}\widehat{\beta}^{(1)}\|^2 + \lambda|\widehat{\beta}^{(1)}|_1 + \|Y - \mathbf{X}\widehat{\beta}^{(2)}\|^2 + \lambda|\widehat{\beta}^{(2)}|_1\right).$$

   Conclude that $\mathbf{X}\widehat{\beta}_\lambda^{(1)} = \mathbf{X}\widehat{\beta}_\lambda^{(2)}$, so the fitted value $\widehat{f}_\lambda$ is unique.

2. Let again $\widehat{\beta}_\lambda^{(1)}$ and $\widehat{\beta}_\lambda^{(2)}$ be two solutions of (4.4) with $\lambda > 0$. From the optimality Condition (4.2), there exists $\widehat{z}^{(1)}$ and $\widehat{z}^{(2)}$, such that

$$-2\mathbf{X}^T(Y - \mathbf{X}\widehat{\beta}_\lambda^{(1)}) + \lambda\widehat{z}^{(1)} = 0 \quad \text{and} \quad -2\mathbf{X}^T(Y - \mathbf{X}\widehat{\beta}_\lambda^{(2)}) + \lambda\widehat{z}^{(2)} = 0.$$

   Check that $\widehat{z}^{(1)} = \widehat{z}^{(2)}$. We write henceforth $\widehat{z}$ for this common value.

3. Set $J = \left\{j : |\widehat{z}_j| = 1\right\}$. Prove that any solution $\widehat{\beta}_\lambda$ to (4.4) fulfills

$$[\widehat{\beta}_\lambda]_{J^c} = 0 \quad \text{and} \quad \mathbf{X}_J^T\mathbf{X}_J[\widehat{\beta}_\lambda]_J = \mathbf{X}_J^T Y - \frac{\lambda}{2}\widehat{z}_J.$$

4. Conclude that

   ”when $\mathbf{X}_J^T\mathbf{X}_J$ is nonsingular, the solution to (4.4) is unique.”

In practice we can check the uniqueness of the Lasso solution by first computing a solution $\widehat{\beta}_\lambda$, then computing $J = \left\{ j : |\mathbf{X}_j^T(Y - \mathbf{X}\widehat{\beta}_\lambda)| = \lambda/2 \right\}$, and finally checking that $\mathbf{X}_J^T\mathbf{X}_J$ is nonsingular.

4.5.2 Support Recovery via the Witness Approach

The goal here is to give some simple conditions for ensuring that the support $\widehat{m}_\lambda$ of the Lasso estimator $\widehat{\beta}_\lambda$ coincides with the support m^* of β^*. It is adapted from Wainwright [126]. The main idea is to compare the solution of Criterion (4.4) with the solution $\widehat{\beta}_\lambda$ of the same minimization problem restricted to the $\beta \in \mathbb{R}^p$, with support in m^*

$$\widetilde{\beta}_\lambda \in \underset{\beta\,:\,\mathrm{supp}(\beta)\subset m^*}{\mathrm{argmin}} \left\{ \|Y - \mathbf{X}\beta\|^2 + \lambda|\beta|_1 \right\}. \tag{4.23}$$

For the sake of simplicity, we assume henceforth that $\mathscr{Q}(\beta) = \|Y - \mathbf{X}\beta\|^2$ is strictly convex ($\mathrm{rank}(\mathbf{X}) = p$), even if the weaker condition $\mathrm{rank}(\mathbf{X}_{m^*}) = |m^*|$ is actually sufficient.

The optimality Condition (4.2) for Problem (4.23) ensures the existence of $\widetilde{z} \in \partial|\widetilde{\beta}_\lambda|_1$, fulfilling $(\nabla\mathscr{Q}(\widetilde{\beta}_\lambda)) + \lambda\widetilde{z})_j = 0$ for all $j \in m^*$. We define $\widehat{z}$ from $\widetilde{z}$ by

$$\widehat{z}_j = \widetilde{z}_j \ \text{ for } j \in m^* \quad \text{ and } \quad \widehat{z}_j = -\frac{1}{\lambda}(\nabla\mathscr{Q}(\widetilde{\beta}_\lambda))_j \ \text{ for } j \notin m^*.$$

A) The witness approach

1. Check that $\nabla\mathscr{Q}(\widetilde{\beta}_\lambda) + \lambda\widehat{z} = 0$.
2. Check that if $|\widehat{z}_j| \leq 1$ for all $j \notin m^*$, then $\widehat{z} \in \partial|\widetilde{\beta}_\lambda|_1$ and $\widetilde{\beta}_\lambda$ is solution of (4.4). Prove that in this case the support $\widehat{m}_\lambda$ of $\widehat{\beta}_\lambda$ is included in m^*.

B) Checking the dual feasibility condition

We assume henceforth that $\mathbf{X}$ fulfills the incoherence condition

$$\left| \mathbf{X}_{(m^*)^c}^T \mathbf{X}_{m^*} (\mathbf{X}_{m^*}^T \mathbf{X}_{m^*})^{-1} \right|_{\ell^\infty \to \ell^\infty} \leq 1 - \gamma, \quad \text{with } 0 < \gamma < 1, \tag{4.24}$$

where $|A|_{\ell^\infty\to\ell^\infty} = \sup_{|x|_\infty=1} |Ax|_\infty$ is the operator norm of the matrix A with respect to the ℓ^∞ norm. We also assume that the columns of $\mathbf{X}$ are normalized: $\|\mathbf{X}_j\| = 1$. We set $\lambda = 4\gamma^{-1}\sigma\sqrt{2(1+A)\log(p)}$, with $A > 0$. We prove in this part that when the incoherence condition (4.24) is met, $|\widehat{z}_j| \leq 1$ for all $j \notin m^*$ with large probability.

For a subset $m \subset \{1, \ldots, p\}$, we write P_m for the orthogonal projection on the linear span $\{\mathbf{X}\beta : \beta \in \mathbb{R}^p \text{ and } \mathrm{supp}(\beta) \subset m\}$, with respect to the canonical scalar product in $\mathbb{R}^p$.

1. Check that $(\widetilde{\beta}_\lambda)_{m^*} = (\mathbf{X}_{m^*}^T\mathbf{X}_{m^*})^{-1}(\mathbf{X}_{m^*}^T Y - \lambda\widetilde{z}_{m^*}/2)$.
2. Check that $\widehat{z}_j = \frac{2}{\lambda}\mathbf{X}_j^T(I - P_{m^*})\varepsilon + \mathbf{X}_j^T\mathbf{X}_{m^*}(\mathbf{X}_{m^*}^T\mathbf{X}_{m^*})^{-1}\widetilde{z}_{m^*}$ for all $j \notin m^*$.

3. Prove that $|\mathbf{X}_j^T\mathbf{X}_{m^*}(\mathbf{X}_{m^*}^T\mathbf{X}_{m^*})^{-1}\widetilde{z}_{m^*}| \leq 1-\gamma$ for all $j \notin m^*$.
4. Prove that with probability at least $1-p^{-A}$ we have $\max_{j\notin m^*}|\frac{2}{\lambda}\mathbf{X}_j^T(I-P_{m^*})\varepsilon| \leq \gamma/2$, and conclude that $|\widehat{z}_j| \leq 1$ for all $j \notin m^*$ with the same probability.

C) Support recovery

1. Check that $(\widetilde{\beta}_\lambda-\beta^*)_{m^*} = (\mathbf{X}_{m^*}^T\mathbf{X}_{m^*})^{-1}(\mathbf{X}_{m^*}^T\varepsilon - \lambda\widetilde{z}_{m^*}/2)$.
2. Prove that with probability at least $1-p^{-A}$

$$\max_{j=1,\ldots,p}\left|(\widetilde{\beta}_\lambda-\beta^*)_j\right| \leq \frac{3\lambda}{4}\left|(\mathbf{X}_{m^*}^T\mathbf{X}_{m^*})^{-1}\right|_{\ell^\infty\to\ell^\infty}.$$

3. Assume that $\min_{j\in m^*}|\beta_j^*| > \frac{3\lambda}{4}\left|(\mathbf{X}_{m^*}^T\mathbf{X}_{m^*})^{-1}\right|_{\ell^\infty\to\ell^\infty}$. Under the hypotheses of Part B, prove by combining the above results that the support $\widehat{m}_\lambda$ of the Lasso estimator $\widehat{\beta}_\lambda$ defined by (4.4) coincides with the support m^* of β^*, with probability at least $1-2p^{-A}$.

4.5.3 Lower Bound on the Compatibility Constant

We will give a simple lower bound on the compatibility constant $\kappa(\beta)$ defined by (4.8). In the following, m refers to the support of β, and $\mathbf{X}_m$ is the matrix obtained by keeping only the column of $\mathbf{X}$ with index in m.

We assume that the norm of the columns $\mathbf{X}_j$ are normalized to one, and we write

$$\theta = \max_{i\neq j}\left|\langle\mathbf{X}_i,\mathbf{X}_j\rangle\right|$$

for the maximum correlation between the columns in $\mathbf{X}$. We prove below that when $|m|$ fulfills $|m| < (11\theta)^{-1}$, the compatibility constant $\kappa(\beta)$ is positive.

1. Considering apart the coordinates in m and the coordinates in m^c, prove that for any $v\in\mathbb{R}^p$ we have

$$\|\mathbf{X}v\|^2 \geq \|\mathbf{X}_m v_m\|^2 - 2\left|v_m\mathbf{X}_m^T\mathbf{X}_{m^c}v_{m^c}\right|.$$

2. Check that $\|\mathbf{X}_m v_m\|^2 \geq \|v_m\|^2 - \theta|v_m|_1^2$.
3. Prove that for $v\in\mathscr{C}(\beta)$, where $\mathscr{C}(\beta)$ is defined in (4.8), we have $\left|v_m\mathbf{X}_m^T\mathbf{X}_{m^c}v_{m^c}\right| \leq 5\theta|v_m|_1^2$.
4. Prove that $\kappa(\beta)^2 \geq 1-11\theta|m|$ and conclude.

4.5.4 On the Group–Lasso

In parts A and B, we prove Theorem 4.4, page 89. The proof follows the same lines as the proof of Theorem 4.1 for the Lasso estimator. In part C, we check the conditions for solving $\alpha\|(A^TA+\alpha I)^{-1}A^TY\| = \lambda/2$ as needed for the block-gradient algorithm described Section 4.3.1.

A) Deterministic bound

We first prove that for $\lambda \geq 3\max_{k=1,\ldots,M} \|\mathbf{X}_{G_k}^T \varepsilon\|$, we have

$$\|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 \leq \inf_{\beta \in \mathbb{R}^p \setminus \{0\}} \left\{ \|\mathbf{X}(\beta - \beta^*)\|^2 + \frac{\lambda^2 \mathrm{card}\,(\mathscr{K}(\beta))}{\kappa_G(\beta)^2} \right\}. \tag{4.25}$$

For any $\mathscr{K} \subset \{1,\ldots,M\}$ and $\beta \in \mathbb{R}^p$, we introduce the notation

$$\|\beta\|_{(\mathscr{K})} = \sum_{k \in \mathscr{K}} \|\beta_{G_k}\|.$$

1. Following the same lines as in the proof of Theorem 4.1, prove that

$$\text{for all } \beta \in \mathbb{R}^p \text{ and for all } z \in \partial \sum_{k=1}^{M} \|\beta_{G_k}\|, \text{ we have}$$
$$2\langle \mathbf{X}(\widehat{\beta}_\lambda - \beta^*), \mathbf{X}(\widehat{\beta}_\lambda - \beta)\rangle \;\leq\; 2\langle \mathbf{X}^T \varepsilon, \widehat{\beta}_\lambda - \beta\rangle - \lambda \langle z, \widehat{\beta}_\lambda - \beta\rangle.$$

2. Let us fix some $\beta \in \mathbb{R}^p$ and write $\mathscr{K} = \mathscr{K}(\beta) = \{k : \beta_{G_k} \neq 0\}$. Prove that for a clever choice of $z \in \partial \sum_k \|\beta_{G_k}\|$, we have

$$-\lambda \langle z, \widehat{\beta}_\lambda - \beta\rangle \leq \lambda \|\widehat{\beta}_\lambda - \beta\|_{(\mathscr{K})} - \lambda \|\widehat{\beta}_\lambda - \beta\|_{(\mathscr{K}^c)}.$$

3. Prove that for $\lambda \geq 3\max_{k=1,\ldots,M} \|\mathbf{X}_{G_k}^T \varepsilon\|$, we have

$$2\langle \mathbf{X}(\widehat{\beta}_\lambda - \beta^*), \mathbf{X}(\widehat{\beta}_\lambda - \beta)\rangle \;\leq\; \frac{\lambda}{3}\left(5\|\widehat{\beta}_\lambda - \beta\|_{(\mathscr{K})} - \|\widehat{\beta}_\lambda - \beta\|_{(\mathscr{K}^c)}\right).$$

4. With Al-Kashi's formula and the Definition (4.18) of $\kappa_G(\beta)$, prove that for $\beta \neq 0$

$$\|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 + \|\mathbf{X}(\widehat{\beta}_\lambda - \beta)\|^2 \leq \|\mathbf{X}(\beta^* - \beta)\|^2 + \frac{2\lambda\sqrt{\mathrm{card}(\mathscr{K})}\,\|\mathbf{X}(\widehat{\beta}_\lambda - \beta)\|}{\kappa_G(\beta)}$$

   and conclude the proof of (4.25).

B) Stochastic control

It remains to prove that with probability at least $1 - e^{-L}$, we have

$$\max_{k=1,\ldots,M} \|\mathbf{X}_{G_k}^T \varepsilon\| \leq \sigma\left(\sqrt{T} + \phi_G\sqrt{2L + 2\log M}\right), \tag{4.26}$$

where $\phi_G = \max_{k=1,\ldots,M} |\mathbf{X}_{G_k}|_{\mathrm{op}}$.

1. Prove that for each $k = 1,\ldots,M$, there exists some exponential random variable ξ_k such that

$$\|\mathbf{X}_{G_k}^T \varepsilon\| \leq \|\mathbf{X}_{G_k}\|_F \sigma + |\mathbf{X}_{G_k}|_{\mathrm{op}} \sigma \sqrt{2\xi_k}.$$

2. Check that $\|\mathbf{X}_{G_k}\|_F^2 = T$ when the columns of $\mathbf{X}$ have norm 1, and check that we have (4.26) with probability at least $1 - e^{-L}$.
3. Conclude the proof of (4.20).

C) On the block descent algorithm

Let λ be a positive real number. The block descent algorithm described in Section 4.3.1, page 87, requires to solve in $\alpha > 0$

$$\alpha \|x_\alpha\| = \lambda/2, \quad \text{where} \quad x_\alpha = (A^T A + \alpha I)^{-1} A^T Y, \tag{4.27}$$

when such a solution exists. In this subsection, we prove that there exists a solution $\alpha > 0$ to the problem (4.27) if and only if $\|A^T Y\| > \lambda/2$.

1. Let $A = \sum_{k=1}^r \sigma_k u_k v_k^T$ be a singular value decomposition of A (see Appendix C for a reminder on the singular value decomposition). Prove that $A^T Y \in \text{span}\{v_1, \ldots, v_r\}$.
2. Check that
$$\alpha x_\alpha = \sum_{k=1}^r \frac{\alpha}{\sigma_k^2 + \alpha} \langle A^T Y, v_k \rangle v_k.$$
3. Check that the map
$$\alpha \to \alpha^2 \|x_\alpha\|^2 = \sum_{k=1}^r \frac{\alpha^2}{(\sigma_k^2 + \alpha)^2} \langle A^T Y, v_k \rangle^2$$
is non decreasing from 0 to $\|A^T y\|^2 = \sum_{k=1}^r \langle A^T Y, v_k \rangle^2$ when α goes from 0 to $+\infty$.
4. Conclude that there exists a solution $\alpha > 0$ to the problem (4.27) if and only if $\|A^T Y\| > \lambda/2$.

4.5.5 Dantzig Selector

The Dantzig selector is a variant of the Lasso that has been proposed by Candès and Tao [40]. It is obtained by solving the minimization problem

$$\widehat{\beta}_\lambda^D \in \underset{\beta : |\mathbf{X}^T (Y - \mathbf{X}\beta)|_\infty \leq \lambda/2}{\operatorname{argmin}} |\beta|_1. \tag{4.28}$$

The Dantzig selector is implemented in the `R` package `flare` available at `http://cran.r-project.org/web/packages/flare/`.

1. Check that the minimization Problem (4.28) is a convex problem that can be recast in a linear program
$$\widehat{\beta}_\lambda^D \in \underset{\beta : -\lambda \mathbf{1} \leq 2\mathbf{X}^T (Y - \mathbf{X}\beta) \leq \lambda \mathbf{1}}{\operatorname{argmin}} \quad \min_{u \in \mathbb{R}_+^p : -u \leq \beta \leq u} \mathbf{1}^T u$$
where $\mathbf{1}$ is a p-dimensional vector with all components equal to 1.

In the following, we will prove that $\widehat{\beta}_\lambda^D$ coincides with the Lasso estimator $\widehat{\beta}_\lambda$ defined by (4.4) when the matrix $(\mathbf{X}^T \mathbf{X})^{-1}$ is diagonal dominant

$$\sum_{k : k \neq j} |(\mathbf{X}^T \mathbf{X})_{kj}^{-1}| < (\mathbf{X}^T \mathbf{X})_{jj}^{-1}, \quad \text{for all } j = 1, \ldots, p.$$

It is adapted from Meinshausen *et al.* [96].

2. Check that the estimator $\widehat{\beta}_\lambda^D$ is a solution of (4.4) if it fulfills
   (a) $2|\mathbf{X}^T(Y-\mathbf{X}\widehat{\beta}_\lambda^D)|_\infty \leq \lambda$
   (b) $2\mathbf{X}_j^T(Y-\mathbf{X}\widehat{\beta}_\lambda^D) = \lambda\,\mathrm{sign}([\widehat{\beta}_\lambda^D]_j)$ for all $j \in \mathrm{supp}(\widehat{\beta}_\lambda^D)$.
3. Assume that the Condition (b) is not satisfied for some $j \in \mathrm{supp}(\widehat{\beta}_\lambda^D)$. Write e_j for the jth vector of the canonical basis in $\mathbb{R}^p$. Prove that the vector $\widetilde{\beta}_\lambda = \widehat{\beta}_\lambda^D - \eta\,\mathrm{sign}([\widehat{\beta}_\lambda^D]_j)(\mathbf{X}^T\mathbf{X})^{-1}e_j$ fulfills Condition (a) for $\eta > 0$ small enough.
4. Prove the inequalities for $\eta > 0$ small enough

$$\begin{aligned} |\widetilde{\beta}_\lambda|_1 &= |[\widetilde{\beta}_\lambda]_{-j}|_1 + |[\widehat{\beta}_\lambda^D]_j| - \eta(\mathbf{X}^T\mathbf{X})^{-1}_{jj} \\ &\leq |\widehat{\beta}_\lambda^D|_1 + \eta \sum_{k:k\neq j} |(\mathbf{X}^T\mathbf{X})^{-1}_{kj}| - \eta(\mathbf{X}^T\mathbf{X})^{-1}_{jj}. \end{aligned}$$

5. Conclude that $\widehat{\beta}_\lambda^D$ must fulfill Condition (b) and finally that $\widehat{\beta}_\lambda^D = \widehat{\beta}_\lambda$.

We emphasize that the Dantzig selector and the Lasso do not coincide in general.

4.5.6 Projection on the ℓ^1-Ball

We present in this exercise a simple algorithm for computing the projection of a vector $\beta \in \mathbb{R}^p$ on the ℓ^1-ball of radius $R > 0$. When $|\beta|_1 \leq R$, the projection is simply β. We assume in the following that $|\beta|_1 > R$. We write $\beta_{(j)}$ for the j^{th}-largest coordinate of β in absolute value, and for $\lambda \geq 0$ we set

$$S_\lambda(\beta) = \left[\beta_j\left(1 - \frac{\lambda}{|\beta_j|}\right)_+\right]_{j=1,\ldots,p}.$$

1. Check that $S_\lambda(\beta) \in \mathrm{argmin}_{\alpha\in\mathbb{R}^p}\left\{\|\beta-\alpha\|^2 + 2\lambda|\alpha|_1\right\}$.
2. Prove that the projection of β on the ℓ^1-ball of radius $R > 0$ is given by $S_{\widehat{\lambda}}(\beta)$, where $\widehat{\lambda} > 0$ is such that $|S_{\widehat{\lambda}}(\beta)|_1 = R$.
3. Let $\widehat{J} \in \{1,\ldots,p\}$ be such that $|\beta_{(\widehat{J}+1)}| \leq \widehat{\lambda} < |\beta_{(\widehat{J})}|$, with the convention $\beta_{(p+1)} = 0$. Check that

$$|S_{\widehat{\lambda}}(\beta)|_1 = \sum_{j\leq\widehat{J}} |\beta_{(j)}| - \widehat{J}\widehat{\lambda}.$$

4. Prove that $\widehat{J}$ then fulfills the two conditions

$$\sum_{j\leq\widehat{J}} |\beta_{(j)}| - \widehat{J}|\beta_{(\widehat{J})}| < R \quad \text{and} \quad \sum_{j\leq\widehat{J}+1} |\beta_{(j)}| - (\widehat{J}+1)|\beta_{(\widehat{J}+1)}| \geq R.$$

5. Conclude that the projection of β on the ℓ^1-ball of radius $R > 0$ is given by $S_{\widehat{\lambda}}(\beta)$, where $\widehat{\lambda}$ is given by

$$\widehat{\lambda} = \widehat{J}^{-1}\left(\sum_{j\leq\widehat{J}} |\beta_{(j)}| - R\right) \quad \text{with} \quad \widehat{J} = \max\left\{J : \sum_{j\leq J} |\beta_{(j)}| - J|\beta_{(J)}| < R\right\}.$$

4.5.7 Ridge and Elastic-Net

We consider the linear model $Y = \mathbf{X}\beta + \varepsilon$, with $Y, \varepsilon \in \mathbb{R}^n$ et $\beta \in \mathbb{R}^p$. We assume that $\mathbf{E}[\varepsilon] = 0$ and $\mathbf{Cov}(\varepsilon) = \sigma^2 I_p$.

A) Ridge regression

For $\lambda > 0$, the ridge estimator $\widehat{\beta}_\lambda$ is defined by

$$\widehat{\beta}_\lambda \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \mathscr{L}_1(\beta) \quad \text{with} \quad \mathscr{L}_1(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2. \tag{4.29}$$

1. Check that $\mathscr{L}_1$ is strictly convex and has a unique minimum.
2. Prove that $\widehat{\beta}_\lambda = A_\lambda Y$ with $A_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T$.
3. Let $\sum_{k=1}^r \sigma_k u_k v_k^T$ be a singular value decomposition of $\mathbf{X}$ (see Theorem (C.1), page 229, in Appendix C). Prove that

$$A_\lambda = \sum_{k=1}^r \frac{\sigma_k}{\sigma_k^2 + \lambda} v_k u_k^T \overset{\lambda \to 0+}{\to} (\mathbf{X}^T)^+$$

   where A^+ is the Moore–Penrose pseudo-inverse of A (see Appendix C for a reminder on the Moore–Penrose pseudo inverse).

4. Check that we have

$$\mathbf{X}\widehat{\beta}_\lambda = \sum_{k=1}^r \frac{\sigma_k^2}{\sigma_k^2 + \lambda} \langle u_k, Y\rangle\, u_k. \tag{4.30}$$

5. Let us denote by $P = \sum_{j=1}^r v_j v_j^T$ the projection on the range of $\mathbf{X}^T$. Check that we have

$$\mathbb{E}\left[\widehat{\beta}_\lambda\right] = \sum_{k=1}^r \frac{\sigma_k^2}{\sigma_k^2 + \lambda} \langle v_k, \beta\rangle v_k,$$

   and

$$\left\|\beta - \mathbb{E}\left[\widehat{\beta}_\lambda\right]\right\|^2 = \|\beta - P\beta\|^2 + \sum_{k=1}^r \left(\frac{\lambda}{\lambda + \sigma_k^2}\right)^2 \langle v_k, \beta\rangle^2.$$

6. Check that the variance of the ridge estimator is given by

$$\operatorname{var}\left(\widehat{\beta}_\lambda\right) = \sigma^2 \operatorname{Tr}(A_\lambda^T A_\lambda) = \sigma^2 \sum_{k=1}^r \left(\frac{\sigma_k}{\sigma_k^2 + \lambda}\right)^2.$$

7. How does the bias and the variance of $\widehat{\beta}_\lambda$ vary when λ increases?

Remark. We notice from (4.30) that the Ridge estimator shrinks Y in the directions u_k where $\sigma_k \ll \lambda$, whereas it leaves Y almost unchanged in the directions u_k where $\sigma_k \gg \lambda$.

B) Elastic-Net

From the Question A)2 we observe that the Ridge regression does not select variables. Actually, the difference between the Lasso estimator and the Ridge estimator is that the ℓ^1 penalty is replaced by a ℓ^2 penalty. We have seen in Section 4.2.1 that the selection property of the Lasso estimator is induced by the non smoothness of the ℓ^1 ball. Since the ℓ^2 ball is smooth, it is not surprising that the Ridge estimator does not select variables.

The Elastic-Net estimator involves both a ℓ^2 and a ℓ^1 penalty. It is meant to improve the Lasso estimator when the columns of $\mathbf{X}$ are strongly correlated. It is defined for $\lambda, \mu \geq 0$ by

$$\widetilde{\beta}_{\lambda,\mu} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \mathscr{L}_2(\beta) \quad \text{with} \quad \mathscr{L}_2(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2 + \mu|\beta|_{\ell^1}.$$

In the following, we assume that the columns of $\mathbf{X}$ have norm 1.

1. Check that the partial derivative of $\mathscr{L}_2$ with respect to $\beta_j \neq 0$ is given by

$$\partial_j \mathscr{L}_2(\beta) = 2\left((1+\lambda)\beta_j - R_j + \frac{\mu}{2}\operatorname{sign}(\beta_j)\right) \quad \text{with} \quad R_j = \mathbf{X}_j^T\left(Y - \sum_{k\neq j}\beta_k\mathbf{X}_k\right).$$

2. Prove that the minimum of $\beta_j \to \mathscr{L}_2(\beta_1, \ldots, \beta_j, \ldots, \beta_p)$ is reached at

$$\beta_j = \frac{R_j}{1+\lambda}\left(1 - \frac{\mu}{2|R_j|}\right)_+.$$

3. What is the difference between the coordinate descent algorithm for the Elastic-Net and the one for the Lasso estimator?

The Elastic-Net procedure is implemented in the `R` package `glmnet` available at `http://cran.r-project.org/web/packages/glmnet/`.

Chapter 5

Estimator Selection

Which modeling and which estimator shall I use? Eternal questions of the statistician when facing data.

The first step is to write down a statistical model suited to analyze the data. This step requires some deep discussions with the specialists of the field where the data come from, some basic data mining to detect some possible key features of the data, and some,... experience. This process is crucial for the subsequent work, yet it seems hardly amenable to a theorization.

The second step is to choose at best an estimator. Assume, for example, that after some pretreatments of the data, you end up with a linear model (2.2). In some cases, you know from expert knowledge the class of structures (coordinate sparsity, group sparsity, variation-sparsity, etc.) hidden in the data. But in many cases, this class is unknown, and we need to select this class from the data. Even if you know the class of structures, say the coordinate sparsity, many estimators have been proposed for this setting, including the Lasso estimator (4.4), the mixing estimator (3.2), the forward–backward estimator of Section 2.5, the Dantzig selector (4.28), the Ridge estimator (4.29), and many others. Which of these estimators shall you use? By the way, all these estimators depend on some "tuning parameter" (λ for the Lasso, β for the mixing, etc.), and we have seen that a suitable choice of these estimators requires the knowledge of the variance σ^2 of the noise. Unfortunately, this variance is usually unknown. So even if you are a Lasso enthusiast, you miss some key information to apply properly the Lasso procedure.

Of course, you should start by removing the estimators whose computational cost exceeds your computational resources. But then you will still have to decide among different estimation schemes and different tuning parameters.

We formalize this issue in Section 5.1, and we then describe different approaches for selecting the estimators and their tuning parameters. The problem of choosing the parameter λ for the Lasso estimator (4.4) will be use as a prototypal example of parameter tuning. We describe three approaches for estimator selection. The two first approaches are quite universal. The first one is based on some data resampling, and the second one is inspired by model selection techniques. The last approach is specifically designed for the estimators introduced in Chapter 4. The theory for analyzing these selection procedures is somewhat more involved than the theory presented in the previous chapters, so we mainly focus in this chapter on the practical description of the procedures, and we refer to the original papers for the details on the theoretical aspects.

5.1 Estimator Selection

Let us formalize the issues discussed above in the regression setting

$$Y_i = f(x^{(i)}) + \varepsilon_i, \quad i = 1, \dots, n,$$

with $f : \mathbb{R}^p \to \mathbb{R}$, for the problem of estimating $f^* = [f(x^{(i)})]_{i=1,\dots,n}$. We want to choose at best one estimator among many different estimators corresponding to different classes of structures (coordinate sparsity, group sparsity, etc.), to different estimation schemes (Lasso, mixing, etc.) and to different values of their tuning parameters.

For an estimation scheme s and a tuning parameter h (possibly multidimensional), we denote by $\widehat{f}_{(s,h)}$ the corresponding estimator of f^*. Let $\mathscr{S}$ be a (finite) family of estimation schemes. At each scheme $s \in \mathscr{S}$, we associate a (finite) set $\mathscr{H}_s$ of tuning parameters. Pooling together all these estimators $\{\widehat{f}_{(s,h)} : s \in \mathscr{S},\ h \in \mathscr{H}_s\}$, our ideal goal is to select among this family an estimator $\widehat{f}_{(s,h)}$ whose risk

$$R\big(\widehat{f}_{(s,h)}\big) = \mathbb{E}\left[\|\widehat{f}_{(s,h)} - f^*\|^2\right]$$

is almost as small as the oracle risk $\min_{\{s \in \mathscr{S}, h \in \mathscr{H}_s\}} R\big(\widehat{f}_{(s,h)}\big)$.

In order to lighten the notations, let us denote by λ the couple (s,h). Setting

$$\Lambda = \bigcup_{s \in \mathscr{S}} \bigcup_{h \in \mathscr{H}_s} \{(s,h)\},$$

this ideal objective can be rephrased as the problem of selecting among $\{\widehat{f}_\lambda,\ \lambda \in \Lambda\}$ an estimator $\widehat{f}_{\widehat{\lambda}}$ whose risk $R(\widehat{f}_{\widehat{\lambda}})$ is almost as small as the oracle risk $\min_{\lambda \in \Lambda} R(\widehat{f}_\lambda)$.

Tuning the Lasso Estimator

The Lasso estimator is a prototypal example of estimator requiring a data-driven selection of its tuning parameter λ. Actually, the choice of λ proposed in Corollary 4.3 is proportional to the standard deviation σ of Y, which is usually unknown. This issue is due to the fact that the Lasso estimator is not invariant by rescaling, as explained below.

Let $\widehat{\beta}(Y, \mathbf{X})$ be any estimator of β^* in the linear model $Y = \mathbf{X}\beta^* + \varepsilon$. Assume that we change the unit of measurement of Y, which amounts to rescale Y by a scaling factor $s > 0$. We then observe $sY = \mathbf{X}(s\beta) + s\varepsilon$ instead of Y. A proper estimator of β should be scale-invariant, which means that it fulfills

$$\widehat{\beta}(sY, \mathbf{X}) = s\widehat{\beta}(Y, \mathbf{X}) \quad \text{for any } s > 0.$$

It turns out that the Lasso estimator (4.4) is not scale-invariant. Actually, let us denote by $\widehat{\beta}_{\lambda}^{\mathrm{lasso}}(Y,\mathbf{X})$ the solution of the minimization problem (4.4) (assuming for simplicity that there is a unique solution). Since

$$\|sY-\mathbf{X}\beta\|^2+\lambda|\beta|_1=s^2\left(\|Y-\mathbf{X}(\beta/s)\|^2+\frac{\lambda}{s}\,|\beta/s|_1\right),$$

we have

$$\widehat{\beta}_{\lambda}^{\mathrm{lasso}}(sY,\mathbf{X})=s\widehat{\beta}_{\lambda/s}^{\mathrm{lasso}}(Y,\mathbf{X}),$$

so the Lasso estimator is not scale-invariant, since we need to rescale the parameter λ by $1/s$. This explains why a sensible choice of the tuning parameter λ of the Lasso estimator (4.4) should be proportional to the standard deviation σ of the response Y.

In practice, to choose the parameter λ, we choose a finite grid Λ of $\mathbb{R}^+$ and select a Lasso estimator $\widehat{f}_{\widehat{\lambda}}=\mathbf{X}\widehat{\beta}_{\widehat{\lambda}}^{\mathrm{lasso}}(Y,\mathbf{X})$ among the collection $\{\widehat{f}_{\lambda}=\mathbf{X}\widehat{\beta}_{\lambda}^{\mathrm{lasso}}(Y,\mathbf{X}):\lambda\in\Lambda\}$ with one of the procedures described below.

5.2 Cross-Validation Techniques

The cross-validation (CV) schemes are nearly universal in the sense that they can be implemented in most statistical frameworks and for most estimation procedures. The principle of the cross-validation schemes is to keep aside a fraction of the data in order to evaluate the prediction accuracy of the estimators. More precisely, the data is split into a *training* set and a *validation* set: the estimators are computed with the *training* set only, and the *validation* set is used for estimating their prediction risk. This training / validation splitting is eventually repeated several times. The cross-validated estimator then selects the estimator with the smallest estimated risk (see below for details). The most popular cross-validation schemes are:

- *Hold-out CV*, which is based on a single split of the data for *training* and *validation*.
- *V-fold CV*, where the data is split into V subsamples. Each subsample is successively removed for *validation*, the remaining data being used for *training*; see Figure 5.1.
- *Leave-one-out*, which corresponds to n-fold CV.
- *Leave-q-out*, (or *delete-q-CV*) where every possible subset of cardinality q of the data is removed for *validation*, the remaining data being used for *training*.

V-fold cross-validation is arguably the most popular cross-validation scheme: It is more stable than the hold-out thanks to the repeated subsampling, and for small V it is computationally much less intensive than the Leave-one-out or Leave-q-out. We describe below the V-fold CV selection procedure in the regression setting of Section 5.1.

train	train	train	train	test
train	train	train	test	train
train	train	test	train	train
train	test	train	train	train
test	train	train	train	train

Figure 5.1 *Recursive data splitting for 5-fold CV*

V-Fold Cross-Validation

Cross-validation schemes are naturally suited to estimators $\widehat{f}_\lambda$ of f^* of the following form. Let us denote by $\mathscr{D} = (y_i, x^{(i)})_{i=1,\ldots,n} \in \mathbb{R}^{(p+1)n}$ our data and assume that there is a map $F_\lambda : \cup_{n\in\mathbb{N}}\mathbb{R}^{(p+1)n} \times \mathbb{R}^p \to \mathbb{R}$ such that $\widehat{f}_\lambda = [F_\lambda(\mathscr{D}, x^{(i)})]_{i=1,\ldots,n}$. In the case of Lasso estimator (4.4), the map F_λ is given by $F_\lambda((Y,\mathbf{X}),x) = \langle \widehat{\beta}_\lambda^{\text{lasso}}(Y,\mathbf{X}), x\rangle$ (with the notations of Section 5.1).

Let $\{1,\ldots,n\} = I_1 \cup \ldots \cup I_V$ be a partition of $\{1,\ldots,n\}$. For a given $k \in \{1,\ldots,V\}$ we denote by $\mathscr{D}_{-I_k}$ the partial data set $\mathscr{D}_{-I_k} = \left\{(y_i, x^{(i)}) : i \in \{1,\ldots,n\} \setminus I_k\right\}$, where we have removed the data points with index in I_k. For each k, the partial data set $\mathscr{D}_{-I_k}$ is used to compute the map $F_\lambda(\mathscr{D}_{-I_k}, \cdot)$, and the remaining data $\mathscr{D}_{I_k} = (y_i, x^{(i)})_{i\in I_k}$ is used to evaluate the accuracy of $F_\lambda(\mathscr{D}_{-I_k}, \cdot)$ for predicting the response y. More precisely, in the setting of Section 5.1, this evaluation will be performed by computing for each $i \in I_k$ the square-prediction error $(y_i - F_\lambda(\mathscr{D}_{-I_k}, x^{(i)}))^2$. Averaging these prediction errors, we obtain the V-fold CV ℓ^2-risk of F_λ

$$\widehat{R}_{CV}[F_\lambda] = \frac{1}{V}\sum_{k=1}^{V} \frac{1}{|I_k|}\sum_{i\in I_k}\left(y_i - F_\lambda(\mathscr{D}_{-I_k}, x^{(i)})\right)^2. \tag{5.1}$$

The V-fold CV procedure then selects $\widehat{\lambda}_{CV}$ by minimizing this estimated risk

$$\widehat{\lambda}_{CV} \in \operatorname*{argmin}_{\lambda\in\Lambda} \widehat{R}_{CV}[F_\lambda].$$

The final estimator of f^* is $\widehat{f}_{\widehat{\lambda}_{CV}}$.

V-fold CV is widely used for estimator selection. It is in particular very popular for tuning procedure such as the Lasso or the group–Lasso. It usually gives good results in practice, yet it suffers from the two following caveats.

Caveat 1. The V-fold CV selection procedure requires to compute V times the maps $\{F(\mathscr{D}_{-I_k}, \cdot) : \lambda \in \Lambda\}$ and also the final estimator $\widehat{f}_{\widehat{\lambda}_{CV}}$. The computational cost of the V-fold CV procedure is therefore roughly $V \times \text{card}(\Lambda)$ times the computational cost of computing a single estimator $\widehat{f}_\lambda$. For large V, this computational cost can overly exceed the available computational ressources. A remedy for this issue is to choose a small V. Yet, for V small, we lose the stabilizing effect of the repeated subsampling, and the V-fold CV procedure can be unstable in this case. A suitable choice of V must then achieve a good balance between the computational complexity and the stability. A common choice in practice is $V = 10$, but some choose smaller values (like $V = 5$) when the computations resources are limited. We point out that when several central processing units (CPU) are available, the maps $\{F(\mathscr{D}_{-I_k}, \cdot) : \lambda \in \Lambda,\ k = 1, \dots, V\}$ can be computed in parallel.

Caveat 2. Despite its wide popularity and its simplicity, there is no general theoretical result that guarantees the validity of the V-fold CV procedure in high dimensional settings. Actually, while it is not hard to identify the expected values of $\widehat{R}_{CV}[F_\lambda]$ in a random design setting (see Exercise 5.6.1), the analysis of the fluctuations of $\widehat{R}_{CV}[F_\lambda]$ around its expectation is much more involved in a high-dimensional setting. In practice, data splitting can be an issue when n is small, since the estimators $F_\lambda(\mathscr{D}_{-I_k}, \cdot)$ are only based on the partial data set $\mathscr{D}_{-I_k}$ and they can be unstable. This instability shrinks when n increases.

5.3 Complexity Selection Techniques

In order to avoid the two above caveats, an alternative point of view is to adapt to estimator selection the ideas of model selection introduced in Chapter 2. Similarly to the model selection estimator (2.9), we introduce a collection $\{S_m,\ m \in \mathscr{M}\}$ of models designed to approximate the estimators $(\widehat{f}_\lambda)_{\lambda \in \Lambda}$ and a probability distribution π on $\mathscr{M}$. We choose exactly the same models and probability distributions as in Chapter 2, depending on the setting under study (coordinate sparsity, group sparsity, variation sparsity, etc.).

Compared to model selection as described in Chapter 2, we have to address three issues. First, for computational efficiency, we cannot explore a huge collection of models. For this reason, we restrict to a subcollection of models $\{S_m,\ m \in \widehat{\mathscr{M}}\}$, possibly data-dependent. The choice of this subcollection depends on the statistical problem; see Section 5.3.1 for the coordinate-sparse setting and 5.3.2 for the group-sparse regression. Second, we have to take into account the fact that a good model S_m for an estimator $\widehat{f}_\lambda$ is a model that achieves a good balance between the approximation error $\|\widehat{f}_\lambda - \text{Proj}_{S_m} \widehat{f}_\lambda\|^2$ and the complexity measured by $\log(1/\pi_m)$. Therefore, the selection criterion involves this approximation error term in addition to the penalty. Finally, the criterion shall not depend on the unknown variance σ^2. For this reason, we replace σ^2 in front of the penalty term by the estimator

$$\widehat{\sigma}_m^2 = \frac{\|Y - \text{Proj}_{S_m} Y\|^2}{n - \dim(S_m)}. \tag{5.2}$$

Combining these three features, we can select the estimator $\widehat{f}_{\widehat{\lambda}}$ by minimizing the criterion

$$\text{Crit}(\widehat{f}_\lambda) = \inf_{m \in \widehat{\mathscr{M}}} \left[\|Y - \text{Proj}_{S_m} \widehat{f}_\lambda\|^2 + \frac{1}{2} \|\widehat{f}_\lambda - \text{Proj}_{S_m} \widehat{f}_\lambda\|^2 + \text{pen}_\pi(m)\, \widehat{\sigma}_m^2 \right], \tag{5.3}$$

where $\widehat{\sigma}_m^2$ is given by (5.2) and $\text{pen}_\pi(m)$ is defined as the unique solution of

$$\mathbb{E}\left[\left(U - \frac{\text{pen}_\pi(m)}{1.1(n-\dim(S_m))}V\right)_+\right] = \pi_m, \tag{5.4}$$

where U and V are two independent chi-square random variables with $\dim(S_m)+1$ and $n-\dim(S_m)-1$ degrees of freedom, respectively. In the cases we will consider here, the penalty $\text{pen}_\pi(m)$ is roughly of the order of $\log(1/\pi_m)$ (see Exercise 5.6.3) and therefore it penalizes S_m according to its complexity $\log(1/\pi_m)$. An implementation of Criterion (5.3) is available in the `R` package `LINselect` `http://cran.r-project.org/web/packages/LINselect/`.

Compared to V-fold CV, this selection criterion enjoys two nice properties.

(i) It does not involve any data splitting. The estimators are then built on the whole data set (which is wise when n is small) and they are computed once.

(ii) The risk of the procedure is controlled by a bound similar to Theorem 2.2 in Chapter 2 for model selection. Deriving such a risk bound involves the same ideas as for the model selection procedure of Chapter 2, but it is somewhat more technical. Therefore, we will not detail these theoretical aspects and we will only focus below on the practical aspects. The references for the theoretical details are given in Section 5.5; see also Exercise 5.6.4.

Caveat. The major caveat of the selection procedure (5.3) is that it is designed for the Gaussian setting and there is no guarantee when the noise does not have a Gaussian distribution.

In the remainder of this section, we describe the instantiation of Criterion (5.3) in the coordinate-sparse setting, in the group-sparse setting, and in a mixed setting.

5.3.1 Coordinate-Sparse Regression

Let us consider the problem of selecting among sparse linear regressors $\{\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda : \lambda \in \Lambda\}$ in the linear regression model $Y = \mathbf{X}\beta^* + \varepsilon$ in the coordinate-sparse setting described in Chapter 2. We consider the collection of models $\{S_m,\ m \in \mathcal{M}\}$ and the probability π designed in Chapter 2 for the coordinate-sparse regression setting. For $\lambda \in \Lambda$, we define $\widehat{m}_\lambda = \text{supp}(\widehat{\beta}_\lambda)$ and

$$\widehat{\mathcal{M}} = \{\widehat{m}_\lambda : \lambda \in \Lambda \text{ and } 1 \leq |\widehat{m}_\lambda| \leq n/(3\log p)\}. \tag{5.5}$$

The estimator $\widehat{f}_{\widehat{\lambda}} = \mathbf{X}\widehat{\beta}_{\widehat{\lambda}}$ is then selected by minimizing (5.3).

As mentioned earlier, the risk of the estimator $\widehat{f}_{\widehat{\lambda}}$ can be compared to the risk of the best estimator in the collection $\{\widehat{f}_\lambda : \lambda \in \Lambda\}$. Let us describe such a result in the case where $\widehat{\beta}_\lambda$ is a Lasso-estimator and $\Lambda = \mathbb{R}^+$.

For $m \subset \{1,\dots,p\}$, we define ϕ_m as the largest eigenvalue of $\mathbf{X}_m^T\mathbf{X}_m$, where $\mathbf{X}_m$ is the matrix derived from $\mathbf{X}$ by only keeping the columns with index in m. The following proposition involves the restricted eigenvalue $\phi^* = \max\{\phi_m : \text{card}(m) \leq n/(3\log p)\}$ and the compatibility constant $\kappa(\beta)$ defined by (4.8) in Chapter 4, page 79.

Proposition 5.1 *There exist positive numerical constants C, C_1, C_2, and C_3, such that for any β^* fulfilling*

$$|\beta^*|_0 \leq C\,\frac{\kappa^2(\beta^*)}{\phi^*} \times \frac{n}{\log(p)},$$

the Lasso estimator $\widehat{f}_{\widehat{\lambda}} = \mathbf{X}\widehat{\beta}_{\widehat{\lambda}}$ selected according to (5.3) fulfills

$$\|\mathbf{X}(\widehat{\beta}_{\widehat{\lambda}} - \beta^*)\|^2 \leq C_3 \inf_{\beta \neq 0}\left\{\|\mathbf{X}(\beta^* - \beta)\|^2 + \frac{\phi^*|\beta|_0 \log(p)}{\kappa^2(\beta)}\,\sigma^2\right\}, \qquad (5.6)$$

with probability at least $1 - C_1 p^{-C_2}$.

The risk Bound (5.6) is similar to (4.13) except that

- it does not require the knowledge of the variance σ^2;
- it requires an upper bound on the cardinality of the support of β^*; and
- it involves two constants C_3, ϕ^* larger than 1.

The proof of this result relies on arguments close to those of Theorem 2.2, yet it is slightly lengthy and we omit it (see Section 5.5 for a reference).

5.3.2 Group-Sparse Regression

Let us now describe the procedure (5.3) in the group-sparse setting of Chapter 2. For simplicity, we restrict to the specific case where all the groups G_k have the same cardinality $T = p/M$ and the columns of $\mathbf{X}$ are normalized to 1. For any $m \subset \{1, \ldots, M\}$, we define the submatrix $\mathbf{X}_{(m)}$ of $\mathbf{X}$ by only keeping the columns of $\mathbf{X}$ with index in $\bigcup_{k\in m} G_k$. We also write $\mathbf{X}_{G_k}$ for the submatrix of $\mathbf{X}$ built from the columns with index in G_k. The collection $\{S_m,\ m \in \mathcal{M}\}$ and the probability π are those defined in Chapter 2 for group sparsity. For a given $\lambda > 0$, similarly to the coordinate-sparse case, we define $\widehat{m}(\lambda) = \{k :\ [\widehat{\beta}_\lambda]_{G_k} \neq 0\}$ and

$$\widehat{\mathcal{M}} = \left\{\widehat{m}(\lambda) :\ \lambda > 0 \ \text{ and } \ 1 \leq |\widehat{m}(\lambda)| \leq \frac{n-2}{2T \vee 3\log(M)}\right\}. \qquad (5.7)$$

The estimator $\widehat{f}_{\widehat{\lambda}} = \mathbf{X}\widehat{\beta}_{\widehat{\lambda}}$ is then selected by minimizing (5.3).

As above, let us give a risk bound for the case where the estimator $\widehat{\beta}_\lambda$ is the group–Lasso estimator (4.17) with $\lambda_1 = \ldots = \lambda_M = \lambda \in \mathbb{R}^+$. For $m \subset \{1, \ldots, M\}$, we define $\phi_{(m)}$ as the largest eigenvalue of $\mathbf{X}_{(m)}^T \mathbf{X}_{(m)}$. The following risk bound involves the cardinality $|\mathcal{K}(\beta)|$ of the set $\mathcal{K}(\beta) = \{k : \beta_{G_k} \neq 0\}$, the group-restricted eigenvalue

$$\phi_G^* = \max\left\{\phi_{(m)} : \quad 1 \leq |m| \leq \frac{n-2}{2T \vee 3\log(M)}\right\}, \qquad (5.8)$$

and the group-compatibility constant $\kappa_G(\beta)$ defined by (4.18) in Chapter 4, page 89.

Proposition 5.2 *There exist positive numerical constants C, C_1, C_2, and C_3, such that when*

$$T \leq (n-2)/4 \quad \text{and} \quad 1 \leq |\mathscr{K}(\beta^*)| \leq C\,\frac{\kappa_G^2(\beta^*)}{\phi_G^*} \times \frac{n-2}{\log(M) \vee T},$$

the group–Lasso estimator $\widehat{f}_{\widehat{\lambda}} = \mathbf{X}\widehat{\beta}_{\widehat{\lambda}}$ selected according to (5.3) fulfills

$$\|\mathbf{X}(\widehat{\beta}_{\widehat{\lambda}} - \beta^*)\|^2 \leq C_3 \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta - \beta^*)\|^2 + \frac{\phi_G^*}{\kappa_G^2(\beta)}\, |\mathscr{K}(\beta)|\, (T + \log(M))\,\sigma^2 \right\}, \tag{5.9}$$

with probability at least $1 - C_1 M^{-C_2}$.

Again, for conciseness, we omit the proof of this result (see Section 5.5 for a reference).

5.3.3 Multiple Structures

When we ignore which class of structures (coordinate sparse, group-sparse, variation sparse, etc.) is hidden in the data, we wish to select the best estimator among a collection of estimators corresponding to various classes of structures. To illustrate this point, let us assume that our family $\{\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda : \lambda \in \Lambda\}$ gathers some estimators $\{\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda : \lambda \in \Lambda^{\text{coord}}\}$ with $\widehat{\beta}_\lambda$ coordinate sparse, and some estimators $\{\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda : \lambda \in \Lambda^{\text{group}}\}$ with $\widehat{\beta}_\lambda$ group sparse. Let us denote by $\{S_m^{\text{coord}} : m \in \mathscr{M}^{\text{coord}}\}$ and π^{coord} (respectively, $\{S_m^{\text{group}} : m \in \mathscr{M}^{\text{group}}\}$ and π^{group}) the models and the probability distribution defined in Chapter 2 for the coordinate-sparse setting (respectively, for the group-sparse setting). Writing $\widehat{\mathscr{M}}^{\text{coord}}$ (respectively, $\widehat{\mathscr{M}}^{\text{group}}$) for the family defined by (5.5) (respectively, by (5.7)), the selection procedure (5.3) can then be applied with the collection of models $\{S_m^{\text{coord}} : m \in \widehat{\mathscr{M}}^{\text{coord}}\} \cup \{S_m^{\text{group}} : m \in \widehat{\mathscr{M}}^{\text{group}}\}$, and with the probability distribution $\pi_m = \pi_m^{\text{coord}}/2$ if $m \in \mathscr{M}^{\text{coord}}$ and $\pi_m = \pi_m^{\text{group}}/2$ if $m \in \mathscr{M}^{\text{group}}$. The theoretical guarantees on the selection procedure (5.3) then ensures that the selected estimator $\widehat{f}_{\widehat{\lambda}}$ is almost as good as the best of the estimators in $\{\widehat{f}_\lambda : \lambda \in \Lambda\}$. Again, we refer to the original papers (see Section 5.5) for the details.

5.4 Scaled-Invariant Criteria

In this section, we describe an alternative to the cross-validation techniques and the complexity selection techniques for the specific problem of choosing the tuning parameter of procedures like the Lasso. In the linear regression setting $Y = \mathbf{X}\beta^* + \varepsilon$, as explained in Section 5.1, a sensible estimator $\widehat{\beta}(Y,\mathbf{X})$ of β^* should be scaled-invariant, which means that $\widehat{\beta}(sY,\mathbf{X}) = s\widehat{\beta}(Y,\mathbf{X})$ for all $s > 0$. Let us consider an estimator $\widehat{\beta}(Y,\mathbf{X})$, obtained by minimizing some function $\beta \to \mathscr{L}(Y,\mathbf{X},\beta)$. The estimator $\widehat{\beta}(Y,\mathbf{X})$ is scaled invariant if

$$\underset{\beta}{\operatorname{argmin}} \mathscr{L}(sY,\mathbf{X},s\beta) = \underset{\beta}{\operatorname{argmin}} \mathscr{L}(Y,\mathbf{X},\beta).$$

Such a condition is met, for example, when there exists some functions $a(Y,\mathbf{X},s)$ and $b(Y,\mathbf{X},s)$, such that

$$\mathscr{L}(sY,\mathbf{X},s\beta) = a(Y,\mathbf{X},s)\mathscr{L}(Y,\mathbf{X},\beta) + b(Y,\mathbf{X},s).$$

The estimators introduced in the Chapter 4 are obtained by minimizing a function

$$\mathscr{L}(Y,\mathbf{X},\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda\Omega(\beta), \tag{5.10}$$

which is the sum of a quadratic term $\|Y - \mathbf{X}\beta\|^2$ and a convex function Ω homogeneous with degree 1. A prototypal example is the Lasso estimator where $\Omega(\beta) = |\beta|_1$. Such estimators are not scale-invariant. We observe that the standard deviation σ of the noise ε is equal to the standard deviation of $Y = \mathbf{X}\beta^* + \varepsilon$. Setting $\lambda = \mu\sigma$, we derive from (5.10) a scale-invariant estimator by minimizing

$$\sigma^{-1}\|Y - \mathbf{X}\beta\|^2 + \mu\Omega(\beta), \quad \text{with } \sigma = \operatorname{sdev}(Y) = \operatorname{sdev}(\varepsilon)$$

instead of (5.10). Yet, the standard deviation σ of the noise is usually unknown, so we cannot compute the above criterion. An idea is to estimate σ with the norm of the residuals, namely $\|Y - \mathbf{X}\beta\|/\sqrt{n}$. Following this idea we obtain for $\mu > 0$ the scaled-invariant criterion

$$\overline{\mathscr{L}}(Y,\mathbf{X},\beta) = \sqrt{n}\|Y - \mathbf{X}\beta\| + \mu\Omega(\beta). \tag{5.11}$$

Since any norm is convex, the function $\overline{\mathscr{L}}(Y,\mathbf{X},\beta)$ is convex and $\overline{\mathscr{L}}(sY,\mathbf{X},s\beta) = s\overline{\mathscr{L}}(Y,\mathbf{X},\beta)$, so the minimizer $\widehat{\beta}(Y,\mathbf{X})$ is scale-invariant.

Square-Root Lasso

Let us investigate the properties of the estimator $\widehat{\beta}$ obtained by minimizing $\overline{\mathscr{L}}(Y,\mathbf{X},\beta)$ when $\Omega(\beta) = |\beta|_1$ (we can follow the same lines with the other criteria). For $\mu > 0$, the resulting estimator is the so-called *square-root Lasso* estimator

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sqrt{n}\|Y - \mathbf{X}\beta\| + \mu|\beta|_1 \right\}. \tag{5.12}$$

The minimization problem (5.12) is scaled-invariant and it can be computed in high-dimensional settings since it is convex. The next theorem shows that the resulting estimator enjoys some nice theoretical properties.

Theorem 5.3 Risk bound for the square-root Lasso

For $L > 0$, we set $\mu = 4\sqrt{\log(p)+L}$. Assume that for some $\delta \in (0,1)$ we have

$$|\beta^*|_0 \leq \delta^2 \, \frac{n\kappa^2(\beta^*)}{64\mu^2}\,, \tag{5.13}$$

with the compatibility constant $\kappa(\beta)$ defined in (4.8). Then, with probability at least $1-e^{-L}-(1+e^2)e^{-n/24}$, the estimator (5.12) fulfills

$$\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 \leq \frac{482(\log(p)+L)}{\delta(1-\delta)\kappa(\beta^*)^2}\, |\beta^*|_0\sigma^2. \tag{5.14}$$

Proof. We first prove a deterministic bound on $\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2$.

Lemma 5.4 Deterministic bound

Let us consider some $\delta \in (0,1)$ and $\mu > 0$. Assume that (5.13) holds and that ε fulfills

$$2|\mathbf{X}^T\varepsilon|_\infty \leq \mu\|\varepsilon\|/\sqrt{n}. \tag{5.15}$$

Then, for $\varepsilon \neq 0$ we have the upper bound

$$\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 \leq \frac{18\mu^2\|\varepsilon\|^2}{\delta(1-\delta)\kappa(\beta^*)^2 n}\, |\beta^*|_0. \tag{5.16}$$

Proof of Lemma 5.4

We set $m_* = \text{supp}(\beta^*)$. We first prove that

$$|(\widehat{\beta}-\beta^*)_{m_*^c}|_1 \leq 3|(\widehat{\beta}-\beta^*)_{m_*}|_1. \tag{5.17}$$

Since $\widehat{\beta}$ minimizes (5.12), we have

$$\begin{aligned}\|Y-\mathbf{X}\widehat{\beta}\| - \|Y-\mathbf{X}\beta^*\| &\leq \frac{\mu}{\sqrt{n}}\left(|\beta^*|_1 - |\widehat{\beta}|_1\right) = \frac{\mu}{\sqrt{n}}\left(|\beta^*_{m^*}|_1 - |\widehat{\beta}_{m_*}|_1 - |\widehat{\beta}_{m_*^c}|_1\right)\\ &\leq \frac{\mu}{\sqrt{n}}\left(|(\widehat{\beta}-\beta^*)_{m^*}|_1 - |(\widehat{\beta}-\beta^*)_{m_*^c}|_1\right). \end{aligned} \tag{5.18}$$

For $\varepsilon \neq 0$, the differential of the map $\beta \to \|Y-\mathbf{X}\beta\|$ at β^* is

$$\nabla_\beta \|Y-\mathbf{X}\beta^*\| = \frac{-\mathbf{X}^T(Y-\mathbf{X}\beta^*)}{\|Y-\mathbf{X}\beta^*\|} = -\frac{\mathbf{X}^T\varepsilon}{\|\varepsilon\|}.$$

Since $\beta \to \|Y-\mathbf{X}\beta\|$ is convex, we have (Lemma D.1, page 235, in Appendix D)

$$\begin{aligned}\|Y-\mathbf{X}\widehat{\beta}\| - \|Y-\mathbf{X}\beta^*\| &\geq -\left\langle \frac{\mathbf{X}^T\varepsilon}{\|\varepsilon\|}, \widehat{\beta}-\beta^*\right\rangle\\ &\geq -\frac{|\mathbf{X}^T\varepsilon|_\infty}{\|\varepsilon\|}\,|\widehat{\beta}-\beta^*|_1 \geq -\frac{\mu}{2\sqrt{n}}\,|\widehat{\beta}-\beta^*|_1, \end{aligned} \tag{5.19}$$

where the last inequality follows from (5.15). Combining (5.19) with (5.18), we get (5.17).

We now prove (5.16). Since Criterion (5.12) is convex, when $Y \neq \mathbf{X}\widehat{\beta}$, the first-order optimality condition (Lemma D.4, page 237, in Appendix D) ensures the existence of $\widehat{z} \in \partial|\widehat{\beta}|_1$, such that

$$\frac{-\sqrt{n}\mathbf{X}^T(Y-\mathbf{X}\widehat{\beta})}{\|Y-\mathbf{X}\widehat{\beta}\|} + \mu\widehat{z} = 0.$$

In particular,

$$-\langle Y-\mathbf{X}\widehat{\beta}, \mathbf{X}(\widehat{\beta}-\beta^*)\rangle + \frac{\mu}{\sqrt{n}}\|Y-\mathbf{X}\widehat{\beta}\|\langle\widehat{z},\widehat{\beta}-\beta^*\rangle = 0.$$

This inequality still holds when $Y = \mathbf{X}\widehat{\beta}$. Since $|\widehat{z}|_\infty \leq 1$ (Lemma D.5 in Appendix D), we have according to (5.15) and (5.17)

$$\begin{aligned}
\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 &\leq \langle\varepsilon, \mathbf{X}(\widehat{\beta}-\beta^*)\rangle + \mu\,\frac{\|Y-\mathbf{X}\widehat{\beta}\|}{\sqrt{n}}\,|\widehat{\beta}-\beta^*|_1 \\
&\leq \left(|\mathbf{X}^T\varepsilon|_\infty + \mu\,\frac{\|Y-\mathbf{X}\widehat{\beta}\|}{\sqrt{n}}\right)|\widehat{\beta}-\beta^*|_1 \\
&\leq 4\mu\left(\frac{\|\varepsilon\|/2 + \|Y-\mathbf{X}\widehat{\beta}\|}{\sqrt{n}}\right)|(\widehat{\beta}-\beta^*)_{m_*}|_1.
\end{aligned}$$

We first observe that $\|Y-\mathbf{X}\widehat{\beta}\| \leq \|\varepsilon\| + \|\mathbf{X}(\widehat{\beta}-\beta^*)\|$. Furthermore, according to (5.17) and the definition of the compatibility constant (4.8), we have the upper bound $|(\widehat{\beta}-\beta^*)_{m_*}|_1 \leq \sqrt{|\beta^*|_0}\,\|\mathbf{X}(\widehat{\beta}-\beta^*)\|/\kappa(\beta^*)$, so

$$\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 \leq 6\mu\,\frac{\|\varepsilon\|}{\sqrt{n}}\,\frac{\sqrt{|\beta^*|_0}\,\|\mathbf{X}(\widehat{\beta}-\beta^*)\|}{\kappa(\beta^*)} + 4\mu\frac{\sqrt{|\beta^*|_0}\,\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2}{\kappa(\beta^*)\sqrt{n}}.$$

For the second right-hand-side term, Condition (5.13) ensures that the factor $4\mu|\beta^*|_0^{1/2}/(\sqrt{n}\kappa(\beta^*))$ is smaller than $\delta/2$. Applying the inequality $ab \leq (2\delta)^{-1}a^2 + \delta b^2/2$ to the first right-hand-side term then gives

$$\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 \leq \frac{18\mu^2\|\varepsilon\|^2}{\delta\kappa(\beta^*)^2 n}\,|\beta^*|_0 + \frac{\delta}{2}\,\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 + \frac{\delta}{2}\,\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2,$$

from which (5.16) follows. □

Let us now conclude the proof of Theorem 5.3. On the event

$$\mathscr{A} = \left\{|\mathbf{X}^T\varepsilon|_\infty^2 \leq 2\sigma^2(\log(p)+L) \text{ and } |\sigma - \|\varepsilon\|/\sqrt{n}| \leq (1-2^{-1/2})\sigma\right\},$$

Condition (5.15) holds with $\mu^2 = 16(\log(p)+L)$, and the Bound (5.16) enforces

Bound (5.14). So all we need is to check that the probability of the event $\mathscr{A}$ is at least $1-e^{-L}-(1+e^2)e^{-n/24}$. According to the Gaussian concentration Inequality (B.2) and Bound (B.4) in Appendix B, we have

$$\begin{aligned}
&\mathbb{P}\left(|\sigma-\|\varepsilon\|/\sqrt{n}|\geq(1-2^{-1/2})\sigma\right)\\
&\leq\mathbb{P}\left(\|\varepsilon\|\geq\mathbb{E}[\|\varepsilon\|]+(\sqrt{n}-\sqrt{n/2})\sigma\right)+\mathbb{P}\left(\|\varepsilon\|\leq\mathbb{E}[\|\varepsilon\|]-(\sqrt{n-4}-\sqrt{n/2})\sigma\right)\\
&\leq e^{-\left(\sqrt{n}-\sqrt{n/2}\right)^2/2}+e^{-\left(\sqrt{n-4}-\sqrt{n/2}\right)^2/2}\leq(1+e^2)e^{-n/24}.
\end{aligned}$$

Finally, the proof of Corollary 4.3 ensures that

$$\mathbb{P}\left(|\mathbf{X}^T\varepsilon|_\infty^2\geq 2\sigma^2(\log(p)+L)\right)\leq e^{-L},$$

so the probability of the event $\mathscr{A}$ is at least $1-e^{-L}-(1+e^2)e^{-n/24}$. The proof of Theorem 5.3 is complete. □

Since Criterion (5.12) has been derived from the Lasso Criterion (4.4), it is worth it to investigate the possible connections between the square-root Lasso $\widehat{\beta}$ and the Lasso $\widehat{\beta}_\lambda^{\text{lasso}}(Y,\mathbf{X})$. We first observe that

$$\sqrt{n}\,\|Y-\mathbf{X}\beta\|=\min_{\sigma>0}\left\{\frac{n\sigma}{2}+\frac{\|Y-\mathbf{X}\beta\|^2}{2\sigma}\right\},$$

and the minimum is achieved for $\sigma=\|Y-\mathbf{X}\beta\|/\sqrt{n}$. As a consequence, the estimator $\widehat{\beta}$ defined by (5.12) and the standard-deviation estimator $\widehat{\sigma}=\|Y-\mathbf{X}\widehat{\beta}\|/\sqrt{n}$ are solution of the convex minimization problem

$$(\widehat{\beta},\widehat{\sigma})\in\underset{(\beta,\sigma)\in\mathbb{R}^p\times\mathbb{R}^+}{\operatorname{argmin}}\left\{\frac{n\sigma}{2}+\frac{\|Y-\mathbf{X}\beta\|^2}{2\sigma}+\mu|\beta|_1\right\}.\tag{5.20}$$

In particular, defining $\widehat{\sigma}$ by (5.20), the square-root Lasso estimator $\widehat{\beta}$ is a solution of $\widehat{\beta}\in\operatorname{argmin}_\beta\left\{\|Y-\mathbf{X}\beta\|^2+2\mu\widehat{\sigma}|\beta|_1\right\}$. In other words, in the notations of Section 5.1, we have

$$\widehat{\beta}=\widehat{\beta}_{2\mu\widehat{\sigma}}^{\text{lasso}}(Y,\mathbf{X}).\tag{5.21}$$

This link between the square-root Lasso estimator and the Lasso estimator has some nice practical and theoretical consequences. From a practical point of view, we observe that Criterion (5.20) is convex in (β,σ), so the estimator $\widehat{\beta}$ can be efficiently computed by alternating minimization in β and σ. For a given σ, the minimization in β amounts to solve a Lasso problem with tuning parameter $2\mu\sigma$, while for a fixed β the minimization in σ has a closed-form solution $\sigma=\|Y-\mathbf{X}\beta\|/\sqrt{n}$. The resulting algorithm is the so-called scaled-Lasso algorithm described below. An implementation of this algorithm is available in the `R` package `scalreg`
`http://cran.r-project.org/web/packages/scalreg/`.

Scaled-Lasso algorithm

```
Initialization: $\widehat{\beta} = 0$

Repeat until convergence
    • $\widehat{\sigma} = \|Y - \mathbf{X}\widehat{\beta}\| / \sqrt{n}$
    • $\widehat{\beta} = \widehat{\beta}^{\text{lasso}}_{2\mu\widehat{\sigma}}(Y, \mathbf{X})$

Output: $\widehat{\beta}$
```

From a theoretical point of view, the link (5.21) allows to improve Bound (5.14) by transposing the results for the Lasso estimator. The only issue is to get a control on the size of $\widehat{\sigma}$. Such a control can be derived from Theorem 5.3. The next corollary provides a risk bound for the square-root Lasso similar to Corollary 4.3, page 81, for the Lasso.

Corollary 5.5 A tighter risk bound for the square-root Lasso

For $L > 0$, let us set $\mu = 3\sqrt{2\log(p) + 2L}$ and assume that β^* fulfills (5.13) with $\delta = 1/5$. Then, with probability at least $1 - e^{-L} - (1 + e^2)e^{-n/24}$, we have

$$\|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 \leq \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta - \beta^*)\|^2 + \frac{202\sigma^2(L + \log(p))}{\kappa(\beta)^2} |\beta|_0 \right\}. \tag{5.22}$$

We refer to Exercise 5.6.2 for a proof of this corollary.

5.5 References and Discussion

5.5.1 Take-Home Message

When analyzing a data set, after writing down a statistical model (say a linear regression model), the statistician faces three issues.

- What is the underlying class of structures hidden in the data? (for example coordinate sparsity, group sparsity, etc.)
- For a given class of structures, which estimator is the best? For example, for coordinate-sparse structures, shall we use the Lasso estimator, the Dantzig estimator, the Elastic-Net? No estimator is universally better than the others, so the choice must be adapted to the data.
- For a given estimator, which tuning parameter should be chosen? Many estimators (like the Lasso) are not scaled-invariant, so any sensible choice of the tuning parameter depends on the variance of the response Y. Since the variance of Y is usually unknown, the choice of the tuning parameter must be adapted to the data. Furthermore, even for scaled-invariant estimators $\mathbf{X}\widehat{\beta}_\mu$ like the square-root Lasso

estimator (5.12), the oracle tuning parameter

$$\mu^*_{\text{oracle}} \in \underset{\mu}{\text{argmin}}\, \mathbb{E}\left[\|\mathbf{X}(\widehat{\beta}_\mu - \beta^*)\|^2\right]$$

depends on $\mathbf{X}\beta^*$ so it is wise to also tune μ according to the data.

These three issues can be handled all together by gathering a collection of estimators $\{\widehat{f}_\lambda : \lambda \in \Lambda\}$ corresponding to different estimation schemes (adapted to various classes of structures) with different tuning parameters. The ideal objective is then to choose $\widehat{f}_{\widehat{\lambda}}$ in this collection with a risk almost as small as the risk of the oracle estimator $\widehat{f}_{\lambda^*_{\text{oracle}}}$, where

$$\lambda^*_{\text{oracle}} \in \underset{\lambda\in\Lambda}{\text{argmin}}\, \mathbb{E}\left[\|\widehat{f}_\lambda - f^*\|^2\right].$$

Different approaches have been developed in order to handle this issue.

A first class of selection procedures is based on subsampling: part of the data is used to compute the estimators, while the remaining data is used to evaluate their prediction accuracy. This process being possibly repeated several times. This approach, including V-fold CV, is very popular and widely used. It usually provides satisfying results in practice, yet it suffers from two caveats. First, the repeated subsampling can lead to intensive computations exceeding the available computing resources. Second, there is no theoretical guarantees in general on the outcome of the selection process in high-dimensional settings. In particular, the subsampling device can suffer from instability for very small sample sizes.

A second estimator selection approach is inspired by the model selection techniques of Chapter 2. This approach has the nice features to avoid data splitting and to enjoy some non asymptotic theoretical guarantees on the outcome of the selection process.Yet, it is limited to the setting where the noise $\varepsilon_1,\ldots,\varepsilon_n$ is i.i.d. with Gaussian distribution.

Finally, for tuning estimators based on the minimization of a non-homogeneous criterion, a third strategy is to modify this criterion in order to obtain scaled-invariant estimators like the square-root Lasso. This approach is computationally efficient and enjoys some nice theoretical properties. Yet, it is limited to the tuning of some specific estimators like those introduced in Chapter 4, and it does not allow to compare different estimation schemes.

There are a few other estimator selection procedures that have been proposed recently in the literature. We point some of them out in the references below. The mathematical analysis of general estimator selection procedures is somewhat more involved than the theory described in the previous chapters. We also observe that the results stated in Proposition 5.1, Proposition 5.2, and Theorem 5.3 involve some extra conditions on the sparsity of the regressor β^* compared to Theorem 4.1 and Theorem 4.4 in Chapter 4 for the Lasso or group–Lasso estimators. For example, in the coordinate-sparse setting, these conditions roughly require that $|\beta^*|_0$ remains small compared to $n/\log(p)$. This unpleasant condition is unfortunately unavoidable when the variance σ^2 is unknown, as has been shown by Verzelen [124].

The theory of estimator selection has to be strengthened, yet some tools are already available, each providing various interesting features.

5.5.2 References

Cross-validation techniques date back at least from the '60s, classical references including the papers by Mosteller and Tukey [97], Devroye and Wagner [51], Geisser [61], Stone [115], and Shao [111]. The asymptotic analysis of V-fold-CV has been carried out in the '80s by Li [89], among others. The non asymptotic analysis of V-fold CV is much more involved. Some non asymptotic results have been derived by Arlot, Célisse, and Lerasle [4, 7, 87] for some specific settings. We refer to Arlot and Celisse [6] for a recent review on the topic.

The complexity selection Criterion (5.3) has been introduced and analyzed by Baraud, Giraud, Huet, and Verzelen [16, 17, 67]. An associated `R` package `LINselect` is available on the CRAN archive network
`http://cran.r-project.org/web/packages/LINselect/`.

The square-root Lasso estimator (5.12) has been introduced and analyzed by Belloni, Chernozhukov, Antoniadis, Sun, and Zhang [22, 3, 116] (see Städler, Bühlmann, and van de Geer [114] for a variant based on a penalized maximum likelihood criterion). An associated `R` package `scalreg` is also available on the CRAN archive network `http://cran.r-project.org/web/packages/scalreg/`. An analog of the square-root Lasso for the group-sparse setting is analyzed in Bunea *et al.* [35].

Some other non asymptotic approaches have been developed for estimator selection in the regression setting with unknown variance σ^2. The *slope-heuristic* developed by Birgé and Massart [30], Arlot and Massart [8], and Arlot and Bach [5] builds on the following idea. Consider $\widehat{f}_{\widehat{\lambda}}$ selected by minimizing a penalized criterion $\mathrm{crit}(\widehat{f}_\lambda) = \|Y-\widehat{f}_\lambda\|^2 + \mathrm{pen}(\widehat{f}_\lambda)$. Assume that there exists some penalty $\mathrm{pen}_{\min}(\widehat{f}_\lambda)$, such that

- when $\mathrm{pen}(\widehat{f}_\lambda) = K\mathrm{pen}_{\min}(\widehat{f}_\lambda)$ with $K > \sigma^2$, we have an oracle risk bound like Theorem 2.2 for $\widehat{f}_{\widehat{\lambda}}$, and
- when $\mathrm{pen}(\widehat{f}_\lambda) = K\mathrm{pen}_{\min}(\widehat{f}_\lambda)$ with $K < \sigma^2$, we have $\widehat{f}_{\widehat{\lambda}} \approx Y$.

For example, in the setting of Exercise 2.9.1 in Chapter 2, the penalty $\mathrm{pen}_{\min}(\widehat{f}_m) = 2|m|\log(p)$ is a minimal penalty. When such a minimal penalty exists, we then observe the following phenomenon. Define $\widehat{\lambda}_K$ by

$$\widehat{\lambda}_K \in \operatorname*{argmin}_{\lambda\in\Lambda}\left\{\|Y-\widehat{f}_\lambda\|^2 + K\mathrm{pen}_{\min}(\widehat{f}_\lambda)\right\}.$$

When $K < \sigma^2$, we have $\widehat{f}_{\widehat{\lambda}_K} \approx Y$, whereas when $K > \sigma^2$, we have $\|Y-\widehat{f}_{\widehat{\lambda}_K}\|$ large. The rough idea is then to track a value $\widehat{K}$ where this transition occurs and use this value as an estimator of σ^2. We refer to the original papers [30, 5] for a more precise description of the slope-heuristic procedure and some non asymptotic bounds.

Another approach based on pairwise test comparisons on a discretized space has been proposed and analyzed by Baraud and Birgé [28, 14, 15]. The procedure cannot be easily sketched in a couple of lines, so we refer to the original papers. This estimator selection technique has the nice feature to be very flexible and it enjoys the property to be able to automatically adapt to the distribution of the noise. This last property is extremely desirable, since we usually do not know the distribution of the noise. Unfortunately, the computational complexity of the selection procedure is generally very high and it cannot be directly implemented even in moderate dimensions.

As mentioned above, there is a need for some efforts to strengthen the theory of estimator selection. We refer to Giraud *et al.* [67] for a recent review on this issue in the regression setting.

5.6 Exercises

5.6.1 Expected V-Fold CV ℓ^2-Risk

A natural setting for analyzing the V-fold CV selection procedure is the random design regression setting, where the observations $(Y_1,X_1),\ldots,(Y_n,X_n)$ are i.i.d. with common distribution $\mathbb{P}$. We assume in the following that the variance of Y is finite, and we keep the notations of Section 5.2. Writing $f(x)$ for (a version of) the conditional expectation $f(x)=\mathbb{E}[Y\,|\,X=x]$, we have $Y_i=f(X_i)+\varepsilon_i$ with the $\varepsilon_1,\ldots,\varepsilon_n$ i.i.d. centered and with finite variance.

For any measurable function $g:\mathbb{R}^p\to\mathbb{R}$, we denote by $\|g\|_{L^2(\mathbb{P}^X)}$ the expectation $\|g\|_{L^2(\mathbb{P}^X)}=\mathbb{E}\left[g(X_1)^2\right]^{1/2}$.

1. Prove that $\mathbb{E}[\varepsilon_1|X_1]=0$ and $\mathbb{E}\left[(Y_1-g(X_1))^2\right]=\|g-f\|^2_{L^2(\mathbb{P}^X)}+\mathrm{var}(\varepsilon_1)$.
2. Prove the equality
$$\mathbb{E}\left[(Y_1-F_\lambda(\mathscr{D}_{-I_1},X_1))^2\right]=\mathbb{E}_{-I_1}\left[\|F_\lambda(\mathscr{D}_{-I_1},\cdot)-f\|^2_{L^2(\mathbb{P}^X)}\right]+\mathrm{var}(\varepsilon_1),$$
where $\mathbb{E}_{-I_1}$ refers to the expectation with respect to $\mathscr{D}_{-I_1}$.
3. Conclude that the expected value of the V-fold CV ℓ^2-risk $\widehat{R}_{CV}[F_\lambda]$ defined by (5.1) is given by
$$\mathbb{E}\left[\widehat{R}_{CV}[F_\lambda]\right]=\mathbb{E}_{-I_1}\left[\|F_\lambda(\mathscr{D}_{-I_1},\cdot)-f\|^2_{L^2(\mathbb{P}^X)}\right]+\mathrm{var}(\varepsilon_1).$$

Remark. Since the variance of ε_1 does not depend on λ, the V-fold CV ℓ^2-risk is equal up to a constant to an unbiased estimator of the integrated risk $\mathbb{E}_{-I_1}\left[\|F_\lambda(\mathscr{D}_{-I_1},\cdot)-f\|^2_{L^2(\mathbb{P}^X)}\right]$. This risk can be viewed as an approximation of the risk $\mathbb{E}\left[\|F_\lambda(\mathscr{D},\cdot)-f\|^2_{L^2(\mathbb{P}^X)}\right]$, which measures the mean L^2-distance between the estimator $F_\lambda(\mathscr{D},\cdot)$ of f and f.

5.6.2 Proof of Corollary 5.5

The proof of Corollary 5.5 builds on the link (5.21) between the Lasso and square-root Lasso estimators. The idea is to bound $\widehat{\sigma} = \|Y - \mathbf{X}\widehat{\beta}\|/\sqrt{n}$ from above and below and apply Theorem 4.1 for the Lasso. To bound $\widehat{\sigma}$, we essentially check that

$$\frac{\|\varepsilon\| - \|\mathbf{X}(\widehat{\beta}-\beta^*)\|}{\sqrt{n}} \leq \widehat{\sigma} \leq \frac{\|\varepsilon\| + \|\mathbf{X}(\widehat{\beta}-\beta^*)\|}{\sqrt{n}} \tag{5.23}$$

and then bound $\|\mathbf{X}(\widehat{\beta}-\beta^*)\|$ with (5.16). In the following, we work on the event

$$\mathscr{A} = \left\{3|\mathbf{X}^T\varepsilon|_\infty \leq \mu\sigma \text{ and } |\sigma - \|\varepsilon\|/\sqrt{n}| \leq (1-2^{-1/2})\sigma\right\}.$$

1. Prove that the event $\mathscr{A}$ has probability at least $1 - e^{-L} - (1+e^2)e^{-n/24}$.
2. Check Inequalities (5.23).
3. From Lemma 5.4, prove that under the hypotheses of Corollary 5.5 we have on the event $\mathscr{A}$ for $\delta \leq 1/5$

$$\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 \leq \frac{18\mu^2|\beta^*|_0}{\delta(1-\delta)\kappa(\beta^*)^2 n}\,\|\varepsilon\|^2 \leq \frac{9\delta}{32(1-\delta)}\,\|\varepsilon\|^2 \leq \left(1-2^{-1/2}\right)^2\|\varepsilon\|^2.$$

4. Check that we have $3|\mathbf{X}^T\varepsilon|_\infty \leq 2\mu\widehat{\sigma}$ on the event $\mathscr{A}$.
5. From Theorem 4.1, prove that on the event $\mathscr{A}$ we have

$$\begin{aligned}\|\mathbf{X}(\widehat{\beta}-\beta^*)\|^2 &\leq \inf_{\beta\neq 0}\left\{\|\mathbf{X}(\beta-\beta^*)\|^2 + \frac{4\mu^2\widehat{\sigma}^2}{\kappa(\beta)^2}\,|\beta|_0\right\}\\ &\leq \inf_{\beta\neq 0}\left\{\|\mathbf{X}(\beta-\beta^*)\|^2 + \frac{202(L+\log(p))\sigma^2}{\kappa(\beta)^2}\,|\beta|_0\right\}\end{aligned}$$

   and conclude the proof of Corollary 5.5.

5.6.3 Some Properties of Penalty (5.4)

For any positive integers D, N, we denote by X_D and X_N two independent positive random variables, such that X_D^2 and X_N^2 are two independent chi-square random variables with, respectively, D and N degrees of freedom. We write

$$p_D(x) = 2^{-D/2}\Gamma(D/2)^{-1}x^{D/2-1}e^{-x/2}\mathbf{1}_{\mathbb{R}^+}(x), \quad x \in \mathbb{R}$$

for the probability distribution function of X_D^2. From (B.4), page 222, in Chapter B we have $\sqrt{D-4} \leq \mathbb{E}[X_D] \leq \sqrt{D}$. In the following, d_m denotes the dimension of the model S_m.

A) Computation of the penalty

We prove in this first part that the penalty $\text{pen}_\pi(m)$ defined by (5.4), page 106, is given by $\text{pen}_\pi(m) = 1.1(n-d_m)x_{n,d_m}$, where x_{n,d_m} is solution in x of

$$\pi_m = (d_m+1)\mathbb{P}\left(F_{d_m+3,n-d_m-1} \geq \frac{n-d_m-1}{d_m+3}x\right) - x(n-d_m-1)\mathbb{P}\left(F_{d_m+1,n-d_m+1} \geq \frac{n-d_m+1}{d_m+1}x\right), \quad (5.24)$$

where $F_{d,r}$ is a Fisher random variable with d and r degrees of freedom.

1. Check that $xp_D(x) = Dp_{D+2}(x)$ for any positive integer D and $x \in \mathbb{R}$.
2. For $\alpha > 0$ prove the equalities

$$\mathbb{E}\left[(X_D^2 - \alpha X_N^2)_+\right] = \int_0^\infty \int_{\alpha y}^\infty xp_D(x)p_N(y)\,dx\,dy - \alpha \int_0^\infty \int_{\alpha y}^\infty yp_D(x)p_N(y)\,dx\,dy$$
$$= D\mathbb{P}\left(X_{D+2}^2 \geq \alpha X_N^2\right) - \alpha N\mathbb{P}\left(X_D^2 \geq \alpha X_{N+2}^2\right).$$

3. Conclude the proof of (5.24).

B) An upper bound on the penalty

In this second part, we prove that for any model m fulfilling

$$\left(\sqrt{d_m+1} + 2\sqrt{\log(8/\pi_m)}\right)^2 \leq n - d_m - 5, \quad (5.25)$$

we have the upper bound

$$\text{pen}_\pi(m) \leq 2.2\frac{n-d_m}{n-d_m-5}\left(\sqrt{d_m+1} + 2\sqrt{\log(8/\pi_m)}\right)^2 \quad (5.26)$$

(we refer to Proposition 4 in Baraud *et al.* [16] for a tighter upper bound).

Let $L > 0$ and N, D be such that

$$t_{D,N,L} = \frac{\sqrt{D} + 2\sqrt{L}}{\sqrt{N-4}} \leq 1.$$

We define the function $F_t : \mathbb{R}^{N+D} \to \mathbb{R}$ by

$$F_t(x_1, \ldots, x_{D+N}) = \sqrt{x_1^2 + \ldots + x_D^2} - t\sqrt{x_{D+1}^2 + \ldots + x_{D+N}^2}\,.$$

1. Prove that $F_{t_{D,N,L}}$ is $\sqrt{2}$-Lipschitz. By considering the variable $F_{t_{D,N,L}}(\varepsilon)$ with ε a standard Gaussian random variable in R^{N+D}, prove with the Gaussian concentration Inequality (B.2) and the Inequalities (B.4) in Appendix B that there exists a standard exponential random variable ξ, such that $X_D - t_{D,N,L}X_N \leq 2\sqrt{\xi} - 2\sqrt{L}$, and thus

$$X_D^2 - 2t_{D,N,L}^2 X_N^2 \leq 8\left(\sqrt{\xi} - \sqrt{L}\right)_+^2.$$

2. Prove that for any $a, b \geq 0$ we have $(a-b)_+^2 \leq (a^2-b^2)_+$ and check the inequality

$$\mathbb{E}\left[\left(X_D^2 - 2t_{D,N,L}^2 X_N^2\right)_+\right] \leq 8\mathbb{E}\left[(\xi - L)_+\right] = 8e^{-L}.$$

3. From Definition (5.4) of $\text{pen}_\pi(m)$ conclude that we have the upper Bound (5.26) when the Condition (5.25) is met.

5.6.4 Selecting the Number of Steps for the Forward Algorithm

We consider the linear regression setting $Y = \mathbf{X}\beta^* + \varepsilon = f^* + \varepsilon$ with $\beta^* \in \mathbb{R}^p$ and the collection of models $\{S_m : m \in \mathcal{M}\}$ defined in Section 2.2, page 32, Chapter 2 for the coordinate-sparse setting. For $\lambda = 1, 2, \ldots$, we denote by $\widehat{m}_\lambda = \{j_1, \ldots, j_\lambda\}$ the forward selection algorithm defined recursively by

$$j_\lambda \in \underset{j=1,\ldots,p}{\text{argmin}} \|Y - \text{Proj}_{S_{\{j_1,\ldots,j_{\lambda-1},j\}}} Y\|^2.$$

For $m \subset \{1, \ldots, p\}$, we set $\pi_m = p^{-d_m}$ with $d_m = \dim(S_m)$. In the following, we restrict to models fulfilling

$$\left(\sqrt{2} + 2\sqrt{\log(8p)}\right)^2 d_m \leq n - d_m - 5. \tag{5.27}$$

We define

$$\widehat{\mathcal{M}} = \left\{\widehat{m}_\lambda : \lambda = 1, 2, \ldots \text{ such that } d_{\widehat{m}_\lambda} \text{ fulfills } (5.27)\right\},$$

$\widehat{\Lambda} = \left\{\lambda = 1, 2, \ldots : \widehat{m}_\lambda \in \widehat{\mathcal{M}}\right\}$, and $\bar{S}_m = S_m + < f^* >$, where $< f^* >$ is the line spanned by f^*.

For $\lambda = 1, 2, \ldots$, let us define $\widehat{f}_\lambda = \text{Proj}_{S_{\widehat{m}_\lambda}} Y$. The goal of this exercise is to prove that there exists a numerical constant $C > 1$, such that when $\widehat{f}_{\widehat{\lambda}}$ is a minimizer of Criterion (5.3) over $\left\{\widehat{f}_\lambda : \lambda = 1, 2, \ldots\right\}$, we have

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{\lambda}} - f^*\|^2\right] \leq C\mathbb{E}\left[\inf_{\lambda \in \widehat{\Lambda}}\left\{\|\widehat{f}_\lambda - f^*\|^2 + d_{\widehat{m}_\lambda}\log(p)\sigma^2 + \sigma^2\right\}\right]. \tag{5.28}$$

1. Prove that the set $\widehat{m}_{\widehat{\lambda}}$ is a solution of

$$\widehat{m}_{\widehat{\lambda}} \in \underset{m \in \widehat{\mathcal{M}}}{\text{argmin}}\left\{\|Y - \text{Proj}_{S_m} Y\|^2 + \text{pen}_\pi(m)\widehat{\sigma}_m^2\right\}.$$

2. Following lines similar to the beginning of the proof of Theorem 2.2, prove that for any $\lambda \in \widehat{\Lambda}$

$$\begin{aligned}\|\widehat{f}_{\widehat{\lambda}} - f^*\|^2 \leq & \|\widehat{f}_\lambda - f^*\|^2 + 2\text{pen}_\pi(\widehat{m}_\lambda)\widehat{\sigma}_{\widehat{m}_\lambda}^2 + (1.1)^{-1}\|\widehat{f}_\lambda - f^*\|^2 + Z(\lambda) \\ & + (1.1)^{-1}\|\widehat{f}_{\widehat{\lambda}} - f^*\|^2 + Z(\widehat{\lambda}),\end{aligned}$$

where $\quad Z(\lambda) = 1.1\|\text{Proj}_{\bar{S}_{\widehat{m}_\lambda}} \varepsilon\|^2 - \text{pen}_\pi(\widehat{m}_\lambda)\widehat{\sigma}_{\widehat{m}_\lambda}^2.$

3. Check that $\|Y - \mathrm{Proj}_{\bar{S}_m} Y\|^2$ is stochastically larger than $\|\varepsilon - \mathrm{Proj}_{\bar{S}_m}\varepsilon\|^2$, and then

$$\begin{aligned}\mathbb{E}\left[\sup_{\lambda=1,2,\ldots} Z(\lambda)\right] &\leq \sum_{m\in\mathcal{M}} \mathbb{E}\left[\left(1.1\|\mathrm{Proj}_{\bar{S}_m}\varepsilon\|^2 - \mathrm{pen}_\pi(m)\widehat{\sigma}_m^2\right)_+\right]\\ &\leq \sum_{m\in\mathcal{M}} \mathbb{E}\left[\left(1.1\|\mathrm{Proj}_{\bar{S}_m}\varepsilon\|^2 - \frac{\mathrm{pen}_\pi(m)}{n-d_m}\|\varepsilon - \mathrm{Proj}_{\bar{S}_m}\varepsilon\|^2\right)_+\right]\\ &\leq (1+p^{-1})^p\sigma^2 \leq e\sigma^2.\end{aligned}$$

4. Prove that when Condition (5.27) is fulfilled, we have $n - d_m \geq 24$ and (5.25) is fulfilled. Conclude that according to (5.26), we then have $\mathrm{pen}_\pi(m) \leq 2.2(n-d_m)$ and

$$\mathrm{pen}_\pi(m) \leq 3\left(\sqrt{2} + 2\sqrt{\log(8p)}\right)^2 d_m.$$

5. Prove that

$$\begin{aligned}\widehat{\sigma}_m^2 &\leq \frac{2}{n-d_m}\left(\|\varepsilon - \mathrm{Proj}_{S_m}\varepsilon\|^2 + \|f^* - \mathrm{Proj}_{S_m} f^*\|^2\right)\\ &\leq \frac{2}{n-d_m}\left(2n\sigma^2 + \left(\|\varepsilon\|^2 - 2n\sigma^2\right)_+ + \|f^* - \mathrm{Proj}_{S_m} Y\|^2\right).\end{aligned}$$

6. By combining the two last questions, prove that when $\lambda \in \widehat{\Lambda}$, we have

$$\begin{aligned}&\mathrm{pen}_\pi(\widehat{m}_\lambda)\widehat{\sigma}^2_{\widehat{m}_\lambda}\\ &\quad\leq 24\left(\sqrt{2} + 2\sqrt{\log(8p)}\right)^2 d_{\widehat{m}_\lambda}\sigma^2 + 5\left(\|\varepsilon\|^2 - 2n\sigma^2\right)_+ + 5\|f^* - \widehat{f}_\lambda\|^2.\end{aligned}$$

7. Conclude the proof of (5.28) by combining questions 2, 3, and 6.

Chapter 6

Multivariate Regression

In the previous chapters, we have focused on a response y that was 1-dimensional. In many cases, we do not focus on a single quantity $y \in \mathbb{R}$, but rather on a T-dimensional vector $y = (y_1, \dots, y_T) \in \mathbb{R}^T$ of measurements. It is, of course, possible to analyze each coordinate y_t independently, but it is usually wise to analyze simultaneously the T coordinates $y_1, \dots, y_T$. Actually, when $y_1, \dots, y_T$ are the outcome of a common process, they usually share some common structures, and handling the T measurements $y_1, \dots, y_T$ together enables to rely on these structures.

In this chapter, we give a special focus on the case where the measurements lie in the vicinity of an (unknown) low-dimensional space. In a linear model, this kind of structure translates in terms of low-rank of the regression matrix. We present in the next sections a theory for estimating low-rank matrices and then we investigate how we can handle simultaneously low-rank structures with some other sparsity structures.

6.1 Statistical Setting

We now consider the problem of predicting a T-dimensional *vector* y from a p-dimensional vector of covariates. Similarly to the examples discussed in Chapter 2, the linear model is the canonical model and many different situations can be recast as a linear model. Henceforth, we consider the following model

$$y^{(i)} = (A^*)^T x^{(i)} + \varepsilon^{(i)}, \quad i = 1, \dots, n, \tag{6.1}$$

where $y^{(i)} \in \mathbb{R}^T$, $x^{(i)} \in \mathbb{R}^p$, A^* is a $p \times T$-matrix, and the $\varepsilon^{(i)}$ are i.i.d. with $\mathcal{N}(0, \sigma^2 I_T)$ Gaussian distribution in $\mathbb{R}^T$. Writing Y and E for the $n \times T$ matrices $Y = [y_t^{(i)}]_{i=1,\dots,n,\ t=1,\dots,T}$ and $E = [\varepsilon_t^{(i)}]_{i=1,\dots,n,\ t=1,\dots,T}$, Model (6.1) can be formulated in a matrix form

$$Y = \mathbf{X}A^* + E\,, \tag{6.2}$$

where $\mathbf{X}$ is defined as in the previous chapters by $\mathbf{X}_{ij} = x_j^{(i)}$ for $i = 1, \dots, n$ and $j = 1, \dots, p$. Let M_k denote the k-th column of a matrix M. If we consider each column of Y independently, then we have T independent linear regressions

$$Y_k = \mathbf{X}A_k^* + E_k\,, \quad k = 1, \dots, T.$$

We can consider each regression independently, yet, as mentioned above, these T regressions may share some common structures, and it is wise in this case to analyze the T regressions simultaneously. Assume, for example, that the vectors A_k^* are coordinate sparse. In many cases, the vectors $A_1^*, \ldots, A_T^*$ share the same sparsity pattern, and then it helps to analyze simultaneously the T regression; see Section 6.4.1. The T vectors $A_1^*, \ldots, A_T^*$ may also (approximately) lie in a common (unknown) low-dimensional space. It then means that the rank of A^* is small. We understand in Section 6.3 how we can capitalize on this property. Finally, we investigate in Section 6.4.2 how we can handle simultaneously low-rank properties with coordinate-sparsity.

6.2 A Reminder on Singular Values

Singular values play a central role in low-rank estimation. This section is a brief reminder on singular values; we refer to the Appendix C for proofs and additional results.

Singular value decomposition

Any $n \times p$ matrix A of rank r can be decomposed as $A = \sum_{j=1}^{r} \sigma_j u_j v_j^T$, where

- $\sigma_1 \geq \ldots \geq \sigma_r > 0$,
- $\left\{\sigma_1^2, \ldots, \sigma_r^2\right\}$ are the nonzero eigenvalues of $A^T A$ and AA^T, and
- $\{u_1, \ldots, u_r\}$ and $\{v_1, \ldots, v_r\}$ are two orthonormal families of $\mathbb{R}^n$ and $\mathbb{R}^p$, such that

$$AA^T u_j = \sigma_j^2 u_j \quad \text{and} \quad A^T A v_j = \sigma_j^2 v_j, \quad \text{for } j = 1, \ldots, r.$$

We refer to Theorem C.1, page 229, in Appendix C for a proof. The values $\sigma_1, \sigma_2, \ldots$ are called the singular values of A.

Some matrix norms

In the following, we denote by $\sigma_1(A) \geq \sigma_2(A) \geq \ldots$ the singular values of a matrix A ranked in decreasing order. For $k > \operatorname{rank}(A)$, we define $\sigma_k(A)$ by $\sigma_k(A) = 0$. The Frobenius (or Hilbert–Schmidt) norm of A is defined by

$$\|A\|_F^2 = \sum_{i,j} A_{ij}^2 = \operatorname{trace}(A^T A) = \sum_{k=1}^{\operatorname{rank}(A)} \sigma_k(A)^2.$$

For any integer $q \geq 1$ the Ky–Fan $(2,q)$-norm is defined by

$$\|A\|_{(2,q)}^2 = \sum_{k=1}^{q} \sigma_k(A)^2.$$

For $q = 0$ we set $\|A\|_{(2,0)} = 0$. We observe that $\|A\|_{(2,q)} \leq \|A\|_F$, so $A \to \|A\|_{(2,q)}$ is 1-Lipschitz with respect to the Frobenius norm. For $q = 1$ the Ky–Fan $(2,1)$-norm corresponds to the operator norm

$$\|A\|_{(2,1)} = \sigma_1(A) = |A|_{\text{op}} = \sup_{x:\|x\| \leq 1} \|Ax\|;$$

see Appendix C. In the next sections, we repeatedly use the two following properties (see Theorem C.5, page 233, in Appendix C for a proof).

1. For any matrices $A, B \in \mathbb{R}^{n \times p}$, we have
$$\langle A, B \rangle_F \leq \|A\|_{(2,r)} \, \|B\|_{(2,r)}, \tag{6.3}$$
where $r = \min(\text{rank}(A), \text{rank}(B))$.
2. For any $q \geq 1$, we have
$$\min_{B:\text{rank}(B) \leq q} \|A - B\|_F^2 = \sum_{k=q+1}^{r} \sigma_k(A)^2 \ \text{ for } q < r = \text{rank}(A) \tag{6.4}$$
and $\min_{B:\text{rank}(B)\leq q} \|A - B\|_F^2 = 0$ for $q \geq r$. The minimum is achieved for
$$(A)_{(q)} = \sum_{k=1}^{q \wedge r} \sigma_k(A) u_k v_k^T. \tag{6.5}$$

The matrix $(A)_{(q)}$ is then a "projection" of the matrix A on the set of matrices with rank not larger than q.

6.3 Low-Rank Estimation

6.3.1 If We Knew the Rank of A^*

In Model (6.2), with $\{E_{it} : i = 1, \ldots, n, \ t = 1, \ldots, T\}$ i.i.d. with $\mathcal{N}(0, \sigma^2)$ Gaussian distribution, the negative log-likelihood of a matrix A is
$$-\text{log-likelihood}(A) = \frac{1}{2\sigma^2} \|Y - \mathbf{X}A\|_F^2 + \frac{T}{2} \log(2\pi\sigma^2),$$
where $\|\cdot\|_F$ is the Frobenius (or Hilbert–Schmidt) norm. If we knew the rank r^* of A^*, then we would estimate A^* by the maximum-likelihood estimator $\widehat{A}_{r^*}$ constrained to have a rank at most r^*, namely
$$\widehat{A}_r \in \underset{\text{rank}(A) \leq r}{\text{argmin}} \|Y - \mathbf{X}A\|_F^2, \tag{6.6}$$
with $r = r^*$. This estimator can be computed easily, as explained below.

The next lemma provides a useful formula for $\mathbf{X}\hat{A}_r$ in terms of the singular value decomposition of PY, where P is the orthogonal projector onto the range of $\mathbf{X}$. We refer to Appendix C for a reminder on the Moore–Penrose pseudo-inverse of a matrix.

Lemma 6.1 Computation of $\mathbf{X}\widehat{A}_r$

Write $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T$ *for the projection onto the range of* $\mathbf{X}$*, with* $(\mathbf{X}^T\mathbf{X})^+$ *the Moore–Penrose pseudo-inverse of* $\mathbf{X}^T\mathbf{X}$*.*

Then, for any $r \geq 1$*, we have* $\mathbf{X}\widehat{A}_r = (PY)_{(r)}$*.*

As a consequence, denoting by $PY = \sum_{k=1}^{\mathrm{rank}(PY)} \sigma_k u_k v_k^T$ *a singular value decomposition of* PY*, we have for any* $r \geq 1$

$$\mathbf{X}\widehat{A}_r = \sum_{k=1}^{r\wedge\mathrm{rank}(PY)} \sigma_k u_k v_k^T .$$

Proof. Pythagore's formula gives $\|Y - \mathbf{X}M\|_F^2 = \|Y - PY\|_F^2 + \|PY - \mathbf{X}M\|_F^2$ for any $p \times T$-matrix M. Since the rank of $\mathbf{X}\widehat{A}_r$ is at most r, we have $\|PY - \mathbf{X}\widehat{A}_r\|_F^2 \geq \|PY - (PY)_{(r)}\|_F^2$, and hence

$$\begin{aligned} \|Y - (PY)_{(r)}\|_F^2 &= \|Y - PY\|_F^2 + \|PY - (PY)_{(r)}\|_F^2 \\ &\leq \|Y - PY\|_F^2 + \|PY - \mathbf{X}\widehat{A}_r\|_F^2 = \|Y - \mathbf{X}\widehat{A}_r\|_F^2 . \end{aligned} \tag{6.7}$$

To conclude the proof of the lemma, we only need to check that we have a decomposition $(PY)_{(r)} = \mathbf{X}\tilde{A}_r$ with $\mathrm{rank}(\tilde{A}_r) \leq r$. From (6.5), we have $(PY)_{(r)} = \sum_{k=1}^{r'} \sigma_k u_k v_k^T$, with $r' = r \wedge \mathrm{rank}(PY)$. Since $u_1, \ldots, u_{r'} \in \mathrm{range}(PY) \subset \mathrm{range}(P)$, we have $Pu_k = u_k$ for $k = 1, \ldots, r'$ and $P(PY)_{(r)} = (PY)_{(r)}$. In particular, we have $(PY)_{(r)} = \mathbf{X}\tilde{A}_r$, with $\tilde{A}_r = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T(PY)_{(r)}$. Since $\mathrm{rank}(\tilde{A}_r) \leq \mathrm{rank}((PY)_{(r)}) \leq r$, according to (6.7), the matrix $\tilde{A}_r$ is a minimizer of (6.6) and finally $\mathbf{X}\widehat{A}_r = \mathbf{X}\tilde{A}_r = (PY)_{(r)}$. □

According to Lemma 6.1, for $r \geq \mathrm{rank}(PY)$, we have $\mathbf{X}\widehat{A}_r = \mathbf{X}\widehat{A}_{\mathrm{rank}(PY)}$. Since $\mathrm{rank}(PY) \leq q \wedge T$, with $q = \mathrm{rank}(\mathbf{X}) \leq n \wedge p$, we only need to consider the collection of estimators $\{\widehat{A}_1, \ldots, \widehat{A}_{q\wedge T}\}$. As a nice consequence of Lemma 6.1, the collection of estimators $\{\mathbf{X}\widehat{A}_1, \ldots, \mathbf{X}\widehat{A}_{q\wedge T}\}$ can be computed from a single singular value decomposition of PY. We also observe from the above proof that the matrix $\widehat{A}_r = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T(PY)_{(r)}$ is a solution of the minimization problem (6.6). Let us now investigate the quadratic risk $\mathbb{E}\left[\|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F^2\right]$ of the estimator $\mathbf{X}\widehat{A}_r$.

We remind the reader that $\sigma_k(M)$ denotes the k-th largest singular value of a matrix M, with $\sigma_k(M) = 0$ for $k > \mathrm{rank}(M)$.

Proposition 6.2 Deterministic bound

We set $q = \mathrm{rank}(\mathbf{X})$*. For any* $r \geq 1$ *and* $\theta > 0$*, we have*

$$\|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F^2 \leq c^2(\theta) \sum_{k>r} \sigma_k(\mathbf{X}A^*)^2 + 2c(\theta)(1+\theta)r|PE|_{\mathrm{op}}^2,$$

with $c(\theta) = 1 + 2/\theta$*.*

Proof. Write $\sum_k \sigma_k(\mathbf{X}A^*)u_k v_k^T$ for a singular value decomposition of $\mathbf{X}A^*$. Following the same lines as in the proof of the previous lemma, we can choose a matrix B of rank at most r, such that

$$\mathbf{X}B = (\mathbf{X}A^*)_{(r)} = \sum_{k=1}^{r} \sigma_k(\mathbf{X}A^*)u_k v_k^T .$$

From the definition of $\widehat{A}_r$, we have that $\|Y - \mathbf{X}\widehat{A}_r\|_F^2 \leq \|Y - \mathbf{X}B\|_F^2$, from which follows that

$$\|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F^2 \leq \|\mathbf{X}B - \mathbf{X}A^*\|_F^2 + 2\langle E, \mathbf{X}\widehat{A}_r - \mathbf{X}B\rangle_F . \tag{6.8}$$

Since $\mathbf{X}B$ and $\mathbf{X}\widehat{A}_r$ have a rank at most r, the matrix $\mathbf{X}\widehat{A}_r - \mathbf{X}B$ has a rank at most $2r$, and according to (6.3), we obtain

$$\langle E, \mathbf{X}\widehat{A}_r - \mathbf{X}B\rangle_F = \langle PE, \mathbf{X}\widehat{A}_r - \mathbf{X}B\rangle_F \leq \|PE\|_{(2,2r)} \|\mathbf{X}\widehat{A}_r - \mathbf{X}B\|_F . \tag{6.9}$$

We have $\|PE\|_{(2,2r)}^2 \leq 2r|PE|_{\text{op}}^2$ and $\|\mathbf{X}\widehat{A}_r - \mathbf{X}B\|_F \leq \|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F + \|\mathbf{X}B - \mathbf{X}A^*\|_F$, so using twice the inequality $2xy \leq ax^2 + y^2/a$ for all $a > 0$, $x, y \geq 0$, we obtain by combining (6.8) and (6.9)

$$(1 - 1/a)\|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F^2 \leq (1 + 1/b)\|\mathbf{X}B - \mathbf{X}A^*\|_F^2 + 2(a+b)r|PE|_{\text{op}}^2 .$$

The proposition follows by choosing $a = 1 + \theta/2$ and $b = \theta/2$. □

In order to obtain an upper bound not depending on the noise matrix E, we need a probabilistic upper bound on $|PE|_{\text{op}}$. Investigating the singular spectrum of random matrices is a very active field in mathematics, with applications in statistical physics, data compression, communication networks, statistics, etc. In particular, the operator norm $|PE|_{\text{op}}^2$ is known to be almost surely equivalent to $(\sqrt{q} + \sqrt{T})^2\sigma^2$, as q and T goes to infinity (we remind the reader that q is the rank of $\mathbf{X}$). We also have a non asymptotic result, which is a nice application of Slepian's lemma [112].

Lemma 6.3 Spectrum of random matrix
The expectated value of the operator norm of PE is upper bounded by

$$\mathbb{E}\left[|PE|_{\text{op}}\right] \leq \left(\sqrt{q} + \sqrt{T}\right)\sigma, \quad \textit{with } q = \text{rank}(\mathbf{X}). \tag{6.10}$$

We refer to Davidson and Szarek [49] for a proof of this result. The map $E \longrightarrow |PE|_{\text{op}}$ is 1-Lipschitz with respect to the Frobenius norm, since $|PE|_{\text{op}} \leq \|PE\|_F \leq \|E\|_F$, so according to the Gaussian concentration Inequality (B.2), page 221, there exists an exponential random variable ξ with parameter 1, such that $|PE|_{\text{op}} \leq \mathbb{E}\left[|PE|_{\text{op}}\right] + \sigma\sqrt{2\xi}$. Since $(a+b)^2 \leq 2(a^2 + b^2)$, we have

$$\mathbb{E}\left[|PE|_{\text{op}}^2\right] \leq 2\mathbb{E}\left[|PE|_{\text{op}}\right]^2 + 4\sigma^2\mathbb{E}[\xi] \leq 3\left(\sqrt{q} + \sqrt{T}\right)^2\sigma^2 . \tag{6.11}$$

According to (6.4), we have

$$\min_{\mathrm{rank}(A)\leq r} \|\mathbf{X}A-\mathbf{X}A^*\|_F^2 \geq \min_{\mathrm{rank}(\mathbf{X}A)\leq r} \|\mathbf{X}A-\mathbf{X}A^*\|_F^2 = \sum_{k>r} \sigma_k(\mathbf{X}A^*)^2,$$

so combining Proposition 6.2 (with $\theta = 1$) with this inequality, and (6.11) gives the following risk bound.

Corollary 6.4 Risk of $\widehat{A}_r$
For any $r \geq 1$, we have the risk bound

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}\widehat{A}_r-\mathbf{X}A^*\|_F^2\right] &\leq 9\sum_{k>r}\sigma_k(\mathbf{X}A^*)^2+36r\left(\sqrt{q}+\sqrt{T}\right)^2\sigma^2,\\ &\leq 36\min_{\mathrm{rank}(A)\leq r}\left\{\|\mathbf{X}A-\mathbf{X}A^*\|_F^2+r\left(\sqrt{q}+\sqrt{T}\right)^2\sigma^2\right\},\end{aligned} \tag{6.12}$$

with $q = \mathrm{rank}(\mathbf{X})$.

Let us comment on this bound. We notice that if $r = r^* = \mathrm{rank}(A^*)$, Bound (6.12) gives

$$\mathbb{E}\left[\|\mathbf{X}\widehat{A}_{r^*}-\mathbf{X}A^*\|_F^2\right] \leq 36r^*\left(\sqrt{q}+\sqrt{T}\right)^2\sigma^2.$$

Conversely, it can be shown (see [65] at the end of Section 3) that there exists a constant $C(\mathbf{X}) > 0$ only depending on the ratio $\sigma_1(\mathbf{X})/\sigma_q(\mathbf{X})$, such that for any $r^* \leq q \wedge T$ with $q = \mathrm{rank}(\mathbf{X})$, the minimax lower bound holds

$$\inf_{\widehat{A}} \sup_{A^*:\,\mathrm{rank}(A^*)=r^*} \mathbb{E}\left[\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2\right] \geq C(\mathbf{X})r^*\left(\sqrt{q}+\sqrt{T}\right)^2\sigma^2,$$

where the infimum is over all the estimators $\widehat{A}$. This means that, up to a constant factor, the estimator $\widehat{A}_r$ estimates matrices of rank r at the minimax rate.

Furthermore, according to (6.4), when A^* has a rank larger than r, any estimator $\widehat{A}$ of rank at most r fulfills

$$\|\mathbf{X}A^*-\mathbf{X}\widehat{A}\|_F^2 \geq \|\mathbf{X}A^*-(\mathbf{X}A^*)_{(r)}\|_F^2 = \sum_{k>r}\sigma_k(\mathbf{X}A^*)^2.$$

Therefore, the term $\sum_{k>r}\sigma_k(\mathbf{X}A^*)^2$ is the minimal bias that an estimator of rank r can have.

6.3.2 When the Rank of A^* Is Unknown

When the rank r^* of A^* is unknown, we would like to use the "oracle" estimator $\widehat{A}_{r_o}$, which achieves the best trade-off in (6.12) between the bias term $\sum_{k>r}\sigma_k(\mathbf{X}A^*)^2$

and the variance term $r\left(\sqrt{q}+\sqrt{T}\right)^2\sigma^2$. Of course, the oracle rank r_o is unknown since it depends on the unknown matrix A^*. Similarly to Chapter 2, we select a rank $\widehat{r}$ according to a penalized criterion and show that the risk of $\widehat{A}_{\widehat{r}}$ is almost as small as the risk of $\widehat{A}_{r_o}$.

For a constant $K > 1$, we select $\widehat{r}$ by minimizing the criterion

$$\widehat{r} \in \operatorname*{argmin}_{r=1,\ldots,q\wedge T}\left\{\|Y-\mathbf{X}\widehat{A}_r\|^2+\sigma^2\text{pen}(r)\right\},\quad \text{with pen}(r)=Kr\left(\sqrt{T}+\sqrt{q}\right)^2. \tag{6.13}$$

According to (6.6), the estimator $\widehat{A}=\widehat{A}_{\widehat{r}}$ is then solution of

$$\widehat{A} \in \operatorname*{argmin}_{A\in\mathbb{R}^{p\times T}}\left\{\|Y-\mathbf{X}A\|^2+K\text{rank}(A)\left(\sqrt{T}+\sqrt{q}\right)^2\sigma^2\right\}, \tag{6.14}$$

since the rank of the solution to (6.14) is not larger than $\text{rank}(PY)\leq q\wedge T$. Let us analyze the risk of $\mathbf{X}\widehat{A}$ with $\widehat{A}=\widehat{A}_{\widehat{r}}$.

Theorem 6.5 Oracle risk bound
For any $K>1$, there exists a constant $C_K>1$ depending only on K, such that the estimator $\widehat{A}$ defined by (6.14) fulfills the risk bound (with $q=\text{rank}(\mathbf{X})$)

$$\mathbb{E}\left[\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2\right]\leq C_K\min_{r=1,\ldots,q\wedge T}\left\{\mathbb{E}\left[\|\mathbf{X}\widehat{A}_r-\mathbf{X}A^*\|_F^2\right]+r(\sqrt{T}+\sqrt{q})^2\sigma^2\right\}. \tag{6.15}$$

If we compare Bounds (6.12) and (6.15), we observe that the risk of the estimator $\mathbf{X}\widehat{A}$ is almost as good as the risk of the best of the estimators $\{\mathbf{X}\widehat{A}_r,\ r=1,\ldots,q\wedge T\}$. Since the estimator $\widehat{A}$ can be computed with a single singular value decomposition, we can adapt in practice to the rank of the matrix A^*. Combining (6.12) and (6.15), we obtain the upper bound for the risk of $\mathbf{X}\widehat{A}$

$$\mathbb{E}\left[\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2\right]\leq C'_K\min_{A\in\mathbb{R}^{p\times T}}\left\{\|\mathbf{X}A-\mathbf{X}A^*\|_F^2+\text{rank}(A)(T+q)\sigma^2\right\}, \tag{6.16}$$

with $C'_K>1$ depending only on $K>1$ (again, the minimum in (6.16) is achieved for A with rank not larger than $q\wedge T$).

Proof of Theorem 6.5.
1- Deterministic bound

Let us fix $r\in\{1,\ldots,q\wedge T\}$. From the definition of $\widehat{r}$, we have that $\|Y-\mathbf{X}\widehat{A}\|^2+\sigma^2\text{pen}(\widehat{r})\leq\|Y-\mathbf{X}\widehat{A}_r\|^2+\sigma^2\text{pen}(r)$, from which follows that

$$\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2\leq\|\mathbf{X}\widehat{A}_r-\mathbf{X}A^*\|_F^2+2\langle PE,\mathbf{X}\widehat{A}-\mathbf{X}\widehat{A}_r\rangle_F+\sigma^2\text{pen}(r)-\sigma^2\text{pen}(\widehat{r}). \tag{6.17}$$

Similarly as in the proof of Theorem 2.2 in Chapter 2, if we prove that for some $a > 1, b > 0$

$$2\langle PE, \mathbf{X}\widehat{A} - \mathbf{X}\widehat{A}_r \rangle_F - \sigma^2 \mathrm{pen}(\widehat{r}) \leq a^{-1} \|\mathbf{X}\widehat{A} - \mathbf{X}A^*\|_F^2 + b^{-1} \|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F^2 + Z_r, \tag{6.18}$$

with Z_r fulfilling $\mathbb{E}[Z_r] \leq c\left(\sqrt{q} + \sqrt{T}\right)^2 r\sigma^2$ for some constant $c > 0$, then we have from (6.17)

$$\frac{a-1}{a}\,\mathbb{E}\left[\|\mathbf{X}\widehat{A} - \mathbf{X}A^*\|_F^2\right] \leq \frac{b+1}{b}\,\mathbb{E}\left[\|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F^2\right] + (K+c)\left(\sqrt{q} + \sqrt{T}\right)^2 r\sigma^2,$$

and (6.15) follows.

Let us prove (6.18). As in the proof of Proposition 6.2, we bound the cross term in terms of the operator norm of PE. The rank of $\mathbf{X}\widehat{A} - \mathbf{X}\widehat{A}_r$ is at most $\widehat{r} + r$, so according to (6.3), we have for any $a, b > 0$

$$\begin{aligned}
&2\langle PE, \mathbf{X}\widehat{A} - \mathbf{X}\widehat{A}_r \rangle_F \\
&\leq 2\|PE\|_{(2,\widehat{r}+r)}\ \|\mathbf{X}\widehat{A} - \mathbf{X}\widehat{A}_r\|_F \\
&\leq 2|PE|_{\mathrm{op}}\sqrt{\widehat{r}+r}\left(\|\mathbf{X}\widehat{A} - \mathbf{X}A^*\|_F + \|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F\right) \\
&\leq (a+b)\,(\widehat{r}+r)\,|PE|_{\mathrm{op}}^2 + a^{-1}\|\mathbf{X}\widehat{A} - \mathbf{X}A^*\|_F^2 + b^{-1}\|\mathbf{X}\widehat{A}_r - \mathbf{X}A^*\|_F^2,
\end{aligned}$$

where the last line is obtained by applying twice the inequality $2xy \leq ax^2 + y^2/a$. We then obtain (6.18) with

$$Z_r = (a+b)\,(\widehat{r}+r)\,|PE|_{\mathrm{op}}^2 - \sigma^2 \mathrm{pen}(\widehat{r}).$$

It only remains to prove that $\mathbb{E}[Z_r] \leq c\left(\sqrt{q} + \sqrt{T}\right)^2 r\sigma^2$ for some $c > 0$.

2- Stochastic control

The map $E \to |PE|_{\mathrm{op}}$ is 1-Lipschitz with respect to the Frobenius norm, so according to the Gaussian concentration Inequality (B.2), page 221, and Bound (6.10), there exists an exponential random variable ξ with parameter 1, such that

$$|PE|_{\mathrm{op}} \leq \sigma\left(\sqrt{T} + \sqrt{q}\right) + \sigma\sqrt{2\xi}\,.$$

Since $\widehat{r} \leq q \wedge T$, taking $a = (3+K)/4$ and $b = (K-1)/4$ we obtain that

$$\begin{aligned}
&\widehat{r}\left((a+b)|PE|_{\mathrm{op}}^2 - K(\sqrt{T} + \sqrt{q})^2\sigma^2\right) \\
&\leq\ \frac{1+K}{2}\,(q \wedge T)\left(|PE|_{\mathrm{op}}^2 - \frac{2K}{1+K}(\sqrt{T} + \sqrt{q})^2\sigma^2\right)_+ \\
&\leq\ \frac{1+K}{2}\,(q \wedge T)\left(\left(\sqrt{T} + \sqrt{q} + \sqrt{2\xi}\right)^2 - \frac{2K}{1+K}(\sqrt{T} + \sqrt{q})^2\right)_+ \sigma^2.
\end{aligned}$$

From the inequality $(\sqrt{T}+\sqrt{q}+\sqrt{2\xi})^2 \leq (1+\alpha)(\sqrt{T}+\sqrt{q})^2+2\xi(1+1/\alpha)$ with $\alpha = (K-1)/(K+1)$, we obtain

$$\mathbb{E}\left[\widehat{r}\left((a+b)|PE|_{\mathrm{op}}^2 - K(\sqrt{T}+\sqrt{q})^2\sigma^2\right)\right] \leq 2\frac{K(1+K)}{K-1}(q\wedge T)\sigma^2\mathbb{E}[\xi] \, .$$

Since $q\wedge T \leq \left(\sqrt{T}+\sqrt{q}\right)^2/4$, combining this bound with (6.11), we get (6.18), with

$$\begin{aligned}\mathbb{E}[Z_r] &\leq 3\frac{K+1}{2}\left(\sqrt{q}+\sqrt{T}\right)^2\sigma^2 r + 2\frac{K(1+K)}{K-1}(q\wedge T)\sigma^2 \\ &\leq \left(3\frac{K+1}{2}+\frac{K(1+K)}{2(K-1)}\right)\left(\sqrt{q}+\sqrt{T}\right)^2\sigma^2 r,\end{aligned}$$

and the proof of Theorem 6.5 is complete. □

6.4 Low Rank and Sparsity

As explained at the beginning of the chapter, the matrix A^* in (6.2) is likely to have (approximately) a low rank, but also to be sparse in some sense (coordinate sparse, group sparse, etc.). Therefore, we want to exploit simultaneously the low rank structures and the sparse structures to improve our estimation of A^*.

As a first step, we start with the case where the matrix is row sparse.

6.4.1 Row-Sparse matrices

A natural assumption is that the response $y^{(i)} = \sum_{j=1}^p A^*_{ji}x^{(i)}_j + \varepsilon^{(i)}$ depends from $x^{(i)}$ only through a subset $\{x^{(i)}_j : j \in J^*\}$ of its coordinates. This means that the rows of A^* are zero except those with index in J^*. Again, the difficulty comes from the fact that the set J^* is unknown.

Estimating a row-sparse matrix simply corresponds to a group-sparse regression as described in Chapter 2. Actually, the matrix structure plays no role in this setting, and we can recast the model in a vectorial form. We can stack the columns of A^* into a vector $\mathrm{vect}(A^*)$ of dimension pT and act similarly with Y and E. Then, we end with a simple group-sparse regression

$$\mathrm{vect}(Y) = \widetilde{\mathbf{X}}\,\mathrm{vect}(A^*) + \mathrm{vect}(E),$$

where

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{X} \end{bmatrix} \in \mathbb{R}^{nT\times pT},$$

and where the p groups $G_k = \{k+pl : l = 0,\ldots,T-1\}$ for $k = 1,\ldots,p$ gather the indices congruent modulo p. We can estimate $\mathrm{vect}(A^*)$ with the group–Lasso estimator

(see Chapter 4, Section 4.3.1, page 87)

$$\operatorname{vect}(\widehat{A}_\lambda) \in \operatorname*{argmin}_{\beta\in\mathbb{R}^{pT}} \left\{ \|\operatorname{vect}(Y) - \widetilde{\mathbf{X}}\beta\|^2 + \lambda \sum_{j=1}^{p} \|\beta_{G_j}\| \right\}.$$

Writing $A_{j:}$ for the j-th row of A, the above minimization problem is equivalent to

$$\widehat{A}_\lambda \in \operatorname*{argmin}_{A\in\mathbb{R}^{p\times T}} \left\{ \|Y - \mathbf{X}A\|_F^2 + \lambda \sum_{j=1}^{p} \|A_{j:}\| \right\}. \tag{6.19}$$

We assume in the following that the columns $\mathbf{X}_1,\ldots,\mathbf{X}_p$ of $\mathbf{X}$ have unit norm. We observe that in this case the operator norm $|\widetilde{\mathbf{X}}_{G_k}|_{\mathrm{op}}$ of

$$\widetilde{\mathbf{X}}_{G_k} = \begin{bmatrix} \mathbf{X}_k & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{X}_k \end{bmatrix} \in \mathbb{R}^{nT\times T}$$

is 1 for every $k = 1,\ldots,p$. We can then lift the risk Bound (4.20) from Chapter 4, page 89.

Theorem 6.6 Risk bound for row-sparse matrices

For $\lambda = 3\sigma\left(\sqrt{T} + 2\sqrt{\log(p)}\right)$*, we have with probability at least* $1 - 1/p$

$$\|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*\|_F^2 \leq \min_{A\in\mathbb{R}^{p\times T}} \left\{ \|\mathbf{X}A - \mathbf{X}A^*\|_F^2 + \frac{18\sigma^2}{\widetilde{\kappa}_G(A)} |J(A)|\,(T + 4\log(p)) \right\}, \tag{6.20}$$

where $J(A) = \left\{ j \in \{1,\ldots,p\} : A_{j:} \neq 0 \right\}$ *and*

$$\widetilde{\kappa}_G(A) = \min_{B\in\mathscr{C}_G(J(A))} \sqrt{\frac{|J(A)|\|\mathbf{X}B\|_F}{\sum_{j\in J(A)} \|B_{j:}\|}} \quad \textit{with } \mathscr{C}_G(J) = \left\{ B : \sum_{j\in J^c} \|B_{j:}\| < 5 \sum_{j\in J} \|B_{j:}\| \right\}.$$

We next investigate whether we can improve upon Theorem 6.5 and Theorem 6.6 by taking simultaneously into account low rank and row sparsity.

6.4.2 Criterion for Row-Sparse and Low-Rank Matrices

We now consider the case where the matrix A^* has a low rank and a small number of nonzero rows. The rank of a matrix is not larger than the number of its nonzero rows, so the estimator (6.19) already has a small rank. But here we have in mind a case where the rank of A^* is much smaller than the number of its nonzero rows, and we want to exploit this feature. The rank of A^* and the location of its nonzero rows are unknown. We first investigate how much we can gain by taking into account simultaneously row sparsity and low rank. In this direction, a model selection estimator will be our benchmark.

Let π be a probability distribution on $\mathscr{P}(\{1,\ldots,p\})$. For example, we can set

$$\pi_J = \left(C_p^{|J|}\right)^{-1} e^{-|J|}(e-1)/(e-e^{-p}) \quad \text{for all} \quad J \subset \{1\ldots,p\}, \tag{6.21}$$

with $C_p^k = p!/(k!(p-k)!)$. For $K>1$, we define $\widehat{A}$ as a minimizer of the criterion

$$\begin{aligned}&\text{Crit}(A) = \|Y-\mathbf{X}A\|_F^2 + \text{pen}(A)\sigma^2,\\ &\quad\text{with} \quad \text{pen}(A) = K\left(\sqrt{r(A)}\left(\sqrt{T}+\sqrt{|J(A)|}\right)+\sqrt{2\log\left(\pi_{J(A)}^{-1}\right)}\right)^2,\end{aligned} \tag{6.22}$$

where $r(A)$ is the rank of A and where $J(A) = \{j \in \{1,\ldots,p\} : A_{j:} \neq 0\}$ is the set of nonzero rows of A. Penalty (6.22) is very similar to Penalty (2.9), page 35, for model selection. Actually, the set $V_{J,r} = \{A \in \mathbb{R}^{p\times T} : J(A) \subset J,\ r(A) = r\}$ is a submanifold of dimension $r(T+J-r)$. Since

$$\sqrt{r}(\sqrt{T}+\sqrt{J}) \geq \sqrt{r(T+J-r)} = \sqrt{\dim(V_{J,r})},$$

Penalty (6.22) can be viewed as an upper bound of the penalty

$$\text{pen}'(A) = K\left(\sqrt{\dim(V_{J,r})}+\sqrt{2\log(\pi_J^{-1})}\right)^2, \quad \text{for } A \in V_{J,r},$$

which has the same form as Penalty (2.9).

The minimization Problem (6.22) has a computational complexity that is prohibitive in high-dimensional settings, since it requires to explore all the subsets $J \subset \{1\ldots,p\}$ in general. Yet, the resulting estimator provides a good benchmark in terms of statistical accuracy. We discuss in the following sections how we can relax (6.22) in order to obtain a convex criterion.

Theorem 6.7 Risk bound for row-sparse and low-rank matrices

For any $K>1$, there exists a constant $C_K>1$ depending only on K, such that the estimator $\widehat{A}$ defined by (6.22) fulfills the risk bound

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2\right]\\ &\quad \leq C_K \min_{A\neq 0}\left\{\|\mathbf{X}A-\mathbf{X}A^*\|_F^2 + r(A)(T+|J(A)|)\sigma^2 + \log\left(\pi_{J(A)}^{-1}\right)\sigma^2\right\}.\end{aligned} \tag{6.23}$$

If we choose π_J as in (6.21), we obtain for some constant $C_K'>1$ depending only on $K>1$

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2\right]\\ &\quad \leq C_K' \min_{A\neq 0}\left\{\|\mathbf{X}A-\mathbf{X}A^*\|_F^2 + r(A)(T+|J(A)|)\sigma^2 + |J(A)|\log\left(ep/|J(A)|\right)\sigma^2\right\}.\end{aligned}$$

We observe that the term $r(A)(T+|J(A)|)$ can be much smaller than $|J(A)|T$ appearing in (6.20) if $r(A)$ is small compared to $\min(|J(A)|,T)$. Similarly, if $|J(A)|$ is small compared to $\operatorname{rank}(\mathbf{X})$, then the above upper bound is small compared to the risk Bound (6.16) for the low-rank case. Thus, with an estimator that takes into account simultaneously row sparsity and low rank, we can get a significant improvement in estimation.

Proof of Theorem 6.7.
1- Deterministic bound
We fix a nonzero matrix $A \in \mathbb{R}^{p\times T}$. Starting from $\operatorname{Crit}(\widehat{A}) \le \operatorname{Crit}(A)$, we obtain

$$\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2 \le \|\mathbf{X}A-\mathbf{X}A^*\|_F^2 + 2\langle E, \mathbf{X}\widehat{A}-\mathbf{X}A\rangle_F + \operatorname{pen}(A)\sigma^2 - \operatorname{pen}(\widehat{A})\sigma^2, \quad (6.24)$$

with

$$\operatorname{pen}(A) \le 4K\left(r(A)(T+|J(A)|)+\log\left(\pi_{J(A)}^{-1}\right)\right).$$

As in the proofs of Theorem 2.2 and Theorem 6.5, if we prove that there exist some constants $c_1<1$, $c_2,c_3>0$ and a random variable $Z(A)$, such that

$$2\langle E, \mathbf{X}\widehat{A}-\mathbf{X}A\rangle_F - \operatorname{pen}(\widehat{A})\sigma^2 \le c_1\|\mathbf{X}\widehat{A}-\mathbf{X}A^*\|_F^2 + c_2\|\mathbf{X}A-\mathbf{X}A^*\|_F^2 + Z(A), \quad (6.25)$$

with $\mathbb{E}[Z(A)] \le c_3 r(A)T\sigma^2$, then (6.23) follows from (6.24) and (6.25).

Let us prove (6.25). For any $J \subset \{1,\ldots,p\}$, we write $\mathbf{X}_J$ for the matrix obtained from $\mathbf{X}$ by keeping only the columns in J, and we define S_J as the orthogonal of $\operatorname{range}(\mathbf{X}A)$ in $\operatorname{range}(\mathbf{X}A)+\operatorname{range}(\mathbf{X}_J)$, so that

$$\operatorname{range}(\mathbf{X}A)+\operatorname{range}(\mathbf{X}_J) = \operatorname{range}(\mathbf{X}A) \bigoplus S_J.$$

In particular, the linear span S_J has a dimension at most $|J|$. For notational simplicity, we will write in the following $\widehat{J}$ for $J(\widehat{A})$, $\widehat{r}$ for the rank of $\widehat{A}$, P_J for the orthogonal projector onto S_J and P_A for the orthogonal projector onto the range of $\mathbf{X}A$. Since $\mathbf{X}(\widehat{A}-A) = P_{\widehat{J}}\mathbf{X}\widehat{A} + P_A\mathbf{X}(\widehat{A}-A)$ with $\operatorname{rank}(P_{\widehat{J}}\mathbf{X}\widehat{A}) \le \widehat{r}$, from (6.3) we get

$$\begin{aligned}
&2\langle E, \mathbf{X}\widehat{A}-\mathbf{X}A\rangle_F\\
&= 2\langle P_{\widehat{J}}E, P_{\widehat{J}}\mathbf{X}\widehat{A}\rangle_F + 2\langle P_A E, P_A\mathbf{X}(\widehat{A}-A)\rangle_F\\
&\le 2\|P_{\widehat{J}}E\|_{(2,\widehat{r})}\|P_{\widehat{J}}\mathbf{X}\widehat{A}\|_F + 2\|P_A E\|_F\,\|P_A\mathbf{X}(\widehat{A}-A)\|_F\\
&\le \frac{K+1}{2}\left[\|P_{\widehat{J}}E\|_{(2,\widehat{r})}^2 + \|P_A E\|_F^2\right] + \frac{2}{K+1}\left[\|P_{\widehat{J}}\mathbf{X}\widehat{A}\|_F^2 + \|P_A\mathbf{X}(\widehat{A}-A)\|_F^2\right],
\end{aligned}$$

where we used twice in the last line the inequality $2xy \le ax^2 + a^{-1}y^2$ with $a=(K+1)/2$. According to Pythagore's formula, we have $\|P_{\widehat{J}}\mathbf{X}\widehat{A}\|_F^2 + \|P_A\mathbf{X}(\widehat{A}-A)\|_F^2 = \|\mathbf{X}(\widehat{A}-A)\|_F^2$. Since $\|\mathbf{X}(\widehat{A}-A)\|_F^2 \le (1+b)\|\mathbf{X}(\widehat{A}-A^*)\|_F^2 + (1+b^{-1})\|\mathbf{X}(A-A^*)\|_F^2$, taking $b=(K-1)/4$, we obtain (6.25) with $c_1 = a^{-1}(1+b) < 1$, $c_2 = a^{-1}(1+b^{-1})$ and

$$Z(A) = \frac{K+1}{2}\|P_A E\|_F^2 + \widehat{\Delta}, \quad \text{where } \widehat{\Delta} = \frac{K+1}{2}\|P_{\widehat{J}}E\|_{(2,\widehat{r})}^2 - \operatorname{pen}(\widehat{A})\sigma^2.$$

We observe that $\mathbb{E}\left[\|P_A E\|_F^2\right] = \mathrm{rank}\,(\mathbf{X}A)\,T\sigma^2 \leq r\,(A)\,T\sigma^2$, so to conclude the proof of (6.23), it only remains to check that

$$\mathbb{E}\left[\widehat{\Delta}\right] \leq \frac{2K(K+1)}{K-1}T\sigma^2. \tag{6.26}$$

2- Stochastic control

We know from Lemma 6.3 that for any $J \subset \{1,\ldots,p\}$ and $r \leq p \wedge T$,

$$\mathbb{E}\left[\|P_J E\|_{(2,r)}\right] \leq \sqrt{r}\,\mathbb{E}\left[\sigma_1(P_J E)\right] \leq \sigma\sqrt{r}\left(\sqrt{\dim(S_J)}+\sqrt{T}\right) \leq \sigma\sqrt{r}\left(\sqrt{|J|}+\sqrt{T}\right).$$

Since the map $E \to \|P_J E\|_{(2,r)}$ is 1-Lipschitz with respect to the Frobenius norm, the Gaussian concentration Inequality (B.2) ensures for each $J \subset \{1,\ldots,p\}$ and $r \in \{1,\ldots,p\wedge T\}$ the existence of a standard exponential random variable $\xi_{J,r}$, such that

$$\|P_J E\|_{(2,r)} \leq \sigma\left(\sqrt{r}\left(\sqrt{|J|}+\sqrt{T}\right)+\sqrt{2\xi_{J,r}}\right). \tag{6.27}$$

As in the proof of Theorem 2.2 in Chapter 2, we observe that for all $J \subset \{1,\ldots,p\}$ and all $r \in \{1,\ldots,p\wedge T\}$, we have

$$\begin{aligned}
&\left(\sqrt{r}\left(\sqrt{|J|}+\sqrt{T}\right)+\sqrt{2\xi_{J,r}}\right)^2 \\
&\leq \left(\sqrt{r}\left(\sqrt{|J|}+\sqrt{T}\right)+\sqrt{2\log(\pi_J^{-1})}+\sqrt{2\left(\xi_{J,r}-\log(\pi_J^{-1})\right)_+}\right)^2 \\
&\leq \frac{2K}{K+1}\left(\sqrt{r}\left(\sqrt{|J|}+\sqrt{T}\right)+\sqrt{2\log(\pi_J^{-1})}\right)^2+\frac{4K}{K-1}\left(\xi_{J,r}-\log(\pi_J^{-1})\right)_+.
\end{aligned} \tag{6.28}$$

Since $\widehat{r} \leq p \wedge T$, combining (6.27) with (6.28), we obtain

$$\begin{aligned}
\widehat{\Delta} &= \frac{K+1}{2}\left(\|P_{\widehat{J}}E\|^2_{(2,\widehat{r})}-\frac{2K}{K+1}\left(\sqrt{\widehat{r}}\left(\sqrt{|\widehat{J}|}+\sqrt{T}\right)+\sqrt{2\log(\pi_{\widehat{J}}^{-1})}\right)^2\sigma^2\right) \\
&\leq \frac{2K(K+1)}{K-1}\sigma^2\left(\xi_{\widehat{J},\widehat{r}}-\log(\pi_{\widehat{J}}^{-1})\right)_+\mathbf{1}_{\widehat{r}\geq 1}.
\end{aligned}$$

To conclude the proof of (6.26), we check that

$$\begin{aligned}
\mathbb{E}\left[\left(\xi_{\widehat{J},\widehat{r}}-\log(\pi_{\widehat{J}}^{-1})\right)_+\mathbf{1}_{\widehat{r}\geq 1}\right] &\leq \sum_{r=1}^{p\wedge T}\sum_{J\subset\{1,\ldots,p\}}\mathbb{E}\left[\left(\xi_{J,r}-\log(\pi_J^{-1})\right)_+\right] \\
&\leq T\sum_{J\subset\{1,\ldots,p\}}\pi_J \;=\; T.
\end{aligned}$$

The proof of Theorem 6.7 is complete. □

The above procedure satisfies a nice risk bound, but it is computationally untractable since we cannot explore all the subsets $J \subset \{1,\dots,p\}$. A natural idea to enforce row sparsity is to use a group penalty as in (6.20). Yet, if we a add a constraint on the rank in (6.20), there is no computationally efficient algorithm for solving exactly this problem. A possible direction for combining sparse and low-rank constraints is to convexify the constraint on the rank. We discuss this issue in the next section.

6.4.3 Convex Criterion for Low Rank Matrices

We emphasize that for the pure low-rank estimation of Section 6.3, there is no need to convexify Criterion (6.14), since it can be minimized efficiently from a single singular value decomposition. The convexification is only needed when we want to combine low-rank properties with some other structures as row sparsity. Yet, as a first step, we start by analyzing the convexification of (6.14).

The main idea underlying the introduction of the Lasso estimator is to replace the constraint on the number of nonzero coordinates of β by a constraint on the sum of their absolute values. Following the same idea, we can replace the constraint on the rank of A, which is the number of nonzero singular values of A, by a constraint on the nuclear norm of A, which is the sum of the singular values of A. This gives the following convex criterion

$$\widehat{A}_\lambda \in \operatorname*{argmin}_{A\in\mathbb{R}^{p\times T}} \left\{ \|Y - \mathbf{X}A\|_F^2 + \lambda |A|_* \right\}, \tag{6.29}$$

where λ is a positive tuning parameter and $|A|_* = \sum_k \sigma_k(A)$ is the nuclear norm of A. Similarly to the Lasso estimator, we can provide a risk bound for this estimator.

Theorem 6.8 Risk bound for the convex mutivariate criterion
Let $K > 1$ and set

$$\lambda = 2K\sigma_1(\mathbf{X})\left(\sqrt{T}+\sqrt{q}\right)\sigma, \quad \textit{with } q = \operatorname{rank}(X).$$

Then, with probability larger than $1 - e^{-(K-1)^2(T+q)/2}$, we have

$$\|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*\|_F^2 \le \inf_A \left\{ \|\mathbf{X}A - \mathbf{X}A^*\|_F^2 + 9K^2\,\frac{\sigma_1(\mathbf{X})^2}{\sigma_q(\mathbf{X})^2}\left(\sqrt{T}+\sqrt{q}\right)^2\sigma^2 \operatorname{rank}(A) \right\}.$$

This risk bound is similar to (6.16), except that there is a constant 1 in front of the bias term (which is good news) and a constant $\sigma_1(\mathbf{X})^2/\sigma_q(\mathbf{X})^2$ in front of the variance term (which is bad news). This last constant can be huge in practice, since data matrices $\mathbf{X}$ tend to have a smallest singular value $\sigma_q(\mathbf{X})$ very small. When this constant remains of a reasonable size, the estimator $\widehat{A}_\lambda$ has properties similar to those of $\widehat{A}$ defined by (6.14).

Proof. The proof is very similar to the proof of (4.13) for the Lasso estimator.

1- Deterministic bound

We first derive a deterministic bound on $\|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*\|_F^2$.

Lemma 6.9

For $\lambda \geq 2\sigma_1(\mathbf{X}^T E)$, we have

$$\|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*\|_F^2 \leq \inf_{A\in\mathbb{R}^{p\times T}} \left\{ \|\mathbf{X}A - \mathbf{X}A^*\|_F^2 + \frac{9\lambda^2}{4\sigma_q(\mathbf{X})^2}\operatorname{rank}(A) \right\}. \tag{6.30}$$

Proof of Lemma 6.9.

Let us introduce the set $\mathbb{A} := \left\{A \in \mathbb{R}^{p\times T} : A = P_{\mathbf{X}^T} A\right\}$, where $P_{\mathbf{X}^T}$ is the orthogonal projector onto the range of $\mathbf{X}^T$. Since we have the orthogonal decomposition $\mathbb{R}^p = \ker(\mathbf{X}) \bigoplus \operatorname{range}(\mathbf{X}^T)$, for all matrices $A \in \mathbb{R}^{p\times T}$, we have

$$\mathbf{X}P_{\mathbf{X}^T}A = \mathbf{X}A \quad \text{and} \quad \operatorname{rank}(P_{\mathbf{X}^T}A) \leq \operatorname{rank}(A).$$

In particular, in (6.30) the infimum over $A \in \mathbb{R}^{p\times T}$ coincides with the infimum over $A \in \mathbb{A}$, and we only need to prove (6.30) with the infimum over $\mathbb{A}$. Similarly, from Inequality (C.3), page 233, in Appendix C, we observe that $|P_{\mathbf{X}^T}\widehat{A}_\lambda|_* \leq |\widehat{A}_\lambda|_*$ with strict inequality if $P_{\mathbf{X}^T}\widehat{A}_\lambda \neq \widehat{A}_\lambda$. Therefore, the estimator $\widehat{A}_\lambda$ belongs to the space $\mathbb{A}$.

The optimality condition (D.3), page 237, in Appendix D for convex functions ensures the existence of a matrix $\widehat{Z} \in \partial|\widehat{A}_\lambda|_*$ such that $-2\mathbf{X}^T(Y - \mathbf{X}\widehat{A}_\lambda) + \lambda\widehat{Z} = 0$. Since $Y = \mathbf{X}A^* + E$, for any $A \in \mathbb{A}$, we have

$$2\langle \mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*, \mathbf{X}\widehat{A}_\lambda - \mathbf{X}A\rangle_F - 2\langle \mathbf{X}^T E, \widehat{A}_\lambda - A\rangle_F + \lambda\langle \widehat{Z}, \widehat{A}_\lambda - A\rangle_F = 0.$$

The subgradient monotonicity of convex functions (D.2), page 236 ensures that for all $Z \in \partial|A|_*$, we have $\langle \widehat{Z}, \widehat{A}_\lambda - A\rangle_F \geq \langle Z, \widehat{A}_\lambda - A\rangle_F$. As a consequence,

$$\begin{aligned}&\text{for all } A \in \mathbb{A} \text{ and for all } Z \in \partial|A|_*, \text{ we have}\\ &2\langle \mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*, \mathbf{X}\widehat{A}_\lambda - \mathbf{X}A\rangle_F \leq 2\langle \mathbf{X}^T E, \widehat{A}_\lambda - A\rangle_F - \lambda\langle Z, \widehat{A}_\lambda - A\rangle_F.\end{aligned} \tag{6.31}$$

Let us denote by $A = \sum_{k=1}^r \sigma_k u_k v_k^T$ the singular value decomposition of A, with $r = \operatorname{rank}(A)$. We write P_u (respectively, P_v) for the orthogonal projector onto $\operatorname{span}\{u_1, \ldots, u_r\}$ (respectively, onto $\operatorname{span}\{v_1, \ldots, v_r\}$). We also set $P_u^\perp = I - P_u$ and $P_v^\perp = I - P_v$. According to Lemma D.6, page 238, in Appendix D, the subdifferential of $|A|_*$ is given by

$$\partial|A|_* = \left\{ \sum_{k=1}^r u_k v_k^T + P_u^\perp W P_v^\perp \ : \ W \in \mathbb{R}^{p\times T} \text{ with } \sigma_1(W) \leq 1 \right\}.$$

Let us set $W = 2\mathbf{X}^T E/\lambda$. Since $\sigma_1(W) = 2\sigma_1(\mathbf{X}^T E)/\lambda \leq 1$, the matrix $Z = \sum_{k=1}^r u_k v_k^T + P_u^\perp W P_v^\perp$ belongs to $\partial|A|_*$. The decomposition

$$\widehat{A}_\lambda - A = P_u(\widehat{A}_\lambda - A) + P_u^\perp(\widehat{A}_\lambda - A)P_v + P_u^\perp(\widehat{A}_\lambda - A)P_v^\perp$$

gives

$$\begin{aligned}
&2\langle \mathbf{X}^T E, \widehat{A}_\lambda - A\rangle_F - \lambda\langle Z, \widehat{A}_\lambda - A\rangle_F \\
&= 2\langle \mathbf{X}^T E, \widehat{A}_\lambda - A\rangle_F - 2\langle P_u^\perp \mathbf{X}^T E P_v^\perp, \widehat{A}_\lambda - A\rangle_F - \lambda\langle \sum_{k=1}^r u_k v_k^T, \widehat{A}_\lambda - A\rangle_F \\
&= 2\langle \mathbf{X}^T E, \widehat{A}_\lambda - A\rangle_F - 2\langle \mathbf{X}^T E, P_u^\perp(\widehat{A}_\lambda - A) P_v^\perp\rangle_F - \lambda\langle \sum_{k=1}^r u_k v_k^T, \widehat{A}_\lambda - A\rangle_F \\
&= 2\langle \mathbf{X}^T E, P_u(\widehat{A}_\lambda - A)\rangle_F + 2\langle \mathbf{X}^T E, P_u^\perp(\widehat{A}_\lambda - A) P_v\rangle_F - \lambda\langle \sum_{k=1}^r u_k v_k^T, P_u(\widehat{A}_\lambda - A) P_v\rangle_F .
\end{aligned}$$

Lemma C.2, page 231, in Appendix C ensures that $\langle A, B\rangle_F \leq \sigma_1(A)|B|_*$, so

$$\begin{aligned}
&2\langle \mathbf{X}^T E, \widehat{A}_\lambda - A\rangle_F - \lambda\langle Z, \widehat{A}_\lambda - A\rangle_F \\
&\leq 2\sigma_1(\mathbf{X}^T E)\left(|P_u(\widehat{A}_\lambda - A)|_* + |P_u^\perp(\widehat{A}_\lambda - A)P_v|_*\right) + \lambda|P_u(\widehat{A}_\lambda - A)P_v|_* .
\end{aligned}$$

Since $2\sigma_1(\mathbf{X}^T E) \leq \lambda$, we obtain

$$\begin{aligned}
&2\langle \mathbf{X}^T E, \widehat{A}_\lambda - A\rangle_F - \lambda\langle Z, \widehat{A}_\lambda - A\rangle_F \\
&\leq \lambda|P_u(\widehat{A}_\lambda - A)|_* + \lambda|P_u^\perp(\widehat{A}_\lambda - A)P_v|_* + \lambda|P_u(\widehat{A}_\lambda - A)P_v|_* \\
&\leq \lambda\sqrt{\mathrm{rank}(A)}\left(\|P_u(\widehat{A}_\lambda - A)\|_F + \|P_u^\perp(\widehat{A}_\lambda - A)P_v\|_F + \|P_u(\widehat{A}_\lambda - A)P_v\|_F\right),
\end{aligned}$$

where we used in the last line $\mathrm{rank}(P_u) = \mathrm{rank}(P_v) = \mathrm{rank}(A)$ and the inequality $|M|_* \leq \sqrt{\mathrm{rank}(M)}\,\|M\|_F$ from Lemma C.2, page 231, in Appendix C. According to Inequalities (C.3) and (C.4), page 233, in Appendix C, the three above Frobenius norms are upper bounded by $\|\widehat{A}_\lambda - A\|_F$, so combining the above bound with (6.31) and Al-Kashi formula

$$2\langle \mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*, \mathbf{X}\widehat{A}_\lambda - \mathbf{X}A\rangle_F = \|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*\|_F^2 + \|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A\|_F^2 - \|\mathbf{X}A - \mathbf{X}A^*\|_F^2,$$

we obtain

$$\|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*\|_F^2 + \|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A\|_F^2 \leq \|\mathbf{X}A - \mathbf{X}A^*\|_F^2 + 3\lambda\sqrt{\mathrm{rank}(A)}\,\|\widehat{A}_\lambda - A\|_F .$$

Let us denote by $\mathbf{X}^+$ the Moore–Penrose pseudo-inverse of $\mathbf{X}$ (see Section C.2 in Appendix C). For any matrix $M \in \mathbb{A}$, we have $M = P_{\mathbf{X}^T}M = \mathbf{X}^+\mathbf{X}M$, and thus according again to Lemma C.2, page 231, we have

$$\|M\|_F \leq |\mathbf{X}^+|_{\mathrm{op}}\|\mathbf{X}M\|_F = \sigma_q(\mathbf{X})^{-1}\|\mathbf{X}M\|_F .$$

Since $A - \widehat{A}_\lambda$ belongs to $\mathbb{A}$, we have $\|\widehat{A}_\lambda - A\|_F \leq \sigma_q(\mathbf{X})^{-1}\|\mathbf{X}(\widehat{A}_\lambda - A)\|_F$, and therefore

$$\begin{aligned}
\|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A^*\|_F^2 &\leq \|\mathbf{X}A - \mathbf{X}A^*\|_F^2 - \|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A\|_F^2 + \frac{3\lambda\sqrt{\mathrm{rank}(A)}}{\sigma_q(\mathbf{X})}\|\mathbf{X}\widehat{A}_\lambda - \mathbf{X}A\|_F \\
&\leq \|\mathbf{X}A - \mathbf{X}A^*\|_F^2 + \frac{9\lambda^2}{4\sigma_q(\mathbf{X})^2}\,\mathrm{rank}(A),
\end{aligned}$$

where we used in the last line the inequality $2ab \leq a^2 + b^2$. The proof of Lemma 6.9 is complete. □

2- Stochastic control

To conclude the proof of Theorem 6.8, it remains to check that

$$\mathbb{P}\left(\sigma_1(\mathbf{X}^T E) \geq K\sigma_1(\mathbf{X})\left(\sqrt{T} + \sqrt{q}\right)\sigma\right) \leq e^{-(K-1)^2(T+q)/2}, \quad \text{for all } K > 1. \quad (6.32)$$

Writing $P_{\mathbf{X}}$ for the projection onto the range of $\mathbf{X}$, we have $\mathbf{X}^T E = \mathbf{X}^T P_{\mathbf{X}} E$, so $\sigma_1(\mathbf{X}^T E) \leq \sigma_1(\mathbf{X})\sigma_1(P_{\mathbf{X}} E)$. As in the proof of Theorem 6.5, combining the Gaussian concentration Inequality (B.2), page 221, and the Bound (6.10), we obtain that

$$\sigma_1(P_{\mathbf{X}} E) \leq \sigma\left(\sqrt{T} + \sqrt{q}\right) + \sigma\sqrt{2\xi},$$

for some exponential random variable ξ with parameter 1. Bound (6.32) follows, and the proof of Theorem 6.8 is complete. □

6.4.4 Convex Criterion for Sparse and Low-Rank Matrices

In order to combine the benefits of coordinate sparsity and low rankness, it is natural to penalize the negative log-likelihood by both the group-ℓ^1 and the nuclear norms

$$\widehat{A}_{\lambda,\mu} \in \operatorname*{argmin}_{A \in \mathbb{R}^{p \times T}} \left\{ \|Y - \mathbf{X}A\|_F^2 + \lambda |A|_* + \mu \sum_{j=1}^{p} \|A_{j:}\| \right\},$$

where $A_{j:}$ denotes the j-th row of A and $\lambda, \mu > 0$. The resulting criterion is convex in A.

It is not difficult to combine the analysis for the Lasso estimator and the analysis of Theorem 6.8 in order to get a risk bound for $\widehat{A}_{\lambda,\mu}$. Yet, to the best of our knowledge, the resulting risk bound does not improve on the results with the nuclear alone or the ℓ^1 norm alone. The issue of proving a risk bound similar to (6.23) for a convex criterion remains an open question.

6.5 Discussion and References

6.5.1 Take-Home Message

We can easily adapt to the rank of the regression matrix A^* by implementing the estimator (6.14). In principle, it is possible to improve significantly the estimator accuracy by taking simultaneously into account low-rank structures with coordinate-sparse structures, as explained in Section 6.4.2. Yet, the procedure (6.22) is computationally intractable. Whether we can achieve such an improvement with a convex criterion remains an open question. We refer to Bunea *et al.* [36] for a careful analysis of some computationally tractable procedures for handling simultaneously rows and rank selection.

6.5.2 References

Most of the material presented in this chapter is lifted or adapted from Bunea, She, and Wegkamp [37, 36]. Lemma 6.3 comes from Davidson and Szarek [49], and Theorem 6.8 is adapted from Koltchinskii, Lounici, and Tsybakov [79].

We refer to Bach [11] for the convex Criterion (6.29) and examples of applications. Finally, Exercise 6.6.3 is adapted from Giraud [65].

6.6 Exercises

6.6.1 Hard-Thresholding of the Singular Values

We consider the estimators $\widehat{A}_r$ defined by (6.6) and for $\lambda > 0$ the selection criterion

$$\widehat{r}_\lambda \in \operatorname*{argmin}_r \left\{ \|Y - \mathbf{X}\widehat{A}_r\|_F^2 + \lambda r \right\}.$$

The selection Criterion (6.13) corresponds to the choice $\lambda = K\left(\sqrt{T} + \sqrt{q}\right)^2 \sigma^2$. With the same notations as those of Lemma 6.1, we write $PY = \sum_k \sigma_k u_k v_k^T$ for the SVD decomposition of PY, with $\sigma_1 \geq \sigma_2 \geq \ldots$.

1. Prove that $\|Y - \mathbf{X}\widehat{A}_r\|_F^2 = \|Y\|_F^2 - \sum_{k=1}^r \sigma_k^2$ for $r \leq \operatorname{rank}(PY)$.
2. Check that $\widehat{r}_\lambda = \max\left\{r : \sigma_r^2 \geq \lambda\right\}$ and conclude that

$$\mathbf{X}\widehat{A}_{\widehat{r}_\lambda} = \sum_k \sigma_k \mathbf{1}_{\sigma_k^2 \geq \lambda}\, u_k v_k^T.$$

6.6.2 Exact Rank Recovery

We denote by r^* the rank of A^*, by $\sigma_1(M) \geq \sigma_2(M) \geq \ldots$ the singular values of M ranked in decreasing order, and we consider the selection procedure (6.13).

1. Prove from the previous exercise that

$$\mathbb{P}(\widehat{r} \neq r^*) = \\ \mathbb{P}\left(\sigma_{r^*+1}(PY) \geq \sqrt{K}(\sqrt{T} + \sqrt{q})\sigma \quad \text{or} \quad \sigma_{r^*}(PY) < \sqrt{K}(\sqrt{T} + \sqrt{q})\sigma\right).$$

2. Deduce from Weyl inequality (Theorem C.6 in Appendix C) that

$$\mathbb{P}(\widehat{r} \neq r^*) \leq \\ \mathbb{P}\left(\sigma_1(PE) \geq \min\left(\sqrt{K}(\sqrt{T} + \sqrt{q})\sigma, \sigma_{r^*}(\mathbf{X}A^*) - \sqrt{K}(\sqrt{T} + \sqrt{q})\sigma\right)\right).$$

3. Assume that $\sigma_{r^*}(\mathbf{X}A^*) \geq 2\sqrt{K}(\sqrt{T} + \sqrt{q})\sigma$. Prove that in this case, the probability to recover the exact rank r^* is lower bounded by

$$\mathbb{P}(\widehat{r} = r^*) \geq 1 - \exp\left(-\frac{(\sqrt{K} - 1)^2}{2}\left(\sqrt{T} + \sqrt{q}\right)^2\right).$$

6.6.3 Rank Selection with Unknown Variance

We consider here the case where both the variance σ^2 and the rank $r^* = \mathrm{rank}(A^*)$ are unknown. A classical selection criterion in this setting is

$$\widehat{r} \in \underset{r<nT/\lambda}{\mathrm{argmin}} \left\{ \frac{\|Y - \mathbf{X}\widehat{A}_r\|_F^2}{nT - \lambda r} \right\}, \quad \text{with } \lambda > 0. \tag{6.33}$$

We notice that we can recast this criterion as

$$\widehat{r} \in \underset{r<nT/\lambda}{\mathrm{argmin}} \left\{ \|Y - \mathbf{X}\widehat{A}_r\|_F^2 + \lambda r \widehat{\sigma}_r^2 \right\} \quad \text{with } \widehat{\sigma}_r^2 = \frac{\|Y - \mathbf{X}\widehat{A}_r\|_F^2}{nT - \lambda r},$$

so it can be viewed as a version of (6.13), with the variance σ^2 replaced by $\widehat{\sigma}_r^2$. In the following, we set $\lambda = K(\sqrt{T} + \sqrt{q})^2$ with $K > 1$, and we assume that $1 \leq r^* = \mathrm{rank}(A^*) \leq nT/(2\lambda)$.

1. Prove that $\|Y - (PY)_{(r^*)}\|_F \leq \|E\|_F$.
2. Deduce from the previous question and Criterion (6.33) that
$$\|Y - (PY)_{(\widehat{r})}\|_F^2 \leq \|E\|_F^2 - \frac{\lambda(\widehat{r} - r^*)}{nT - \lambda r_*}\|E\|_F^2.$$
3. Prove from the above inequality and (6.3) that for any $\alpha > 0$
$$\begin{aligned}\|\mathbf{X}\widehat{A}_{\widehat{r}} - \mathbf{X}A^*\|_F^2 &\leq 2\langle \mathbf{X}\widehat{A}_{\widehat{r}} - \mathbf{X}A^*, E\rangle_F - \frac{\lambda(\widehat{r} - r^*)}{nT - \lambda r_*}\|E\|_F^2 \\ &\leq \alpha\|PE\|_{(2,r^*+\widehat{r})}^2 + \alpha^{-1}\|\mathbf{X}\widehat{A}_{\widehat{r}} - \mathbf{X}A^*\|_F^2 - \frac{\lambda(\widehat{r} - r^*)}{nT - \lambda r_*}\|E\|_F^2.\end{aligned}$$
4. Check that for $\alpha > 1$
$$\frac{\alpha - 1}{\alpha}\|\mathbf{X}\widehat{A}_{\widehat{r}} - \mathbf{X}A^*\|_F^2 \leq \lambda r^* \frac{\|E\|_F^2}{nT - \lambda r^*} + \alpha r^* |PE|_{\mathrm{op}}^2 + \widehat{r}\left(\alpha |PE|_{\mathrm{op}}^2 - \lambda \frac{\|E\|_F^2}{nT}\right).$$
5. For $\alpha > 1$ and $\delta > 0$, such that $K \geq \alpha(1+\delta)/(1-\delta)^2$, combining the bound
$$\widehat{r}\left(\alpha|PE|_{\mathrm{op}}^2 - \lambda\frac{\|E\|_F^2}{nT}\right) \leq$$
$$(q \wedge T)\left(\alpha |PE|_{\mathrm{op}}^2 - \lambda(1-\delta)^2\sigma^2\right)_+ + \lambda(q \wedge T)\left((1-\delta)^2\sigma^2 - \frac{\|E\|_F^2}{nT}\right)_+$$
with the Gaussian concentration Inequality (B.2), page 221, prove that
$$\begin{aligned}\mathbb{E}\left[\widehat{r}\left(\alpha|PE|_{\mathrm{op}}^2 - \lambda\frac{\|E\|_F^2}{nT}\right)\right] &\leq 2(1+\delta^{-1})(q \wedge T)\sigma^2 + 4KnTe^{-\delta^2 nT/2}\sigma^2 \\ &\leq 2(1+\delta^{-1})(q \wedge T)\sigma^2 + 8K\delta^{-2}e^{-1}\sigma^2.\end{aligned}$$
6. Conclude that there exists a constant $C_K > 1$ depending only on $K > 1$, such that
$$\mathbb{E}\left[\|\mathbf{X}\widehat{A}_{\widehat{r}} - \mathbf{X}A^*\|_F^2\right] \leq C_K r^*(T + q)\sigma^2.$$

   Compare this bound with (6.16).

Chapter 7

Graphical Models

Graphical modeling is a convenient theory for encoding the conditional dependencies between p random variables $X_1, \dots, X_p$ by a graph g. Graphical models are used in many different frameworks (image analysis, physics, economics, etc.), and they have been proposed for investigating biological regulation networks, brain connections, etc.

The concept of conditional dependence is more suited than the concept of dependence in order to catch "direct links" between variables, as explained in Figure 7.1 below.

When there is a snow storm in Paris, we observe both huge traffic jams and plenty of snowmen in parks. So there is a strong correlation (and thus dependence) between the size of the parisian traffic jams and the number of snowmen in parks in Paris.

Of course, snowmen do not cause traffic jams. Traffic jams and snowmen are correlated only because they are both induced by snow falls. These causality relationships are represented in the side picture by edges.

Conditional dependencies better reflect these relationships. Actually, conditionally in the snow falls, the size of the traffic jams and the number of snowmen are likely to be independent.

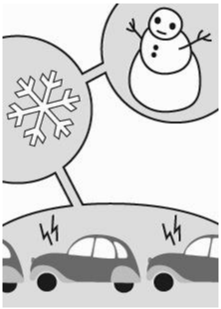


Figure 7.1 *Difference between dependence and conditional dependence.*

There are mainly two types of graphical models, which encode conditional dependencies in two different ways. We briefly present these two types based on directed and nondirected graphs in Section 7.2. Our main goal in this chapter will be to learn the graph of conditional dependencies between $(X_1, \dots, X_p)$ from an n-sample of $(X_1, \dots, X_p)$, with a special focus on the case where the graph has few edges and the sample size n is smaller than the number p of variables. As explained below, it is a very hard and nonparametric problem in general, and we will mainly investigate the case where $(X_1,\dots,X_p)$ follows a Gaussian distribution with (unknown) covariance Σ. In this case, the conditional dependencies are encoded in the precision matrix Σ^{-1} and our problem mainly amounts to estimate the locations of the nonzero entries of a sparse precision matrix.

7.1 Reminder on Conditional Independence

We remind the reader that two random variables X and Y are independent conditionally on a variable Z (we write $X \perp\!\!\!\perp Y \mid Z$) if their conditional laws fulfill

$$\mathrm{law}((X,Y)|Z) = \mathrm{law}(X|Z) \otimes \mathrm{law}(Y|Z).$$

In particular, if the distribution of (X,Y,Z) has a positive density f with respect to a σ-finite product measure μ, then

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z \quad &\Longleftrightarrow \quad f(x,y|z) = f(x|z)f(y|z) \quad \mu\text{-a.e.} \\ &\Longleftrightarrow \quad f(x,y,z) = f(x,z)f(y,z)/f(z) \quad \mu\text{-a.e.} \\ &\Longleftrightarrow \quad f(x|y,z) = f(x|z) \quad \mu\text{-a.e.}, \end{aligned}$$

where $f(x,y|z)$ (resp. $f(x|z)$) represents the conditional density of (x,y) (resp. x) given z.

We also recall that for any measurable function h, we have the property

$$X \perp\!\!\!\perp (Y,W) \mid Z \Longrightarrow X \perp\!\!\!\perp Y \mid (Z, h(W)). \tag{7.1}$$

In order to avoid unnecessary technicalities, we assume in the remainder of this chapter that the distribution of $(X_1,\dots,X_p)$ has a positive continuous density with respect to σ-finite product measure in $\mathbb{R}^p$.

7.2 Graphical Models

7.2.1 Directed Acyclic Graphical Models

Let us consider a directed graph g (made of nodes and arrows, with the arrows linking some nodes) with p nodes labelled from 1 to p, as in Figure 7.2. We will assume that g is acyclic, which means that no sequence of arrows forms a loop in graph g. We call parents of a nodes b, such that there exists an arrow $b \to a$, and we denote by $\mathrm{pa}(a)$ the set of parents of a. We call descendent of a the nodes that can be reached from a by following some sequence of arrows (a included), and we denote by $\mathrm{de}(a)$ the set of descendants of a. For example, in Figure 7.2, the descendents of 11 are $\{11,12,18\}$ and its parents are $\{5,8,10\}$.

Directed acyclic graphical model

Let g be a Directed Acyclic Graph (DAG). The distribution of the random variable $X = (X_1 \dots, X_p)$ is a graphical model according to g if it fulfills the property

$$\text{for all } a: \quad X_a \perp\!\!\!\perp \{X_b,\ b \notin \text{de}(a)\} \;\Big|\; \{X_c,\ c \in \text{pa}(a)\}.$$

We write $\mathscr{L}(X) \sim g$ when this property is met.

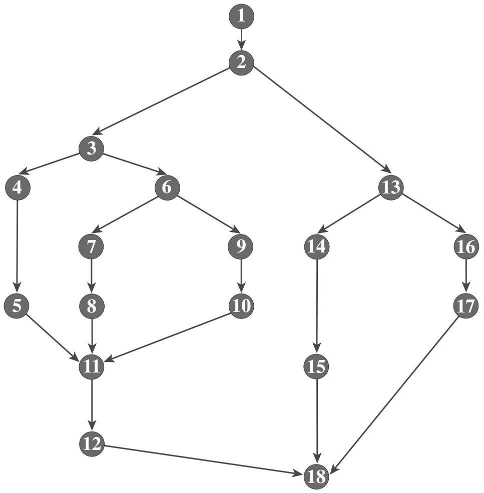


Figure 7.2 *Directed Acyclic Graphical models:* $X_a \perp\!\!\!\perp \{X_b,\ b \notin \text{de}(a)\} \mid \{X_c,\ c \in \text{pa}(a)\}$.

Let us consider two DAGs g and g', with g a subgraph of g' (denoted by $g \subset g'$) and a random variable X, such that $\mathscr{L}(X) \sim g$. Writing $\text{pa}'(a)$ (respectively, $\text{de}'(a)$) for the parents (respectively, the descendants) of the node a in g', we observe that, due to the acyclicity,

$$\text{pa}(a) \subset \text{pa}'(a) \subset \{1, \dots, p\} \setminus \text{de}'(a) \subset \{1, \dots, p\} \setminus \text{de}(a).$$

Accordingly, it follows from $\mathscr{L}(X) \sim g$ and the property (7.1) of the conditional independence, that for any $a = 1, \dots, p$

$$X_a \perp\!\!\!\perp \left\{X_b : b \notin \text{de}'(a)\right\} \mid \left\{X_c : c \in \text{pa}'(a)\right\}.$$

As a consequence for $g \subset g'$, we have $\mathscr{L}(X) \sim g \implies \mathscr{L}(X) \sim g'$. In particular, there is in general no unique DAG g, such that $\mathscr{L}(X) \sim g$. Yet, we may wonder whether there exists a unique DAG g^* minimal for the inclusion, such that $\mathscr{L}(X) \sim g^*$. It is unfortunately not the case, as can be seen in the following simple example. Consider $X_1, \dots, X_p$ generated by the autoregressive model

$$X_{i+1} = \alpha X_i + \varepsilon_i \quad \text{with} \quad X_0 = 0, \quad \alpha \neq 0 \quad \text{and} \quad \varepsilon_i \ \text{i.i.d.}$$

Since $(X_1,\dots,X_p)$ is a Markov chain, we have

$$(X_1,\dots,X_i) \perp\!\!\!\perp (X_i,\dots,X_p) \mid X_i \quad \text{for all } i=1,\dots,p.$$

As a consequence, the two graphs

$$1 \to 2 \to \dots \to p \quad \text{and} \quad 1 \leftarrow 2 \leftarrow \dots \leftarrow p$$

are minimal graphs for this model.

⚠ Be careful with the interpretation of directed graphical models!

Our objective in this chapter is to learn from data a (minimal) graph of conditional dependencies between p variables. Since there is no unique minimal DAG g, such that $\mathscr{L}(X) \sim g$, the problem of estimating "the" minimal acyclic graph of a distribution $\mathbb{P}$ is an ill-posed problem. Yet, it turns out that the minimal DAGs associated with a given distribution only differ by the direction of (some of) their arrows. So instead of trying to estimate a minimal DAG associated to a distribution $\mathbb{P}$, we can try to estimate the locations (but not the directions) of the arrows of the minimal DAGs associated to $\mathbb{P}$. This problem can be solved efficiently under some conditions (see Spirtes *et al.* [113] and Kalisch and Bühlmann [77, 78]). An alternative is to consider another notion of graphical models based on nondirected graphs, more suited for our problem. We will follow this alternative in the remainder of this chapter.

Before moving to nondirected graphical models, we emphasize that directed acyclic graphical models are powerful tools for modeling, defining a distribution $\mathbb{P}$ and computing it. These nice features mainly rely on the factorization formula for distributions $\mathbb{P}$ with a positive density f with respect to a σ-finite product measure in $\mathbb{R}^p$. Actually, in such a case, for any DAG g such that $\mathbb{P} \sim g$, we have the factorization formula

$$f(x_1,\dots,x_p) = \prod_{i=1}^{p} f(x_i|x_{\mathrm{pa}(i)}), \tag{7.2}$$

where $f(x_i|x_{\mathrm{pa}(i)})$ is the conditional density of x_i given $x_{\mathrm{pa}(i)}$. This factorization is easily proved by induction; see Exercise 7.6.1. A common way to use (7.2) is to start from a known graph g representing known causality relationships between the variables, and then build the density f from the conditional densities $f(x_i|x_{\mathrm{pa}(i)})$ according to Formula (7.2).

7.2.2 Nondirected Models

We consider now nondirected graphs g (made of nodes and edges, with the edges linking some nodes) with p nodes labelled from 1 to p, as in Figure 7.3. Graph g induces a symmetric relation on $\{1,\dots,p\}$ by

$$a \overset{g}{\sim} b \quad \Longleftrightarrow \quad \text{there is an edge between } a \text{ and } b \text{ in graph } g.$$

We call neighbors of a, the nodes in $\mathrm{ne}(a) = \left\{b : b \overset{g}{\sim} a\right\}$, and we set $\mathrm{cl}(a) = \mathrm{ne}(a) \cup \{a\}$.

Nondirected graphical model

The distribution of the random variable $X = (X_1 \ldots, X_p)$ is a graphical model according to graph g if it fulfills the property

$$\text{for all } a : \quad X_a \perp\!\!\!\perp \{X_b,\ b \notin \mathrm{cl}(a)\} \bigm| \{X_c,\ c \in \mathrm{ne}(a)\}.$$

We write $\mathscr{L}(X) \sim g$ when this property is met.

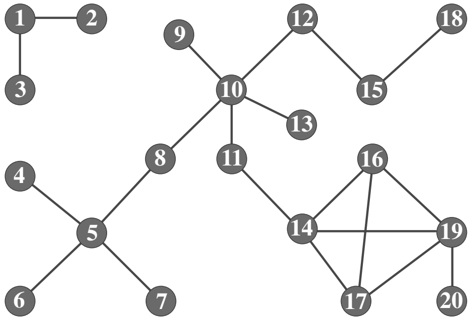


Figure 7.3 *Nondirected graphical models:* $X_a \perp\!\!\!\perp \{X_b : b \nsim a\} \mid \{X_c : c \sim a\}$.

Again, we check that if $\mathscr{L}(X) \sim g$ and $g \subset g'$ then $\mathscr{L}(X) \sim g'$, so there is no unique graph g, such that $\mathscr{L}(X) \sim g$. In particular, if $g_\#$ represents the complete graph (where all the nodes are connected all together), then $\mathscr{L}(X) \sim g_\#$. Yet, when X has a positive continuous density with respect to some σ-finite product measure, there exists a unique minimal graph g_* (for inclusion), such that $\mathscr{L}(X) \sim g_*$. We will prove this result in the Gaussian setting in Section 7.3, and we refer to Lauritzen [81] Chapter 3 for the general case. In the following, we call simply "graph of X" the minimal graph g_*, such that $\mathscr{L}(X) \sim g_*$.

There is a simple connection between directed and nondirected graphical models. Let g be a *directed* graph such that $\mathscr{L}(X) \sim g$. We associate to g the so-called moral graph, which is the *nondirected* graph g_m obtained as follows:

1. For each node, set an edge between its parents in g.
2. Replace all arrows by edges.

We refer to Figure 7.4 for an illustration.

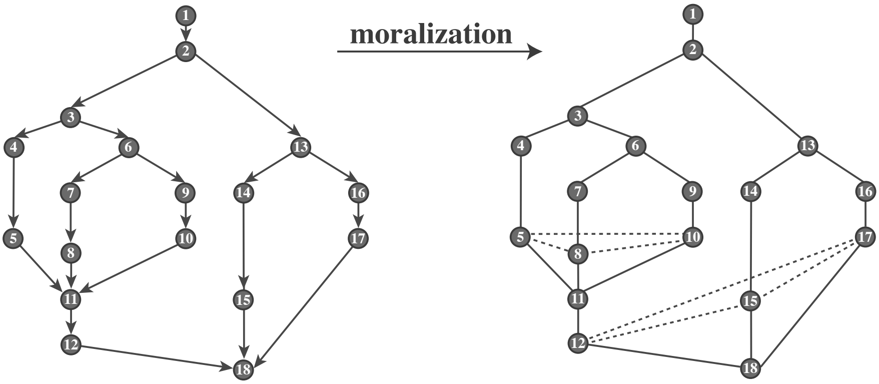


Figure 7.4 *Moralization of a Directed Acyclic Graph. Left: the original graph. Right: the associated moral graph. The extra edges linking the parents are represented with dashed lines.*

Lemma 7.1 From directed to nondirected graphical models

Let g be a directed graph, and write g_m for its (nondirected) moral graph defined above. Then, if X has a positive density with respect to a σ-finite product measure,

$$\mathscr{L}(X) \sim g \implies \mathscr{L}(X) \sim g_m .$$

We refer to Exercise 7.6.2 for a proof of this lemma. We emphasize that the moral graph g_m may not coincide with the minimal graph g_* associated to X.

In the following, our goal is to estimate from an n-sample the minimal graph g_*, such that $\mathscr{L}(X) \sim g_*$. In general, it is a very hard problem in high-dimensional settings. Actually, a result due to Hammersley and Clifford [72] ensures that a distribution with positive continuous density f is a graphical model with respect to a nondirected graph g if and only if f fulfills the factorization

$$f(x) = \prod_{c \in \text{cliques}(g)} f_c(x_c), \tag{7.3}$$

where the f_c are positive functions and the cliques of g are the subgraphs of g that are completely connected (all the nodes of a clique are linked altogether; for example, the nodes 14, 16 ,17, 19 form a clique in Figure 7.3). We refer again to Lauritzen [81] Chapter 3 for a proof of this result. Trying to infer such a minimal decomposition of the density f is a very hard problem in general. Yet, it is tractable in some cases, for example, when $(X_1, \dots, X_p)$ follows a Gaussian distribution. In the following, we will focus on this issue of estimating g_* from an n-sample of a Gaussian distribution.

7.3 Gaussian Graphical Models (GGM)

We assume in the remainder of this chapter that X follows a Gaussian $\mathcal{N}(0,\Sigma)$ distribution, with Σ unknown and nonsingular. Our goal will be to estimate from an n-sample $X^{(1)},\ldots,X^{(n)}$ of X the nondirected graph g_*, minimal for the inclusion, such that $\mathscr{L}(X) \sim g_*$. We have in mind that the sample size n can be smaller than the dimension p of X.

We will denote by $\mathbf{X}$ the $n \times p$ matrix with rows given by the transpose of $X^{(1)},\ldots,X^{(n)}$ and we write $\widehat{\Sigma}$ for the empirical covariance matrix

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)})(X^{(i)})^T = \frac{1}{n}\mathbf{X}^T\mathbf{X}.$$

7.3.1 Connection with the Precision Matrix and the Linear Regression

A nice feature of the Gaussian $\mathcal{N}(0,\Sigma)$ distribution, with Σ nonsingular, is that the minimal graph g_* is encoded in the precision matrix $K = \Sigma^{-1}$. Actually, let us define graph g with nodes labelled by $\{1,\ldots,p\}$ according to the symmetric relation for $a \neq b$

$$a \overset{g}{\sim} b \iff K_{a,b} \neq 0. \tag{7.4}$$

The next lemma shows that g is the minimal (nondirected) graph, such that $\mathscr{L}(X) \sim g$.

Lemma 7.2 GGM and precision matrix

For graph g defined by (7.4), we have

1. *$\mathscr{L}(X) \sim g$ and g is the minimal graph fulfilling this property.*
2. *For any $a \in \{1,\ldots,p\}$, there exists $\varepsilon_a \sim \mathcal{N}(0, K_{aa}^{-1})$ independent of $\{X_b : b \neq a\}$, such that*

$$X_a = -\sum_{b\in\mathrm{ne}(a)} \frac{K_{ab}}{K_{aa}} X_b + \varepsilon_a\,. \tag{7.5}$$

Proof.

1. We write $\mathrm{nn}(a) = \{1,\ldots,p\} \setminus \mathrm{cl}(a)$ for the nodes of g non-neighbor to a. Let us consider the two sets $A = \{a\} \cup \mathrm{nn}(a)$ and $B = \mathrm{ne}(a)$. The precision matrix restricted to A is of the form

$$K_{AA} = \begin{pmatrix} K_{aa} & 0 \\ 0 & K_{\mathrm{nn}(a)\,\mathrm{nn}(a)} \end{pmatrix}.$$

Lemma A.4, page 215, in Appendix A ensures that the distribution of $X_{\{a\}\cup\mathrm{nn}(a)}$ given $X_{\mathrm{ne}(a)}$ is Gaussian with covariance matrix

$$(K_{AA})^{-1} = \begin{pmatrix} K_{aa}^{-1} & 0 \\ 0 & (K_{\mathrm{nn}(a)\,\mathrm{nn}(a)})^{-1} \end{pmatrix}.$$

Since independence is equivalent to zero covariance for Gaussian random variables, the variables X_a and $X_{\text{nn}(a)}$ are independent conditionally on $X_{\text{ne}(a)}$. We note that this property is no longer true if we remove an edge between a and one of its neighbor in g, since in this case the off-diagonal blocks of $(K_{AA})^{-1}$ are nonzero.

2. The second point simply rephrases Formula (A.3), page 215, in Appendix A.

□

In the next subsections, we build on Lemma 7.2 in order to derive procedures for estimating g_*.

7.3.2 Estimating g by Multiple Testing

Corollary A.4 in Appendix A ensures that, for $a \neq b$, the conditional correlation (also called partial correlation) of X_a and X_b given $\{X_c : c \neq a,b\}$ is given by

$$\text{cor}(X_a, X_b | X_c : \ c \neq a,b) = \frac{-K_{ab}}{\sqrt{K_{aa}K_{bb}}}. \tag{7.6}$$

As a consequence, there is an edge in the graph g_* if and only if the conditional correlation $\text{cor}(X_a, X_b | X_c : c \neq a,b)$ is nonzero. A natural idea is then to estimate g_* by testing if these conditional correlations are nonzero. When $n > p-2$, the conditional correlations can be estimated by replacing K in (7.6) by the inverse of the empirical covariance matrix $\widehat{\Sigma}^{-1}$

$$\widehat{\text{cor}}(X_a, X_b | X_c : c \neq a,b) = \widehat{\rho}_{ab} = \frac{-\left[\widehat{\Sigma}^{-1}\right]_{ab}}{\sqrt{\left[\widehat{\Sigma}^{-1}\right]_{aa}\left[\widehat{\Sigma}^{-1}\right]_{bb}}}.$$

When $\text{cor}(X_a, X_b | X_c : c \neq a,b) = 0$, we have (see Anderson [2], Chapter 4.3)

$$\widehat{t}_{ab} = \sqrt{n-2-p} \times \frac{\widehat{\rho}_{ab}}{\sqrt{1-\widehat{\rho}_{ab}^2}} \sim \text{Student}(n-p-2).$$

For each $a < b$, let us denote by $\widehat{p}_{ab}$ the p-value associated to the test statistic $|\widehat{t}_{ab}|$. We can then estimate the set of nonzero conditional correlations by applying a multiple testing procedure (see Chapter 8) to the set of p-values $\{\widehat{p}_{ab}, 1 \leq a < b \leq p\}$. This procedure makes perfect sense when n is large, much larger than p. But when $n-p$ is small, the empirical conditional correlation $\widehat{\rho}_{ab}$ is very unstable and the procedure leads to poor results. When $p > n$, the empirical covariance matrix $\widehat{\Sigma}$ is not invertible, so the procedure cannot be implemented.

There have been several propositions to circumvent this issue. One proposition is to work with $\widehat{\text{cor}}(X_a, X_b | X_c : c \in S)$, with S a small subset of $\{1, \ldots, p\} \setminus \{a,b\}$ (like pairs), instead of $\widehat{\text{cor}}(X_a, X_b | X_c : c \neq a,b)$; see, e.g., Wille and Bühlmann [127], and Castelo and Roverato [41]. While the empirical conditional correlation

$\widehat{\mathrm{cor}}(X_a, X_b | X_c : c \in S)$ does not suffer from instability when the cardinality of S is small compared to n, it is unclear what we estimate at the end in general.

If we look carefully at the definition of $\widehat{\mathrm{cor}}(X_a, X_b | X_c : c \neq a, b)$, we observe that the instability for large p comes from the fact that we estimate K by $\widehat{\Sigma}^{-1}$, which is unstable. An idea is then to build a more stable estimator of K.

7.3.3 Sparse Estimation of the Precision Matrix

We have in mind the case where the underlying graph g is sparse (it has a few edges). Since $K_{ab} = 0$ when there is no edge between a and b in g_*, the sparsity of g_* translates into coordinate sparsity for the precision matrix K. Exploiting this sparsity can significantly improve the estimation of K upon $\widehat{\Sigma}^{-1}$. Following the ideas of Chapter 4, we can then estimate K by minimizing the negative log-likelihood penalized by the ℓ^1 norm of K. Let us first derive the likelihood of a $p \times p$ positive symmetric matrix $K \in \mathscr{S}_p^+$

$$\begin{aligned}\text{Likelihood}(K) &= \prod_{i=1}^{n} \sqrt{\frac{\det(K)}{(2\pi)^p}} \exp\left(-\frac{1}{2}(X^{(i)})^T K X^{(i)}\right) \\ &= \left(\frac{\det(K)}{(2\pi)^p}\right)^{n/2} \exp\left(-\frac{n}{2}\langle \widehat{\Sigma}, K\rangle_F\right),\end{aligned}$$

where the last equality follows from $\langle X^{(i)}, KX^{(i)}\rangle = \langle X^{(i)}(X^{(i)})^T, K\rangle_F$. Removing the terms not depending on K, we obtain that the negative-log-likelihood is given (up to a constant) by

$$-\frac{n}{2}\log(\det(K)) + \frac{n}{2}\langle K, \widehat{\Sigma}\rangle_F .$$

Since $-\log(\det(K))$ is convex (Exercise 7.6.3), the above function is convex. Similarly, as for the Lasso estimator, minimizing the negative log-likelihood penalized by the ℓ^1 norm of K will induce a coordinate sparse estimator. The resulting estimator

$$\widehat{K}_\lambda = \underset{K \in \mathscr{S}_p^+}{\operatorname{argmin}} \left\{ -\frac{n}{2}\log(\det(K)) + \frac{n}{2}\langle K, \widehat{\Sigma}\rangle_F + \lambda \sum_{a \neq b} |K_{ab}| \right\}, \tag{7.7}$$

is usually called the "graphical-Lasso" or simply "glasso" estimator. We point out that we do not penalize the diagonal elements, since they are not expected to be zero. From a theoretical point of view, risk bounds similar to those for the Lasso have been derived for the glasso estimator by Ravikumar *et al.* [100], under some "compatibility conditions" on the true precision matrix, which are hard to interpret. From a practical point of view, minimizing the convex Criterion (7.7) is more challenging than for the Lasso Criterion (4.4), but a powerful numerical scheme has been developed in Friedman *et al.* [59], and the glasso estimator can be computed in quite large dimensions (at least a few thousands). This scheme is implemented in the `R` package `glasso` available at `http://cran.r-project.org/web/packages/glasso/` . Nevertheless, some poor results have been reported in practice by Villers *et al.* [125] when $n \leq p$ compared to some other estimation schemes.

7.3.4 Estimation of g by Regression

From the second part of Lemma 7.2, we observe that the sparsity of K induces some coordinate-sparsity in the linear regressions (7.5) for $a = 1, \ldots, p$. We can then recast our problem in a multivariate regression setting.

Let us define the matrix θ by $\theta_{ab} = -K_{ab}/K_{bb}$ for $b \neq a$ and $\theta_{aa} = 0$. Equation (7.5) ensures that $\mathbb{E}[X_a|X_b : b \neq a] = \sum_b \theta_{ba} X_b$, so the vector $[\theta_{ba}]_{b:b\neq a}$ minimizes

$$\mathbb{E}\left[\left(X_a - \sum_{b:b\neq a} \beta_b X_b\right)^2\right]$$

over the vectors $\beta \in \mathbb{R}^{p-1}$. Writing Θ for the space of $p \times p$ matrices with zero on the diagonal, after summing on a, we obtain

$$\begin{aligned} \theta & \in \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}\left[\sum_{a=1}^{p}\left(X_a - \sum_{b:b\neq a} \theta_{ba} X_b\right)^2\right] \\ & = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}\left[\|X - \theta^T X\|^2\right] \\ & = \operatorname*{argmin}_{\theta\in\Theta} \|\Sigma^{1/2}(I-\theta)\|_F^2 , \end{aligned} \qquad (7.8)$$

where the last equality comes from $\mathbb{E}[\|AX\|^2] = \langle A, A\Sigma\rangle_F$ for $X \sim \mathcal{N}(0,\Sigma)$. For $a \neq b$, since $K_{ab} \neq 0$ if and only if $\theta_{ab} \neq 0$, an alternative idea to estimate g_* is to build a coordinate-sparse estimator $\widehat{\theta}$ of the matrix θ. If we replace Σ by $\widehat{\Sigma}$ in (7.8), we obtain

$$\|\widehat{\Sigma}^{1/2}(I-\theta)\|_F = \frac{1}{n}\langle (I-\theta), \mathbf{X}^T\mathbf{X}(I-\theta)\rangle_F = \frac{1}{n}\|\mathbf{X}(I-\theta)\|_F^2 .$$

Adding a penalty $\Omega(\theta)$ enforcing coordinate sparsity, we end with the estimator

$$\widehat{\theta}_\lambda \in \operatorname*{argmin}_{\theta\in\Theta}\left\{\frac{1}{n}\|\mathbf{X}-\mathbf{X}\theta\|_F^2 + \lambda\Omega(\theta)\right\}.$$

This corresponds to a special case of coordinate-sparse multivariate regression. We discuss below two choices of penalty for enforcing sparsity: the classical ℓ^1 penalty as in the Lasso estimator and a ℓ^1/ℓ^2 penalty as in group–Lasso estimator.

With the ℓ^1 penalty

If we choose for $\Omega(\theta)$ the ℓ^1 norm of θ, we obtain

$$\widehat{\theta}_\lambda^{\ell^1} \in \operatorname*{argmin}_{\theta\in\Theta}\left\{\frac{1}{n}\|\mathbf{X}-\mathbf{X}\theta\|_F^2 + \lambda\sum_{a\neq b}|\theta_{ab}|\right\}. \qquad (7.9)$$

Since we have

$$\frac{1}{n}\|\mathbf{X}-\mathbf{X}\theta\|_F^2+\lambda\sum_{a\neq b}|\theta_{ab}|=\sum_{a=1}^{p}\left(\frac{1}{n}\|\mathbf{X}_a-\sum_{b:b\neq a}\theta_{ba}\mathbf{X}_b\|^2+\lambda\sum_{b:b\neq a}|\theta_{ba}|\right),$$

we can split the minimization (7.9) on $\theta\in\Theta$ into p minimization problems in $\mathbb{R}^{p-1}$

$$\left[\widehat{\theta}_\lambda^{\ell^1}\right]_{-a,a}\in\operatorname*{argmin}_{\beta\in\mathbb{R}^{p-1}}\left\{\frac{1}{n}\|\mathbf{X}_a-\sum_b\beta_b\mathbf{X}_b\|^2+\lambda|\beta|_{\ell^1}\right\}\quad\text{for } a=1,\ldots,p,$$

with the notation $\theta_{-a,a}=[\theta_{ba}]_{b:b\neq a}$. We recognize p Lasso estimators that can be computed efficiently; see Section 4.2.4 in Chapter 4.

Unfortunately, there is a difficulty in order to define a nondirected graph $\widehat{g}$ from $\widehat{\theta}^{\ell^1}$. Actually, whereas the zeros of θ are symmetric with respect to the diagonal, no constraint enforces that $\widehat{\theta}_{ab}^{\ell^1}\neq 0$ when $\widehat{\theta}_{ba}^{\ell^1}\neq 0$ and conversely. So we have to choose an arbitrary decision rule in order to build a graph $\widehat{g}$ from $\widehat{\theta}^{\ell^1}$. For example, we may decide to set an edge between a and b in $\widehat{g}$ when either $\widehat{\theta}_{ab}^{\ell^1}\neq 0$ or $\widehat{\theta}_{ba}^{\ell^1}\neq 0$. Another example is to set an edge $a\sim b$ in $\widehat{g}$ when both $\widehat{\theta}_{ab}^{\ell^1}\neq 0$ and $\widehat{\theta}_{ba}^{\ell^1}\neq 0$. In order to avoid these unsatisfactory rules, we can modify the penalty in order to enforce the symmetry of the zeros of $\widehat{\theta}$.

With the ℓ^1/ℓ^2 penalty

The idea is to regroup the nondiagonal indices (a,b) by symmetric pairs $\{(a,b),(b,a)\}$ and apply a group–Lasso penalty, namely

$$\widehat{\theta}_\lambda^{\ell^1/\ell^2}\in\operatorname*{argmin}_{\theta\in\Theta}\left\{\frac{1}{n}\|\mathbf{X}-\mathbf{X}\theta\|_F^2+\lambda\sum_{a<b}\sqrt{\theta_{ab}^2+\theta_{ba}^2}\right\}.\qquad(7.10)$$

This estimator has the nice property to be coordinate sparse with symmetric zeros. So there is no ambiguity in order to define a graph $\widehat{g}$ from $\widehat{\theta}_\lambda^{\ell^1/\ell^2}$.

The minimization problem (7.10) is convex in $\mathbb{R}^{p\times p}$, unfortunately, it cannot be split into p subproblems. It is then computationally more intensive to compute $\widehat{\theta}_\lambda^{\ell^1/\ell^2}$ than $\widehat{\theta}_\lambda^{\ell^1}$ in large dimensions. We refer to Exercise 7.6.4 for the description of a block gradient descent algorithm for minimizing (7.10).

Theoretical results

We cannot apply directly the results of Chapter 4 on the Lasso and group–Lasso estimators, since the design matrix $\mathbf{X}$ is random. Nevertheless, we can work conditionally on the design and then integrate the result. For the sake of illustration, we give an example of the kind of results that we can obtain for the estimator $\widehat{\theta}_\lambda^{\ell^1}$.

We introduce the population restricted eigenvalue

$$\phi_*=\max\left\{|\Sigma_{mm}|_{\text{op}}:m\subset\{1,\ldots,p\},\ \text{card}(m)\leq n/(3\log p)\right\}$$

and define the degree of a matrix A by $d(A) = \max\{|A_{:a}|_0 : a = 1,\dots,p\}$. Note that for $A = \theta$, the degree $d(\theta)$ is equal to the degree of the graph g_*, which is defined by $\text{degree}(g_*) = \max\{\text{ne}(a) : a = 1,\dots,p\}$.

Theorem 7.3 Risk bound for $\widehat{\theta}^{\ell^1}$

Assume that the diagonal entries of Σ are equal to one and set $\text{cor}(\Sigma) = \max_{a\neq b}|\Sigma_{a,b}|$. *Then there exist some constants $C, C', C_1, C_2, C_3 > 0$, such that when*

$$1 \leq d(\theta) \leq C\,\frac{n}{\phi_* \log(p)} \bigwedge \frac{1}{\text{cor}(\Sigma)} \tag{7.11}$$

for $\lambda = C'\phi_*\sqrt{\log(p)}$, the estimator $\widehat{\theta}^{\ell^1}_\lambda$ fulfills with probability at least $1 - C_1\left(p^{-C_2} + p^2 e^{-C_2 n}\right)$ *the risk bound*

$$\|\Sigma^{1/2}(\widehat{\theta}^{\ell^1}_\lambda - \theta)\|_F^2 \leq C_3 \frac{\phi_*^2 \log(p)}{n}\,|\theta|_0. \tag{7.12}$$

The above theorem provides a control of $\|\Sigma^{1/2}(\widehat{\theta}^{\ell^1} - \theta)\|_F^2$. This quantity does not directly quantify the accuracy of the estimation of g_*, but rather the accuracy of $\widehat{\theta}^{\ell^1}$ for predicting the variables X_a with $(\widehat{\theta}^{\ell^1})^T X$. Actually, assume that we obtain a new observation $X^{\text{new}} \sim \mathcal{N}(0,\Sigma)$ independent of the n-sample used to compute $\widehat{\theta}^{\ell^1}$. Then

$$\begin{aligned}\sum_{a=1}^{p} \mathbb{E}^{\text{new}}\left[\left(X_a^{\text{new}} - \sum_{b:b\neq a} \widehat{\theta}^{\ell^1}_{ba} X_b^{\text{new}}\right)^2\right] &= \mathbb{E}^{\text{new}}\left[\|X^{\text{new}} - (\widehat{\theta}^{\ell^1})^T X^{\text{new}}\|^2\right] \\ &= \|\Sigma^{1/2}(I - \widehat{\theta}^{\ell^1})\|_F^2,\end{aligned}$$

where $\mathbb{E}^{\text{new}}$ represents the expectation with respect to X^{new}. Since θ is the orthogonal projection of I onto Θ with respect to the scalar product $\langle \cdot, \Sigma\,\cdot\rangle_F$, Pythagore's formula gives

$$\sum_{a=1}^{p} \mathbb{E}^{\text{new}}\left[\left(X_a^{\text{new}} - \sum_{b:b\neq a} \widehat{\theta}^{\ell^1}_{ba} X_b^{\text{new}}\right)^2\right] = \|\Sigma^{1/2}(I-\theta)\|_F^2 + \|\Sigma^{1/2}(\theta - \widehat{\theta}^{\ell^1})\|_F^2.$$

The first term represents the minimal error for predicting the X_a from the $\{X_b : b \neq a\}$ and the second term corresponds to the stochastic error due to the estimation procedure. Theorem 7.3 gives a bound for this last term. Condition (7.11) requires that the degree of the graph g_* remains small compared to $n/\log(p)$. It has been shown that this condition on the degree is unavoidable (see Verzelen [124]), and the degree of the graph seems to be a notion of sparsity well-suited to characterize the statistical difficulty for estimation in the Gaussian graphical model. The second constraint on the product $d(\theta)\text{cor}(\Sigma)$ is due to the use of the lower bound of Exercise 4.5.3 on the restricted eigenvalues, and it is suboptimal.

Proof. The proof mainly relies on Theorem 4.1 in Chapter 4 for the Lasso estimator. Yet, the adaptation to the random design setting is slightly technical and lengthy, and we only sketch the main lines.
The main idea is to start from the formula

$$n^{-1/2}\mathbf{X}_a = \sum_{b\neq a} n^{-1/2}\mathbf{X}_b\theta_{ab} + n^{-1/2}\varepsilon_a \quad \text{for } a = 1,\ldots,p,$$

with $\varepsilon_a \sim \mathcal{N}(0, K_{aa}^{-1}I_n)$ and work conditionally on $\{\mathbf{X}_b : b \neq a\}$ for each a. Then, for

$$\lambda \geq 3\max_{a\neq b}|\mathbf{X}_b^T\varepsilon_a/n|,$$

combining the risk Bound (4.13) with Exercise 4.5.3, we obtain (after summing the p bounds)

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}(\widehat{\theta}_\lambda^{\ell^1} - \theta)\|_F^2 &\leq \inf_{A\in\Theta}\left\{\frac{1}{n}\|\mathbf{X}(A-\theta)\|_F^2 + \frac{\lambda^2|A|_0}{\left(1-11d(A)\max_{b\neq a}|\mathbf{X}_a^T\mathbf{X}_b|/n\right)_+}\right\}\\ &\leq \frac{\lambda^2|\theta|_0}{\left(1-11d(\theta)\max_{b\neq a}|\mathbf{X}_a^T\mathbf{X}_b|/n\right)_+}. \end{aligned} \tag{7.13}$$

To conclude, all we need is to be able to replace $n^{-1/2}\mathbf{X}$ by $\Sigma^{1/2}$ in the above bound. In a high-dimensional setting, where $p > n$, we do not have $\|\mathbf{X}v\|^2/n \approx \|\Sigma^{1/2}v\|^2$ for all vector $v \in \mathbb{R}^p$, yet, the next lemma shows that it is true with high probability simultaneously for all sparse vector v. More precisely, there exist some constants $0 < c_- < c_+ < +\infty$, such that we have with high probability

$$c_-\|\Sigma^{1/2}v\|^2 \leq \frac{1}{n}\|\mathbf{X}v\|^2 \leq c_+\|\Sigma^{1/2}v\|^2$$

for all v fulfilling $|v|_0 \leq n/(3\log(p))$.

Lemma 7.4 Restricted isometry constants

For $1 \leq d < n \wedge p$, there exists an exponential random variable ξ with parameter 1, such that

$$\inf_{\beta:|\beta|_0\leq d}\frac{n^{-1/2}\|\mathbf{X}\beta\|}{\|\Sigma^{1/2}\beta\|} \geq 1 - \frac{\sqrt{d}+\sqrt{2\log(C_p^d)}+\delta_d+\sqrt{2\xi}}{\sqrt{n}}, \tag{7.14}$$

where $\delta_d = \log\left(4\pi\log(C_p^d)\right)/\sqrt{8\log(C_p^d)}$.

Similarly, there exists an exponential random variable ξ' with parameter 1, such that

$$\sup_{\beta:|\beta|_0\leq d}\frac{n^{-1/2}\|\mathbf{X}\beta\|}{\|\Sigma^{1/2}\beta\|} \leq 1 + \frac{\sqrt{d}+\sqrt{2\log(C_p^d)}+\delta_d+\sqrt{2\xi'}}{\sqrt{n}}. \tag{7.15}$$

Proof. We refer to Exercise 7.6.8 for a proof of this lemma. □

To conclude the proof of Theorem 7.3, it remains to check that under the hypotheses of Theorem 7.3, with probability at least $1-C_1(p^{-C_2}+p^2e^{-C_2n})$, we have

1. $\max_{a\neq b}|\mathbf{X}_b^T\varepsilon_a/n|\leq C\sqrt{\log(p)/n}$ and $\max_{a\neq b}|\mathbf{X}_a^T\mathbf{X}_b|/n\leq C_4\text{cor}(\Sigma)$, and
2. $d(\widehat{\theta}_\lambda^{\ell^1})\leq n/(6\log(p))$ when $\lambda\geq C'\phi^*\sqrt{\log(p)/n}$,
3. $n^{-1}\|\mathbf{X}(\widehat{\theta}_\lambda^{\ell^1}-\theta)\|_F^2\geq c_-\|\Sigma^{1/2}(\widehat{\theta}_\lambda^{\ell^1}-\theta)\|_F^2$.

The first point ensures that under the assumptions of the theorem, we have with large probability $\lambda\geq 3\max_{a\neq b}|\mathbf{X}_b^T\varepsilon_a/n|$ and $d(\theta)\max_{b\neq a}|\mathbf{X}_a^T\mathbf{X}_b|/n\leq 1/22$. The two last points allows to bound $\|\Sigma^{1/2}(\widehat{\theta}_\lambda^{\ell^1}-\theta)\|^2$ by a constant times $n^{-1}\|\mathbf{X}(\widehat{\theta}_\lambda^{\ell^1}-\theta)\|^2$. Plugging this bounds in (7.13), we get the risk Bound (7.12)

Checking the above 3 points is somewhat lengthy; we only point out the main arguments. The first point can be proved with the Gaussian concentration Inequality (B.2) and by noticing that $\mathbf{X}_b^T\varepsilon_a=\|\mathbf{X}_b\|N_{a,b}$, where $N_{a,b}$ is independent of $\|\mathbf{X}_b\|$ and follows a Gaussian distribution with variance at most 1. The second point can be proved by combining Lemma 3.2 in the Appendix of Giraud *et al.* [67], Exercise 4.5.3, and Lemma 7.4. The last point is obtained by combining Lemma 7.4 with $d(\widehat{\theta}_\lambda^{\ell^1}-\theta)\leq d(\widehat{\theta}_\lambda^{\ell^1})+d(\theta)$ and the second point. □

Theorem 7.3 describes the prediction performance of the estimator $\widehat{\theta}^{\ell^1}$ but gives little information on the recovery of g_*. Various conditions have been proposed to ensure the recovery of graph g_* by the estimator $\widehat{\theta}^{\ell^1}$; see, for example, Wainwright [126] or Meinshausen and Buhlmann [95]. Unfortunately, these conditions are not likely to be met in high-dimensional settings, and the best we can expect in practice is a partial recovery of graph g_*.

7.4 Practical Issues

Hidden variables

It may happen that we do not observe some variables that have a strong impact on conditional dependencies. For example, if we investigate the regulation between genes, we may lack the measurements for a key gene that regulates many genes of interest. In such a case, even if the graph of conditional dependencies between all the variables (observed and nonobserved) is sparse, the graph of conditional dependencies between the sole observed variables will not be sparse in general; see Exercise 7.6.5. Therefore, the presence of hidden variables can strongly impact the inference in graphical models. We refer to Exercise 7.6.5 for details on this issue and an estimation procedure taking into account hidden variables.

Non-Gaussian variables

Our data may not have a Gaussian distribution. As explained above, inferring the conditional dependencies of a general distribution is unrealistic in high-dimensional settings, even when this distribution is a graphical model with respect to a sparse graph.

In particular, trying to infer graph g and the functions f_c in the decomposition (7.3) is hopeless in a high-dimensional setting without additional structural assumptions. We need to restrict either to some simple classes of graph g or to some special classes of densities f. A possible approach is to assume that some transformations of the data is Gaussian. For example, write F_j for the cumulative distribution function of X_j and Φ for the cumulative distribution of the standard Gaussian distribution. Then, the variable $Z_j = \Phi^{-1}(F_j(X_j))$ follows a standard Gaussian distribution. A structural hypothesis is to *assume* that in addition $(Z_1,\ldots,Z_p)$ has a Gaussian distribution with nonsingular covariance matrix. Then, since the minimal graphs of $(X_1,\ldots,X_p)$ and $(Z_1,\ldots,Z_p)$ coincide, we are back to the Gaussian setting. The point is that in practice we do not know F_j. A first approach is to estimate F_j with some nonparametric estimator $\widehat{F}_j$ and work with the transformed data $(\Phi^{-1}(\widehat{F}_1(X_1)),\ldots,\Phi^{-1}(\widehat{F}_p(X_p)))$ as if they were distributed according to a Gaussian distribution.

A more powerful approach amounts to estimate the correlation matrix of $(Z_1,\ldots,Z_p)$ from the ranked statistics of $X_1,\ldots,X_p$ and then replace in the procedures described above the empirical covariance matrix $\widehat{\Sigma}$ by this estimated correlation matrix; see Exercise 7.6.7 for details and Section 7.5.2 for references. This procedure is implemented in the `R` package `huge` available at `http://cran.r-project.org/web/packages/huge/`.

7.5 Discussion and References

7.5.1 Take-Home Message

Graphical modeling is a nice theory for encoding the conditional dependencies between random variables. Estimating the minimal graph depicting the conditional dependencies is in general out of reach in high-dimensional settings. Yet, it is possible for Gaussian random variables, since in this case the minimal graph simply corresponds to the zero entries of the inverse of the covariance matrix of the random variables. The estimation of the graph can even be achieved with a sample size smaller than the number of variables, as long as the degree of the graph remains small enough.

In practice, we should not expect more than a partial recovery of the graph when the sample size is smaller than the number of variables. Furthermore, the variables are not likely to be Gaussian, and we may also have to deal with some hidden variables. Handling these two issues requires some additional work, as explained Section 7.4.

7.5.2 References

We refer to Lauritzen [81] for an exhaustive book on graphical models. The grapical-Lasso procedure has been proposed by Banerjee, El Ghaoui, and d'Aspremont [13] and the numerical optimization has been improved by Friedman, Hastie, and Tibshirani [59]. The regression procedure with the ℓ^1 penalty has been proposed and analyzed by Meinshausen and Bühlmann [95]. Theorem 7.3 is a simple combination of Proposition 4.1 in Chapter 4, Lemma 3.2 in Appendix of Giraud *et al.* [67], and Lemma 1 in Giraud [63].

Similarly to Chapter 5, we have at our disposal many different estimation procedures, which all depend on at least one tuning parameter. Therefore, we have a large family $\widehat{\mathscr{G}}$ of estimated graphs, and we need a criterion in order to select "at best" a graph among this family. GGMselect [66] is a `R` package that has been designed for this issue `http://cran.r-project.org/web/packages/GGMselect/`. It selects a graph $\widehat{g}$ among $\widehat{\mathscr{G}}$ by minimizing a penalized criterion closely linked to Criterion (5.3) for estimator selection. The resulting estimator satisfies some oracle-like inequality similar to (2.12).

Exercise 7.6.5 on graphical models with hidden variables is mainly based on Chandrasekaran, Parrilo and Willsky [45]. Exercise 7.6.6 is derived from Cai *et al.* [39], and Exercise 7.6.7 on Gaussian copula graphical models is adapted from Lafferty *et al.* [80], Liu *et al.* [90], and Xue and Zou [128].

7.6 Exercises

7.6.1 Factorization in Directed Models

We will prove the factorization formula (7.2) by induction. With no loss of generality, we can assume that node p is a leaf in graph g (which means that it has no descendent).

1. Prove that $f(x_1,\ldots,x_p) = f(x_p|x_{\mathrm{pa}(p)})f(x_1,\ldots,x_{p-1})$.
2. Prove the factorization formula (7.2).

7.6.2 Moralization of a Directed Graph

Let us consider a directed graph g. The moral graph g^m of the directed graph g is the nondirected graph obtained by linking all the parents of each node together and by replacing directed edges by nondirected edges. Assume that $X = (X_1,\ldots,X_p)$ has a positive density with respect to a σ-finite product measure. We will prove that if $\mathscr{L}(X) \sim g$ then $\mathscr{L}(X) \sim g^m$.

1. Starting from the factorization formula (7.2), prove that there exists two functions g_1 and g_2, such that

$$f(x) = g_1(x_a, x_{\mathrm{ne}^m(a)})g_2(x_{\mathrm{nn}^m(a)}, x_{\mathrm{ne}^m(a)}),$$

   where $\mathrm{ne}^m(a)$ represents the nodes neighbor to a in g^m and $\mathrm{nn}^m(a) = \{1,\ldots,p\} \setminus \mathrm{cl}^m(a)$, with $\mathrm{cl}^m(a) = \mathrm{ne}^m(a) \cup \{a\}$.

2. Prove that $f(x|x_{\mathrm{ne}^m(a)}) = f(x_a|x_{\mathrm{ne}^m(a)})f(x_{\mathrm{nn}^m(a)}|x_{\mathrm{ne}^m(a)})$.
3. Conclude that $\mathscr{L}(X) \sim g^m$.

7.6.3 Convexity of $-\log(\det(\mathbf{K}))$

We will prove the convexity of $K \to -\log(\det(K))$ on the set of symmetric positive definite matrices. Let K and S be two symmetric positive definite matrices. Since $K^{-1/2}SK^{-1/2}$ is symmetric, there exist an orthogonal matrix U and a diagonal matrix D, such that $K^{-1/2}SK^{-1/2} = UDU^T$. We set $Q = K^{1/2}U$.

1. Check that $K = QQ^T$ and $S = QDQ^T$.
2. For $\lambda \in [0,1]$, prove that
$$-\log(\det(\lambda S + (1-\lambda)K) = -\log(\det(K)) - \log(\det(\lambda D + (1-\lambda)I)).$$
3. From the convexity of $x \to -\log(x)$, conclude that
$$\begin{aligned}-\log(\det(\lambda S + (1-\lambda)K) &\leq -\log(\det(K)) - \lambda \log(\det(D)) \\ &= -(1-\lambda)\log(\det(K)) - \lambda \log(\det(S)).\end{aligned}$$

7.6.4 Block Gradient Descent with the ℓ^1/ℓ^2 Penalty

We fix $\lambda > 0$, and we consider the estimator $\widehat{\theta}$ defined by (7.10). We will assume that the columns $\mathbf{X}_a$ of $\mathbf{X}$ fulfills $n^{-1}\mathbf{X}_a^T\mathbf{X}_a = 1$.

1. When $\|(\theta_{ab}, \theta_{ba})\| \neq 0$, check that the partial gradient ∇_{ab} of Criterion (7.10) according to $(\theta_{ab}, \theta_{ba})$ is
$$\nabla_{ab} = -\frac{2}{n}\begin{pmatrix}\mathbf{X}_a^T(\mathbf{X}_b - \sum_{k\neq b}\theta_{kb}\mathbf{X}_k) \\ \mathbf{X}_b^T(\mathbf{X}_a - \sum_{k\neq a}\theta_{ka}\mathbf{X}_k)\end{pmatrix} + \frac{\lambda}{\|(\theta_{ab}, \theta_{ba})\|}\begin{pmatrix}\theta_{ab} \\ \theta_{ba}\end{pmatrix}$$
2. We define $\Delta = \begin{pmatrix}\Delta_{ab} \\ \Delta_{ba}\end{pmatrix}$ with
$$\Delta_{ab} = \frac{1}{n}\mathbf{X}_a^T(\mathbf{X}_b - \sum_{k\neq a,b}\theta_{kb}\mathbf{X}_k).$$
Prove that minimizing (7.10) in the variables $(\theta_{ab}, \theta_{ba})$ gives
$$\begin{pmatrix}\widehat{\theta}_{ab} \\ \widehat{\theta}_{ba}\end{pmatrix} = \left(1 - \frac{\lambda}{2\|\Delta\|}\right)_+ \begin{pmatrix}\Delta_{ab} \\ \Delta_{ba}\end{pmatrix}.$$
3. Propose an algorithm in order to compute the solution $\widehat{\theta}$ of (7.10).
4. Show that if we add a second penalty $\gamma\|\theta\|_F^2$ to Criterion (7.10), the only change in the above algorithm is that the update is divided by $(1+\gamma)$.

7.6.5 Gaussian Graphical Models with Hidden Variables

We assume that $\{1,\ldots,p\} = O \cup H$, with $O \cap H = \emptyset$. We consider a Gaussian random variable $X = \begin{pmatrix}X_O \\ X_H\end{pmatrix} \sim \mathcal{N}(0,\Sigma)$ with $\Sigma = \begin{pmatrix}\Sigma_{OO} & \Sigma_{OH} \\ \Sigma_{HO} & \Sigma_{HH}\end{pmatrix}$. In particular, the variable X_O follows a $\mathcal{N}(0,\Sigma_{OO})$ distribution. We set $\tilde{K}_O = (\Sigma_{OO})^{-1}$.

1. Prove that $\tilde{K}_O = K_{OO} - K_{OH}(K_{HH})^{-1}K_{HO}$, where $K := \Sigma^{-1} = \begin{pmatrix} K_{OO} & K_{OH} \\ K_{HO} & K_{HH} \end{pmatrix}$.
2. Let g_O be the minimal (nondirected) graph such that $\mathscr{L}(X_O) \sim g_O$. Is the graph g_O a subgraph of the minimal graph g, such that $\mathscr{L}(X) \sim g$?
3. We assume that $\mathscr{L}(X) \sim g$ with g sparse. What can we say about the sparsity of K_{OO}? What can we say about the rank of $K_{OH}(K_{HH})^{-1}K_{HO}$ compared to $h = \text{card}(H)$?
4. Very often, we have variables $(X_1, \ldots, X_p)$, with conditional dependencies depicted by a sparse graph g, but we cannot observe all the variables. Furthermore, we even do not know the actual number of unobserved variables. In other words, we only observe an n-sample of $X_O \sim \mathscr{N}(0, \Sigma_{OO})$ and we have no information on X_H and the size of H. Nevertheless, we want to reconstruct on the basis of these observations graph g_O^* with nodes labelled by O and defined by

$$a \overset{g_O^*}{\sim} b \Longleftrightarrow a \overset{g}{\sim} b .$$

   Explain why it is equivalent to estimate the location of the nonzero coordinates of K_{OO}.
5. We have in mind that h is small. We have seen that $\tilde{K}_O = K_{OO} - L$, with K_{OO} sparse and $\text{rank}(L) \leq h$. Propose an estimation procedure inspired by the glasso.
6. Propose another procedure inspired by the regression approach.

7.6.6 Dantzig Estimation of Sparse Gaussian Graphical Models

Let $X^{(1)}, \ldots, X^{(n)}$ be a i.i.d. sample of a Gaussian distribution $\mathscr{N}(0, \Sigma)$ in $\mathbb{R}^p$, with Σ nonsingular and $K = \Sigma^{-1}$ sparse. In this exercise, we investigate the estimation of K with a matrix version of the Dantzig selector (see Exercise 4.5.5, page 96)

$$\widehat{K} \in \underset{B \in \mathbb{R}^{p \times p} : |\widehat{\Sigma} B - I|_\infty \leq \lambda}{\text{argmin}} |B|_{1,\infty}, \tag{7.16}$$

where $\widehat{\Sigma} = n^{-1} \sum_{i=1}^n X^{(i)} (X^{(i)})^T$ is the empirical covariance matrix, λ is non-negative and

$$|A|_\infty = \max_{i,j} |A_{ij}| \quad \text{and} \quad |A|_{1,\infty} = \max_j \sum_i |A_{ij}|.$$

Henceforth, A_j will refer to the j-th column of the matrix A.

1. Check that for any matrices of appropriate size, we have $|AB|_\infty \leq |A|_\infty |B|_{1,\infty}$ and also $|AB|_\infty \leq |A|_{1,\infty} |B|_\infty$ when A is symmetric.
2. We define the matrix $\widetilde{K}$ by $\widetilde{K}_{ij} = \widehat{\beta}_i^{(j)}$, where $\widehat{\beta}^{(j)}$ is solution of

$$\widehat{\beta}^{(j)} \in \underset{\beta \in \mathbb{R}^p : |\widehat{\Sigma} \beta - e_j|_\infty \leq \lambda}{\text{argmin}} |\beta|_1 ,$$

   with e_j the j-th vector of the canonical basis of $\mathbb{R}^p$. Prove that $\widetilde{K}$ is a solution of (7.16).

In the following, $\widehat{K}$ refers to the solution $\widetilde{K}$ defined above. This solution can be computed very efficiently, since it simply amounts to compute p Dantzig selectors.

A) Deterministic bound on $|\widehat{\mathbf{K}}-\mathbf{K}|_\infty$

In this part, we consider λ fulfilling

$$\lambda \geq |K|_{1,\infty}|\widehat{\Sigma}-\Sigma|_\infty. \tag{7.17}$$

1. Prove that when (7.17) is met, we have $|\widehat{\Sigma}K-I|_\infty \leq \lambda$, and therefore $|\widehat{K}_j|_1 \leq |K_j|_1$ for all $j=1,\ldots,p$.
2. Prove the inequalities
$$\begin{aligned}|\widehat{K}-K|_\infty &\leq |K|_{1,\infty}\left(|\widehat{\Sigma}\widehat{K}-I|_\infty+|(\Sigma-\widehat{\Sigma})\widehat{K}|_\infty\right)\\ &\leq |K|_{1,\infty}\left(\lambda+|K|_{1,\infty}|\widehat{\Sigma}-\Sigma|_\infty\right).\end{aligned}$$
3. Conclude that when (7.17) is met, we have
$$|\widehat{K}-K|_\infty \leq 2\lambda|K|_{1,\infty}. \tag{7.18}$$

B) Probabilistic bound on $|\widehat{\Sigma}-\Sigma|_\infty$

We assume henceforth that $\Sigma_{aa}=1$ for $a=1,\ldots,p$. Since Σ is nonsingular, we have $|\Sigma_{ab}|<1$ for all $a\neq b$.

1. Let X be a $\mathcal{N}(0,\Sigma)$ Gaussian random variable. We set $Z_1=(2(1+\Sigma_{12}))^{-1/2}(X_1+X_2)$ and $Z_2=(2(1-\Sigma_{12}))^{-1/2}(X_1-X_2)$. Prove that Z_1 and Z_2 are i.i.d., with $\mathcal{N}(0,1)$ Gaussian distribution and
$$X_1X_2=\frac{1}{4}\left(2(1+\Sigma_{12})Z_1^2-2(1-\Sigma_{12})Z_2^2\right).$$
2. Check that for $0\leq x\leq 1/2$, we have $-\log(1-x)\leq x+x^2$ and $-\log(1+x)\leq -x+x^2/2$ and prove the bound for $0\leq s\leq 1/4$
$$\mathbb{E}\left[e^{sX_1X_2}\right]=((1-(1+\Sigma_{12})s)(1+(1-\Sigma_{12})s))^{-1/2}\leq \exp(\Sigma_{12}s+2s^2).$$
Check that this bound still holds when $|\Sigma_{12}|=1$.
3. For any $a,b\in\{1,\ldots,p\}$, $0\leq s\leq 1/4$ and $t>0$, prove that
$$\mathbb{P}\left(\widehat{\Sigma}_{ab}-\Sigma_{ab}>t\right)\leq e^{-snt}e^{2s^2n}.$$
4. For $0<t\leq 1$, setting $s=t/4$, prove that
$$\mathbb{P}\left(|\widehat{\Sigma}-\Sigma|_\infty>t\right)\leq p(p+1)e^{-nt^2/8}.$$
5. For $\log(p)\leq n/32$, prove that
$$\mathbb{P}\left(|\widehat{\Sigma}-\Sigma|_\infty>4\sqrt{\frac{2\log(p)}{n}}\right)\leq\frac{2}{p^2}.$$

C) Bounds in sup norm and Frobenius norm

We define $d = \max_{j=1,\ldots,p} |K_j|_0$, which corresponds to the degree of the minimal graph associated to the Gaussian distribution $\mathcal{N}(0,\Sigma)$.

1. For $\lambda = 4|K|_{1,\infty}\sqrt{2\log(p)/n}$, by combining the results of parts A and B, prove that with probability at least $1-2/p^2$, we have
$$|\widehat{K}-K|_\infty \leq 8|K|_{1,\infty}^2\sqrt{\frac{2\log(p)}{n}}\,. \tag{7.19}$$
2. For any $J \subset \{1,\ldots,p\}^2$, we define the matrix K^J by $K^J_{ij} = K_{ij}\mathbf{1}_{(i,j)\in J}$. In the following, we set $J = \left\{(i,j) : |\widehat{K}_{ij}| > |\widehat{K}-K|_\infty\right\}$. Prove that for all $j \in \{1,\ldots,p\}$, we have
$$|\widehat{K}_j^{J^c}|_1 = |\widehat{K}_j|_1 - |\widehat{K}_j^J|_1 \leq |K_j|_1 - |\widehat{K}_j^J|_1 \leq |K_j - \widehat{K}_j^J|_1\,,$$
and hence $|\widehat{K}^{J^c}|_{1,\infty} \leq |K-\widehat{K}^J|_{1,\infty}$.
3. Prove that $K_{ij} \neq 0$ for all $(i,j) \in J$, and then
$$|\widehat{K}-K|_{1,\infty} \leq 2|\widehat{K}^J - K|_{1,\infty} \leq 2d|\widehat{K}^J - K|_\infty \leq 4d|\widehat{K}-K|_\infty\,.$$
4. Conclude that
$$\|\widehat{K}-K\|_F^2 \leq p|\widehat{K}-K|_{1,\infty}|\widehat{K}-K|_\infty \leq 4pd|\widehat{K}-K|_\infty^2.$$
5. Prove that when $\lambda = 4|K|_{1,\infty}\sqrt{2\log(p)/n}$, with probability at least $1-2/p^2$, we have
$$\|\widehat{K}-K\|_F^2 \leq 512pd|K|_{1,\infty}^4\,\frac{\log(p)}{n}\,.$$

The estimation procedure (7.16) is implemented in the `R` package `flare` available at `http://cran.r-project.org/web/packages/flare/`.

7.6.7 Gaussian Copula Graphical Models

We consider here a case where $X = (X_1,\ldots,X_p)$ is not a Gaussian random variable. We assume, as in Section 7.4, that there exists a random variable $Z = (Z_1,\ldots,Z_p)$, with $\mathcal{N}(0,\Sigma^Z)$ Gaussian distribution and p increasing differentiable functions $f_a : \mathbb{R}\to\mathbb{R}$, for $a = 1,\ldots,p$, such that $(X_1,\ldots,X_p) = (f_1(Z_1),\ldots,f_p(Z_p))$. We also assume that Σ^Z is nonsingular, and $\Sigma^Z_{aa} = 1$ for $a = 1,\ldots,p$. We define $K^Z = (\Sigma^Z)^{-1}$. We remind the reader of the classical equality when $\Sigma^Z_{aa} = 1$ for $a = 1,\ldots,p$
$$\mathbb{P}(Z_a > 0, Z_b > 0) = \frac{1}{4}\left(1+\frac{2}{\pi}\arcsin(\Sigma^Z_{ab})\right) \quad \text{for any } a,b \in \{1,\ldots,p\}\,.$$

1. Prove that the minimal graph g^* associated to $(X_1,\ldots,X_p)$ coincides with the minimal graph associated to $(Z_1,\ldots,Z_p)$. In particular, for $a \neq b$ there is an edge between a and b in g^* if and only if $K^Z_{ab} \neq 0$.

2. Let $\widetilde{Z}$ be an independent copy of Z and define $\widetilde{X} = (f_1(\widetilde{Z}_1), \ldots, f_p(\widetilde{Z}_p))$. Prove that for any $a, b \in \{1, \ldots, p\}$

$$\begin{aligned} \tau_{ab} &= \mathbb{E}\left[\operatorname{sign}\left((X_a - \widetilde{X}_a)(X_b - \widetilde{X}_b)\right)\right] \\ &= \mathbb{E}\left[\operatorname{sign}\left((Z_a - \widetilde{Z}_a)(Z_b - \widetilde{Z}_b)\right)\right] = \frac{2}{\pi}\arcsin(\Sigma^Z_{ab}), \end{aligned} \tag{7.20}$$

where $\operatorname{sign}(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x\leq 0}$.

A) Kendall's tau

We assume now that we only observe an i.i.d. sample $X^{(1)}, \ldots, X^{(n)}$ of X. In particular, the functions $f_1, \ldots, f_p$ and the covariance Σ^Z are unknown. Since the minimal graph g^* associated to $(X_1, \ldots, X_p)$ can be read on the precision matrix $K^Z = (\Sigma^Z)^{-1}$, our goal here is to estimate K^Z from the observations $X^{(1)}, \ldots, X^{(n)}$. The main idea is to build on the equation (7.20) in order to construct an estimator $\widehat{\Sigma}^Z$ of Σ^Z based on $X^{(1)}, \ldots, X^{(n)}$, and then apply the procedure (7.16) with $\widehat{\Sigma}$ replaced by $\widehat{\Sigma}^Z$.

For any $a, b \in \{1, \ldots, p\}$, we define $\widehat{\Sigma}^Z_{ab} = \sin(\pi\widehat{\tau}_{ab}/2)$ where

$$\widehat{\tau}_{ab} = \frac{2}{n(n-1)}\sum_{i<j}\operatorname{sign}\left((X^{(i)}_a - X^{(j)}_a)(X^{(i)}_b - X^{(j)}_b)\right).$$

1. Check that the function $F : \mathbb{R}^{2n} \to [-1, 1]$ defined by

$$F\big((x_1, y_1), \ldots, (x_n, y_n)\big) = \frac{2}{n(n-1)}\sum_{i<j}\operatorname{sign}\left((x_i - x_j)(y_i - y_j)\right)$$

fulfills

$$\left|F\big((x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_n,y_n)\big) - F\big((x_1,y_1),\ldots,(x'_i,y'_i),\ldots,(x_n,y_n)\big)\right| \leq \frac{4}{n}$$

for any $x_1, \ldots, x_n, y_1, \ldots, y_n, x'_i, y'_i \in \mathbb{R}$.

2. From McDiarmid concentration inequality (Theorem B.4, page 219, in Appendix B), prove that for $a < b$ and $t > 0$

$$\mathbb{P}\left(|\widehat{\tau}_{ab} - \tau_{ab}| > t\right) \leq 2e^{-nt^2/8} \quad \text{and} \quad \mathbb{P}\left(|\widehat{\Sigma}^Z_{ab} - \Sigma^Z_{ab}| > t\right) \leq 2e^{-nt^2/(2\pi^2)}.$$

3. Conclude that for any $t > 0$, we have

$$\mathbb{P}\left(|\widehat{\Sigma}^Z - \Sigma^Z|_\infty > t\right) \leq p^2 e^{-nt^2/(2\pi^2)} \text{ and } \mathbb{P}\left(|\widehat{\Sigma}^Z - \Sigma^Z|_\infty > 2\pi\sqrt{\frac{2\log(p)}{n}}\right) \leq \frac{1}{p^2}.$$

B) Graph estimation

Let $\widehat{K}^Z$ be the matrix obtained by solving the minimization problem (7.16), with $\widehat{\Sigma}$ replaced by $\widehat{\Sigma}^Z$. From $\widehat{K}^Z$, we build the graph $\widehat{g}$ by setting an edge between a and b if and only if both $\widehat{K}^Z_{ab}$ and $\widehat{K}^Z_{ba}$ are nonzero. We use below the same notations as in Exercise 7.6.6.

1. From (7.18), prove that when $\lambda = 2\pi|K^Z|_{1,\infty}\sqrt{2\log(p)/n}$, with probability at least $1-1/p^2$, we have
$$|\widehat{K}^Z - K^Z|_\infty \leq 4\pi|K^Z|_{1,\infty}^2\sqrt{\frac{2\log(p)}{n}}\,.$$

2. Assume that all the entries K^Z_{ab} of K^Z are either 0 or larger in absolute value than $4\pi|K^Z|_{1,\infty}^2\sqrt{2\log(p)/n}$. Prove that $g^* \subset \widehat{g}$, with probability at least $1-1/p^2$.

7.6.8 Restricted Isometry Constant for Gaussian Matrices

We prove in the following the upper Bound (7.15). The proof of the lower bound follows exactly the same lines. For a linear span $V \subset \mathbb{R}^p$ and an $n \times p$ matrix Z, we introduce the notation
$$\lambda_V(Z) = \sup_{v\in V\setminus\{0\}} \frac{n^{-1/2}\|Zv\|}{\|v\|}.$$

1. Define the matrix Z by $Z = \mathbf{X}\Sigma^{-1/2}$. Check that the Z_{ij} are i.i.d. with $\mathcal{N}(0,1)$ Gaussian distribution.

2. We write $\mathcal{M}$ for the set gathering all the subsets of $\{1,\ldots,p\}$. For $m \in \mathcal{M}$, we define S_m as the linear space spanned by $\{e_j : j \in m\}$, where $e_1,\ldots,e_p$ is the canonical basis of $\mathbb{R}^p$. To each linear span S_m, we associate the linear span $V_m = \Sigma^{1/2}S_m$. Check that for $d \leq n \wedge p$
$$\sup_{\beta:|\beta|_0\leq d} \frac{n^{-1/2}\|\mathbf{X}\beta\|}{\|\Sigma^{1/2}\beta\|} = \sup_{m\in\mathcal{M}:|m|=d} \lambda_{V_m}(Z). \tag{7.21}$$

In the following, we prove that for any collection $V_1,\ldots,V_N$ of d dimensional linear spaces, we have
$$\sup_{i=1,\ldots,N} \lambda_{V_i}(Z) \leq 1 + \frac{\sqrt{d}+\sqrt{2\log(N)}+\delta_N+\sqrt{2\xi}}{\sqrt{n}}, \tag{7.22}$$

where ξ is an exponential random variable and
$$\delta_N = \frac{\log\left(\frac{1}{N}+\sqrt{4\pi\log(N)}\right)}{\sqrt{2\log(N)}}.$$

The upper Bound (7.15) then follows by simply combining (7.21) with (7.22).

3. Let P_{V_i} be the orthogonal projector onto V_i. From Lemma 6.3 in Chapter 6, we know that the largest singular value $\sigma_1(ZP_{V_i})$ of ZP_{V_i} fulfills the inequality $\mathbb{E}[\sigma_1(ZP_{V_i})] \leq \sqrt{n}+\sqrt{d}$. Applying the Gaussian concentration inequality and noticing that $\sqrt{n}\lambda_{V_i}(Z) \leq \sigma_1(ZP_{V_i})$, prove that there exist some exponential random variables $\xi_1,\ldots,\xi_N$, such that for all $i = 1,\ldots,N$
$$\lambda_{V_i}(Z) \leq \mathbb{E}[\lambda_{V_i}(Z)] + \sqrt{2\xi_i/n} \leq 1+\sqrt{d/n}+\sqrt{2\xi_i/n}.$$

4. Applying again the Gaussian concentration inequality, prove that there exists an exponential random variable ξ, such that

$$\sup_{i=1,\dots,N} \lambda_{V_i}(Z) \leq 1 + \sqrt{d/n} + \mathbb{E}\left[\max_{i=1,\dots,N} \sqrt{2\xi_i/n}\right] + \sqrt{2\xi/n}.$$

5. Prove that for any $s > 0$,

$$\mathbb{E}\left[\max_{i=1,\dots,N} \sqrt{2\xi_i}\right] \leq s^{-1} \log\left(\sum_{i=1}^{N} \mathbb{E}\left[e^{s\sqrt{2\xi_i}}\right]\right).$$

6. Check that

$$\mathbb{E}\left[e^{s\sqrt{2\xi_i}}\right] = \int_0^\infty e^{su-u^2/2} u\,du \leq s\sqrt{2\pi}\,e^{s^2/2} + 1.$$

7. Choosing $s = \sqrt{2\log(N)}$, prove that $\mathbb{E}\left[\max_{i=1,\dots,N} \sqrt{2\xi_i}\right] \leq \sqrt{2\log(N)} + \delta_N$ and conclude the proof of (7.22).

Chapter 8

Multiple Testing

In this chapter, we switch from the estimation problem to the test problem, and we explain some possible ways to handle the impact of high-dimensionality in this context. More precisely, we will focus on the problem of performing simultaneously a large number of tests. This issue is of major importance in practice: Many scientific experiments seek to determine if a given factor has an impact on various quantities of interest. For example, we can seek for the possible side effects (headache, stomach pain, drowsiness, etc.) induced by a new drug. From a statistical perspective, this amounts to test *simultaneously* for each quantity of interest the hypothesis "the factor has no impact on this quantity" against "the factor has an impact on this quantity." As we have seen in the Chapter 1, considering simultaneously many different tests induces a loss in our ability to discriminate between the two hypotheses. We present in this chapter the theoretical basis for reducing at best this deleterious effect. We start by illustrating the issue on a simple example, and then we introduce the bases of False Discovery Rate control.

8.1 An Introductory Example

8.1.1 Differential Expression of a Single Gene

Let us assume that we have r measurements for the expression of a gene g in two different conditions A and B (corresponding, for example, to some normal cells and some cancerous cells).

Conditions	Measurements
A	$X_1^A,\ldots,X_r^A$
B	$X_1^B,\ldots,X_r^B$

We want to know if there is a difference in the expression of this gene between these two conditions A and B. In the formalism of test theory, we want to discriminate between the two hypotheses:

- $\mathscr{H}_0$: "the means of the X_i^A and the X_i^B are the same"
- $\mathscr{H}_1$: "the means of the X_i^A and the X_i^B are different"

A classical test statistic

Setting $Z_i = X_i^A - X_i^B$ for $i = 1,\dots,r$ we can reject $\mathscr{H}_0$ when

$$\widehat{S} := \frac{|\overline{Z}|}{\sqrt{\widehat{\sigma}^2/r}} \geq s = \text{some threshold,} \tag{8.1}$$

with $\overline{Z}$ the empirical mean of the Z_i and $\widehat{\sigma}^2$ the empirical variance of the Z_i. The threshold s_α is chosen, such that the probability to wrongly reject $\mathscr{H}_0$ is not larger than α. At the end, the test is simply $T = \mathbf{1}_{\widehat{S} \geq s_\alpha}$.

Case of Gaussian measurements

In the special case where

$$X_i^A \overset{\text{i.i.d.}}{\sim} \mathscr{N}(\mu_A, \sigma_A^2) \quad \text{and} \quad X_i^B \overset{\text{i.i.d.}}{\sim} \mathscr{N}(\mu_B, \sigma_B^2),$$

the hypothesis $\mathscr{H}_0$ corresponds to "$\mu_A = \mu_B$" and the hypothesis $\mathscr{H}_1$ corresponds to "$\mu_A \neq \mu_B$." In addition, if the variables $X_1^A, \dots, X_r^A, X_1^B, \dots, X_r^B$ are independent, then the $Z_i = X_i^A - X_i^B$ are i.i.d. with $\mathscr{N}(\mu_A - \mu_B, \sigma_A^2 + \sigma_B^2)$-Gaussian distribution. In this case, the statistic $\widehat{S}$ defined by (8.1) is distributed under the null hypothesis as the absolute value of a student random variable $\mathscr{T}(r-1)$, with $r-1$ degrees of freedom. The threshold s_α is then defined for $0 < \alpha < 1$ as the solution of

$$\mathbb{P}(|\mathscr{T}(r-1)| \geq s_\alpha) = \alpha.$$

It is common in the scientific literature to display the p-value of a test rather than the outcome of the test. In our case, the p-value of the test is simply the value $\widehat{\alpha}$, such that $s_{\widehat{\alpha}} = \widehat{S}$. Since $\alpha \to s_\alpha$ is decreasing, the outcome of a test of level α can be directly read from the p-value:

- If the p-value $\widehat{\alpha}$ is larger than α, then $s_\alpha > \widehat{S}$ and the hypothesis $\mathscr{H}_0$ is not rejected.
- If the p-value $\widehat{\alpha}$ is not larger than α, then $s_\alpha \leq \widehat{S}$ and the hypothesis $\mathscr{H}_0$ is rejected.

8.1.2 Differential Expression of Multiple Genes

DNA microarrays and the Next Generation Sequencing technologies allow us to measure the expression level of thousands of genes simultaneously. Our statistical objective is then to test *simultaneously* for all genes $g \in \{1,\dots,m\}$

- $\mathscr{H}_{0,g}$: "the mean expression levels of the gene g in conditions A and B are the same"
- $\mathscr{H}_{1,g}$: "the mean expression levels of the gene g in conditions A and B are different"

If we reject $\mathscr{H}_{0,g}$ when the p-value $\widehat{p}_g$ is not larger than α (or equivalently $\widehat{S}_g \geq s_\alpha$), then for each *individual* gene g, the probability to reject wrongly $\mathscr{H}_{0,g}$ is at most α. Nevertheless, if we consider the m genes *simultaneously* the number of hypotheses

$\mathscr{H}_{0,g}$ wrongly rejected (called false positives) can be high. Actually, the mean number of false positives is

$$\mathbb{E}\,[\text{False Positives}] \;=\; \sum_{g\,:\,\mathscr{H}_{0,g}\,\text{true}} \mathbb{P}_{\mathscr{H}_{0,g}}(\widehat{S}_g \geq s_\alpha) = \text{card}\left\{g : \mathscr{H}_{0,g} \text{ is true}\right\} \times \alpha$$

if the threshold s_α is such that $\mathbb{P}_{\mathscr{H}_{0,g}}(\widehat{S}_g \geq s_\alpha) = \alpha$ for every g. For example, for typical values like $\alpha = 5\%$ and card $\left\{g : \mathscr{H}_{0,g} \text{ is true}\right\} = 10000$, we obtain on average 500 false positives.

From a biological point of view, a gene that presents a differential expression between the two conditions A and B is a gene that is suspected to be involved in the response to the change of "environment" between A and B. Further experiments must be carried out in order to validate (or not) this "discovery." If in our list of genes suspected to present a differential expression there are 500 false positives, it means that we have 500 false discoveries, and a lot of time and money will be spent in useless experiments. Therefore, biologists ask for a list of genes that contains as few false positives as possible. Of course, the best way to avoid false positives is to declare no gene positive, but in this case, there is no discovery and the data are useless. In this chapter, we will present a couple of procedures designed in order to control the number of false discoveries in multiple testing settings while not losing too much in terms of power.

8.2 Statistical Setting

In the remainder of this chapter, we consider the following setting. We have m families of probability distribution $\{\mathbb{P}_\theta : \theta \in \Theta_i\}$ with $i = 1, \ldots, m$, and we consider simultaneously the m tests

$$\mathscr{H}_{0,i} : \theta \in \Theta_{0,i} \quad \text{against} \quad \mathscr{H}_{1,i} : \theta \in \Theta_{1,i} \qquad \text{for } i = 1, \ldots, m,$$

where $\Theta_{0,i}$ and $\Theta_{1,i}$ are two disjointed subsets of Θ_i. We assume that for each $i \in \{1, \ldots, m\}$, we have a test of the form $\mathbf{1}_{\widehat{S}_i \geq s_i}$, where $\widehat{S}_i$ is some observed statistic and s_i is some threshold value.

8.2.1 p-Values

For $\theta \in \Theta_i$, let us denote by $T_\theta(s) = \mathbb{P}_\theta(\widehat{S}_i \geq s)$ the tail distribution of the statistic $\widehat{S}_i$ under $\mathbb{P}_\theta$. The p-value associated to the statistic $\widehat{S}_i$ for the test i is

$$\widehat{p}_i = \sup_{\theta \in \Theta_{0,i}} T_\theta(\widehat{S}_i). \tag{8.2}$$

It corresponds to the maximum probability under the null hypothesis to observe a value for our statistic not smaller than the value $\widehat{S}_i$ that we have actually observed.

Th next proposition states that the p-values are stochastically larger than a uniform random variable under the null hypothesis.

Proposition 8.1 Distribution of p-values

The p-value $\widehat{p}_i$ defined by (8.2) fulfills the distributional property

$$\sup_{\theta \in \Theta_{0,i}} \mathbb{P}_\theta(\widehat{p}_i \le u) \le u, \quad \textit{for all } u \in [0,1]. \tag{8.3}$$

We say that under the null hypothesis, the p-values are stochastically larger than a uniform random variable.

Proof. For any $\theta \in \Theta_{0,i}$ and $u \in [0,1]$, we have

$$\mathbb{P}_\theta(\widehat{p}_i \le u) = \mathbb{P}_\theta\left(\sup_{\theta \in \Theta_{0,i}} T_\theta(\widehat{S}_i) \le u\right) \le \mathbb{P}_\theta\left(T_\theta(\widehat{S}_i) \le u\right). \tag{8.4}$$

For $u \in [0,1]$, we define $T_\theta^{-1}(u) = \inf\{s \in \mathbb{R} : T_\theta(s) \le u\}$. Since T_θ is nonincreasing, we have

$$]T_\theta^{-1}(u), +\infty[\subset \{s \in \mathbb{R} : T_\theta(s) \le u\} \subset [T_\theta^{-1}(u), +\infty[\,.$$

Let us consider apart the cases where $T_\theta(T_\theta^{-1}(u)) \le u$ and $T_\theta(T_\theta^{-1}(u)) > u$.

- When $T_\theta(T_\theta^{-1}(u)) \le u$, we have $\{s \in \mathbb{R} : T_\theta(s) \le u\} = [T_\theta^{-1}(u), +\infty[$ and then

$$\mathbb{P}_\theta\left(T_\theta(\widehat{S}_i) \le u\right) = \mathbb{P}_\theta\left(\widehat{S}_i \ge T_\theta^{-1}(u)\right) = T_\theta\left(T_\theta^{-1}(u)\right) \le u.$$

- When $T_\theta(T_\theta^{-1}(u)) > u$, we have $\{s \in \mathbb{R} : T_\theta(s) \le u\} =]T_\theta^{-1}(u), +\infty[$, and therefore

$$\mathbb{P}_\theta\left(T_\theta(\widehat{S}_i) \le u\right) = \mathbb{P}_\theta\left(\widehat{S}_i > T_\theta^{-1}(u)\right) = \lim_{\varepsilon \searrow 0} T_\theta\left(T_\theta^{-1}(u) + \varepsilon\right) \le u,$$

  where the last inequality comes from $T_\theta^{-1}(u) + \varepsilon \in \{s \in \mathbb{R} : T_\theta(s) \le u\}$ for all $\varepsilon > 0$.

Combining the two last displays with (8.4), we have proved that $\mathbb{P}_\theta(\widehat{p}_i \le u) \le u$ for all $\theta \in \Theta_{0,i}$ and $u \in [0,1]$. □

8.2.2 Multiple Testing Setting

A multiple testing procedure is a procedure that takes as input the vector of p-values $(\widehat{p}_1, \ldots, \widehat{p}_m)$ corresponding to the m tests and returns a set of indices $\widehat{R} \subset I = \{1, \ldots, m\}$, which gives the set of the null hypotheses $\left\{\mathscr{H}_{0,i} : i \in \widehat{R}\right\}$ that are rejected. Writing I_0 for the set

$$I_0 = \{i \in \{1, \ldots, m\} : \mathscr{H}_{0,i} \text{ is true}\}, \tag{8.5}$$

we call *false positive* (FP) the indices $i \in \widehat{R} \cap I_0$ and *true positive* (TP) the indices $i \in \widehat{R} \setminus I_0$. In the following, we will use the notations

$$\mathsf{FP} = \mathrm{card}(\widehat{R} \cap I_0) \quad \text{and} \quad \mathsf{TP} = \mathrm{card}(\widehat{R} \setminus I_0).$$

Ideally, we would like a procedure that selects $\widehat{R}$ in such a way that FP is small and TP is large. Of course, there is a balance to find between these two terms, since a severe control of the number of false positives usually induces a small number of true positives.

8.2.3 Bonferroni Correction

The Bonferroni correction provides a severe control of the number of false positives. It is designed in order to control the probability of existence of false positives $\mathbb{P}(\mathsf{FP} > 0)$. It is defined by

$$\widehat{R}_{\text{Bonf}} = \{i : \widehat{p}_i \leq \alpha/m\}.$$

Let us denote by m_0 the cardinality of I_0. According to Proposition 8.1, we have

$$\mathbb{P}(\mathsf{FP} > 0) = \mathbb{P}(\exists i \in I_0 : \widehat{p}_i \leq \alpha/m) \leq \sum_{i \in I_0} \sup_{\theta \in \Theta_{0,i}} \mathbb{P}_\theta(\widehat{p}_i \leq \alpha/m) \leq m_0\alpha/m \leq \alpha.$$

The probability of existence of false positives is thus smaller than α. The Bonferroni procedure avoids false positives but produces only a few true positives in general. Actually, it amounts to using the level α/m for each test, which can be very conservative when m is large.

In the next section, we will describe some procedures that control the (mean) proportion of false positives among $\widehat{R}$ instead of the absolute number of false positives.

8.3 Controlling the False Discovery Rate

The False Discovery Proportion (FDP) corresponds to the proportion $\mathsf{FP}/(\mathsf{FP}+\mathsf{TP})$ of false positives among the positives (with the convention $0/0 = 0$). The False Discovery Rate (FDR) is defined as the mean False Discovery Proportion

$$\mathsf{FDR} = \mathbb{E}\left[\frac{\mathsf{FP}}{\mathsf{FP}+\mathsf{TP}}\mathbf{1}_{\{\mathsf{FP}+\mathsf{TP}\geq 1\}}\right].$$

This quantity was introduced in the ’90s, and it is now widely used in science, especially in biostatistics.

8.3.1 Heuristics

Let us try to guess what could be a procedure that controls the FDR. To start with, we notice that if we want to have FP as small as possible and TP as large as possible, then the rejected p-values should correspond to the smallest p-values. So the only issue is to determine how many p-values can be rejected while keeping the FDR lower than α. Therefore, we will focus on rejection sets $\widehat{R}$ of the form

$$\widehat{R} = \{i \in I : \widehat{p}_i \leq t(\widehat{p}_1, \ldots, \widehat{p}_m)\}, \tag{8.6}$$

with $t : [0,1]^m \to [0,1]$. In the following, we will seek for some functions t that prevent from an FDR larger than α, while maximizing the size of $\widehat{R}$.

Let us investigate informally this point. According to (8.3), for a given threshold $\tau > 0$, the number FP of false positives in the rejection set $\widehat{R} = \{i : \widehat{p}_i \leq \tau\}$ fulfills

$$\mathbb{E}[\mathsf{FP}] = \mathbb{E}\left[\sum_{i\in I_0} \mathbf{1}_{\{\widehat{p}_i \leq \tau\}}\right] \leq \sum_{i\in I_0} \sup_{\theta\in\Theta_{0,i}} \mathbb{P}_\theta(\widehat{p}_i \leq \tau) \leq \mathrm{card}(I_0)\,\tau \leq m\tau. \qquad (8.7)$$

Let us denote by $\widehat{p}_{(1)} \leq \ldots \leq \widehat{p}_{(m)}$ the p-values ranked in a nondecreasing order. For any $\tau \in [\widehat{p}_{(k)}, \widehat{p}_{(k+1)}[$, we have $\{i : \widehat{p}_i \leq \tau\} = \left\{i : \widehat{p}_i \leq \widehat{p}_{(k)}\right\}$. Therefore, we only have to focus on a threshold $t(\widehat{p}_1, \ldots, \widehat{p}_m)$ in (8.6) of the form $t(\widehat{p}_1, \ldots, \widehat{p}_m) = \widehat{p}_{(\widehat{k})}$, with $\widehat{k} = \widehat{k}(\widehat{p}_1, \ldots, \widehat{p}_m)$. With this choice, we notice that $\mathrm{card}(\widehat{R}) = \widehat{k}$, so according to (8.7), we expect to have for $t(\widehat{p}_1, \ldots, \widehat{p}_m) = \widehat{p}_{(\widehat{k})}$

$$\frac{\mathsf{FP}}{\mathsf{FP}+\mathsf{TP}} = \frac{\mathsf{FP}}{\mathrm{card}(\widehat{R})} \overset{?}{\leq} \frac{m\widehat{p}_{(\widehat{k})}}{\widehat{k}} \quad \text{"on average."}$$

This nonrigorous computation suggests that we should have an FDR upper bounded by $\alpha > 0$, as soon as the integer $\widehat{k}$ fulfills

$$\widehat{p}_{(\widehat{k})} \leq \alpha\widehat{k}/m. \qquad (8.8)$$

Since we want to have the cardinality of $\widehat{R}$ as large as possible, and since $\mathrm{card}(\widehat{R}) = \widehat{k}$, we then choose $\widehat{k} = \max\left\{k : \widehat{p}_{(k)} \leq \alpha k/m\right\}$. The rejection set suggested by the above discussion is then $\widehat{R} = \left\{i \in I : \widehat{p}_i \leq \widehat{p}_{(\widehat{k})}\right\}$ or equivalently

$$\widehat{R} = \left\{i \in I : \widehat{p}_i \leq \alpha\widehat{k}/m\right\}, \quad \text{with } \widehat{k} = \max\left\{k : \widehat{p}_{(k)} \leq \alpha k/m\right\}.$$

Such a rejection set is illustrated in Figure 8.1.

8.3.2 Step-Up Procedures

The previous informal discussion suggests to choose a set of rejected hypotheses $\widehat{R}$ in the form

$$\widehat{R} = \left\{i \in I : \widehat{p}_i \leq \alpha\beta(\widehat{k})/m\right\} \quad \text{with} \quad \widehat{k} = \max\left\{k \in I : \widehat{p}_{(k)} \leq \alpha\beta(k)/m\right\} \qquad (8.9)$$

where $\beta : \{1, \ldots, m\} \to \mathbb{R}^+$ is some nondecreasing function, and $\widehat{p}_{(1)} \leq \ldots \leq \widehat{p}_{(m)}$ are the p-values ranked in nondecreasing order. When $\left\{k \in I : \widehat{p}_{(k)} \leq \alpha\beta(k)/m\right\} = \emptyset$, we set $\widehat{R} = \emptyset$. Figure 8.1 gives an illustration of the choice of $\widehat{k}$ for the function $\beta(k) = k$ suggested by the previous discussion. The next theorem provides a bound on the FDR of the procedure (8.9).

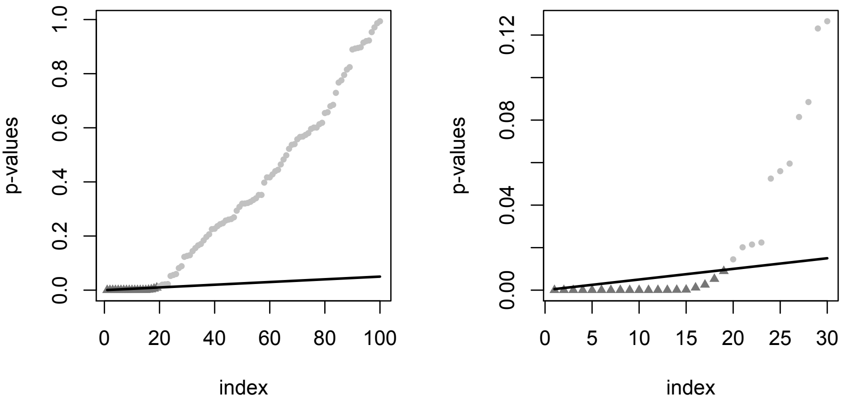


Figure 8.1 *Some p-values $\widehat{p}_{(1)} \leq \ldots \leq \widehat{p}_{(m)}$ ranked in increasing order and their position relative to the line $k \to \alpha k/m$ (in black). The right picture is a zoom-in of the left one. Dark gray triangles: p-values corresponding to rejected hypotheses. Light gray dots: p-values corresponding to nonrejected hypotheses.*

Theorem 8.2 General control of the FDR

Let $\beta : \{1, \ldots, m\} \to \mathbb{R}^+$ be a nondecreasing function and for $\alpha > 0$ define $\widehat{R}$ by (8.9), with the convention that $\widehat{R} = \emptyset$ when $\left\{k \in I : \widehat{p}_{(k)} \leq \alpha\beta(k)/m\right\} = \emptyset$.

Writing $m_0 = \operatorname{card}(I_0)$ with I_0 defined by (8.5), we have the following upper bound on the FDR *of the procedure*

$$\mathsf{FDR}(\widehat{R}) \leq \alpha\,\frac{m_0}{m} \sum_{j \geq 1} \frac{\beta(j \wedge m)}{j(j+1)}\,. \tag{8.10}$$

Proof. We use below the convention that $\widehat{k} = 0$ when $\widehat{R} = \emptyset$. We observe from the Definition (8.9) that $\widehat{k} = \operatorname{card}(\widehat{R})$. From the definition of the FDR, we have

$$\begin{aligned}
\mathsf{FDR}(\widehat{R}) &= \mathbb{E}\left[\frac{\operatorname{card}\left\{i \in I_0 : \widehat{p}_i \leq \alpha\beta(\widehat{k})/m\right\}}{\widehat{k}} \mathbf{1}_{\widehat{k} \geq 1}\right] \\
&= \sum_{i \in I_0} \mathbb{E}\left[\mathbf{1}_{\widehat{p}_i \leq \alpha\beta(\widehat{k})/m} \frac{\mathbf{1}_{\widehat{k} \geq 1}}{\widehat{k}}\right].
\end{aligned}$$

For $\widehat{k} \geq 1$, we have

$$\frac{1}{\widehat{k}} = \sum_{j \geq 1} \frac{\mathbf{1}_{j \geq \widehat{k}}}{j(j+1)}\,,$$

so applying first Fubini, then $\beta(\widehat{k}) \leq \beta(j \wedge m)$ for $j \geq \widehat{k}$, and finally Proposition 8.1, we obtain

$$\begin{aligned}
\mathsf{FDR}(\widehat{R}) &\overset{\text{Fubini}}{=} \sum_{i\in I_0}\sum_{j\geq 1}\frac{1}{j(j+1)}\underbrace{\mathbb{E}\left[\mathbf{1}_{j\geq\widehat{k}}\,\mathbf{1}_{\widehat{p}_i\leq\alpha\beta(\widehat{k})/m}\,\mathbf{1}_{\widehat{k}\geq 1}\right]}_{\leq\,\mathbb{P}\left(\widehat{p}_i\leq\alpha\beta(j\wedge m)/m\right)}\\
&\overset{\text{Prop. 8.1}}{\leq} \sum_{i\in I_0}\sum_{j\geq 1}\frac{1}{j(j+1)}\times\alpha\beta(j\wedge m)/m\\
&\leq \alpha\,\frac{m_0}{m}\sum_{j\geq 1}\frac{\beta(j\wedge m)}{j(j+1)}\,.
\end{aligned}$$

The proof of Theorem 8.2 is complete. □

We point out that the upper Bound (8.10) is sharp in the sense that for α small enough, there exist some distributions of the p-values $(\widehat{p}_1,\ldots,\widehat{p}_m)$, such that (8.10) is an equality; see Theorem 5.2 in Guo and Rao [71].

If we choose β in such a way that the sum fulfills the condition

$$\sum_{j\geq 1}\frac{\beta(j\wedge m)}{j(j+1)}\leq 1, \tag{8.11}$$

then the FDR of the procedure is less than α. We notice that for our guess $\beta(k)=k$ suggested by the informal discussion, the upper Bound (8.10) is equal to $\alpha m_0 H_m/m$ with $H_m = 1+1/2+\ldots+1/m$. In particular, this choice of β does not meet Condition (8.11). A popular choice of function β fulfilling Condition (8.11) is the linear function $\beta(k)=k/H_m$. The procedure (8.9) with this choice of β is called the Benjamini–Yekutieli procedure.

Corollary 8.3 FDR control of Benjamini–Yekutieli procedure

We set $H_m = 1+1/2+\ldots+1/m$. The multiple testing procedure defined by $\widehat{R}=\emptyset$ when $\left\{k\in I:\widehat{p}_{(k)}\leq\alpha k/(mH_m)\right\}=\emptyset$ and otherwise

$$\widehat{R}=\left\{i\in I:\widehat{p}_i\leq\alpha\widehat{k}/(mH_m)\right\},\quad \textit{with}\quad \widehat{k}=\max\left\{k\in I:\widehat{p}_{(k)}\leq\alpha k/(mH_m)\right\} \tag{8.12}$$

has an FDR *upper bounded by* α.

Another example of function β fulfilling (8.11) is

$$\beta(k)=\alpha\,\frac{k(k+1)}{2m},\quad \text{for } k=1,\ldots,m.$$

For this choice of β, we have $\beta(k)\geq k/H_m$ when $k+1\geq 2m/H_m$, so the resulting procedure tends to reject more hypotheses than the Benjamini–Yekutieli procedure when there are many positives.

8.3.3 FDR Control under the WPRDS Property

In the Benjamini–Yekutieli procedure, function β is linear with a slope $1/H_m \sim 1/\log(m)$. The smaller the slope, the less true positives we have. Since (8.10) is sharp for some distributions of the p-values, we know that we cannot choose a larger slope in general. Yet, we may wonder if in some cases we can choose a linear function β with a larger slope while keeping the FDR smaller than α.

The discussion in Section 8.3.1 suggests that for some distributions of the p-values, we may expect some FDR control with the choice $\beta(k) = k$ instead of $\beta(k) = k/H_m$ in Procedure (8.9). The choice $\beta(k) = k$ corresponds to the Benjamini–Hochberg procedure defined by

$$\widehat{R} = \left\{ i \in I : \widehat{p}_i \leq \alpha \widehat{k}/m \right\} \quad \text{where} \quad \widehat{k} = \max\left\{ k \in I : \widehat{p}_{(k)} \leq \alpha k/m \right\}, \tag{8.13}$$

with the convention $\widehat{R} = \emptyset$ when $\left\{ k \in I : \widehat{p}_{(k)} \leq \alpha k/m \right\} = \emptyset$. We give below a simple distributional condition on the p-values which ensures an FDR control at level α for the Benjamini–Hochberg procedure.

In the next definition, we will say that a function $g : [0,1]^m \to \mathbb{R}^+$ is nondecreasing if for any $p, q \in [0,1]^m$ such that $p_i \geq q_i$ for all $i = 1, \ldots, m$, we have $g(p) \geq g(q)$.

Weak Positive Regression Dependency (WPRDS)

The distribution of the p-values $(\widehat{p}_1, \ldots, \widehat{p}_m)$ is said to fulfill the Weak Positive Regression Dependency Property (WPRDS) if for any bounded measurable nondecreasing function $g : [0,1]^m \to \mathbb{R}^+$ and for all $i \in I_0$, the function

$$u \to \mathbb{E}\left[g(\widehat{p}_1, \ldots, \widehat{p}_m) \,|\, \widehat{p}_i \leq u \right] \quad \text{is nondecreasing} \tag{8.14}$$

on the interval $\{u \in [0,1] : \mathbb{P}(\widehat{p}_i \leq u) > 0\}$.

The set of distributions fulfilling the WPRDS property includes the independent distributions.

Lemma 8.4 Independent p-values fulfills the WPRDS
Assume that the $(\widehat{p}_i)_{i \in I_0}$ are independent random variables and that the $(\widehat{p}_i)_{i \in I \setminus I_0}$ are independent from the $(\widehat{p}_i)_{i \in I_0}$. Then, the distribution of $(\widehat{p}_1, \ldots, \widehat{p}_m)$ fulfills the WPRDS property.

Proof. Let us consider some $i \in I_0$ and some bounded measurable nondecreasing function $g : [0,1]^m \to \mathbb{R}^+$. With no lost of generality, we can assume (for notational simplicity) that $i = 1$. The random variable $\widehat{p}_1$ is independent of $(\widehat{p}_2, \ldots, \widehat{p}_m)$, so for any u such that $\mathbb{P}(\widehat{p}_1 \leq u) > 0$, we have

$$\mathbb{E}\left[g(\widehat{p}_1, \ldots, \widehat{p}_m) \,|\, \widehat{p}_1 \leq u \right] = \int_{(x_2, \ldots, x_m) \in [0,1]^{m-1}} \mathbb{E}\left[g(\widehat{p}_1, x_2, \ldots, x_m) \,|\, \widehat{p}_1 \leq u \right] \mathbb{P}(\widehat{p}_2 \in dx_2, \ldots, \widehat{p}_m \in dx_m).$$

To prove the lemma, we only need to check that $u \to \mathbb{E}[g(\widehat{p}_1,x_2,\ldots,x_m)\,|\,\widehat{p}_1 \leq u]$ is nondecreasing for all $x_2,\ldots,x_m \in [0,1]$. Since the function g is nondecreasing, the function $g_1 : x_1 \to g(x_1,x_2,\ldots,x_m)$ is also nondecreasing. Writing $g_1^{-1}(t) = \inf\{x \in [0,1] : g_1(x) \geq t\}$ for the right inverse of g_1, we have

$$\begin{aligned}
\mathbb{E}[g(\widehat{p}_1,x_2,\ldots,x_m)\,|\,\widehat{p}_1 \leq u] &= \mathbb{E}[g_1(\widehat{p}_1)\,|\,\widehat{p}_1 \leq u] \\
&= \int_0^{+\infty} \mathbb{P}(g_1(\widehat{p}_1) \geq t\,|\,\widehat{p}_1 \leq u)\,dt \\
&= \int_0^{+\infty} \mathbb{P}(\widehat{p}_1 \geq g_1^{-1}(t)\,|\,\widehat{p}_1 \leq u)\,dt.
\end{aligned}$$

To conclude, we simply notice that

$$u \to \mathbb{P}(\widehat{p}_1 \geq g_1^{-1}(t)\,|\,\widehat{p}_1 \leq u) = \left(1 - \frac{\mathbb{P}(\widehat{p}_1 < g_1^{-1}(t))}{\mathbb{P}(\widehat{p}_1 \leq u)}\right)_+$$

is nondecreasing for all $t \in \mathbb{R}^+$. □

Another example of p-values fulfilling the WPRDS property is the p-values associated to some $(\widehat{S}_1,\ldots,\widehat{S}_m)$ distributed according to a $\mathcal{N}(0,\Sigma)$ Gaussian distribution, with $\Sigma_{ij} \geq 0$ for all $i,j = 1,\ldots,m$; see Exercise 8.6.3. We refer to Benjamini and Yekutieli [24] for some other examples.

The next theorem ensures that the FDR of the Benjamini–Hochberg procedure (8.13) is upper bounded by α when the WPRDS property is met.

Theorem 8.5 FDR control of Benjamini–Hochberg procedure

When the distribution of the p-values fulfills the WPRDS property, the multiple testing procedure defined by $\widehat{R} = \emptyset$ *when* $\{k \in I : \widehat{p}_{(k)} \leq \alpha k/m\} = \emptyset$ *and otherwise*

$$\widehat{R} = \left\{i \in I : \widehat{p}_i \leq \alpha\widehat{k}/m\right\}, \quad \textit{with} \quad \widehat{k} = \max\left\{k \in I : \widehat{p}_{(k)} \leq \alpha k/m\right\},$$

has an FDR *upper bounded by* α.

Proof. We use again the convention $\widehat{k} = 0$ when $\widehat{R} = \emptyset$. Since $\widehat{k} = \text{card}(\widehat{R})$, we have

$$\begin{aligned}
\mathrm{FDR}(\widehat{R}) &= \mathbb{E}\left[\frac{\text{card}\left\{i \in I_0 : \widehat{p}_i \leq \alpha\widehat{k}/m\right\}}{\widehat{k}} \mathbf{1}_{\widehat{k}\geq 1}\right] \\
&= \sum_{i \in I_0} \mathbb{E}\left[\mathbf{1}_{\widehat{p}_i \leq \alpha\widehat{k}/m} \frac{\mathbf{1}_{\widehat{k}\geq 1}}{\widehat{k}}\right] \\
&= \sum_{i \in I_0} \sum_{k=1}^{m} \frac{1}{k} \mathbb{P}\left(\widehat{k} = k \text{ and } \widehat{p}_i \leq \alpha k/m\right) \\
&= \sum_{i \in I_0} \sum_{k=k_i^*}^{m} \frac{1}{k} \mathbb{P}\left(\widehat{k} = k\,|\,\widehat{p}_i \leq \alpha k/m\right) \mathbb{P}(\widehat{p}_i \leq \alpha k/m),
\end{aligned}$$

where $k_i^* = \inf\{k \in \mathbb{N} : \mathbb{P}(\widehat{p}_i \le \alpha k/m) > 0\}$ and with the convention that the sum from k_i^* to m is zero if $m < k_i^*$. By Proposition 8.1, we have $\mathbb{P}(\widehat{p}_i \le \alpha k/m) \le \alpha k/m$, so we obtain

$$\begin{aligned}\mathsf{FDR}(\widehat{R}) &\le \sum_{i\in I_0}\sum_{k=k_i^*}^{m} \frac{\alpha}{m}\,\mathbb{P}\left(\widehat{k}=k\,|\,\widehat{p}_i \le \alpha k/m\right)\\ &\le \frac{\alpha}{m}\sum_{i\in I_0}\sum_{k=k_i^*}^{m}\left[\mathbb{P}\left(\widehat{k}\le k\,|\,\widehat{p}_i \le \alpha k/m\right) - \mathbb{P}\left(\widehat{k}\le k-1\,|\,\widehat{p}_i \le \alpha k/m\right)\right].\end{aligned}$$

The function

$$g(\widehat{p}_1,\ldots,\widehat{p}_m) = \mathbf{1}_{\left\{\max\left\{j:\widehat{p}_{(j)}\le \alpha j/m\right\}\le k\right\}} = \mathbf{1}_{\{\widehat{k}\le k\}}$$

is nondecreasing with respect to $(\widehat{p}_1,\ldots,\widehat{p}_m)$, so the WPRDS property ensures that for $k \ge k_i^*$

$$\mathbb{P}\left(\widehat{k}\le k\,|\,\widehat{p}_i \le \alpha k/m\right) \le \mathbb{P}\left(\widehat{k}\le k\,|\,\widehat{p}_i \le \alpha(k+1)/m\right).$$

We then obtain a telescopic sum and finally

$$\mathsf{FDR}(\widehat{R}) \le \frac{\alpha}{m}\sum_{i\in I_0}\mathbf{1}_{k_i^*\le m}\,\mathbb{P}\left(\widehat{k}\le m\,\middle|\,\widehat{p}_i \le \frac{\alpha(m+1)}{m}\right) \le \frac{m_0}{m}\alpha \le \alpha.$$

The proof of Theorem 8.5 is complete. □

8.4 Illustration

We illustrate the implementation of the Benjamini–Hochberg procedure on a microarray data set from Golub *et al.* [70]. We first load the data that are in the `multtest` package available on the website `http://www.bioconductor.org`.

```
library(multtest)
library(stats)
data(golub)     # load the data
```

The golub data set is a 3051×38 matrix. Each row reports the expression level for $m = 3051$ genes. The first 27 columns correspond to patients with leukemia of type “ALL,” the last 11 columns correspond to patients with leukemia of type “AML.” Our goal is to find genes that have a differential expression between these two conditions. Therefore, for each gene we perform a t-test and we record the corresponding p-value.

```
golub1<-golub[,1:27]   # data for Leukemia ALL
golub2<-golub[,28:38]  # data for Leukemia AML
m<-3051
p<-rep(0,m)
# compute the p-values with a two-sample t-test
for (i in 1:m)  p[i]<-t.test(golub1[i,],golub2[i,])$p.value
```

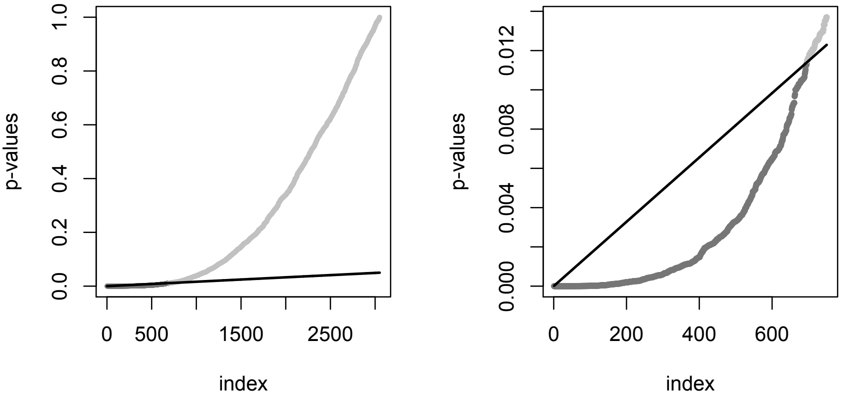


Figure 8.2 *The p-values ranked in increasing order. The dark-gray p-values correspond to rejected hypotheses, and the light-gray ones correspond to those not rejected. The black line represents the map $k \longrightarrow \alpha k/m$. The right picture is a zoom-in of the left one.*

We then compute the number $\widehat{k}$ of p-values rejected according to Formula (8.13) and the rejection set $\widehat{R}$. We also print the names of the genes for which a differential expression has been detected.

```
k<-sum(sort(p)<=0.05*(1:m)/m) # number of p-values rejected
R<-(1:m)[p<=0.05*k/m]         # rejection set
print(golub.gnames[R,2])      # print the names of the genes
```

Finally, we plot the rejected p-values; see Figure 8.2.

```
par(mfrow=c(1,2))
plot(1:m,sort(p),col=c(rep(2,k),rep(3,m-k)),type="p",pch=20,cex=0.7,
xlab="index",ylab="p-values")
points(1:m,0.05*(1:m)/m,col=1,type="l",lwd=2)
plot(1:750,sort(p)[1:750],type="p",col=c(rep(2,k),rep(3,750-k)),
pch=20,cex=0.7,xlab="index",ylab="p-values")
points(1:750,0.05*(1:750)/m,col=1,type="l",lwd=2)
```

We refer the interested reader to the paper by Dudoit, Fridlyand, and Speed [55] for a careful analysis of this data set.

8.5 Discussion and References

8.5.1 Take-Home Message

When we perform simultaneously m tests of level α the average number of false positives (tests for which $\mathscr{H}_0$ is wrongly rejected) is αm_0, where m_0 is the number of tests for which $\mathscr{H}_0$ is true. When m_0 is large, this number of false positives can be large, even larger than the number of true positives (tests for which $\mathscr{H}_0$ has been correctly rejected). In order to avoid this deleterious effect of multiple testing, we must apply a multiple testing selection procedure on the p-values of the m tests.

A first possibility is to control the probability of existence of false positives. This can be performed by the Bonferroni procedure which simply rejects the $\mathscr{H}_0$-hypotheses for which the p-values are not larger than α/m. The main drawback of this procedure is its lack of power for m large.

An alternative is to control the FDR, which is the mean ratio of the number of false positives by the total number of rejected hypotheses $\mathscr{H}_0$. The Benjamini–Yekutieli procedure offers such a control, but it can be more conservative than the Bonferroni procedure in some cases; see Exercise 8.6.1. The Benjamini–Hochberg procedure is a more powerful procedure, but it offers a control on the FDR only under some distributional hypotheses on the p-values. This last procedure is widely used in the scientific literature, especially in biology and medicine.

8.5.2 References

The FDR concept and the procedure (8.13) have been introduced in the seminal paper of Benjamini and Hochberg [23], while the procedure (8.12) has been introduced and analyzed by Benjamini and Yekutieli [24]. The proofs of Theorem 8.2 and Theorem 8.5 presented in these notes are adapted from Blanchard and Roquain [31].

It turns out that the control of the FDR obtained in Theorem 8.2 and Theorem 8.5 are of level $m_0\alpha/m$ instead of α. Since m_0 is unknown, we cannot directly correct this level. A lot of work has been done in order to achieve a better level, mainly by trying to estimate m_0. We refer the interested reader to the survey by Roquain [106] for references on this topic and many other issues related to FDR control. We finally refer to Goeman and Solari [68], Dudoit and van der Laan [56], and Dickhaus [52] for a recent review and two detailed books on multiple testing for genomics and life sciences.

8.6 Exercises

8.6.1 FDR versus FWER

The Family Wise Error Rate (FWER) is defined as $\mathsf{FWER} = \mathbb{P}(\mathsf{FP} > 0)$.

1. Prove that $\mathsf{FDR} \leq \mathsf{FWER}$.
2. What control of the FDR offers the Bonferroni procedure?

3. Prove that the number of hypotheses rejected by the Bonferroni procedure is smaller than the number of hypotheses rejected by the Benjamini–Hochberg procedure (8.13).
4. Check that the Benjamini–Yekutieli procedure (8.12) rejects more hypothesis than the Bonferroni procedure only when $\widehat{k} \geq H_m$, with $\widehat{k}$ and H_m defined in Corollary 8.3.

8.6.2 WPRDS Property

We prove in this exercise that if the distribution of the p-values $(\widehat{p}_1, \ldots, \widehat{p}_m)$ fulfills Property (8.14) for any nondecreasing indicator function $g = \mathbf{1}_\Gamma$, then it fulfills the WPRDS property. Below, $g : [0,1]^m \to \mathbb{R}^+$ denotes a bounded measurable nondecreasing function.

1. Prove that for any $i \in \{1, \ldots, m\}$ and u, such that $\mathbb{P}(\widehat{p}_i \leq u) > 0$, we have
$$\mathbb{E}\left[g(\widehat{p}_1, \ldots, \widehat{p}_m) \,|\, \widehat{p}_i \leq u\right] = \int_0^{+\infty} \mathbb{P}\left(g(\widehat{p}_1, \ldots, \widehat{p}_m) \geq t \,|\, \widehat{p}_i \leq u\right) dt\,.$$
2. Check that the indicator function $(\widehat{p}_1, \ldots, \widehat{p}_m) \to \mathbf{1}_{g^{-1}([t,+\infty[)}(\widehat{p}_1, \ldots, \widehat{p}_m)$ is nondecreasing for all $t \geq 0$, and conclude that $(\widehat{p}_1, \ldots, \widehat{p}_m)$ fulfills the WPRDS property.

8.6.3 Positively Correlated Normal Test Statistics

Assume that $(\widehat{S}_1, \ldots, \widehat{S}_m)$ is distributed according to a $\mathcal{N}(\mu, \Sigma)$-Gaussian distribution, with $\Sigma_{ij} \geq 0$ for all $i, j = 1, \ldots, m$. We want to test $\mathscr{H}_{0,i}$: "$\mu_i = 0$" against $\mathscr{H}_{1,i}$: "$\mu_i > 0$." We consider the tests $\mathbf{1}_{\widehat{S}_i \geq s_i}$ for $i = 1, \ldots, m$. The associated p-values are $\widehat{p}_i = T_i(\widehat{S}_i)$ with $T_i(s) = \mathbb{P}(\varepsilon_i \geq s)$, where ε_i has a $\mathcal{N}(0, \Sigma_{ii})$-Gaussian distribution.

For any vector $v \in \mathbb{R}^m$ and any $1 \leq i \leq m$, we define $v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_m)$. For two subsets A, B of $\{1, \ldots, p\}$, we denote by $\Sigma_{A,B}$ the matrix $[\Sigma_{ij}]_{i \in A, j \in B}$.

1. Check with Lemma A.4 in Appendix A that the conditional distribution of $\widehat{S}_{-i}$ given $\widehat{S}_i = x$ is the Gaussian distribution with mean $\mu_{-i} + \Sigma_{-i,i}(x - \mu_i)/\Sigma_{i,i}$ and covariance matrix $\Sigma_{-i,-i} - \Sigma_{-i,i}\Sigma_{i,-i}/\Sigma_{i,i}$.
2. For $u > 0$, prove that to any bounded measurable nondecreasing function $g : [0,1]^m \to \mathbb{R}^+$, we can associate a bounded measurable nonincreasing function $f : \mathbb{R}^m \to \mathbb{R}^+$, such that
$$\mathbb{E}\left[g(\widehat{p}_1, \ldots, \widehat{p}_m) \,|\, \widehat{p}_i \leq u\right] = \mathbb{E}\left[f(\widehat{S}_i, \widehat{S}_{-i}) \,|\, \widehat{S}_i \geq T_i^{-1}(u)\right],$$
with $T_i^{-1}(u) = \inf\{s \in \mathbb{R} : T_i(s) \leq u\}$.
3. We define $\phi(x) = \mathbb{E}\left[f(x, \mu_{-i} + \Sigma_{-i,i}(x - \mu_i)/\Sigma_{i,i} + \varepsilon_{-i})\right]$, where the random variable ε_{-i} follows a $\mathcal{N}(0, \Sigma_{-i,-i} - \Sigma_{-i,i}\Sigma_{i,-i}/\Sigma_{i,i})$ Gaussian distribution. Prove the

equalities

$$\begin{aligned}\mathbb{E}\left[g(\widehat{p}_1,\ldots,\widehat{p}_m)\,|\,\widehat{p}_i \leq u\right] &= \mathbb{E}\left[\phi(\widehat{S}_i)\,|\,\widehat{S}_i \geq T_i^{-1}(u)\right] \\ &= \int_0^{+\infty} \mathbb{P}\left(\phi(\widehat{S}_i) \geq t\,|\,\widehat{S}_i \geq T_i^{-1}(u)\right) dt\,.\end{aligned}$$

4. Check that ϕ is nonincreasing, and prove that the p-values associated with the tests $\mathbf{1}_{\widehat{S}_i \geq s_i}$ fulfill the WPRDS property.

Chapter 9

Supervised Classification

The goal of automatic classification is to predict at best the class y of an object x from some observations. A typical example is the spam filter of our mailbox, which predicts (more or less fairly) whether a mail is a spam or not. It is omnipresent in our daily life, by filtering the spams in our mailbox, reading automatically post-code on our postal letters, or recognizing faces on photos that we post on social networks. It is also very important in sciences, e.g., in medicine for early diagnosis of diseases from high-throughput data and in the industry, e.g., for detecting potential customers from their profiles.

In this chapter, we consider the setting of supervised classification: We have a data set recording the label (or class) y of n observed points (or objects) $x \in \mathscr{X}$, and we want to build from these data a function $h(x)$ that predicts the label of any point $x \in \mathscr{X}$. The function h is called a *classifier*. When we want to predict a label y with a function $h(x)$, we fall into the regression setting. Yet, the discrete nature of the labels allows us to strongly weaken the statistical modeling of the pair (x,y) by merely requiring that the observations $(X_i,Y_i)_{i=1,\ldots,n}$ are i.i.d.

We focus in this chapter on binary classification (only two classes). In Section 9.2, we present Vapnik's theory, which can be viewed as an analog of the model selection problem of Chapter 2 in the context of supervised classification. Due to the discrete nature of the labels y, the theory relies heavily on some combinatorial arguments. Similarly as for model selection, the classifier of Section 9.2 enjoys some good statistical properties, but suffers from a prohibitive computational cost. As in Chapter 4, some practical procedures can be built from a convex relaxation of the procedure of Section 9.2. We describe two convex relaxations in Section 9.3 leading to the popular Adaboost and Support Vector Machine (SVM) algorithms.

9.1 Statistical Modeling

9.1.1 Bayes Classifier

For the sake of simplicity, we restrict ourself in this chapter to the case where we only have two classes (as for the spam filter) labelled by -1 and $+1$. The problem of automatic classification can then be modeled as follows. Let $\mathscr{X}$ be some measurable space. Each outcome $X \in \mathscr{X}$ has a label $Y \in \{-1,+1\}$. We only observe the points

$X \in \mathscr{X}$, and our aim is to find a (measurable) function $h : \mathscr{X} \to \{-1,+1\}$, called *classifier*, such that $h(X)$ predicts at best the label Y.

Let us first quantify the prediction accuracy of a classifier h. Assume that the couple $(X,Y) \in \mathscr{X} \times \{-1,+1\}$ is sampled from a distribution $\mathbb{P}$. For a classifier $h : \mathscr{X} \to \{-1,+1\}$, the probability of misclassification is

$$L(h) = \mathbb{P}(Y \neq h(X)).$$

In the following, we will quantify the quality of a classifier h by its probability $L(h)$ of misclassification. This measure of quality is natural, yet we point out that some other measures can be more suited in some specific contexts. For example, while we do not care if a spam ends from time to time in our main mailbox, we definitely want to avoid that a personal mail ends in our spam box. In such a case, there is an asymmetry between the two different types of error, and another measure of quality should be considered.

Ideally, we would like to classify the data according to the classifier h_* minimizing the probability $L(h)$ of misclassification. Since $|Y - h(X)| \in \{0,2\}$, we have

$$L(h) = \frac{1}{4}\mathbb{E}\left[(Y-h(X))^2\right] = \frac{1}{4}\mathbb{E}\left[(Y-\mathbb{E}[Y|X])^2\right] + \frac{1}{4}\mathbb{E}\left[(\mathbb{E}[Y|X]-h(X))^2\right].$$

Therefore, $L(h)$ is minimal for the Bayes classifier

$$h_*(X) = \mathrm{sign}(\mathbb{E}[Y|X]) \quad \text{where} \quad \mathrm{sign(x)} = \mathbf{1}_{x>0} - \mathbf{1}_{x\leq 0} \quad \text{for } x \in \mathbb{R}. \tag{9.1}$$

If the distribution $\mathbb{P}$ were known, we would simply use the Bayes classifier h_* in order to have the smallest possible probability of misclassification. Unfortunately, the distribution $\mathbb{P}$ is usually unknown, so we cannot compute the Bayes classifier h_*.

In practice, we only have access to some training data $(X_i,Y_i)_{i=1,\ldots,n}$ i.i.d. with distribution $\mathbb{P}$, and our goal is to build from this training data a classifier $\widehat{h} : \mathscr{X} \to \{-1,+1\}$, such that $L(\widehat{h}) - L(h_*)$ is as small as possible.

9.1.2 Parametric Modeling

A first approach is to assume that the distribution $\mathbb{P}$ belongs to a parametric family of distributions. Conditioning on Y, we have

$$\mathbb{P}(X \in dx, Y=k) = \pi_k \mathbb{P}(X \in dx|Y=k), \quad \text{with } \pi_k = \mathbb{P}(Y=k) \text{ for } k \in \{-1,1\}.$$

So in order to parametrize the distribution $\mathbb{P}$, we only need to parametrize the two conditional distributions $\mathbb{P}(X \in dx|Y=1)$ and $\mathbb{P}(X \in dx|Y=-1)$.

A popular model when $\mathscr{X} = \mathbb{R}^d$ is to assume that the conditional distributions $\mathbb{P}(X \in dx|Y=k)$ are Gaussian with mean μ_k and covariance Σ_k. When we have $\Sigma_1 = \Sigma_{-1} = \Sigma$, the Bayes classifier is given by

$$h_*(x) = \mathrm{sign}\left(\left\langle \Sigma^{-1}(\mu_1 - \mu_{-1}), x - \frac{\mu_1+\mu_{-1}}{2}\right\rangle + \log(\pi_1/\pi_{-1})\right);$$

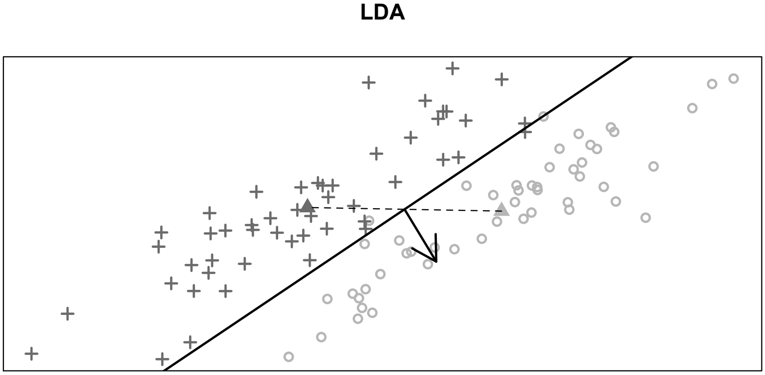


Figure 9.1 *Linear discriminant analysis. The triangles represent the means μ_1 and μ_{-1}, the arrow represents the vector $\widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_{-1})$, and the black line represents the frontier between $\{\widehat{h}_{LDA} = 1\}$ and $\{\widehat{h}_{LDA} = -1\}$.*

see Exercise 9.5.1. A point x is then classified according to its position relative to the affine hyperplane orthogonal to $\Sigma^{-1}(\mu_1 - \mu_{-1})$ and with offset $\log(\pi_1/\pi_{-1}) - (\mu_1 - \mu_{-1})^T\Sigma^{-1}(\mu_1 + \mu_{-1})/2$. In practice, we can classify the data with

$$\widehat{h}_{\mathrm{LDA}}(x) = \mathrm{sign}\left(\left\langle \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_{-1}), x - \frac{\widehat{\mu}_1 + \widehat{\mu}_{-1}}{2} \right\rangle + \log(\widehat{\pi}_1/\widehat{\pi}_{-1})\right), \qquad (9.2)$$

where $\widehat{\pi}_k$ is the empirical proportion of the label k, the mean $\widehat{\mu}_k$ is the empirical mean of the X_i, such that $Y_i = k$, and $\widehat{\Sigma}$ is the empirical covariance of the data. This leads to the *Linear Discriminant Analysis* (LDA). We refer to Exercise 9.5.1 for more details, and to Figure 9.1 for an illustration.

The parametric modeling is powerful when the model is correct, but in many cases, we do not know the distribution of the points X given the labels Y, and an incorrect modeling can lead to poor results; see Figure 9.2 for an illustration where the LDA fails.

9.1.3 Semi-Parametric Modeling

Since the Bayes classifier (9.1) only depends on the conditional distribution of Y given X, we can avoid to model the distribution of X as above. From the formula

$$\mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X) - \mathbb{P}(Y = -1|X) = 2\mathbb{P}(Y = 1|X) - 1,$$

we obtain that the Bayes classifier $h_*(x)$ is given by

$$h_*(x) = \mathrm{sign}\big(\mathbb{P}(Y = 1|X = x) - 1/2\big), \quad \text{for } x \in \mathscr{X}.$$

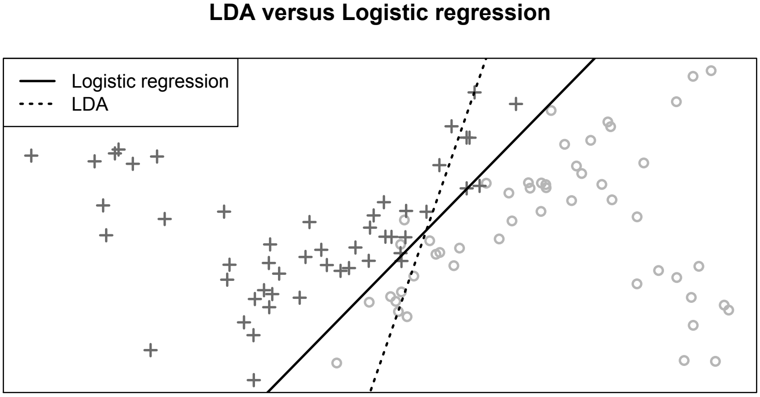


Figure 9.2 *The black line represents the separating hyperplane of the logistic regression. The dashed line represents the separating hyperplane of the LDA.*

A classical approach is to assume a parametric model for the conditional probability $\mathbb{P}(Y=1|X=x)$. The most popular model in $\mathbb{R}^d$ is probably the *logistic model*, where

$$\mathbb{P}(Y=1|X=x)=\frac{\exp(\alpha+\langle\beta,x\rangle)}{1+\exp(\alpha+\langle\beta,x\rangle)} \quad \text{for all } x\in\mathscr{X}, \tag{9.3}$$

with $\alpha\in\mathbb{R}$ and $\beta\in\mathbb{R}^d$. In this case, we have $\mathbb{P}(Y=1|X=x)>1/2$ if and only if $\exp(\alpha+\langle\beta,x\rangle)>1$, so the Bayes classifier h_* has the simple form

$$h_*(x)=\operatorname{sign}\big(\alpha+\langle\beta,x\rangle\big) \quad \text{for all } x\in\mathscr{X}.$$

We observe that the frontier between $\{h_*=1\}$ and $\{h_*=-1\}$ is again an affine hyperplane, with orthogonal direction β and offset α.

We can estimate the parameters (α,β) by maximizing the conditional likelihood of Y given X

$$(\widehat{\alpha},\widehat{\beta})\in \underset{(\alpha,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmax}}\left\{\prod_{i:Y_i=1}\frac{\exp(\alpha+\langle\beta,X_i\rangle)}{1+\exp(\alpha+\langle\beta,X_i\rangle)}\prod_{i:Y_i=-1}\frac{1}{1+\exp(\alpha+\langle\beta,X_i\rangle)}\right\},$$

and compute the classifier $\widehat{h}_{\mathsf{logistic}}(x)=\operatorname{sign}\big(\widehat{\alpha}+\langle\widehat{\beta},x\rangle\big)$ for all $x\in\mathscr{X}$. We emphasize that even if the Bayes classifiers have the same shape in the LDA and in the logistic modeling, the two procedures do not lead to the same classifier in general. In particular, if the conditional distribution of the X given Y is far from Gaussian, the LDA can produce some very poor results, while the logistic model will work as long as the modeling (9.3) remains valid; see Figure 9.2 for such a case.

9.1.4 Nonparametric Modeling

We may wish to weaken further our hypotheses on the distribution of (X,Y) and adopt a nonparametric point of view. Instead of assuming that the distribution of (X,Y) belongs to some parametric or semi parametric set of distributions (as in Sections 9.1.2 and 9.1.3), we will rather assume that h_* is "smooth" in some sense (suited to the classification setting).

A classical approach in nonparametric estimation is to replace the ideal risk minimization by some constrained empirical risk minimization. In our case, we cannot minimize $h \to L(h)$, since $L(h)$ is unknown, but we can minimize instead the empirical probability of misclassification

$$\widehat{L}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)} = \widehat{\mathbb{P}}_n(Y \neq h(X)), \tag{9.4}$$

where $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$. There is in general no unique minimizer of $\widehat{L}_n$, and even if an unconstrained minimizer of the empirical risk perfectly classifies the labels in the data set, it produces in general a very poor prediction for a new point. We must then restrict the minimization of $h \to \widehat{L}_n(h)$ to a set $\mathscr{H}$ of classifiers with limited "flexibility"

$$\widehat{h}_{\mathscr{H}} \in \operatorname*{argmin}_{h \in \mathscr{H}} \widehat{L}_n(h), \quad \text{with } \widehat{L}_n \text{ defined by (9.4).} \tag{9.5}$$

As we will see in the next section, the appropriate notion of "flexibility" in this context corresponds to some combinatorial complexity of the set $\mathscr{H}$ measuring the classification flexibility offered by the classifiers in $\mathscr{H}$. We stress that the set $\mathscr{H}$ of classifiers, usually called *dictionary*, plays the same role as the model S in the model selection setting of Chapter 2. In particular, we face the same issues as in Chapter 2: How does $\widehat{h}_{\mathscr{H}}$ behave compared to h_*? Which dictionary $\mathscr{H}$ should be chosen? We investigate these two issues in the next section.

9.2 Empirical Risk Minimization

We analyze in this section the classifier $\widehat{h}_{\mathscr{H}}$ defined by (9.5), and we explain how we can handle the problem of the choice of $\mathscr{H}$. Decomposing the difference between the misclassification probabilities $L(\widehat{h}_{\mathscr{H}})$ and $L(h_*)$, we find

$$0 \leq L(\widehat{h}_{\mathscr{H}}) - L(h_*) = \underbrace{\min_{h \in \mathscr{H}} L(h) - L(h_*)}_{\text{approximation error}} + \underbrace{L(\widehat{h}_{\mathscr{H}}) - \min_{h \in \mathscr{H}} L(h)}_{\text{stochastic error}}.$$

The first term is a bias term that measures the ability of the classifiers $h \in \mathscr{H}$ to produce a classification as good as the Bayes classifier h_*. This approximation error is purely deterministic, and enlarging the dictionary $\mathscr{H}$ can only reduce it. The second term measures the error made by minimizing over $h \in \mathscr{H}$ the empirical misclassification probability $\widehat{L}_n(h)$ instead of the true misclassification probability $L(h)$. This term

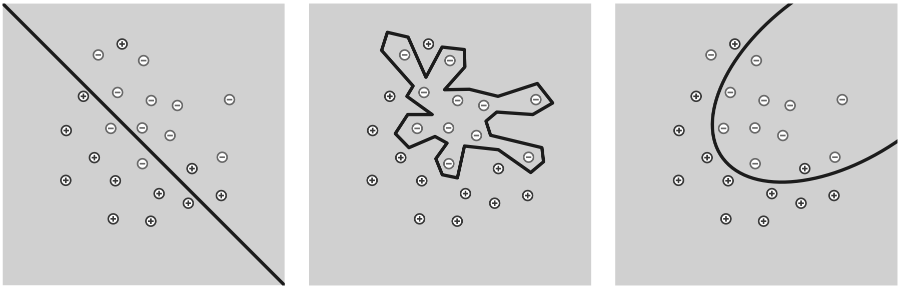


Figure 9.3 *Examples of classification produced by different dictionaries. Left: dictionary of linear classifiers $\mathscr{H}_{\mathsf{lin}}$. Center: dictionary of polygon classifiers $\mathscr{H}_{\mathsf{poly}}$. Right: dictionary of quadratic forms.*

is stochastic, and it tends to increase when $\mathscr{H}$ increases. This phenomenon is illustrated in Figure 9.3. In this illustration, with $\mathscr{X}=\mathbb{R}^2$, the classifiers of the dictionary $\mathscr{H}_{\mathsf{lin}}=\{h(x)=\text{sign}(\langle w,x\rangle):\|w\|=1\}$ are not flexible enough and they produce a poor classification. In this case, the approximation error is large. On the other hand, the classifiers of the dictionary $\mathscr{H}_{\mathsf{poly}}=\{h(x)=2\mathbf{1}_A(x)-1: A \text{ polygon in } \mathscr{X}\}$ are very flexible and can always classify exactly the data $(X_i,Y_i)_{i=1,\dots,n}$ when the X_i are distinct. The empirical error $\widehat{L}_n(\widehat{h}_{\mathscr{H}_{\mathsf{poly}}})$ is then 0, but $\widehat{h}_{\mathscr{H}_{\mathsf{poly}}}$ tends to produce a poor classification of new data (X,Y), and the stochastic term $L(\widehat{h}_{\mathscr{H}})-\min_{h\in\mathscr{H}}L(h)$ is large. The last example, based on a less flexible set of quadratic classifiers, produces a better result, even though its empirical error is larger than the one of $\widehat{h}_{\mathscr{H}_{\mathsf{poly}}}$.

To choose a good dictionary $\mathscr{H}$, we shall then find a good balance between the approximation properties of $\mathscr{H}$ and its size. The first step toward a procedure for selecting the dictionary $\mathscr{H}$ is to assess the misclassification probability of the empirical risk minimizer $\widehat{h}_{\mathscr{H}}$.

9.2.1 Misclassification Probability of the Empirical Risk Minimizer

As mentioned above, increasing the size of $\mathscr{H}$ tends to increase the stochastic error $L(\widehat{h}_{\mathscr{H}})-\min_{h\in\mathscr{H}}L(h)$. Actually, it is not really the size of the dictionary that matters, but rather its flexibility in terms of classification. For example, we cannot classify correctly the three labeled points $\{((0,1),+1),((1,1),-1),((1,0),+1)\}$ with a classifier in $\mathscr{H}_{\mathsf{lin}}$. Conversely, for any set of labeled points $(x_i,y_i)_{i=1,\dots,n}$ with distinct $x_1,\dots,x_n$, there exists $h\in\mathscr{H}_{\mathsf{poly}}$ such that $h(x_i)=y_i$.

In order to capture this classification flexibility, we introduce the shattering coefficient

$$\mathbb{S}_n(\mathscr{H})=\max_{(x_1,\dots,x_n)\in\mathscr{X}^n}\text{card}\{(h(x_1),\dots,h(x_n)):h\in\mathscr{H}\}\,, \tag{9.6}$$

which gives the maximal number of different labeling of n points that the classifiers in $\mathscr{H}$ can produce. For example, since n distinct points can be arbitrarily labelled with classifiers in $\mathscr{H}_{\mathsf{poly}}$, we have $\mathbb{S}_n(\mathscr{H}_{\mathsf{poly}}) = 2^n$. On the contrary, the number of possible labeling of n points with classifiers in $\mathscr{H}_{\mathsf{lin}}$ is more limited. Actually, Proposition 9.6 in Section 9.2.2 ensures that $\mathbb{S}_n(\mathscr{H}_{\mathsf{lin}}) \leq (n+1)^2$ in dimension 2. The next theorem provides an upper bound on the stochastic error and a confidence interval for the misclassification probability $L(\widehat{h}_{\mathscr{H}})$ in terms of the shattering coefficient (9.6). Again, the proof relies on a concentration inequality.

Theorem 9.1 Control of the stochastic error

For any $t > 0$, with probability at least $1 - e^{-t}$, we have

$$L(\widehat{h}_{\mathscr{H}}) - \min_{h\in\mathscr{H}} L(h) \leq 4\sqrt{\frac{2\log(2\mathbb{S}_{\mathscr{H}}(n))}{n}} + \sqrt{\frac{2t}{n}} \tag{9.7}$$

and

$$\left|L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}})\right| \leq 2\sqrt{\frac{2\log(2\mathbb{S}_{\mathscr{H}}(n))}{n}} + \sqrt{\frac{t}{2n}}\,. \tag{9.8}$$

Proof. We split the proof of Theorem 9.1 into three lemmas. The first lemma shows that the left-hand terms in (9.7) and (9.8) can be upper bounded in terms of the maximum difference over $\mathscr{H}$ between the empirical misclassification probability and the true misclassification probability

$$\widehat{\Delta}_n(\mathscr{H}) = \sup_{h\in\mathscr{H}} \left|\widehat{L}_n(h) - L(h)\right|. \tag{9.9}$$

The second lemma ensures that the difference between $\widehat{\Delta}_n(\mathscr{H})$ and its expectation is smaller than $\sqrt{\xi/(2n)}$, with ξ a standard exponential random variable. Finally, the third lemma upper bounds the expectation of $\widehat{\Delta}_n(\mathscr{H})$ by $2\sqrt{2\log(2\mathbb{S}_{\mathscr{H}}(n))/n}$.

Lemma 9.2

We have the upper bounds

$$L(\widehat{h}_{\mathscr{H}}) - \min_{h\in\mathscr{H}} L(h) \leq 2\widehat{\Delta}_n(\mathscr{H}) \quad \text{and} \quad \left|L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}})\right| \leq \widehat{\Delta}_n(\mathscr{H}).$$

Proof of Lemma 4.12. For any $h \in \mathscr{H}$, we have $\widehat{L}_n(\widehat{h}_{\mathscr{H}}) \leq \widehat{L}_n(h)$, and therefore

$$\begin{aligned} L(\widehat{h}_{\mathscr{H}}) - L(h) &= L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}}) + \widehat{L}_n(\widehat{h}_{\mathscr{H}}) - L(h) \\ &\leq L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}}) + \widehat{L}_n(h) - L(h) \\ &\leq 2\widehat{\Delta}_n(\mathscr{H}). \end{aligned}$$

Since this inequality is true for any $h \in \mathscr{H}$, the first bound of Lemma 9.2 follows. The second bound is obvious. □

In order to prove Theorem 9.1, it remains to prove that

$$\widehat{\Delta}_n(\mathscr{H}) \leq 2\sqrt{\frac{2\log(2\mathbb{S}_{\mathscr{H}}(n))}{n}} + \sqrt{\frac{t}{2n}},$$

with probability at least $1-e^{-t}$. The first step is a concentration inequality for $\widehat{\Delta}_n(\mathscr{H})$.

Lemma 9.3

With probability at least $1-e^{-t}$, *we have*

$$\widehat{\Delta}_n(\mathscr{H}) \leq \mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] + \sqrt{\frac{t}{2n}}\,.$$

Proof of Lemma 9.3. We have $\widehat{\Delta}_n(\mathscr{H}) = F((X_1,Y_1),\ldots,(X_n,Y_n))$, with

$$\begin{array}{rccl} F & : & (\mathscr{X}\times\{-1,+1\})^n & \to \mathbb{R} \\ & & ((x_1,y_1),\ldots,(x_n,y_n)) & \mapsto \dfrac{1}{n}\sup_{h\in\mathscr{H}}\left|\sum_{i=1}^n \mathbf{1}_{y_i\neq h(x_i)} - L(h)\right|. \end{array}$$

For any $(x_1,y_1),\ldots,(x_n,y_n),(x_i',y_i') \in \mathscr{X}\times\{-1,+1\}$, we have

$$\left|F((x_1,y_1),\ldots,(x_i',y_i'),\ldots,(x_n,y_n)) - F((x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_n,y_n))\right| \leq \frac{1}{n},$$

so according to McDiarmid concentration inequality (Theorem B.4, page 219, in Appendix B), with probability at least $1-e^{-2ns^2}$, we have $\widehat{\Delta}_n(\mathscr{H}) \leq \mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] + s$. Lemma 9.3 follows by setting $s = \sqrt{t/(2n)}$. □

It remains to bound the expectation of $\widehat{\Delta}_n(\mathscr{H})$ in terms of $\mathbb{S}_{\mathscr{H}}(n)$.

Lemma 9.4

For any dictionary $\mathscr{H}$, *we have the upper bound*

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] \leq 2\sqrt{\frac{2\log(2\mathbb{S}_{\mathscr{H}}(n))}{n}}\,.$$

Proof of Lemma 9.4. The proof of Lemma 9.4 is divided into two parts. The first part **(i)** is based on a classical and elegant symmetrization argument. The second part **(ii)** is a classical application of Jensen inequality.

(i) The first step is to upper bound $\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right]$ with the symmetrization lemma

(Theorem B.7, page 223, in Appendix B). Let $\sigma_1, \dots, \sigma_n$ be n i.i.d. random variables uniformly distributed on $\{-1, +1\}$ and independent of $(X_i, Y_i)_{i=1,\dots,n}$. According to (B.5), with $Z_i = (X_i, Y_i)$ for $i = 1, \dots, n$ and $f(Z_i) = \mathbf{1}_{Y_i \neq h(X_i)}$, we obtain

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] \leq 2\mathbb{E}\mathbb{E}_\sigma\left[\sup_{h\in\mathscr{H}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{Y_i\neq h(X_i)}\right|\right],$$

where $\mathbb{P}_\sigma$ refers to the expectation with respect to $\sigma_1, \dots, \sigma_n$. This term can be simply upper bounded by

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] \leq 2 \max_{y\in\{-1,+1\}^n} \max_{x\in\mathscr{X}^n} \mathbb{E}_\sigma\left[\sup_{h\in\mathscr{H}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{y_i\neq h(x_i)}\right|\right].$$

At this point, we notice that we have replaced an expectation with respect to the unknown probability distribution $\mathbb{P}$ by an expectation with respect to the known probability distribution $\mathbb{P}_\sigma$.

(ii) For any $(x,y) \in \mathscr{X}^n \times \{-1,+1\}^n$, let us define the set

$$\mathscr{V}_{\mathscr{H}}(x,y) = \left\{(\mathbf{1}_{y_1\neq h(x_1)}, \dots, \mathbf{1}_{y_n\neq h(x_n)}) : h \in \mathscr{H}\right\}.$$

The last upper bound on $\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right]$ can be written as

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] \leq \frac{2}{n} \times \max_{y\in\{-1,+1\}^n} \max_{x\in\mathscr{X}^n} \mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}_{\mathscr{H}}(x,y)} |\langle\sigma, v\rangle|\right],$$

where $\langle x,y\rangle$ is the canonical scalar product on $\mathbb{R}^n$. We notice that for any $y \in \{-1,1\}^n$ there is a bijection between $\mathscr{V}_{\mathscr{H}}(x,y)$ and the set $\{(h(x_1), \dots, h(x_n)) : h \in \mathscr{H}\}$. As a consequence, we have the upper bound

$$\max_{y\in\{-1,+1\}^n} \max_{x\in\mathscr{X}^n} \operatorname{card}(\mathscr{V}_{\mathscr{H}}(x,y)) \leq \mathbb{S}_n(\mathscr{H}).$$

In view of the last two inequalities, in order to conclude the proof of Lemma 9.4, it simply remains to prove the following result.

Lemma 9.5

For $\sigma_1, \dots, \sigma_n$ i.i.d. with $\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2$, we have

$$\mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}} |\langle\sigma, v\rangle|\right] \leq \sqrt{2n\log(2\operatorname{card}(\mathscr{V}))}, \quad \textit{for any finite } \mathscr{V} \subset [-1,1]^n. \tag{9.10}$$

Proof of Lemma 9.5. Writing $\mathscr{V}^\# = \mathscr{V} \cup -\mathscr{V}$, Jensen inequality ensures that for any

$s > 0$

$$\begin{aligned}\mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}}|\langle\sigma,v\rangle|\right] &= \mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}^\#}\langle\sigma,v\rangle\right] \leq \frac{1}{s}\log\mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}^\#}e^{s\langle\sigma,v\rangle}\right] \\ &\leq \frac{1}{s}\log\left(\sum_{v\in\mathscr{V}^\#}\mathbb{E}_\sigma\left[e^{s\langle\sigma,v\rangle}\right]\right). \qquad (9.11)\end{aligned}$$

Combining the facts that the σ_i are independent, $(e^x + e^{-x}) \leq 2e^{x^2/2}$ for all $x \in \mathbb{R}$ and $v_i^2 \leq 1$ for all $v \in \mathscr{V}^\#$, we have

$$\mathbb{E}_\sigma\left[e^{s\langle\sigma,v\rangle}\right] = \prod_{i=1}^n \mathbb{E}_\sigma\left[e^{sv_i\sigma_i}\right] = \prod_{i=1}^n \frac{1}{2}(e^{sv_i}+e^{-sv_i}) \leq \prod_{i=1}^n e^{s^2v_i^2/2} \leq e^{ns^2/2}.$$

Plugging this inequality in (9.11), we obtain

$$\mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}}|\langle\sigma,v\rangle|\right] \leq \frac{\log(\mathrm{card}(\mathscr{V}^\#))}{s} + \frac{ns}{2} \qquad \text{for any } s > 0.$$

The right-hand side is minimal for $s = \sqrt{2\log(\mathrm{card}(\mathscr{V}^\#))/n}$, which gives the upper bound

$$\mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}}|\langle\sigma,v\rangle|\right] \leq \sqrt{2n\log(\mathrm{card}(\mathscr{V}^\#))}\,.$$

We finally obtain (9.10) by noticing that $\mathrm{card}(\mathscr{V}^\#) \leq 2\,\mathrm{card}(\mathscr{V})$. □

The proof of Lemma 9.4 is complete, and Bounds (9.7) and (9.8) are obtained by combining Lemma 9.2, Lemma 9.3, and Lemma 9.4. □

Theorem 9.1 provides a control of the misclassification probability in terms of the shattering coefficient (9.6). The shattering coefficient offers a good notion of complexity for a set $\mathscr{H}$ of classifiers, but its computation can be tricky in practice. In the next section, we prove that a nice combinatorial property of the shattering coefficients provides a simple upper bound on $\mathbb{S}_n(\mathscr{H})$, depending on $\mathscr{H}$ only through a single quantity, the Vapnik–Chervonenkis dimension of $\mathscr{H}$.

9.2.2 Vapnik–Chervonenkis Dimension

By convention, we set $\mathbb{S}_0(\mathscr{H}) = 1$. From the definition (9.6) of $\mathbb{S}_n(\mathscr{H})$, we have $\mathbb{S}_n(\mathscr{H}) \leq 2^n$ for all $n \in \mathbb{N}$. We call Vapnik–Chervonenkis (VC) dimension of $\mathscr{H}$ the integer $d_{\mathscr{H}}$ defined by

$$d_{\mathscr{H}} = \sup\left\{d \in \mathbb{N} : \mathbb{S}_d(\mathscr{H}) = 2^d\right\} \in \mathbb{N}\cup\{+\infty\}. \qquad (9.12)$$

It corresponds to the maximum number of points in $\mathscr{X}$ that can be arbitrarily classified by the classifiers in $\mathscr{H}$. The next proposition gives an upper bound on the shattering coefficient $\mathbb{S}_n(\mathscr{H})$ in terms of the VC dimension $d_{\mathscr{H}}$.

Proposition 9.6 Sauer's lemma

Let $\mathscr{H}$ be a set of classifiers with finite VC dimension $d_{\mathscr{H}}$. For any $n \in \mathbb{N}$, we have

$$\mathbb{S}_n(\mathscr{H}) \leq \sum_{i=0}^{d_{\mathscr{H}}} C_n^i \leq (n+1)^{d_{\mathscr{H}}} \quad \textit{with} \quad C_n^i = \begin{cases} \frac{n!}{i!\,(n-i)!} & \textit{for } n \geq i \\ 0 & \textit{for } n < i. \end{cases}$$

Proof. We first prove by induction on k the inequality

$$\mathbb{S}_k(\mathscr{H}) \leq \sum_{i=0}^{d_{\mathscr{H}}} C_k^i \tag{9.13}$$

for any $\mathscr{H}$ with finite VC dimension $d_{\mathscr{H}}$.

Let us consider the case $k=1$. If $d_{\mathscr{H}}=0$, then $\mathbb{S}_1(\mathscr{H})<2$ and so $\mathbb{S}_1(\mathscr{H})=1=C_1^0$. If $d_{\mathscr{H}} \geq 1$, we have $\mathbb{S}_1(\mathscr{H})=2$, which is also equal to $C_1^0+C_1^1$.

Assume now that (9.13) is true for all $k \leq n-1$. Let us consider $\mathscr{H}$, with finite VC dimension $d_{\mathscr{H}}$. When $d_{\mathscr{H}}=0$, no set of point can be shattered, so all points can only be labelled in one way. Therefore, $\mathbb{S}_k(\mathscr{H})=1$ and (9.13) is true for all k. We assume now that $d_{\mathscr{H}} \geq 1$. Let $x_1,\ldots,x_n$ be n points in $\mathscr{X}$ and define

$$\mathscr{H}(x_1,\ldots,x_n) = \{(h(x_1),\ldots,h(x_n)) : h \in \mathscr{H}\}.$$

We want to prove that

$$\operatorname{card}(\mathscr{H}(x_1,\ldots,x_n)) \leq \sum_{i=0}^{d_{\mathscr{H}}} C_k^i. \tag{9.14}$$

The set $\mathscr{H}(x_1,\ldots,x_n)$ depends only on the values of $h \in \mathscr{H}$ on $\{x_1,\ldots,x_n\}$, so we can replace $\mathscr{H}$ by $\mathscr{F} = \left\{h|_{\{x_1,\ldots,x_n\}} : h \in \mathscr{H}\right\}$ in the definition of $\mathscr{H}(x_1,\ldots,x_n)$. Since $d_{\mathscr{F}}$ is not larger than $d_{\mathscr{H}}$, it is enough to prove (9.14) for $\mathscr{F}$. Therefore, we assume (with no loss of generality) that $\mathscr{X} = \{x_1,\ldots,x_n\}$ and $\mathscr{H}=\mathscr{F}$. Let us consider the set

$$\mathscr{H}' = \left\{h \in \mathscr{H} : h(x_n)=1 \text{ and } h' = h - 2 \times \mathbf{1}_{\{x_n\}} \in \mathscr{H}\right\}.$$

Since $\mathscr{H}(x_1,\ldots,x_n) = \mathscr{H}'(x_1,\ldots,x_n) \cup (\mathscr{H} \setminus \mathscr{H}')(x_1,\ldots,x_n)$, we have

$$\operatorname{card}(\mathscr{H}(x_1,\ldots,x_n)) \leq \operatorname{card}\left(\mathscr{H}'(x_1,\ldots,x_n)\right) + \operatorname{card}\left((\mathscr{H} \setminus \mathscr{H}')(x_1,\ldots,x_n)\right). \tag{9.15}$$

Let us bound apart the cardinality of $\mathscr{H}'(x_1,\ldots,x_n)$ and the cardinality of $(\mathscr{H} \setminus \mathscr{H}')(x_1,\ldots,x_n)$.

1. We note that $\operatorname{card}(\mathscr{H}'(x_1,\ldots,x_n)) = \operatorname{card}(\mathscr{H}'(x_1,\ldots,x_{n-1}))$ since $h(x_n)=1$ for all $h \in \mathscr{H}'$. Let us check that the VC dimension $d_{\mathscr{H}'}$ of $\mathscr{H}'$ is at most $d_{\mathscr{H}}-1$. Actually, if d points $x_{i_1},\ldots,x_{i_d}$ of $\mathscr{X} = \{x_1,\ldots,x_n\}$ are shattered by $\mathscr{H}'$, then $x_n \notin \left\{x_{i_1},\ldots,x_{i_d}\right\}$ since $h(x_n)=1$ for all $h \in \mathscr{H}'$. Furthermore, the set

$\{x_{i_1},\dots,x_{i_d},x_n\}$ is shattered by $\mathscr{H}' \cup \{h' = h - 2\times \mathbf{1}_{\{x_n\}} : h \in \mathscr{H}\}$, which is included in $\mathscr{H}$ according to the definition of $\mathscr{H}'$. So, $d+1 \le d_{\mathscr{H}}$, which implies $d_{\mathscr{H}'} \le d_{\mathscr{H}} - 1$. Applying (9.13) with $k = n-1$, we obtain

$$\mathrm{card}\left(\mathscr{H}'(x_1,\dots,x_n)\right) = \mathrm{card}\left(\mathscr{H}'(x_1,\dots,x_{n-1})\right) \le \sum_{i=0}^{d_{\mathscr{H}}-1} C^i_{n-1}. \qquad (9.16)$$

2. When $h,h' \in \mathscr{H} \setminus \mathscr{H}'$ fulfill $h(x_i) = h'(x_i)$ for $i = 1,\dots,n-1$, they also fulfill $h(x_n) = h'(x_n)$, otherwise either h or h' would belong to $\mathscr{H}'$. Therefore, we have as above $\mathrm{card}\left((\mathscr{H}\setminus\mathscr{H}')(x_1,\dots,x_n)\right) = \mathrm{card}\left((\mathscr{H}\setminus\mathscr{H}')(x_1,\dots,x_{n-1})\right)$. Furthermore, $d_{\mathscr{H}\setminus\mathscr{H}'}$ is not larger than $d_{\mathscr{H}}$, since $\mathscr{H}\setminus\mathscr{H}' \subset \mathscr{H}$, so Equation (9.13) with $k = n-1$ gives

$$\mathrm{card}\left((\mathscr{H}\setminus\mathscr{H}')(x_1,\dots,x_n)\right) = \mathrm{card}\left((\mathscr{H}\setminus\mathscr{H}')(x_1,\dots,x_{n-1})\right) \le \sum_{i=0}^{d_{\mathscr{H}}} C^i_{n-1}\,. \qquad (9.17)$$

Combining (9.15), (9.16), and (9.17), we obtain

$$\mathrm{card}\left(\mathscr{H}(x_1,\dots,x_n)\right) \le \sum_{i=1}^{d_{\mathscr{H}}} C^{i-1}_{n-1} + \sum_{i=0}^{d_{\mathscr{H}}} C^i_{n-1} = \sum_{i=0}^{d_{\mathscr{H}}} C^i_n\,,$$

since $C^i_{n-1} + C^{i-1}_{n-1} = C^i_n$ for $i \ge 1$. As a consequence, (9.13) is true for $k = n$, and the induction is complete.

The second upper bound of the proposition is obtained by

$$\sum_{i=0}^{d} C^i_n \le \sum_{i=0}^{d} \frac{n^i}{i!} \le \sum_{i=0}^{d} C^i_d n^i = (1+n)^d.$$

The proof of Proposition 9.6 is complete. □

Remark. The reader may check (by induction) that for $d \le n$, we also have

$$\sum_1^d C^i_n \le \left(\frac{en}{d}\right)^d,$$

which improves the bound $\sum_1^d C^i_n \le (n+1)^d$ when $3 \le d \le n$.

Let us give some examples of VC dimension for some simple dictionaries on $\mathscr{X} = \mathbb{R}^d$. The proofs are let as exercises.

Example 1: linear classifiers.
The VC dimension of the set $\mathscr{H} = \{h(x) = \mathrm{sign}(\langle w,x\rangle) : \|w\| = 1\}$ of linear classifiers is d (see Exercise 9.5.2).

Example 2: affine classifiers.
The VC dimension of the set $\mathscr{H} = \{h(x) = \mathrm{sign}(\langle w,x\rangle + b) : \|w\| = 1,\ b \in \mathbb{R}\}$ of affine classifiers is $d+1$ (see Exercise 9.5.2).

Example 3: hyper-rectangle classifiers.
The VC dimension of the set $\mathscr{H} = \left\{h(x) = 2\mathbf{1}_A(x) - 1 : A \text{ hyper-rectangle of } \mathbb{R}^d\right\}$ of hyper-rectangle classifiers is $2d$.

Example 4: convex polygon classifiers.
The VC dimension of the set $\mathscr{H} = \left\{h(x) = 2\mathbf{1}_A(x) - 1 : A \text{ convex polygon of } \mathbb{R}^d\right\}$ of convex polygon classifiers is $+\infty$ (consider n points on the unit sphere: For any subset of these points, you can choose their convex hull as convex polygon).

Finally, we can state the following corollary of Theorem 9.1.

Corollary 9.7 Control of the stochastic error for Vapnick dictionaries

For $\mathscr{H}$ with VC dimension $1 \leq d_{\mathscr{H}} < +\infty$, for any $t > 0$, we have the upper bound

$$L(\widehat{h}_{\mathscr{H}}) \leq \min_{h \in \mathscr{H}} L(h) + 4\sqrt{\frac{2d_{\mathscr{H}} \log(2n+2))}{n}} + \sqrt{\frac{2t}{n}}$$

with probability at least $1 - e^{-t}$.

Let us now investigate the problem of the choice of the dictionary $\mathscr{H}$.

9.2.3 Dictionary Selection

Let us consider a collection $\{\mathscr{H}_1, \ldots, \mathscr{H}_M\}$ of dictionaries. Similarly to Chapter 2, we would like to select among this collection, the dictionary $\mathscr{H}_o$ with the smallest misclassification probability $L(\widehat{h}_{\mathscr{H}_o})$. The so-called *oracle* dictionary $\mathscr{H}_o$ depends on the unknown distribution $\mathbb{P}$, so it is not accessible to the statistician. In the following, we will build on Theorem 9.1 in order to design a data-driven procedure for selecting a dictionary $\mathscr{H}_{\widehat{m}}$ among the collection $\{\mathscr{H}_1, \ldots, \mathscr{H}_M\}$, with performances similar to those of $\mathscr{H}_o$.

The oracle dictionary $\mathscr{H}_o$ is obtained by minimizing the misclassification probability $L(\widehat{h}_{\mathscr{H}})$ over $\mathscr{H} \in \{\mathscr{H}_1, \ldots, \mathscr{H}_M\}$. A first idea is to select $\mathscr{H}_{\widehat{m}}$ by minimizing over the collection $\{\mathscr{H}_1, \ldots, \mathscr{H}_M\}$ the empirical misclassification probability $\widehat{L}_n(\widehat{h}_{\mathscr{H}})$. This selection procedure will not give good results, since for any $\mathscr{H} \subset \mathscr{H}'$ we always have $\widehat{L}_n(\widehat{h}_{\mathscr{H}'}) \leq \widehat{L}_n(\widehat{h}_{\mathscr{H}})$ by Definition (9.5). So the procedure will tend to select the largest dictionary. For designing a good selection procedure, we have to take into account the fluctuations of $\widehat{L}_n(\widehat{h}_{\mathscr{H}})$ around $L(\widehat{h}_{\mathscr{H}})$, as in Chapter 2. The Bound (9.8) in Theorem 9.1 gives us a control of these fluctuations. Building on this bound, we have the following result.

Theorem 9.8 Dictionary selection

Let us consider the dictionary selection procedure

$$\widehat{m} \in \operatorname*{argmin}_{m=1,\ldots,M} \left\{ \widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) + \text{pen}(\mathscr{H}_m) \right\}, \quad \textit{with} \quad \text{pen}(\mathscr{H}) \geq 2\sqrt{\frac{2\log(2\mathbb{S}_n(\mathscr{H}))}{n}}.$$

Then, for any $t > 0$, with probability at least $1 - e^{-t}$, we have

$$L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \leq \min_{m=1,\ldots,M} \left\{ \inf_{h \in \mathscr{H}_m} L(h) + 2\,\text{pen}(\mathscr{H}_m) \right\} + \sqrt{\frac{2\log(M) + 2t}{n}}. \tag{9.18}$$

Before proving Theorem 9.8, let us comment on Bound (9.18). Since $\min_{h \in \mathscr{H}} L(h) \leq L(\widehat{h}_{\mathscr{H}})$, we obtain with probability $1 - e^{-t}$

$$L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \leq L(\widehat{h}_{\mathscr{H}_o}) + 2\,\text{pen}(\mathscr{H}_o) + \sqrt{\frac{2\log(M) + 2t}{n}}.$$

In particular, we can compare the misclassification probability of the selected classifier with the misclassification probability of the best classifier among the collection $\{\widehat{h}_{\mathscr{H}_1}, \ldots, \widehat{h}_{\mathscr{H}_M}\}$.

We also notice that the second term of Bound (9.18) increases as $\sqrt{2\log(M)/n}$ with the number M of candidate dictionaries.

Proof of Theorem 9.8. We recall the notation $\widehat{\Delta}_n(\mathscr{H}) = \sup_{h \in \mathscr{H}} \left| \widehat{L}_n(h) - L(h) \right|$. The Lemma 9.2 ensures that

$$L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \leq \widehat{L}_n(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) + \widehat{\Delta}_n(\mathscr{H}_{\widehat{m}}).$$

According to Lemma 9.3 and Lemma 9.4, we have for $s > 0$ and $m \in \{1, \ldots, M\}$

$$\mathbb{P}\left(\widehat{\Delta}_n(\mathscr{H}_m) > \text{pen}(\mathscr{H}_m) + \sqrt{s/(2n)} \right) \leq e^{-s}.$$

For $s = \log(M) + t$, the union bound ensures that

$$\mathbb{P}\left(\widehat{\Delta}_n(\mathscr{H}_m) \leq \text{pen}(\mathscr{H}_m) + \sqrt{\frac{\log(M) + t}{2n}}, \quad \text{for all } m = 1, \ldots, M \right) \geq 1 - e^{-t} \tag{9.19}$$

Therefore, according to the definition of the selection criterion, we have with probability at least $1 - e^{-t}$

$$\begin{aligned} L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) &\leq \widehat{L}_n(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) + \text{pen}(\mathscr{H}_{\widehat{m}}) + \sqrt{\frac{\log(M) + t}{2n}} \\ &\leq \min_{m=1,\ldots,M} \left\{ \widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) + \text{pen}(\mathscr{H}_m) \right\} + \sqrt{\frac{\log(M) + t}{2n}}. \end{aligned} \tag{9.20}$$

To conclude, we only need to control the size of $\widehat{L}_n(\widehat{h}_{\mathscr{H}_m})$ in terms of $\inf_{h\in\mathscr{H}_m} L(h)$. This can be done directly by combining (9.7) and (9.8), but the resulting bound is not tight.

In order to compare $\widehat{L}_n(\widehat{h}_{\mathscr{H}_m})$ to $\inf_{h\in\mathscr{H}_m} L(h)$, let us notice that for any $h \in \mathscr{H}_m$, we have

$$\widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) \leq \widehat{L}_n(h) \leq L(h) + \widehat{\Delta}_n(\mathscr{H}_m),$$

so taking the infimum over $h \in \mathscr{H}_m$, we obtain for all $m = 1,\ldots,M$

$$\widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) \leq \inf_{h\in\mathscr{H}_m} L(h) + \widehat{\Delta}_n(\mathscr{H}_m).$$

Combining this bound with (9.19) and (9.20), we obtain with probability at least $1-e^{-t}$

$$L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \leq \min_{m=1,\ldots,M}\left\{\inf_{h\in\mathscr{H}_m} L(h) + 2\,\mathrm{pen}(\mathscr{H}_m)\right\} + 2\sqrt{\frac{\log(M)+t}{2n}}.$$

The proof of Theorem 9.8 is complete. □

Remark. Combining (9.19) and Lemma 9.2, we obtain the confidence interval for the misclassification probability

$$\mathbb{P}\left(L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \in \left[\widehat{L}_n(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) - \delta(\widehat{m},t), \widehat{L}_n(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) + \delta(\widehat{m},t)\right]\right) \geq 1 - e^{-t},$$

$$\text{with} \quad \delta(\widehat{m},t) = \mathrm{pen}(\mathscr{H}_{\widehat{m}}) + \sqrt{\frac{\log(M)+t}{2n}}.$$

9.3 From Theoretical to Practical Classifiers

9.3.1 Empirical Risk Convexification

The empirical risk minimization classifier analyzed in the previous section has some very nice statistical properties, but it cannot be used in practice because of its computational cost. Actually, there is no efficient way to minimize (9.5), since neither $\mathscr{H}$ nor $\widehat{L}_n$ are convex. The situation is very similar to the situation met in Chapter 2 for the model selection procedure (2.9). As in Chapter 4, we will derive some practical classifiers from (9.5) by a convex relaxation of the minimization (9.5). Some of the most popular classification algorithms are obtained by following this principle. The empirical misclassification probability $\widehat{L}_n$ will be replaced by some convex surrogate and the set of classifiers $\mathscr{H}$ will be replaced by some convex functional set $\mathscr{F} \subset \mathbb{R}^{\mathscr{X}}$.

Let us consider some convex set $\mathscr{F}$ of functions from $\mathscr{X}$ to $\mathbb{R}$. A function $f \in \mathscr{F}$ is not a classifier, but we can use it for classification by classifying the data points according to the sign of f. In other words, we can associate to f the classifier $\mathrm{sign}(f)$. The empirical misclassification probability of this classifier can be written as

$$\widehat{L}_n(\mathrm{sign}(f)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i\,\mathrm{sign}(f)(X_i)<0\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i f(X_i)<0\}}.$$

Let us replace this empirical misclassification probability $\widehat{L}_n$ by some convex surrogate, which is more amenable to numerical computations. A simple and efficient way to obtain a convex criterion is to replace the loss function $z \to \mathbf{1}_{z<0}$ by some convex function $z \to \ell(z)$. Building on this simple idea, we will focus in the following on classifiers obtained by the procedure

$$\widehat{h}_{\mathscr{F}} = \operatorname{sign}(\widehat{f}_{\mathscr{F}}),$$
$$\text{where} \quad \widehat{f}_{\mathscr{F}} \in \operatorname*{argmin}_{f\in\mathscr{F}} \widehat{L}_n^{\ell}(f), \quad \text{with} \quad \widehat{L}_n^{\ell}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i f(X_i)). \tag{9.21}$$

This classifier can be computed efficiently, since both $\mathscr{F}$ and $\widehat{L}_n^{\ell}$ are convex. Many classical classifiers are obtained by solving (9.21), with some specific choices of $\mathscr{F}$ and ℓ; see Section 9.3.3 and 9.3.4 for some examples.

Some popular convex loss ℓ

It is natural to consider a convex loss function ℓ, which is nonincreasing and nonnegative. Usually, we also ask that $\ell(z) \geq \mathbf{1}_{z<0}$ for all $z \in \mathbb{R}$, since in this case we can give an upper bound on the misclassification probability; see Theorem 9.10. Some classical loss functions are

- the exponential loss $\ell(z) = e^{-z}$,
- the logit loss $\ell(z) = \log_2(1+e^{-z})$, and
- the hinge loss $\ell(z) = (1-z)_+$, with $(x)_+ = \max(0,x)$.

A plot of these three functions is given in Figure 9.4.

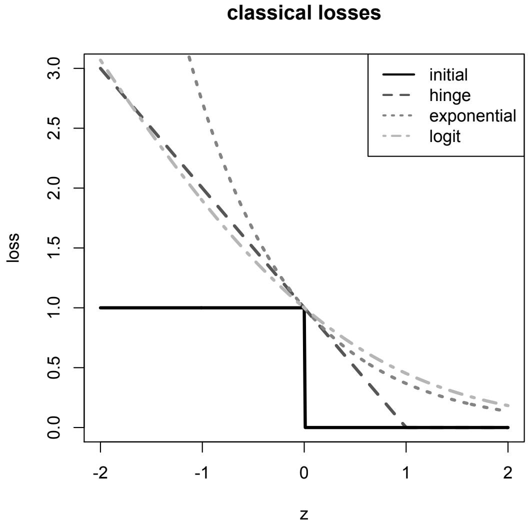


Figure 9.4 *Plot of the exponential, hinge, and logit losses.*

Some classical functional sets $\mathscr{F}$

The main popular convex functional sets $\mathscr{F}$ can be grouped into two classes.

A first popular class of sets $\mathscr{F}$ is obtained by taking a linear combination of a finite family $\mathscr{H} = \{h_1, \ldots, h_p\}$ of classifiers

$$\mathscr{F} = \left\{ f : f(x) = \sum_{j=1}^{p} \beta_j h_j(x) \text{ with } \beta_j \in \mathscr{C} \right\}, \tag{9.22}$$

where $\mathscr{C}$ is a convex subset of $\mathbb{R}^p$. Typical choices for $\mathscr{C}$ are the ℓ^1-ball $\{\beta \in \mathbb{R}^p : |\beta|_1 \leq R\}$, the simplex $\left\{\beta \in \mathbb{R}^p : \beta_j \geq 0,\ \sum_{j=1}^{p} \beta_j \leq 1\right\}$, or the whole space $\mathbb{R}^p$. This choice appears, for example, in boosting methods; see Section 9.3.4. The basic classifiers $\{h_1, \ldots, h_p\}$ are often called *weak learners*. A popular choice of weak learners in $\mathbb{R}^d$ is $h_j(x) = \text{sign}(x_j - t_j)$ with $t_j \in \mathbb{R}$.

A second popular class of sets $\mathscr{F}$ is obtained by taking a ball of a Reproducing Kernel Hilbert Space (RKHS). We refer to Appendix E for a brief introduction to RKHS. Since the smoothness of a function in an RKHS is driven by its norm (see, e.g., Formula (E.5), page 241, in Appendix E), a ball of an RKHS corresponds to a set of smooth functions. Let $\mathscr{F}_k$ be an RKHS with reproducing kernel k, and write $\|f\|_{\mathscr{F}}$ for the Hilbert norm of $f \in \mathscr{F}_k$. For notational simplicity, in the following, we simply write $\mathscr{F}$ for $\mathscr{F}_k$. Minimizing $\widehat{L}_n^{\ell}$ over the ball $\{f \in \mathscr{F} : \|f\|_{\mathscr{F}} \leq R\}$ is equivalent to minimizing over $\mathscr{F}$ the dual Lagrangian problem

$$\widehat{f}_{\mathscr{F}} \in \underset{f \in \mathscr{F}}{\text{argmin}}\, \widetilde{L}_n^{\ell}(f), \quad \text{with} \quad \widetilde{L}_n^{\ell}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i f(X_i)) + \lambda \|f\|_{\mathscr{F}}^2, \tag{9.23}$$

for some $\lambda > 0$. This kind of classifier appears, for example, in Support Vector Machine algorithms, presented in Section 9.3.3. The solution of (9.23) fulfills the following representation formula.

Proposition 9.9 Representation formula

The minimization problem (9.23) is equivalent to $\widehat{f}_{\mathscr{F}} = \sum_{j=1}^{n} \widehat{\beta}_j k(X_j, .)$, *with*

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^n}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\sum_{j=1}^{n} \beta_j Y_i k(X_j, X_i) \right) + \lambda \sum_{i,j=1}^{n} \beta_i \beta_j k(X_i, X_j) \right\}. \tag{9.24}$$

Proof. Let V be the linear space spanned by $k(X_1, .), \ldots, k(X_n, .)$, where $k(x, .)$ refers to the map $y \to k(x, y)$. Decomposing $f = f_V + f_{V^\perp}$ according to the orthogonal decomposition $\mathscr{F} = V \bigoplus V^\perp$, we have by the reproducing property (E.2), page 239,

$$f(X_i) = \langle f, k(X_i, .) \rangle_{\mathscr{F}} = \langle f_V, k(X_i, .) \rangle_{\mathscr{F}} = f_V(X_i).$$

Combining this formula with the Pythagorean formula, we obtain

$$\widetilde{L}_n^\ell(f_V + f_{V^\perp}) = \frac{1}{n}\sum_{i=1}^n \ell\big(Y_i f_V(X_i)\big) + \lambda \|f_V\|_{\mathscr{F}}^2 + \lambda \|f_{V^\perp}\|_{\mathscr{F}}^2 .$$

Since λ is positive, any minimizer $\widehat{f}$ of $\widetilde{L}_n^\ell$ must fulfill $\widehat{f}_{V^\perp} = 0$, so it is of the form

$$\widehat{f}_{\mathscr{F}} = \sum_{i=1}^n \widehat{\beta}_i k(X_i, .).$$

Furthermore, the reproducing property (E.2) ensures again that $\langle k(X_i, .), k(X_j, .)\rangle_{\mathscr{F}} = k(X_i, X_j)$, so

$$\left\| \sum_{j=1}^n \beta_j k(X_j, .) \right\|_{\mathscr{F}}^2 = \sum_{i,j=1}^n \beta_i \beta_j k(X_i, X_j).$$

The proof of Proposition 9.9 is complete. □

The representation formula is of major importance in practice, since it reduces the infinite-dimensional minimization problem (9.23) into an n-dimensional convex minimization problem (9.24) that can be solved efficiently. In the Section 9.3.3 on Support Vector Machines, we will give a more precise description of the solution of this problem when ℓ is the hinge loss.

Another important feature of the representation formula is that we only need to know the positive definite kernel k in order to compute the classifier $\widehat{h}_{\mathscr{F}}$. In particular, we do not need to identify the RKHS associated to k in order to define and compute the estimator (9.24). The RKHS $\mathscr{F}$ is only used implicitly in order to understand the nature of the classifier $\widehat{h}_{\mathscr{F}}$.

9.3.2 Statistical Properties

The classifier $\widehat{h}_{\mathscr{F}}$ given by (9.21) with $\mathscr{F}$ and ℓ convex has the nice feature to have a low computational cost, but does it have some good statistical properties?

Link with the bayes classifier

The empirical risk minimizer $\widehat{h}_{\mathscr{H}}$ of Section 9.2 minimizes the empirical version $\widehat{L}_n(h)$ of the misclassification probability $L(h) = \mathbb{P}(Y \neq h(X))$ over some set $\mathscr{H}$ of classifiers. The function $\widehat{f}_{\mathscr{F}}$ minimizes instead the empirical version $\widehat{L}_n^\ell(h)$ of $\widehat{L}^\ell(h) = \mathbb{E}\left[\ell(Yf(X))\right]$ over some functional set $\mathscr{F}$. The classifier $\widehat{h}_{\mathscr{H}}$ can then be viewed as an empirical version of the bayes classifier h_* which minimizes $\mathbb{P}(Y \neq h(X))$ over the set of measurable functions $h : \mathscr{X} \to \{-1, +1\}$, whereas the function $\widehat{f}_{\mathscr{F}}$ is an empirical version of the function f_*^ℓ, which minimizes $\mathbb{E}\left[\ell(Yf(X))\right]$ over the set of measurable functions $f : \mathscr{X} \to \mathbb{R}$. A first point is to understand the link between the bayes classifier h_* and the sign of the function f_*^ℓ. It turns out that under some weak assumptions on ℓ, the sign of f_*^ℓ exactly coincides with the bayes classifier h_*, so

$\text{sign}(f_*^\ell)$ minimizes the misclassification probability $\mathbb{P}(Y \neq h(X))$. Let us check this point.

Conditioning on X, we have

$$\begin{aligned}\mathbb{E}\left[\ell(Yf(X)\right] &= \mathbb{E}\left[\mathbb{E}\left[\ell(Yf(X)|X\right]\right] \\ &= \mathbb{E}\left[\ell(f(X))\mathbb{P}(Y=1|X) + \ell(-f(X))(1-\mathbb{P}(Y=1|X))\right].\end{aligned}$$

Assume that ℓ is decreasing, differentiable, and strictly convex (e.g., exponential or logit loss). The above expression is minimum for $f_*^\ell(X)$ solution of

$$\frac{\ell'(-f(X))}{\ell'(f(X))} = \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)},$$

when such a solution exists. Since ℓ is strictly convex, we then have $f(X) > 0$ if and only if $\ell'(-f(X))/\ell'(f(X)) > 1$, so

$$f_*^\ell(X) > 0 \iff \mathbb{P}(Y=1|X) > 1/2 \iff \mathbb{E}[Y|X] = 2\mathbb{P}(Y=1|X) - 1 > 0.$$

Since $h_*(X) = \text{sign}(\mathbb{E}\left[Y|X\right])$ (see Section 9.1), we obtain $\text{sign}(f_*^\ell) = h_*$. This equality also holds true for the hinge loss ℓ (check it!).

To sum up the above discussion, the target function f_*^ℓ approximated by $\widehat{f}_{\mathscr{F}}$ does perfectly make sense for the classification problem, since, under some weak assumptions, its sign coincides with the best possible classifier h_* (the Bayes classifier).

Upper-bound on the misclassification probability

We focus now on the misclassification probability $L(\widehat{h}_{\mathscr{F}})$ of the classifier $\widehat{h}_{\mathscr{F}} = \text{sign}(\widehat{f}_{\mathscr{F}})$ given by (9.21). In practice, it is important to have an upper bound on the misclassification probability $L(\widehat{h}_{\mathscr{F}})$, which can be computed from the data. The next theorem provides such an upper bound for some typical examples of set $\mathscr{F}$.

Theorem 9.10 Confidence bound on $L(\widehat{h}_{\mathscr{F}})$

For any $R > 0$, we set $\Delta\ell(R) = |\ell(R) - \ell(-R)|$. We assume here that the loss-function ℓ is convex, nonincreasing, non-negative, α-Lipschitz on $[-R, R]$ and fulfills $\ell(z) \geq \mathbf{1}_{z<0}$ for all z in $\mathbb{R}$. We consider the classifier $\widehat{h}_{\mathscr{F}}$ given by (9.21).

(a) *When $\mathscr{F}$ is of the form (9.22), with $\mathscr{C} = \{\beta \in \mathbb{R}^p : |\beta|_1 \leq R\}$, we have with probability at least $1-e^{-t}$*

$$L(\widehat{h}_{\mathscr{F}}) \leq \widehat{L}_n^\ell(\widehat{f}_{\mathscr{F}}) + 4\alpha R\sqrt{\frac{2\log(2p)}{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}. \tag{9.25}$$

(b) *Let $\mathscr{F}$ be the ball of radius R of an RKHS with kernel k fulfilling $k(x,x) \leq 1$ for all $x \in \mathscr{X}$. Then, we have with probability at least $1-e^{-t}$*

$$L(\widehat{h}_{\mathscr{F}}) \leq \widehat{L}_n^\ell(\widehat{f}_{\mathscr{F}}) + \frac{4\alpha R}{\sqrt{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}. \tag{9.26}$$

Proof. We first prove a general upper bound for $L(\widehat{h}_{\mathscr{F}})$, similar to Theorem 9.1.

Lemma 9.11

Assume that $\sup_{f\in\mathscr{F}}|f(x)| \leq R < +\infty$. *For any loss ℓ fulfilling the hypotheses of Theorem 9.10, we have with probability at least* $1-e^{-t}$

$$L(\widehat{h}_{\mathscr{F}}) \leq \widehat{L}_n^{\ell}(\widehat{f}_{\mathscr{F}}) + \frac{4\alpha}{n}\max_{x\in\mathscr{X}^n}\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] + \Delta\ell(R)\sqrt{\frac{t}{2n}}, \qquad (9.27)$$

where $\sigma_1,\ldots,\sigma_n$ are i.i.d. random variables with distribution $\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2$.

Proof of Lemma 9.11. The proof of this lemma relies on the same arguments as the proof of Theorem 9.1. We set

$$\widehat{\Delta}_n^{\ell}(\mathscr{F}) = \sup_{f\in\mathscr{F}}\left|\widehat{L}_n^{\ell}(f) - L^{\ell}(f)\right| \quad \text{with } L^{\ell}(f) = \mathbb{E}\left[\ell(Yf(X))\right].$$

The first point is to notice that since $\ell(z) \geq \mathbf{1}_{z<0}$ for all real z, we have

$$\begin{aligned} L(\widehat{h}_{\mathscr{F}}) = \mathbb{P}(Y\widehat{f}_{\mathscr{F}}(X) < 0) &\leq \mathbb{E}\left[\ell(Y\widehat{f}_{\mathscr{F}}(X))\right] \\ &\leq \widehat{L}_n^{\ell}(\widehat{f}_{\mathscr{F}}) + \widehat{\Delta}_n^{\ell}(\mathscr{F}). \end{aligned}$$

As in Lemma 9.3, the McDiarmid concentration inequality (Theorem B.4, page 219, in Appendix B) ensures that with probability at least $1-e^{-t}$, we have

$$\widehat{\Delta}_n^{\ell}(\mathscr{F}) \leq \mathbb{E}\left[\widehat{\Delta}_n^{\ell}(\mathscr{F})\right] + \Delta\ell(R)\sqrt{\frac{t}{2n}}.$$

To conclude the proof of the lemma, it only remains to prove that

$$\mathbb{E}\left[\widehat{\Delta}_n^{\ell}(\mathscr{F})\right] \leq \frac{4\alpha}{n}\max_{x\in\mathscr{X}^n}\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right]. \qquad (9.28)$$

Following exactly the same lines as in the proof of Lemma 9.4 (replacing $\mathbf{1}_{Y_i\neq h(X_i)}$ by $\ell(Y_i f(X_i)) - \ell(0)$), we obtain

$$\mathbb{E}\left[\widehat{\Delta}_n^{\ell}(\mathscr{F})\right] \leq \frac{2}{n}\max_{y\in\{-1,+1\}^n}\max_{x\in\mathscr{X}^n}\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i(\ell(y_i f(x_i)) - \ell(0))\right|\right].$$

We finally use the α-Lipschitz property of ℓ to conclude: According to the Contraction principle (Proposition B.8, page 224, in Appendix B, with $\varphi(z) = \ell(z) - \ell(0)$ and $\mathscr{Z} = \{[y_i f(x_i)]_{i=1,\ldots,n} : f \in \mathscr{F}\}$), we have

$$\begin{aligned} \mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i(\ell(y_i f(x_i)) - \ell(0))\right|\right] &\leq 2\alpha\,\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i y_i f(x_i)\right|\right] \\ &= 2\alpha\,\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right], \end{aligned}$$

where we used in the last line that $(\sigma_1,\ldots,\sigma_n)$ has the same distribution as $(y_1\sigma_1,\ldots,y_n\sigma_n)$ for any $y\in\{-1,1\}^n$. Combining the last two bounds gives (9.28), and the proof of Lemma 9.11 is complete. □

(a) Let us prove now Bound (9.25). The map $\beta\to\sum_{i=1}^n\sigma_i\sum_{j=1}^p\beta_jh_j(x_i)$ is linear, so it reaches its maximum and minimum on the ℓ^1-ball $\mathscr{C}$ at one of the vertices of $\mathscr{C}$. Therefore, we have

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n\sigma_if(x_i)\right|\right]=R\,\mathbb{E}_\sigma\left[\max_{j=1,\ldots,p}\left|\sum_{i=1}^n\sigma_ih_j(x_i)\right|\right].$$

It remains to apply Inequality (9.10) with $\mathscr{V}=\left\{(h_j(x_1),\ldots,h_j(x_n)):j=1,\ldots,p\right\}$, whose cardinality is at most p in order to obtain

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n\sigma_if(x_i)\right|\right]\leq R\,\sqrt{2n\log(2p)}.$$

Bound (9.25) then follows from Lemma 9.11.

(b) We now turn to the second Bound (9.26) and write $\|.\|_{\mathscr{F}}$ for the norm in the RKHS. According to the reproducing formula (E.2), page 239, and the Cauchy–Schwartz inequality, we have

$$\left|\sum_{i=1}^n\sigma_if(x_i)\right|=\left|\left\langle f,\sum_{i=1}^n\sigma_ik(x_i,.)\right\rangle_{\mathscr{F}}\right|\leq\|f\|_{\mathscr{F}}\left\|\sum_{i=1}^n\sigma_ik(x_i,.)\right\|_{\mathscr{F}}.$$

Since $\|f\|_{\mathscr{F}}\leq R$, applying Jensen inequality, we obtain

$$\begin{aligned}\mathbb{E}_\sigma\left[\sup_{\|f\|_{\mathscr{F}}\leq R}\left|\sum_{i=1}^n\sigma_if(x_i)\right|\right] &\leq R\,\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n\sigma_ik(x_i,.)\right\|_{\mathscr{F}}\right]\\ &\leq R\sqrt{\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n\sigma_ik(x_i,.)\right\|_{\mathscr{F}}^2\right]}\\ &= R\sqrt{\sum_{i,j=1}^n k(x_i,x_j)\mathbb{E}_\sigma[\sigma_i\sigma_j]}\,,\end{aligned}$$

where we have used $\langle k(x_i,.),k(x_j,.)\rangle_{\mathscr{F}}=k(x_i,x_j)$. Since $k(x,x)\leq 1$ and $\mathbb{E}\left[\sigma_i\sigma_j\right]=0$ for $i\neq j$, we get

$$\mathbb{E}_\sigma\left[\sup_{\|f\|_{\mathscr{F}}\leq R}\left|\sum_{i=1}^n\sigma_if(x_i)\right|\right]\leq R\sqrt{\sum_{i=1}^n k(x_i,x_i)\mathbb{E}_\sigma[\sigma_i^2]}\;\leq\;R\sqrt{n}.$$

Combining again the reproducing property with the Cauchy–Schwartz inequality, we obtain

$$|f(x)|=|\langle f,k(x,.)\rangle_{\mathscr{F}}|\leq R\sqrt{k(x,x)}\leq R.$$

So $\mathscr{F}$ fulfills the hypotheses of the Lemma 9.11, which gives

$$L(\widehat{h}_{\mathscr{F}}) \leq \widehat{L}_n^{\ell}(\widehat{f}_{\mathscr{F}}) + \frac{4\alpha R}{\sqrt{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}\,.$$

The proof of Theorem 9.10 is complete. □

It is possible to derive risk bounds similar to (9.7) for $L(\widehat{h}_{\mathscr{H}})$; we refer to Bousquet, Boucheron, and Lugosi [32] for a review of such results. In the remainder of this chapter, we will describe two very popular classification algorithms: the Support Vector Machines and AdaBoost.

9.3.3 Support Vector Machines

The Support Vector Machines (SVM) algorithm corresponds to the estimator (9.23) with the hinge loss $\ell(z) = (1-z)_+$. The final classification is performed according to $\widehat{h}_{\mathscr{F}}(x) = \text{sign}(\widehat{f}_{\mathscr{F}}(x))$. We stress that there is a unique solution to (9.23) when ℓ is the convex loss since (9.23) is strictly convex when $\lambda > 0$. It turns out that there is a very nice geometrical interpretation of the solution $\widehat{f}_{\mathscr{F}}$, from which originates the name "Support Vector Machines."

Proposition 9.12 Support Vectors

The solution of (9.23) is of the form $\widehat{f}_{\mathscr{F}}(x) = \sum_{i=1}^n \widehat{\beta}_i k(X_i, x)$, *with*

$$\begin{cases} \widehat{\beta}_i = 0 & \textit{if} \quad Y_i\widehat{f}_{\mathscr{F}}(X_i) > 1 \\ \widehat{\beta}_i = Y_i/(2\lambda n) & \textit{if} \quad Y_i\widehat{f}_{\mathscr{F}}(X_i) < 1 \\ 0 \leq Y_i\widehat{\beta}_i \leq 1/(2\lambda n) & \textit{if} \quad Y_i\widehat{f}_{\mathscr{F}}(X_i) = 1\,. \end{cases}$$

The vectors X_i *with index i, such that* $\widehat{\beta}_i \neq 0$, *are called support vectors.*

Proof. Writing K for the matrix $[k(X_i,X_j)]_{i,j=1,\ldots,n}$, we know from the representation Formula (9.24) that the solution of (9.23) is of the form $\widehat{f}_{\mathscr{F}} = \sum_{j=1}^n \widehat{\beta}_j k(X_j,.)$, with

$$\widehat{\beta} \in \underset{\beta\in\mathbb{R}^n}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n \left(1 - Y_i[K\beta]_i\right)_+ + \lambda\beta^T K\beta \right\}.$$

The above minimization problem is not smooth, so we introduce some slack variables $\widehat{\xi}_i = (1 - Y_i[K\widehat{\beta}]_i)_+$ and rewrite the minimization problem as

$$(\widehat{\beta},\widehat{\xi}) \in \underset{\substack{\beta,\xi\,\in\,\mathbb{R}^n \text{ such that} \\ \xi_i \geq 1 - Y_i[K\beta]_i \\ \xi_i \geq 0}}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n \xi_i + \lambda\beta^T K\beta \right\}. \tag{9.29}$$

This problem is now smooth and convex, and the Karush–Kuhn–Tucker conditions for the Lagrangian dual problem

$$(\widehat{\beta},\widehat{\xi}) \in \operatorname*{argmin}_{\beta,\xi\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\beta^T K\beta - \sum_{i=1}^{n}\left(\alpha_i(\xi_i - 1 + Y_i[K\beta]_i) + \gamma_i\xi_i\right) \right\} \qquad (9.30)$$

gives the formulas for $i,j = 1,\ldots,n$

$$\text{first-order conditions:} \quad 2\lambda[K\widehat{\beta}]_j = \sum_{i=1}^{n} K_{ij}\alpha_i Y_i \quad \text{and} \quad \alpha_j + \gamma_j = \frac{1}{n},$$

$$\text{slackness conditions:} \quad \min(\alpha_i, \widehat{\xi}_i - 1 + Y_i[K\widehat{\beta}]_i) = 0 \quad \text{and} \quad \min(\gamma_i, \widehat{\xi}_i) = 0.$$

The first first-order condition is fulfilled with $\widehat{\beta}_i = \alpha_i Y_i/(2\lambda)$. Since $\widehat{f}_{\mathscr{F}}(X_i) = [K\widehat{\beta}]_i$, the first slackness condition enforces that $\widehat{\beta}_i = 0$ if $Y_i\widehat{f}_{\mathscr{F}}(X_i) > 1$. The second slackness condition, together with the second first-order optimality condition, enforces that $\widehat{\beta}_i = Y_i/(2\lambda n)$ if $\widehat{\xi}_i > 0$ and $0 \leq Y_i\widehat{\beta}_i \leq 1/(2\lambda n)$ otherwise. To conclude the proof of the proposition, we notice that when $\widehat{\xi}_i > 0$, we have $\widehat{\beta}_i$ and α_i nonzero, and therefore $Y_i\widehat{f}_{\mathscr{F}}(X_i) = 1 - \widehat{\xi}_i < 1$ according to the first slackness condition. □

We observe that when the matrix $K = [k(X_i,X_j)]_{i,j=1,\ldots,n}$ is nonsingular, there is a unique solution $\widehat{\beta}$ to (9.24). We refer to Exercise 9.5.5 for the computation of $\widehat{\beta}$. An implementation of the SVM is available, e.g., in the `R` package `kernlab` at `http://cran.r-project.org/web/packages/kernlab/`.

Let us now interpret geometrically Proposition 9.12.

Geometrical interpretation: linear kernel

We start with the simplest kernel $k(x,y) = \langle x,y\rangle$ for all $x,y \in \mathbb{R}^d$. The associated RKHS is the space of linear forms $\mathscr{F} = \left\{\langle w,.\rangle : w \in \mathbb{R}^d\right\}$. In this case,

$$\widehat{f}_{\mathscr{F}}(x) = \sum_{i=1}^{n}\widehat{\beta}_i\langle X_i,x\rangle = \langle\widehat{w},x\rangle \quad \text{with} \quad \widehat{w} = \sum_{i=1}^{n}\widehat{\beta}_i X_i,$$

so the classifier $\widehat{h}_{\mathscr{F}}(x) = \text{sign}(\langle\widehat{w},x\rangle)$ assigns labels to points according to their position relative to the hyperplane $\left\{x \in \mathbb{R}^d : \langle\widehat{w},x\rangle = 0\right\}$. The normal $\widehat{w}$ to the hyperplane is a linear combination of the support vectors, which are the data points X_i, such that $Y_i\langle\widehat{w},X_i\rangle \leq 1$. They are represented by squares in Figure 9.5. The hyperplanes $\left\{x \in \mathbb{R}^d : \langle\widehat{w},x\rangle = +1\right\}$ and $\left\{x \in \mathbb{R}^d : \langle\widehat{w},x\rangle = -1\right\}$ are usually called margin-hyperplanes.

We notice the following important property of the SVM. If we add to the learning data set a point X_{n+1}, which fulfills $Y_{n+1}\langle\widehat{w},X_{n+1}\rangle > 1$, then the vector $\widehat{w}$ and the classifier $\widehat{h}_{\mathscr{F}}$ do not change. In other words, only data points that are wrongly classified or classified with not enough margin (i.e., $Y_i\langle\widehat{w},X_i\rangle \leq 1$) do influence the separating hyperplane $\left\{x \in \mathbb{R}^d : \langle\widehat{w},x\rangle = 0\right\}$. This property is in contrast with the LDA classifier (9.2), where all the points have an equal weight in the definition of the $\widehat{\mu}_k$ and $\widehat{\Sigma}$.

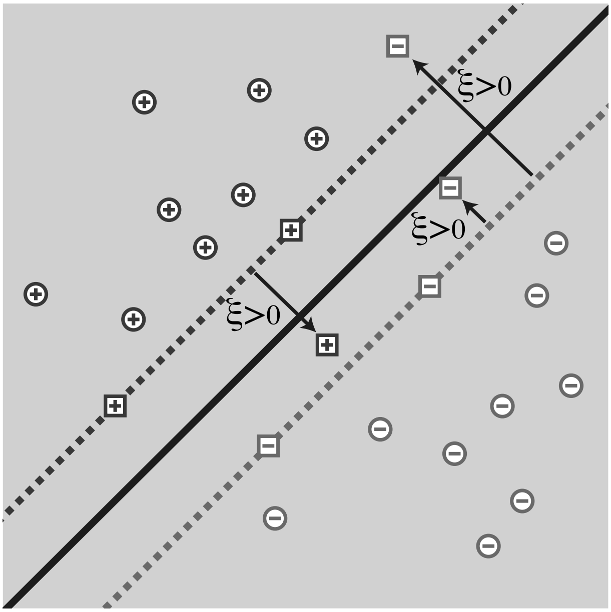


Figure 9.5 *Classification with a linear SVM: The separating hyperplane* $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = 0\right\}$ *is represented in black, and the two margin hyperplanes* $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = +1\right\}$ *and* $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = -1\right\}$ *are represented in dotted blue and red, respectively. The support vectors are represented by squares.*

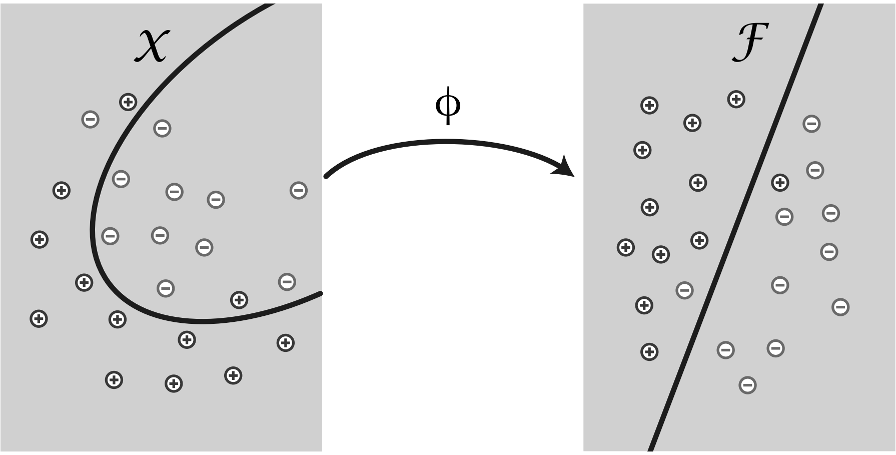


Figure 9.6 *Classification with a nonlinear kernel: The linear classification in* $\mathscr{F}$ *produces a nonlinear classification in* $\mathscr{X}$ *via the reciprocal image of* ϕ.

Geometrical interpretation: arbitrary positive definite kernels

Let us denote by $\phi : \mathscr{X} \to \mathscr{F}$ the map $\phi(x) = k(x,.)$, usually called the "feature map." According to the reproducing property (E.2), page 239, and Proposition 9.12, we have

$$\widehat{f}_{\mathscr{F}}(x) = \langle \widehat{f}_{\mathscr{F}}, \phi(x) \rangle_{\mathscr{F}} = \left\langle \sum_{i=1}^{n} \widehat{\beta}_i \phi(X_i), \phi(x) \right\rangle_{\mathscr{F}} .$$

A point $x \in \mathscr{X}$ is classified by $\widehat{h}_{\mathscr{F}}$ according to the sign of the above scalar product. Therefore, the points $\phi(x) \in \mathscr{F}$ are classified according to the linear classifier on $\mathscr{F}$

$$f \mapsto \operatorname{sign}\left(\langle \widehat{w}_{\phi}, f \rangle_{\mathscr{F}}\right), \quad \text{where } \widehat{w}_{\phi} = \sum_{i=1}^{n} \widehat{\beta}_i \phi(X_i).$$

The separating frontier $\left\{x \in \mathscr{X} : \widehat{f}_{\mathscr{F}}(x) = 0\right\}$ of the classifier $\widehat{h}_{\mathscr{F}}$ is therefore the reciprocal image by ϕ of the intersection of the hyperplane $\left\{f \in \mathscr{F} : \langle \widehat{w}_{\phi}, f \rangle_{\mathscr{F}} = 0\right\}$ in $\mathscr{F}$ with the range of ϕ, as represented in Figure 9.6. We observe that the kernel k then delinearizes the SVM, in the sense that it produces a nonlinear classifier $\widehat{h}_{\mathscr{F}}$ with almost the same computational cost as a linear one in $\mathbb{R}^n$.

You can observe SVM in action with the following recreative applet:
`http://cs.stanford.edu/people/karpathy/svmjs/demo/`.

Why are RKHS useful?

There are mainly two major reasons for using RKHS. The first reason is that using RKHS allows to delinearize some algorithms by mapping $\mathscr{X}$ in $\mathscr{F}$ with $\phi : x \to k(x,.)$, as represented in Figure 9.6. It then provides nonlinear algorithms with almost the same computational complexity as a linear one.

The second reason is that it allows us to apply to any set $\mathscr{X}$ some algorithms that are defined for vectors. Assume, for example, that we want to classify some proteins or molecules according to their therapeutic properties. Let $\mathscr{X}$ represents our set of molecules. For any $x,y \in \mathscr{X}$, let us represent by $k(x,y)$ some measure of similarity between x and y. If the kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is positive definite, then we can directly apply the SVM algorithm in order to classify them; see Figure 9.7. Of course, the key point in this case is to properly design the kernel k. Usually, the kernel $k(x,y)$ is designed according to some properties of x,y that are known to be relevant for the classification problem. For example, the number of common short sequences is a useful index of similarity between two proteins. The computational complexity for evaluating $k(x,y)$ is also an issue that is crucial in many applications with complex data.

We point out that RKHS can be used in many other statistical settings in order to either delinearize an algorithm or to apply a vectorial algorithm to nonvectorial data. In principle, it can be used with any algorithm relying only on scalar products $\langle x,y \rangle$ by replacing these scalar products by the kernel evaluation $k(x,y)$. Some popular examples are the kernel-PCA (for a PCA with nonvectorial data; see Exercise 9.5.6) or the kernel-smoothing in regression.

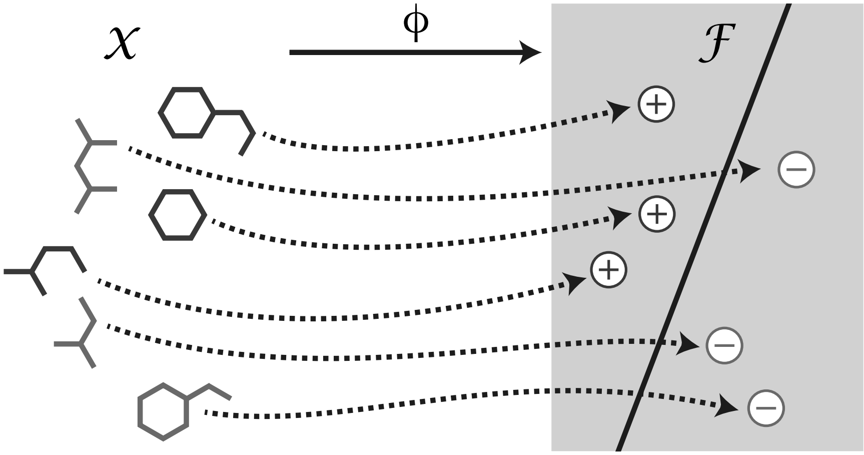


Figure 9.7 *Classification of molecules with an SVM.*

Finally, we stress again that as long as we have a representation formula as (9.24), we do not need to identify the RKHS associated to a positive definite kernel k in order to implement an algorithm based on RKHS. This property is usually referred as the "kernel trick."

9.3.4 AdaBoost

AdaBoost is an algorithm that computes an approximation of the estimator (9.21) with the exponential loss $\ell(z) = e^{-z}$ and the functional space $\mathscr{F} = \text{span}\{h_1, \ldots, h_p\}$, where $h_1, \ldots, h_p$ are p arbitrary classifiers.

The principle of the AdaBoost algorithm is to perform a greedy minimization of

$$\widehat{f}_{\mathscr{F}} \in \underset{f \in \text{span}\{h_1,\ldots,h_p\}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \exp(-Y_i f(X_i)) \right\}.$$

More precisely, it computes a sequence of functions $\widehat{f}_m$ for $m = 0, \ldots, M$ by starting from $\widehat{f}_0 = 0$ and then solving for $m = 1, \ldots, M$

$$\widehat{f}_m = \widehat{f}_{m-1} + \beta_m h_{j_m},$$
$$\text{where} \quad (\beta_m, j_m) \in \underset{\substack{j = 1, \ldots, p \\ \beta \in \mathbb{R}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-Y_i\big(\widehat{f}_{m-1}(X_i) + \beta h_j(X_i)\big)\right).$$

The final classification is performed according to $\widehat{h}_M(x) = \text{sign}(\widehat{f}_M(x))$, which is an approximation of $\widehat{h}_{\mathscr{H}}$ defined by (9.21).

The exponential loss allows us to compute (β_m, j_m) very efficiently. Actually, setting $w_i^{(m)} = n^{-1}\exp(-Y_i\widehat{f}_{m-1}(X_i))$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\exp\left(-Y_i\big(\widehat{f}_{m-1}(X_i)+\beta h_j(X_i)\big)\right) = (e^{\beta}-e^{-\beta})\sum_{i=1}^{n} w_i^{(m)}\mathbf{1}_{h_j(X_i)\neq Y_i} + e^{-\beta}\sum_{i=1}^{n} w_i^{(m)}.$$

When no initial classifier h_j perfectly classifies the data $(X_i, Y_i)_{i=1,\ldots,n}$, so that the condition

$$\mathrm{err}_m(j) = \frac{\sum_{i=1}^{n} w_i^{(m)}\mathbf{1}_{h_j(X_i)\neq Y_i}}{\sum_{i=1}^{n} w_i^{(m)}} < 1 \quad \text{for all } j = 1,\ldots,p$$

is met, the minimizers (β_m, j_m) are given by

$$j_m = \operatorname*{argmin}_{j=1,\ldots,p} \mathrm{err}_m(j) \quad \text{and} \quad \beta_m = \frac{1}{2}\log\left(\frac{1-\mathrm{err}_m(j_m)}{\mathrm{err}_m(j_m)}\right).$$

Noticing that $-Y_i h(X_i) = 2\mathbf{1}_{Y_i\neq h(X_i)} - 1$, we obtain the standard formulation of the AdaBoost algorithm.

```
AdaBoost
Init: $w_i^{(1)} = 1/n$, for $i = 1,\ldots,n$

Iterate: For $m = 1,\ldots,M$ do

    $j_m = \operatorname*{argmin}_{j=1,\ldots,p} \mathrm{err}_m(j)$
    $2\beta_m = \log(1-\mathrm{err}_m(j_m)) - \log(\mathrm{err}_m(j_m))$
    $w_i^{(m+1)} = w_i^{(m)}\exp(2\beta_m\mathbf{1}_{h_{j_m}(X_i)\neq Y_i} - \beta_m)$,    for $i = 1,\ldots,n$

Output: $\widehat{f}_M(x) = \sum_{m=1}^{M}\beta_m h_{j_m}(x)$.
```

We notice that the AdaBoost algorithm gives more and more weight in $\mathrm{err}_m(j)$ to the data points X_i, which are wrongly classified at the stage m.

You can observe AdaBoost in action (with half-plane weak-learners h_j) with the following recreative applet: `http://cseweb.ucsd.edu/~yfreund/adaboost/`.

9.3.5 Classifier Selection

The practical implementation of the SVM classifier or the AdaBoost classifier requires us to choose different quantities: the kernel k and the tuning parameter λ for the SVM, the weak learners $\{h_1,\ldots,h_p\}$ and the integer M for AdaBoost. The most popular technique for choosing these quantities is to apply a V-fold cross-validation scheme (see Chapter 5). Cross-validation can also be used to compare the SVM to the boosting, the Linear Discriminant Analysis, the logistic regression, etc.

9.4 Discussion and References

9.4.1 Take-Home Message

The discrete nature of the labels y enables us to relax the modeling of the data, by merely assuming that the data are i.i.d. For this reason, some may believe that the theory of classification presented Section 9.2 is "model-free." It is not really the case, since Estimator (9.5) relies on a dictionary $\mathscr{H}$, which must provide a good approximation of the Bayes classifier h_* in the sense that $\min_{h\in\mathscr{H}} L(h) - L(h_*)$ must be small. Otherwise the misclassification probability of Estimator (9.5) will be large and the theory is useless. There are then some implicit assumptions on the distribution of the data that are hidden in the choice of the dictionary $\mathscr{H}$ (or of the collection of dictionaries in Section 9.2.3). Yet, Estimator (9.5) enjoys some very nice statistical properties under very weak assumptions. Its main drawback is its high computational complexity that is prohibitive when n is larger than a few tens. As in Chapter 4, a powerful strategy is to relax the minimization problem (9.5) in order to obtain classifiers that are both computationally efficient and statistically grounded. Combining this strategy with the use of kernels provides a very flexible theory of classification that can handle data such as text, graph, images, etc. As usual, the resulting classifiers depend on some tuning parameters and, again, cross-validation is a useful technique for selecting them.

9.4.2 References

The mathematical foundations of supervised classification date back to the seminal work of Vapnik and Chervonenkis [121, 122] in the '70s, and this topic has since attracted a lot of effort. For the reader interested in going beyond the basic concepts presented in this chapter, we refer to the book by Devroye, Györfi, and Lugosi [50] and the survey by Boucheron, Bousquet, and Lugosi [32] for recent developments on the topic and a comprehensive bibliography. For more practical consideration, we refer to the book by Hastie, Tibshirani, and Friedman [73], where many practical algorithms are described and discussed. Finally, we point out that the concepts introduced here arise also for the ranking problem (rank at best some data, as your favorite search engine does); see, e.g., Clémençon, Lugosi and Vayatis [47].

9.5 Exercises

9.5.1 Linear Discriminant Analysis

Let us consider a couple of random variables (X,Y) with values in $\mathbb{R}^d \times \{-1,1\}$, and distribution

$$\mathbb{P}(Y=k) = \pi_k > 0 \quad \text{et} \quad \mathbb{P}(X \in dx | Y=k) = g_k(x)\,dx, \;\; k \in \{-1,1\},\; x \in \mathbb{R}^d,$$

where $\pi_{-1} + \pi_1 = 1$ and g_{-1}, g_1 are two densities in $\mathbb{R}^d$.

1. What is the distribution of X?

2. Check that the Bayes classifier is given by $h_*(x) = \text{sign}\,(\pi_1 g_1(x) - \pi_{-1} g_{-1}(x))$ for $x \in \mathbb{R}^d$.

We assume henceforth that

$$g_k(x) = (2\pi)^{-d/2} \sqrt{\det(\Sigma_k^{-1})} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right), \qquad k = -1, 1,$$

for two nonsingular matrices Σ_{-1}, Σ_1 and $\mu_{-1}, \mu_1 \in \mathbb{R}^d$, with $\mu_{-1} \neq \mu_1$.

3. Prove that when $\Sigma_{-1} = \Sigma_1 = \Sigma$, the condition $\pi_1 g_1(x) > \pi_{-1} g_{-1}(x)$ is equivalent to
$$(\mu_1 - \mu_{-1})^T \Sigma^{-1}\left(x - \frac{\mu_1 + \mu_{-1}}{2}\right) > \log(\pi_{-1}/\pi_1).$$
4. What is the nature of the frontier between $\{h_* = 1\}$ and $\{h_* = -1\}$ in this case?
5. We assume in addition that $\pi_1 = \pi_{-1}$. Prove that
$$\mathbb{P}(h_*(X) = 1 | Y = -1) = \Phi(-d(\mu_1, \mu_{-1})/2),$$
where Φ is the standard Gaussian cumulative function and $d(\mu_1, \mu_{-1})$ is the Mahalanobis distance associated to Σ defined by $d(\mu_1, \mu_{-1}) = \|\Sigma^{-1/2}(\mu_1 - \mu_{-1})\|$.
6. When $\Sigma_1 \neq \Sigma_{-1}$, what is the nature of the frontier between $\{h_* = 1\}$ and $\{h_* = -1\}$?

9.5.2 VC Dimension of Linear Classifiers in $\mathbb{R}^d$

For any $w \in \mathbb{R}^d$, we denote by $h_w : \mathbb{R}^d \to \{-1, 1\}$, the classifier

$$h_w(x) = \text{sign}(\langle w, x \rangle), \text{ for } x \in \mathbb{R}^d.$$

We compute below the VC dimension of $\mathscr{H} = \{h_w : w \in \mathbb{R}^d\}$.

1. We write $e_1, \ldots, e_d$ for the canonical basis of $\mathbb{R}^d$. Prove that for any $\delta \in \{-1, 1\}^d$ there exists $w_\delta \in \mathbb{R}^d$, such that $h_{w_\delta}(e_i) = \delta_i$ for $i = 1, \ldots, d$. Give a lower bound on $d_{\mathscr{H}}$.
2. For any $x_1, \ldots, x_{d+1} \in \mathbb{R}^d$, there exists $\lambda \in \mathbb{R}^{d+1}$ nonzero, such that $\sum_{i=1}^{d+1} \lambda_i x_i = 0$. We can assume that there exists j, such that $\lambda_j > 0$, by changing λ in $-\lambda$ if necessary. We define δ by $\delta_i = \text{sign}\,(\lambda_i)$. By considering the sum $\sum_{i=1}^{d+1} \lambda_i \langle w_\delta, x_i \rangle$, prove that there exists no $w_\delta \in \mathbb{R}^d$, such that $h_{w_\delta}(x_i) = \delta_i$, for $i = 1, \ldots, d+1$. Conclude that $d_{\mathscr{H}} = d$.
3. We define $\tilde{\mathscr{H}} = \{h_{w,b} : w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$, where $h_{w,b}(x) = \text{sign}\,(\langle w, x \rangle - b)$. Prove that the VC dimension of $\tilde{\mathscr{H}}$ is $d + 1$.

9.5.3 Linear Classifiers with Margin Constraints

Assume that $\mathscr{F}$ is an Hilbert space, and consider $\{x_1, \ldots, x_M\} \subset \mathscr{F}$ with $\|x_i\|_{\mathscr{F}} \leq A$ for any $i = 1, \ldots, M$. For $r, R > 0$ we set

$$\mathscr{H} = \left\{h_w : w \in \mathscr{F},\ \|w\|_{\mathscr{F}} \leq R \text{ and } |\langle w, x_i \rangle_{\mathscr{F}}| \geq r \text{ for } i = 1, \ldots, M\right\},$$

where $h_w : \{x_1,\ldots,x_M\} \to \{-1,1\}$ is defined by $h_w(x) = \mathrm{sign}(\langle w,x\rangle_{\mathscr{F}})$ for any $x \in \{x_1,\ldots,x_M\}$. We assume henceforth that $M > A^2R^2/r^2$. We will prove that $d_{\mathscr{H}} \leq A^2R^2/r^2$.

1. For $n \leq M$ and $\sigma_1,\ldots,\sigma_n$ i.i.d. uniform on $\{-1,1\}$, prove that
$$\mathbb{E}\left[\Big\|\sum_{i=1}^{n}\sigma_i x_i\Big\|_{\mathscr{F}}^2\right] = \sum_{i=1}^{n}\mathbb{E}\left[\|\sigma_i x_i\|_{\mathscr{F}}^2\right] \leq nA^2.$$
2. Conclude that there exists $y \in \{-1,1\}^n$, such that $\left\|\sum_{i=1}^n y_i x_i\right\|_{\mathscr{F}}^2 \leq nA^2$.
3. Assume there exists $w \in \mathscr{F}$, such that $h_w \in \mathscr{H}$ and $y_i\langle w,x_i\rangle_{\mathscr{F}} \geq r$ for $i = 1,\ldots,n$. Prove that
$$nr \leq \left\langle w, \sum_{i=1}^{n} y_i x_i \right\rangle_{\mathscr{F}} \leq RA\sqrt{n}.$$
4. Show that $\mathbb{S}_n(\mathscr{H}) < 2^n$ for $R^2A^2/r^2 < n \leq M$ and conclude.

9.5.4 Spectral Kernel

The spectral kernel is a classical kernel for classifying "words" from a finite alphabet $\mathscr{A}$. For $x \in \bigcup_{n\geq q}\mathscr{A}^n$ and $s \in \mathscr{A}^q$, we set
$$N_s(x) = \text{number of occurence of } s \text{ in } x.$$

The spectral kernel is then defined by
$$k(x,y) = \sum_{s\in\mathscr{A}^q} N_s(x)N_s(y)$$
for all $x,y \in \bigcup_{n\geq q}\mathscr{A}^n$. It counts the number of common sequences of length q in x and y.

1. Prove that k is a positive definite kernel on $\bigcup_{n\geq q}\mathscr{A}^n$.
2. Check that the computational complexity for computing $k(x,y)$ is at most of order $\ell(x)+\ell(y)$, where $\ell(x)$ is the length of x.

9.5.5 Computation of the SVM Classifier

We consider the SVM classifier $\widehat{h}_{\mathrm{SVM}} = \mathrm{sign}\left(\sum_{j=1}^n \widehat{\beta}_j k(X_j,.)\right)$, with $\widehat{\beta}$ solution of (9.29). We assume that $K = [k(X_i,X_j)]_{i,j=1,\ldots,n}$ is nonsingular.

1. From the Lagrangian problem (9.30) and the Karush–Kuhn–Tucker conditions proves that
$$\widehat{\beta} \in \underset{\beta\in\mathbb{R}^n}{\mathrm{argmin}}\left\{\lambda\beta^T K\beta - \sum_{i=1}^{n}\widehat{\alpha}_i(Y_i(K\beta)_i - 1)\right\}$$
for some $\widehat{\alpha} \in \mathbb{R}^n$, fulfilling $0 \leq \widehat{\alpha}_i \leq 1/n$ for all $i = 1,\ldots,n$.

2. By Lagrangian duality, we know that $\widehat{\alpha}$ is solution of

$$\widehat{\alpha} \in \underset{0\leq\alpha_i\leq 1/n}{\operatorname{argmax}} \ \min_{\beta\in\mathbb{R}^n} \left\{ \lambda\beta^T K\beta - \sum_{i=1}^n \alpha_i (Y_i (K\beta)_i - 1) \right\}.$$

Proves that $\widehat{\beta}_i = \widehat{\alpha}_i Y_i/(2\lambda)$, where

$$\widehat{\alpha} \in \underset{0\leq\alpha_i\leq 1/n}{\operatorname{argmax}} \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{4\lambda} \sum_{i,j=1}^n K_{ij} y_i y_j \alpha_i \alpha_j \right\}.$$

In particular, the SVM classifier $\widehat{h}_{\text{SVM}}$ can be computed from a simple constrained quadratic maximization.

9.5.6 Kernel Principal Component Analysis (KPCA)

As discussed above, RKHS allows us to "delinearize" some linear algorithms. We give an example here with the Principal Component Analysis (see Exercise 1.6.4, page 23). Assume that we have n points $X^{(1)}, \ldots, X^{(n)} \in \mathscr{X}$, and let us consider an RKHS $\mathscr{F}$ associated to a positive definite kernel k on $\mathscr{X}$. We denote by ϕ the map from $\mathscr{X}$ to $\mathscr{F}$ defined by $\phi : x \to k(x,.)$. The principle of KPCA is to perform a PCA on the points $\phi(X^{(1)}), \ldots, \phi(X^{(n)})$ mapped by ϕ in the RKHS. We then seek for the space $\mathscr{V}_d \subset \mathscr{F}$ fulfilling

$$\mathscr{V}_d \in \underset{\dim(\mathscr{V})\leq d}{\operatorname{argmin}} \sum_{i=1}^n \|\phi(X^{(i)}) - \mathscr{P}_{\mathscr{V}}\phi(X^{(i)})\|_{\mathscr{F}}^2,$$

where the infimum is taken over all the subspaces $\mathscr{V} \subset \mathscr{F}$ with dimension not larger than d, and $\mathscr{P}_{\mathscr{V}}$ denotes the orthogonal projection onto $\mathscr{V}$ with respect to the Hilbert norm $\|.\|_{\mathscr{F}}$ on $\mathscr{F}$. In the following, we denote by $\mathscr{L}$ the linear map

$$\begin{aligned} \mathscr{L} : \mathbb{R}^n &\to \mathscr{F} \\ \alpha \to \mathscr{L}\alpha &= \sum_{i=1}^n \alpha_i \phi(X^{(i)}). \end{aligned}$$

1. By adapting the arguments of Proposition 9.9, prove that $\mathscr{V}_d = \mathscr{L}V_d$, with V_d fulfilling

$$V_d \in \underset{\dim(V)\leq d}{\operatorname{argmin}} \sum_{i=1}^n \|\phi(X^{(i)}) - \mathscr{P}_{\mathscr{L}V}\phi(X^{(i)})\|_{\mathscr{F}}^2,$$

where the infimum is taken over all the subspaces $V \subset \mathbb{R}^n$ with dimension not larger than d.

2. We denote by K the $n \times n$ matrix with entries $K_{i,j} = k(X^{(i)}, X^{(j)})$ for $i, j = 1, \ldots, n$. We assume in the following that K is nonsingular. Prove that $\|\mathscr{L}K^{-1/2}\alpha\|_{\mathscr{F}}^2 = \|\alpha\|^2$ for any $\alpha \in \mathbb{R}^n$.

3. Let V be subspace of $\mathbb{R}^n$ of dimension d and denote by $(b_1,\ldots,b_d)$ an orthonormal basis of the linear span $K^{1/2}V$. Prove that $(\mathscr{L}K^{-1/2}b_1,\ldots,\mathscr{L}K^{-1/2}b_d)$ is an orthonormal basis of $\mathscr{L}V$.
4. Prove the identities
$$\begin{aligned}\mathscr{P}_{\mathscr{L}V}\mathscr{L}\alpha &= \sum_{k=1}^{d}\langle \mathscr{L}K^{-1/2}b_k,\mathscr{L}\alpha\rangle_{\mathscr{F}}\mathscr{L}K^{-1/2}b_k\\ &= \mathscr{L}K^{-1/2}\mathrm{Proj}_{K^{1/2}V}K^{1/2}\alpha,\end{aligned}$$
where $\mathrm{Proj}_{K^{1/2}V}$ denotes the orthogonal projector onto $K^{1/2}V$ in $\mathbb{R}^n$.
5. Let us denote by $(e_1,\ldots,e_n)$ the canonical basis of $\mathbb{R}^n$. Check that
$$\begin{aligned}\sum_{i=1}^{n}\|\phi(X^{(i)})-\mathscr{P}_{\mathscr{L}V}\phi(X^{(i)})\|_{\mathscr{F}}^2 &= \sum_{i=1}^{n}\|\mathscr{L}e_i-\mathscr{L}K^{-1/2}\mathrm{Proj}_{K^{1/2}V}K^{1/2}e_i\|_{\mathscr{F}}^2\\ &= \sum_{i=1}^{n}\|K^{1/2}e_i-\mathrm{Proj}_{K^{1/2}V}K^{1/2}e_i\|^2\\ &= \|K^{1/2}-\mathrm{Proj}_{K^{1/2}V}K^{1/2}\|_F^2,\end{aligned}$$
where $\|.\|_F$ denotes the Frobenius norm $\|A\|_F^2=\sum_{i,j}A_{ij}^2$.
6. With Theorem C.5, page 233, in Appendix C, prove that $V_d=\mathrm{span}\{v_1,\ldots,v_d\}$, where $v_1,\ldots,v_d$ are eigenvectors of K associated to the d largest eigenvalues of K.
7. We set $f_k=\mathscr{L}K^{-1/2}v_k$ for $k=1,\ldots,d$. Check that $(f_1,\ldots,f_d)$ is an orthonormal basis of $\mathscr{V}_d$ and
$$\mathscr{P}_{\mathscr{V}}\phi(X^{(i)})=\sum_{k=1}^{n}\langle v_k,K^{1/2}e_i\rangle f_k.$$
So, in the basis $(f_1,\ldots,f_d)$, the coordinates of the orthogonal projection of the point $\phi(X^{(i)})$ onto $\mathscr{V}$ are $\big(\langle v_1,K^{1/2}e_i\rangle,\ldots,\langle v_d,K^{1/2}e_i\rangle\big)$.

Appendix A

Gaussian Distribution

A.1 Gaussian Random Vectors

A random vector $Y \in \mathbb{R}^d$ is distributed according to the $\mathcal{N}(m,\Sigma)$ Gaussian distribution, with $m \in \mathbb{R}^d$ and $\Sigma \in \mathscr{S}_d^+$ (the set of all $d \times d$ symmetric positive semidefinite matrix), when

$$\mathbb{E}\left[e^{\mathrm{i}\langle\lambda,Y\rangle}\right] = \exp\left(\mathrm{i}\langle\lambda,m\rangle - \frac{1}{2}\lambda^T\Sigma\lambda\right), \quad \text{for all } \lambda \in \mathbb{R}^d. \tag{A.1}$$

When matrix Σ is nonsingular (i.e., positive definite), the $\mathcal{N}(m,\Sigma)$ Gaussian distribution has a density with respect to the Lebesgue measure on $\mathbb{R}^d$ given by

$$\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(y-m)^T\Sigma^{-1}(y-m)\right).$$

Affine transformations of Gaussian distribution are still Gaussian.

Lemma A.1 Affine transformation
Let $Y \in \mathbb{R}^d$ be a random vector with $\mathcal{N}(m,\Sigma)$ Gaussian distribution. Then for any $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$,

$$AY + b \sim \mathcal{N}(Am+b, A\Sigma A^T).$$

In particular, for $a \in \mathbb{R}^d$,

$$\langle a,Y\rangle \sim \mathcal{N}(\langle m,a\rangle, a^T\Sigma a).$$

Proof. The first identity is obtained by computing the characteristic function of $AY + b$

$$\mathbb{E}\left[e^{\mathrm{i}\langle\lambda,AY+b\rangle}\right] = \mathbb{E}\left[e^{\mathrm{i}\langle A^T\lambda,Y\rangle+\mathrm{i}\langle\lambda,b\rangle}\right] = \exp\left(\mathrm{i}\langle A^T\lambda,m\rangle - \frac{1}{2}(A^T\lambda)^T\Sigma A^T\lambda\right)e^{\mathrm{i}\langle\lambda,b\rangle}$$
$$= \exp\left(\mathrm{i}\langle\lambda,Am+b\rangle - \frac{1}{2}\lambda^T A\Sigma A^T\lambda\right).$$

The second identity is obtained with $A = a^T$ and $b = 0$. □

Lemma A.2 Orthogonal projections onto subspaces
Let $Y \in \mathbb{R}^d$ be a random vector with $\mathcal{N}(m,\Sigma)$ Gaussian distribution, and let S and V be two linear spans of $\mathbb{R}^d$ orthogonal with respect to the scalar product induced by Σ. Then the variables $\mathrm{Proj}_S Y$ and $\mathrm{Proj}_V Y$ are independent and follow, respectively, the $\mathcal{N}(\mathrm{Proj}_S m, \mathrm{Proj}_S \Sigma \mathrm{Proj}_S)$ and $\mathcal{N}(\mathrm{Proj}_V m, \mathrm{Proj}_V \Sigma \mathrm{Proj}_V)$ Gaussian distribution.

Proof. Since the projection matrices Proj_S and Proj_V are symmetric, we obtain that the joint characteristic function of $\mathrm{Proj}_S Y$ and $\mathrm{Proj}_V Y$ is

$$\begin{aligned}
\mathbb{E}\left[e^{\mathrm{i}\langle \lambda, \mathrm{Proj}_S Y\rangle + \mathrm{i}\langle \gamma, \mathrm{Proj}_V Y\rangle}\right] &= \mathbb{E}\left[e^{\mathrm{i}\langle \mathrm{Proj}_S \lambda + \mathrm{Proj}_V \gamma, Y\rangle}\right] \\
&= \exp\left(\mathrm{i}\langle \mathrm{Proj}_S \lambda + \mathrm{Proj}_V \gamma, m\rangle - \frac{1}{2}(\mathrm{Proj}_S \lambda + \mathrm{Proj}_V \gamma)^T \Sigma (\mathrm{Proj}_S \lambda + \mathrm{Proj}_V \gamma)\right) \\
&= \exp\left(\mathrm{i}\langle \lambda, \mathrm{Proj}_S m\rangle - \frac{1}{2}\lambda^T \mathrm{Proj}_S \Sigma \mathrm{Proj}_S \lambda\right) \\
&\quad \times \exp\left(\mathrm{i}\langle \gamma, \mathrm{Proj}_V m\rangle - \frac{1}{2}\gamma^T \mathrm{Proj}_V \Sigma \mathrm{Proj}_V \gamma\right) \\
&= \mathbb{E}\left[e^{\mathrm{i}\langle \lambda, \mathrm{Proj}_S Y\rangle}\right] \mathbb{E}\left[e^{\mathrm{i}\langle \gamma, \mathrm{Proj}_V Y\rangle}\right].
\end{aligned}$$

We conclude with Lemma A.1. □

A.2 Chi-Square Distribution

Let $Y \in \mathbb{R}^n$ be a random vector with $\mathcal{N}(0, I_n)$ Gaussian distribution. The χ^2 distribution with n degrees of freedom, corresponds to the distribution of $\|Y\|^2$. In particular, the mean of a $\chi^2(n)$ distribution is

$$\mathbb{E}\left[\|Y\|^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[Y_i^2\right] = n.$$

Lemma A.3 Norms of projections
Let $Y \in \mathbb{R}^n$ be a random vector with $\mathcal{N}(0, I_n)$ Gaussian distribution, and let S be a linear subspace of $\mathbb{R}^n$ with dimension d. Then, the variable $\mathrm{Proj}_S Y$ follows the $\mathcal{N}(0, \mathrm{Proj}_S)$ Gaussian distribution and the square-norm $\|\mathrm{Proj}_S Y\|^2$ follows a χ^2-distribution of degree d.

In particular, $\mathbb{E}\left[\|\mathrm{Proj}_S Y\|^2\right] = \dim(S)$.

Proof. The projection Proj_S is symmetric, so $\mathrm{Proj}_S \mathrm{Proj}_S^T = \mathrm{Proj}_S$ and $\mathrm{Proj}_S Y$ follows a $\mathcal{N}(0, \mathrm{Proj}_S)$ Gaussian distribution according to Lemma A.1.

Let $u_1, \ldots, u_d$ be an orthonormal basis of S and set $U = [u_1, \ldots, u_d]$. Since $U^T U = I_d$, the vector $U^T Y$ follows a $\mathcal{N}(0, I_d)$-distribution and

$$\|\mathrm{Proj}_S Y\|^2 = \sum_{k=1}^{d} (u_k^T Y)^2 = \|U^T Y\|^2$$

follows a χ^2 distribution of degree d. □

A.3 Gaussian Conditioning

We provide in this section a few useful results on Gaussian conditioning.

Lemma A.4
We consider two sets $A = \{1, \ldots, k\}$ and $B = \{1, \ldots, p\} \setminus A$, and a Gaussian random vector $X = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \in \mathbb{R}^p$ with $\mathcal{N}(0, \Sigma)$ distribution. We assume that Σ is nonsingular and write $K = \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix}$ for its inverse.

In the next formulas, K_{AA}^{-1} will refer to the inverse $(K_{AA})^{-1}$ of K_{AA} (and not to $(K^{-1})_{AA} = \Sigma_{AA}$).

Then, the conditional distribution of X_A given X_B is the Gaussian $\mathcal{N}\left(-K_{AA}^{-1} K_{AB} X_B, K_{AA}^{-1}\right)$ distribution. In others words, we have the decomposition

$$X_A = -K_{AA}^{-1} K_{AB} X_B + \varepsilon_A, \quad \textit{where } \varepsilon_A \sim \mathcal{N}\left(0, K_{AA}^{-1}\right) \textit{ is independent of } X_B. \tag{A.2}$$

Proof. We write $g(x_A, x_B)$, respectively, $g(x_A|x_B)$ and $g(x_B)$, for the density of the distribution of X, respectively, of X_A given $X_B = x_B$ and X_B. We have

$$\begin{aligned} g(x_A|x_B) &= g(x_A, x_B)/g(x_B) \\ &= \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2} x_A^T K_{AA} x_A - x_A^T K_{AB} x_B - \frac{1}{2} x_B^T \left(K_{BB} - \Sigma_{BB}^{-1}\right) x_B\right), \end{aligned}$$

with Σ_{BB} the covariance matrix of X_B. Since $\Sigma_{BB}^{-1} = K_{BB} - K_{BA} K_{AA}^{-1} K_{AB}$, we have

$$g(x_A|x_B) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}(x_A + K_{AA}^{-1} K_{AB} x_B)^T K_{AA} (x_A + K_{AA}^{-1} K_{AB} x_B)\right).$$

We recognize the density of the Gaussian $\mathcal{N}\left(-K_{AA}^{-1} K_{AB} x_B, K_{AA}^{-1}\right)$ distribution. □

Corollary A.5 *For any $a \in \{1, \ldots, p\}$, we have*

$$X_a = -\sum_{b: b \neq a} \frac{K_{ab}}{K_{aa}} X_b + \varepsilon_a, \quad \textit{where } \varepsilon_a \sim \mathcal{N}(0, K_{aa}^{-1}) \textit{ is independent of } \{X_b : b \neq a\}. \tag{A.3}$$

Proof. We apply the previous lemma with $A = \{a\}$ and $B = A^c$. □

Finally, we derive from (A.2) the following simple formula for the conditional correlation of X_a and X_b given $\{X_c :\ c \neq a, b\}$, which is defined by

$$\operatorname{cor}(X_a, X_b | X_c : c \neq a, b) = \frac{\operatorname{cov}(X_a, X_b | X_c : c \neq a, b)}{\sqrt{\operatorname{var}(X_a | X_c : c \neq a, b)\, \operatorname{var}(X_b | X_c : c \neq a, b)}}.$$

Corollary A.6 *For any $a, b \in \{1, \ldots, p\}$, we have*

$$\operatorname{cor}(X_a, X_b | X_c : c \neq a, b) = \frac{-K_{ab}}{\sqrt{K_{aa} K_{bb}}}. \tag{A.4}$$

Proof. The previous lemma with $A = \{a, b\}$ and $B = A^c$ gives

$$\operatorname{cov}(X_A | X_B) = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ab} & K_{bb} \end{pmatrix}^{-1} = \frac{1}{K_{aa}K_{bb} - K_{ab}^2} \begin{pmatrix} K_{bb} & -K_{ab} \\ -K_{ab} & K_{aa} \end{pmatrix}.$$

Plugging this formula in the definition of the conditional correlation, we obtain Formula (A.4). □

Appendix B

Probabilistic Inequalities

B.1 Basic Inequalities

Markov inequality plays a central role in the control of the fluctuations of random variables.

Lemma B.1 Markov inequality

For any nondecreasing positive function $\varphi : \mathbb{R} \to \mathbb{R}^+$ and any real valued random variable X, we have

$$\mathbb{P}(X \geq t) \leq \frac{1}{\varphi(t)} \mathbb{E}\left[\varphi(X)\right] \quad \textit{for all} \quad t \in \mathbb{R}.$$

In particular, for any $\lambda > 0$, we have

$$\mathbb{P}(X \geq t) \leq e^{-\lambda t} \mathbb{E}\left[e^{\lambda X}\right] \quad \textit{for all} \quad t \in \mathbb{R}.$$

Proof. Since φ is positive and nondecreasing, we have

$$\mathbb{P}(X \geq t) \leq \mathbb{E}\left[\frac{\varphi(X)}{\varphi(t)} \mathbf{1}_{X \geq t}\right] \leq \frac{1}{\varphi(t)} \mathbb{E}\left[\varphi(X)\right].$$

□

Jensen inequality is another important inequality that controls expectations.

Lemma B.2 Jensen inequality

For any convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ and any random variable X in $\mathbb{R}^d$, such that $\varphi(X)$ is integrable, we have

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}\left[\varphi(X)\right].$$

Proof. Let us denote by $\mathscr{L}_\varphi$ the set of affine functions from $\mathbb{R}^d$ to $\mathbb{R}$, such that $L(x) \leq \varphi(x)$ for all $x \in \mathbb{R}^d$. Since

$$\varphi(x) = \sup_{L \in \mathscr{L}_\varphi} L(x),$$

the linearity of the expectation gives

$$\mathbb{E}[\varphi(X)] = \mathbb{E}\left[\sup_{L\in\mathscr{L}_\varphi} L(X)\right] \geq \sup_{L\in\mathscr{L}_\varphi} \mathbb{E}[L(X)] = \sup_{L\in\mathscr{L}_\varphi} L(\mathbb{E}[X]) = \varphi(\mathbb{E}[X]).$$

□

Lemma B.3 Tail of the Gaussian distribution

Let Z be a standard Gaussian random variable. For any $x \geq 0$, we have

$$\mathbb{P}(|Z| \geq x) \leq e^{-x^2/2}.$$

Proof. The function

$$\varphi(x) = e^{-x^2/2} - \mathbb{P}(|Z| \geq x) = e^{-x^2/2} - \sqrt{\frac{2}{\pi}}\int_x^\infty e^{-t^2/2}\,dt, \quad x \geq 0,$$

takes value 0 at $x = 0$ and its derivative $\varphi'(x) = \left(\sqrt{2/\pi} - x\right) e^{-x^2/2}$ is positive for $x \leq \sqrt{2/\pi}$, so it is positive on $[0, \sqrt{2/\pi}]$. Furthermore, for $x \geq \sqrt{2/\pi}$ we have

$$\sqrt{\frac{2}{\pi}}\int_x^\infty e^{-t^2/2}\,dt \leq \int_x^\infty t e^{-t^2/2}\,dt = e^{-x^2/2},$$

so φ is positive on $\mathbb{R}^+$. □

B.2 Concentration Inequalities

Concentration inequalities provide bounds on the fluctuation of functions of independent random variables around their means. They are central tools for designing and analyzing statistical procedures. We refer to the books by Ledoux [84] and by Boucheron, Lugosi, and Massart [33] for detailed accounts on this topic.

B.2.1 McDiarmid Inequality

McDiarmid concentration Inequality [94] is adapted to the setting where the variables are bounded, as in supervised classification.

Theorem B.4 McDiarmid inequality

Let $\mathscr{X}$ be some measurable set and $F : \mathscr{X}^n \to \mathbb{R}$ be a measurable function, such that there exists $\delta_1, \ldots, \delta_n$, fulfilling

$$\left|F(x_1,\ldots,x_i',\ldots,x_n) - F(x_1,\ldots,x_i,\ldots,x_n)\right| \leq \delta_i, \quad \textit{for all} \quad x_1,\ldots,x_n,x_i' \in \mathscr{X},$$

for all $i = 1,\ldots,n$. Then, for any $t > 0$ and any independent random variables $X_1,\ldots,X_n$, with values in $\mathscr{X}$, we have

$$\mathbb{P}\big(F(X_1,\ldots,X_n) > \mathbb{E}\left[F(X_1,\ldots,X_n)\right] + t\big) \leq \exp\left(-\frac{2t^2}{\delta_1^2 + \ldots + \delta_n^2}\right).$$

In other words, under the assumptions of Theorem B.4, there exists a random variable ξ with exponential distribution of parameter 1, such that

$$F(X_1,\ldots,X_n) \leq \mathbb{E}\left[F(X_1,\ldots,X_n)\right] + \sqrt{\frac{\delta_1^2 + \ldots + \delta_n^2}{2}\,\xi}\,.$$

We give here the original proof, combining Markov inequality with a martingale argument due to Azuma [10]. We refer to Boucheron, Lugosi, and Massart [33], Chapter 6, for a more conceptual proof based on the entropy method.

Proof. Let us denote by $\mathscr{F}_k$ the σ-field $\sigma(X_1,\ldots,X_k)$ with $\mathscr{F}_0 = \{\emptyset, \Omega\}$. For simplicity, we write in the following F for $F(X_1,\ldots,X_n)$, and we define for $k = 1,\ldots,n$

$$\Delta_k = \mathbb{E}\left[F|\mathscr{F}_k\right] - \mathbb{E}\left[F|\mathscr{F}_{k-1}\right].$$

Let us fix some $\lambda, t > 0$. The principle of the proof is to start from the Markov inequality

$$\begin{aligned} \mathbb{P}(F > \mathbb{E}[F] + t) &\leq e^{-\lambda t}\,\mathbb{E}\left[e^{\lambda(F - \mathbb{E}[F])}\right] \\ &= e^{-\lambda t}\,\mathbb{E}\left[\prod_{k=1}^{n} e^{\lambda \Delta_k}\right], \end{aligned} \qquad \text{(B.1)}$$

and then apply repeatedly the following lemma.

Lemma B.5

For any $\lambda > 0$ and $k \in \{1,\ldots,n\}$, we have

$$\mathbb{E}\left[e^{\lambda \Delta_k} | \mathscr{F}_{k-1}\right] \leq e^{\lambda^2 \delta_k^2/8}.$$

Proof of the lemma. Let us define the function F_k by $F_k(X_1,\ldots,X_k) = \mathbb{E}\left[F|\mathscr{F}_k\right]$ and the variables S_k and I_k by

$$\begin{aligned} S_k &= \sup_{x\in\mathscr{X}} F_k(X_1,\ldots,X_{k-1},x) - \mathbb{E}\left[F|\mathscr{F}_{k-1}\right] \\ \text{and} \quad I_k &= \inf_{x\in\mathscr{X}} F_k(X_1,\ldots,X_{k-1},x) - \mathbb{E}\left[F|\mathscr{F}_{k-1}\right]. \end{aligned}$$

We have almost surely $I_k \leq \Delta_k \leq S_k$ and $0 \leq S_k - I_k \leq \delta_k$. The convexity of $x \to e^{\lambda x}$ ensures that

$$e^{\lambda\Delta_k} \leq \frac{\Delta_k - I_k}{S_k - I_k}\, e^{\lambda S_k} + \frac{S_k - \Delta_k}{S_k - I_k}\, e^{\lambda I_k}.$$

Since I_k, S_k are $\mathscr{F}_{k-1}$-measurable and $\mathbb{E}\left[\Delta_k|\mathscr{F}_{k-1}\right] = 0$, we obtain

$$\begin{aligned} \mathbb{E}\left[e^{\lambda\Delta_k}\middle|\mathscr{F}_{k-1}\right] &\leq \mathbb{E}\left[\frac{\Delta_k - I_k}{S_k - I_k}\, e^{\lambda S_k} + \frac{S_k - \Delta_k}{S_k - I_k}\, e^{\lambda I_k}\middle|\mathscr{F}_{k-1}\right] \\ &= \frac{-I_k}{S_k - I_k}\, e^{\lambda S_k} + \frac{S_k}{S_k - I_k}\, e^{\lambda I_k} = e^{\phi(\lambda(S_k - I_k))}, \end{aligned}$$

with

$$\phi(x) = \frac{I_k x}{S_k - I_k} + \log\left(\frac{S_k - I_k e^x}{S_k - I_k}\right).$$

Since $0 \leq S_k - I_k \leq \delta_k$, we only have to check that $\phi(x) \leq x^2/8$ for all $x \geq 0$, in order to conclude the proof of the lemma. Straightforward computations give

$$\phi(0) = 0, \quad \phi'(0) = 0, \quad \text{and} \quad \phi''(x) = \frac{-S_k I_k e^x}{(S_k - I_k e^x)^2} \leq \frac{1}{4},$$

where the last inequality follows from $4ab \leq (a+b)^2$. Taylor inequality then ensures that

$$\phi(x) \leq \frac{x^2}{2} \sup_{t\in[0,x]} \phi''(t) \leq \frac{x^2}{8}, \quad \text{for all } x \geq 0,$$

which conclude the proof of Lemma B.5. □

Let us now prove the McDiarmid inequality. Applying repeatedly Lemma B.5, we obtain

$$\begin{aligned} \mathbb{E}\left[\prod_{k=1}^{n} e^{\lambda\Delta_k}\right] &= \mathbb{E}\left[\prod_{k=1}^{n-1} e^{\lambda\Delta_k}\mathbb{E}\left[e^{\lambda\Delta_n}|\mathscr{F}_{n-1}\right]\right] \\ &\leq e^{\lambda^2\delta_n^2/8}\mathbb{E}\left[\prod_{k=1}^{n-1} e^{\lambda\Delta_k}\right] \leq \ldots \leq e^{\lambda^2\sum_{k=1}^{n}\delta_k^2/8}. \end{aligned}$$

Combining this inequality with (B.1), we get

$$\mathbb{P}(F > \mathbb{E}[F] + t) \leq e^{-\lambda t}\, e^{\lambda^2\sum_{k=1}^{n}\delta_k^2/8}.$$

For $\lambda = 4t\left(\sum_{k=1}^{n}\delta_k^2\right)^{-1}$, it gives the McDiarmid inequality. □

B.2.2 Gaussian Concentration Inequality

Lipschitz functions of Gaussian random variables also fulfill a good concentration around their mean.

Theorem B.6 *Assume that $F : \mathbb{R}^d \to \mathbb{R}$ is 1-Liptschitz and Z has a Gaussian $\mathcal{N}(0, \sigma^2 I_d)$ distribution.*

Then, there exists a variable ξ with exponential distribution of parameter 1, such that

$$F(Z) \le \mathbb{E}[F(Z)] + \sigma\sqrt{2\xi}\,. \tag{B.2}$$

A striking point with the Gaussian concentration is that the size of the fluctuation $\sigma\sqrt{2\xi}$ does not depend on the dimension d. We give here a short proof based on Itô calculus due to Ibragimov, Sudakov, and Tsirel'son [75]. We refer to Ledoux [84] for some more classical and elementary proofs and further references.

Proof. We can assume in the following that $Z \sim \mathcal{N}(0, I_d)$, since $F(Z) = \sigma\widetilde{F}(Z/\sigma)$, where $\widetilde{F}(x) = \sigma^{-1}F(\sigma x)$ is 1-Lipschitz and Z/σ has a $\mathcal{N}(0, I_d)$ distribution. Let $(W_t)_{t\ge 0}$ be a standard brownian motion in $\mathbb{R}^d$ and set for $x \in \mathbb{R}^d$ and $t \in [0,1]$

$$G(t,x) = \mathbb{E}\left[F(x + W_{1-t})\right].$$

The function G is continuous on $[0,1] \times \mathbb{R}^d$, differentiable in t, and infinitely differentiable in x for any $t \in [0,1[$. Furthermore, since the infinitesimal generator of the standard brownian motion is $\frac{1}{2}\Delta$, we have for all $(t,x) \in]0,1[\times\mathbb{R}^d$

$$\frac{\partial}{\partial t}G(t,x) = -\frac{1}{2}\Delta_x G(t,x), \tag{B.3}$$

where Δ_x represents the Laplacian in the variable x. Let $(B_t)_{t\ge 0}$ be another standard brownian motion in $\mathbb{R}^d$, independent of W. Itô's formula (see Revuz-Yor [101], Chapter 4, Theorem 3.3) gives

$$\begin{aligned}\underbrace{G(1,B_1)}_{=F(B_1)} &= G(0,0) + \int_0^1 \frac{\partial}{\partial t}G(s,B_s)\,ds + \int_0^1 \nabla_x G(s,B_s)\cdot dB_s + \frac{1}{2}\int_0^1 \Delta_x G(s,B_s)\,ds\\ &= \underbrace{G(0,0)}_{=\mathbb{E}[F(W_1)]} + \int_0^1 \nabla_x G(s,B_s)\cdot dB_s,\end{aligned}$$

where the last equality follows from Equation (B.3). The stochastic integral $\int_0^t \nabla_x G(s,B_s)\cdot dB_s$ defines a continuous martingale starting from 0, so according to Dubins–Schwarz's representation of continuous martingales ([101], Chapter 5, Theorem 1.6), there exists a 1-dimensional standard brownian motion β, such that

$$\int_0^1 \nabla_x G(s,B_s)\cdot dB_s = \beta_{T_1}\,, \quad \text{where} \quad T_1 = \int_0^1 \|\nabla_x G(s,B_s)\|^2 ds\,.$$

Since F is 1-Lipschitz, the function $x \to G(t,x)$ is 1-Lipschitz for all $t \in [0,1]$, and therefore $\|\nabla_x G(t,x)\| \leq 1$ for all $(t,x) \in [0,1[\times\mathbb{R}^d$. As a consequence, we have that $T_1 \leq 1$ a.s. and

$$\left|\int_0^1 \nabla_x G(s,B_s)\cdot dB_s\right| \leq \sup_{t\in[0,1]} \beta_t \quad \text{a.s.}$$

Finally, the random variable $\sup_{t\in[0,1]} \beta_t$ has the same distribution as the absolute value of a standard Gaussian random variable ([101], Chapter 3, Theorem 3.7), so

$$\begin{aligned}\mathbb{P}\left(F(B_1) \geq \mathbb{E}\left[F(W_1)\right] + x\right) &\leq \mathbb{P}\left(\sup_{t\in[0,1]} \beta_t \geq x\right)\\ &\leq 2\int_x^{+\infty} e^{-t^2/2}\,\frac{dt}{\sqrt{2\pi}} \leq e^{-x^2/2}.\end{aligned}$$

We conclude by noticing that B_1 and W_1 have a Gaussian $\mathcal{N}(0,I_d)$ distribution. □

We note that replacing F by $-F$, we also have $F(Z) \geq \mathbb{E}[F(Z)] - \sigma\sqrt{2\xi'}$ for some ξ' with exponential distribution of parameter 1.

Remark 1. A typical example of use of the Gaussian concentration inequality is with the norm $\|Z\|$ of a $\mathcal{N}(0,\sigma^2 I_d)$ Gaussian random variable. Since the norm is 1-Liptschitz, and since we have $\mathbb{E}[\|Z\|] \leq \sqrt{\mathbb{E}[\|Z\|^2]} = \sigma\sqrt{d}$, the norm $\|Z\|$ fulfills the inequality

$$\|Z\| \leq \sigma\sqrt{d} + \sigma\sqrt{2\xi}\,,$$

where ξ is an exponential random variable of parameter 1.

Remark 2. Another consequence of the Gaussian concentration is that the variance of the norm $\|Z\|$ of a $\mathcal{N}(0,\sigma^2 I_d)$ Gaussian random variable can be bounded independently of the dimension d. Actually, there exists two standard exponential random variable ξ and ξ', such that

$$\mathbb{E}[\|Z\|] - \sigma\sqrt{2\xi'} \leq \|Z\| \leq \mathbb{E}[\|Z\|] + \sigma\sqrt{2\xi}\,.$$

Accordingly, we have the upper bound

$$\begin{aligned}\mathrm{var}[\|Z\|] &= \mathbb{E}\left[(\|Z\| - \mathbb{E}[\|Z\|])_+^2\right] + \mathbb{E}\left[(\mathbb{E}[\|Z\|] - \|Z\|)_+^2\right]\\ &\leq 2\mathbb{E}[\xi]\sigma^2 + 2\mathbb{E}[\xi']\sigma^2 = 4\sigma^2.\end{aligned}$$

As a consequence, we have the following bounds on the expectation of $\|Z\|$

$$(d-4)\,\sigma^2 \leq \mathbb{E}\left[\|Z\|^2\right] - \mathrm{var}[\|Z\|] = \mathbb{E}\left[\|Z\|\right]^2 \leq \mathbb{E}\left[\|Z\|^2\right] = d\,\sigma^2. \tag{B.4}$$

B.3 Symmetrization and Contraction Lemmas

B.3.1 Symmetrization Lemma

The symmetrization lemma is a simple bound, very useful for bounding the expectation of suprema of empirical processes, as, for example, in the proof of Theorem 9.1 in Chapter 9.

Theorem B.7 Symmetrization lemma

Let $\mathscr{Z}$ be a measurable set, and $\mathscr{F}$ be a set of integrable functions $f : \mathscr{Z} \to \mathbb{R}$. Let $Z_1, \ldots, Z_n$ be i.i.d. random variables in $\mathscr{Z}$, and $\sigma_1, \ldots, \sigma_n$ be i.i.d. random variables independent of $Z_1, \ldots, Z_n$, with uniform distribution on $\{-1, +1\}$. Then, we have

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-\mathbb{E}[f(Z_i)])\right|\right] \leq 2\mathbb{E}\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right], \qquad \text{(B.5)}$$

where $\mathbb{E}$ refers to the expectation with respect to $Z_1, \ldots, Z_n$, and $\mathbb{E}_\sigma$ refers to the expectation with respect to $\sigma_1, \ldots, \sigma_n$.

Proof. Let $(\widetilde{Z}_i)_{i=1,\ldots,n}$ be an independent copy of $(Z_i)_{i=1,\ldots,n}$. We first observe that

$$\mathbb{E}[f(Z_i)] = \widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{Z}_i)\right],$$

where $\widetilde{\mathbb{E}}$ refers to the expectation with respect to the random variables $(\widetilde{Z}_i)_{i=1,\ldots,n}$. According to Jensen inequality, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-\mathbb{E}[f(Z_i)])\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{Z}_i)\right]\right|\right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathscr{F}}\widetilde{\mathbb{E}}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \frac{1}{n}\sum_{i=1}^{n} f(\widetilde{Z}_i)\right|\right]\right] \\ &\leq \mathbb{E}\widetilde{\mathbb{E}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(f(Z_i)-f(\widetilde{Z}_i)\right)\right|\right].\end{aligned}$$

By symmetry, we notice that $\left(\sigma_i\big(f(Z_i)-f(\widetilde{Z}_i)\big)\right)_{i=1,\ldots,n}$ has the same distribution as

$\Big(f(Z_i) - f(\widetilde{Z}_i)\Big)_{i=1,\dots,n}$, so the previous bound and the triangular inequality give

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-\mathbb{E}\left[f(Z_i)\right])\right|\right]\\
&\leq \mathbb{E}\widetilde{\mathbb{E}}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(f(Z_i)-f(\widetilde{Z}_i)\right)\right|\right]\\
&\leq \mathbb{E}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right]+\widetilde{\mathbb{E}}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\widetilde{Z}_i)\right|\right]\\
&\leq 2\mathbb{E}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right].
\end{aligned}$$

The proof of Theorem B.7 is complete. □

B.3.2 Contraction Principle

The Contraction Principle of Ledoux and Talagrand (Theorem 4.12 in [85]) is a useful tool for analyzing empirical processes, as, for example, in the proof of Theorem 9.10 in Chapter 9.

Theorem B.8 Contraction Principle

Let $\mathscr{Z}$ be a bounded subset of $\mathbb{R}^n$, and $\varphi:\mathbb{R}\to\mathbb{R}$ be an α-Lipschitz function fulfilling $\varphi(0)=0$. For $\sigma_1,\dots,\sigma_n$ i.i.d. random variables with distribution $\mathbb{P}_{\sigma}(\sigma_i=1)=\mathbb{P}_{\sigma}(\sigma_i=-1)=1/2$, we have

$$\mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left|\sum_{i=1}^{n}\sigma_i\varphi(z_i)\right|\right]\leq 2\alpha\,\mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left|\sum_{i=1}^{n}\sigma_i z_i\right|\right]. \tag{B.6}$$

Proof. We will actually prove the following stronger result: For any function $g:\mathbb{R}^n\to\mathbb{R}$ and any integer $d\leq n$,

$$\mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left(g(z)+\sum_{i=1}^{d}\sigma_i\varphi(z_i)\right)_+\right]\leq \alpha\,\mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left(g(z)+\sum_{i=1}^{d}\sigma_i z_i\right)_+\right]. \tag{B.7}$$

Let us first check that (B.6) follows from (B.7). Since $|x|=(x)_+ + (-x)_+$, we have

$$\mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left|\sum_{i=1}^{n}\sigma_i\varphi(z_i)\right|\right]\leq \mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left(\sum_{i=1}^{n}\sigma_i\varphi(z_i)\right)_+\right]+\mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left(-\sum_{i=1}^{n}\sigma_i\varphi(z_i)\right)_+\right].$$

The random variable $-\sigma$ has the same distribution as σ, so applying (B.7) with $g=0$

and $d = n$, we obtain

$$\begin{aligned}\mathbb{E}_\sigma\left[\sup_{z\in\mathscr{Z}}\left|\sum_{i=1}^n \sigma_i\varphi(z_i)\right|\right] &\leq 2\alpha\,\mathbb{E}_\sigma\left[\sup_{z\in\mathscr{Z}}\left(\sum_{i=1}^n \sigma_i z_i\right)_+\right]\\ &\leq 2\alpha\,\mathbb{E}_\sigma\left[\sup_{z\in\mathscr{Z}}\left|\sum_{i=1}^n \sigma_i z_i\right|\right],\end{aligned}$$

which gives (B.6).

To conclude the proof of Theorem B.8, it remains to prove (B.7). Replacing φ by φ/α, we can assume that φ is 1-Lipschitz. The following technical lemma is the key of the proof.

Lemma B.9

Let $\mathscr{Z}$ be a bounded subset of $\mathbb{R}^n$, and consider $\varphi : \mathbb{R} \to \mathbb{R}$ a 1-Lipschitz function fulfilling $\varphi(0) = 0$. Then for any function $g : \mathbb{R}^n \to \mathbb{R}$, we have

$$\sup_{z,z'\in\mathscr{Z}}\left\{(g(z)+\varphi(z_1))_+ + (g(z')-\varphi(z'_1))_+\right\} \leq \sup_{z,z'\in\mathscr{Z}}\left\{(g(z)+z_1)_+ + (g(z')-z'_1)_+\right\}$$

where z_1 denotes the first coordinate of z.

Proof of Lemma B.9. For any $z, z' \in \mathscr{Z}$, we have for any $\varphi : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$

$$\begin{aligned}&(g(z)+\varphi(z_1))_+ + (g(z')-\varphi(z'_1))_+\\ &\quad= \max\left\{g(z)+\varphi(z_1)+g(z')-\varphi(z'_1), g(z)+\varphi(z_1), g(z')-\varphi(z'_1), 0\right\}. \quad \text{(B.8)}\end{aligned}$$

Let us bound each term in the right-hand side maximum.

1. Let us bound the first term. Since φ is 1-Lipschitz, we have

$$\begin{aligned}g(z)+\varphi(z_1)+g(z')-\varphi(z'_1) &\leq g(z)+g(z')+|z_1-z'_1|\\ &\leq \sup_{z,z'\in\mathscr{Z}}\left\{g(z)+g(z')+|z_1-z'_1|\right\}.\end{aligned}$$

   Due to the symmetry in z and z', we have

$$\sup_{z,z'\in\mathscr{Z}}\left\{g(z)+g(z')+|z_1-z'_1|\right\} = \sup_{z,z'\in\mathscr{Z}}\left\{g(z)+g(z')+z_1-z'_1\right\},$$

   and therefore

$$\begin{aligned}g(z)+\varphi(z_1)+g(z')-\varphi(z'_1) &\leq \sup_{z,z'\in\mathscr{Z}}\left\{g(z)+g(z')+z_1-z'_1\right\}\\ &\leq \sup_{z,z'\in\mathscr{Z}}\left\{(g(z)+z_1)_+ + (g(z')-z'_1)_+\right\}. \quad \text{(B.9)}\end{aligned}$$

2. Let us bound the two other terms in the right-hand side of (B.8). Since φ is 1-Lipschitz and $\varphi(0) = 0$, we have $|\varphi(z_1)| \leq |z_1|$ for all $z_1 \in \mathbb{R}$. This inequality

gives

$$\begin{aligned} g(z) \pm \varphi(z_1) &\leq g(z) + |z_1| \\ &\leq (g(z)+z_1)_+ + (g(z)-z_1)_+ \\ &\leq \sup_{z,z' \in \mathscr{Z}} \left\{ (g(z)+z_1)_+ + (g(z')-z'_1)_+ \right\}. \end{aligned} \tag{B.10}$$

Combining (B.8) with (B.9) and (B.10) completes the proof of Lemma B.9. □

We now prove (B.7) by induction on d.

- For $d = 1$, we have

$$\begin{aligned} \mathbb{E}_{\sigma_1}\left[\sup_{z\in\mathscr{Z}} (g(z)+\sigma_1\varphi(z_1))_+\right] &= \frac{1}{2}\sup_{z\in\mathscr{Z}} (g(z)+\varphi(z_1))_+ + \frac{1}{2}\sup_{z'\in\mathscr{Z}} \big(g(z')-\varphi(z'_1)\big)_+ \\ &= \frac{1}{2}\sup_{z,z'\in\mathscr{Z}} \Big\{ (g(z)+\varphi(z_1))_+ + \big(g(z')-\varphi(z'_1)\big)_+ \Big\}. \end{aligned}$$

Lemma B.9 ensures that

$$\begin{aligned} \mathbb{E}_{\sigma_1}\left[\sup_{z\in\mathscr{Z}} (g(z)+\sigma_1\varphi(z_1))_+\right] &\leq \frac{1}{2}\sup_{z,z'\in\mathscr{Z}} \left\{ (g(z)+z_1)_+ + (g(z')-z'_1)_+ \right\} \\ &= \mathbb{E}_{\sigma_1}\left[\sup_{z\in\mathscr{Z}} (g(z)+\sigma_1 z_1)_+\right], \end{aligned}$$

which proves (B.7) for $d = 1$.

- Assume now that Inequality (B.7) has been proved up to $d = k-1$, with $k \geq 2$. Below, we will denote by $\mathbb{E}_{\sigma_k}$ the expectation under the variable σ_k and by $\mathbb{E}_{\sigma_{-k}}$ the expectation under the variables $\sigma_1, \dots, \sigma_{k-1}$. Applying successively Inequality (B.7) with $d = k-1$ and $d = 1$, we obtain

$$\begin{aligned} &\mathbb{E}_{\sigma}\left[\sup_{z\in\mathscr{Z}}\left(g(z)+\sum_{i=1}^{k}\sigma_i\varphi(z_i)\right)_+\right] \\ &= \mathbb{E}_{\sigma_k}\left[\mathbb{E}_{\sigma_{-k}}\left[\sup_{z\in\mathscr{Z}}\left(\big(g(z)+\sigma_k\varphi(z_k)\big)+\sum_{i=1}^{k-1}\sigma_i\varphi(z_i)\right)_+\right]\right] \quad \text{(Fubini)} \\ &\leq \mathbb{E}_{\sigma_k}\left[\mathbb{E}_{\sigma_{-k}}\left[\sup_{z\in\mathscr{Z}}\left(\big(g(z)+\sigma_k\varphi(z_k)\big)+\sum_{i=1}^{k-1}\sigma_i z_i\right)_+\right]\right] \quad \text{(Ineq. (B.7) with } d=k-1) \\ &= \mathbb{E}_{\sigma_{-k}}\left[\mathbb{E}_{\sigma_k}\left[\sup_{z\in\mathscr{Z}}\left(\Big(g(z)+\sum_{i=1}^{k-1}\sigma_i z_i\Big)+\sigma_k\varphi(z_k)\right)_+\right]\right] \quad \text{(Fubini)} \\ &\leq \mathbb{E}_{\sigma_{-k}}\left[\mathbb{E}_{\sigma_k}\left[\sup_{z\in\mathscr{Z}}\left(g(z)+\sum_{i=1}^{k}\sigma_i z_i\right)_+\right]\right] \quad \text{(Inequality (B.7) with } d=1), \end{aligned}$$

which gives (B.7) for $d = k$. By induction, this completes the proof of (B.7) for any $d \leq n$. □

B.4 Birgé's Inequality

Birgé's inequality [27] is very useful for deriving minimax lower bounds, as, for example, in Theorem 2.3 and Exercise 2.9.6. We state here a simple version of the inequality, and we refer to Corollary 2.18 in Massart's Lecture Notes [93] for a stronger result.

In the following, we denote by

$$\mathscr{K}(\mathbb{P},\mathbb{Q}) = \begin{cases} \int \log \frac{d\mathbb{P}}{d\mathbb{Q}}\, d\mathbb{P} & \text{when } \mathbb{P} \ll \mathbb{Q} \\ +\infty & \text{else} \end{cases}$$

the Kullback–Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$

Theorem B.10 Birgé's Inequality
Let us consider a family $(A_i)_{i=1,\dots,N}$ of disjointed events, and a collection $(\mathbb{P}_i)_{i=1,\dots,N}$ of probability measures. Then, we have

$$\min_{i=1,\dots,N} \mathbb{P}_i(A_i) \leq \frac{2e}{2e+1} \bigvee \frac{\max_{i\neq j} \mathscr{K}(\mathbb{P}_i,\mathbb{P}_j)}{\log(N)}. \tag{B.11}$$

Proof. We can assume that $\mathbb{P}_i \ll \mathbb{P}_j$ for all $i \neq j$, otherwise $\max_{i\neq j} \mathscr{K}(\mathbb{P}_i,\mathbb{P}_j) = +\infty$ and (B.11) is obvious. The proof is mainly based on the following simple inequality.

Lemma B.11

Let X be a bounded random variable and $\mathbb{P}_1, \mathbb{P}_2$ be two probability measures, such that $\mathbb{P}_1 \ll \mathbb{P}_2$ and $\mathbb{P}_2 \ll \mathbb{P}_1$. Then, we have

$$\mathbb{E}_2[X] - \log \mathbb{E}_1\left[e^X\right] \leq \mathscr{K}(\mathbb{P}_2,\mathbb{P}_1), \tag{B.12}$$

where $\mathbb{E}_i$ denotes the expectation with respect to $\mathbb{P}_i$.

Proof of the lemma. Since $-\log$ is convex, Jensen inequality ensures that

$$\begin{aligned} -\log \mathbb{E}_1\left[e^X\right] &= -\log\left(\int e^X \frac{d\mathbb{P}_1}{d\mathbb{P}_2}\, d\mathbb{P}_2\right) \\ &\leq \int -\log\left(e^X \frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) d\mathbb{P}_2 = -\mathbb{E}_2[X] + \mathscr{K}(\mathbb{P}_2,\mathbb{P}_1), \end{aligned}$$

which concludes the proof of Lemma B.11. □

We set $m = \min_{i=1,\dots,N} \mathbb{P}_i(A_i)$, and for $i \leq N-1$, we define $X = \mathbf{1}_{A_i} \log(m/q)$ with $q = (1-m)/(N-1)$. Inequality (B.12) gives

$$\mathbb{P}_i(A_i)\log(m/q) - \log \mathbb{E}_N\left[\left(\frac{m}{q}\right)^{\mathbf{1}_{A_i}}\right] \leq \mathscr{K}(\mathbb{P}_i,\mathbb{P}_N).$$

Computing the expectation, we find

$$\begin{aligned}\log \mathbb{E}_N\left[\left(\frac{m}{q}\right)^{\mathbf{1}_{A_i}}\right] &= \log\left(\mathbb{P}_N(A_i)\left(\frac{m}{q}-1\right)+1\right)\\ &\leq \mathbb{P}_N(A_i)\left(\frac{m}{q}-1\right) \leq \frac{m}{q}\,\mathbb{P}_N(A_i).\end{aligned}$$

We have $\mathbb{P}_i(A_i) \geq m$, so averaging over $i \in \{1,\ldots,N-1\}$ gives

$$m\log(m/q) - \frac{m}{q(N-1)}\sum_{i=1}^{N-1}\mathbb{P}_N(A_i) \leq \overline{\mathscr{K}} = \frac{1}{N-1}\sum_{i=1}^{N-1}\mathscr{K}\left(\mathbb{P}_i,\mathbb{P}_N\right).$$

Since

$$\sum_{i=1}^{N-1}\mathbb{P}_N(A_i) \leq 1-\mathbb{P}_N(A_N) \leq 1-m \quad \text{and} \quad q=\frac{1-m}{N-1},$$

we finally obtain

$$m\log\left(\frac{m(N-1)}{e(1-m)}\right) \leq \overline{\mathscr{K}}.$$

To conclude the proof of (B.11), we simply check that for $m \geq 2e/(2e+1)$, we have

$$\log\left(\frac{m(N-1)}{e(1-m)}\right) \geq \log(2(N-1)) \geq \log(N),$$

and that $\overline{\mathscr{K}} \leq \max_{i\neq j}\mathscr{K}(\mathbb{P}_i,\mathbb{P}_j)$. □

Appendix C

Linear Algebra

C.1 Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is a matrix decomposition that is very useful in many fields of applied mathematics. In the following, we will use that, for any $n \times p$ matrix A, the matrices $A^T A$ and AA^T are symmetric positive semidefinite.

Theorem C.1 Singular value decomposition
Any $n \times p$ matrix A of rank r can be decomposed as

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T, \tag{C.1}$$

where

- $r = \mathrm{rank}(A)$,
- $\sigma_1 \geq \ldots \geq \sigma_r > 0$,
- $\{\sigma_1^2, \ldots, \sigma_r^2\}$ *are the nonzero eigenvalues of $A^T A$ (which are also the nonzero eigenvalues of AA^T), and*
- $\{u_1, \ldots, u_r\}$ *and* $\{v_1, \ldots, v_r\}$ *are two orthonormal families of $\mathbb{R}^n$ and $\mathbb{R}^p$, such that*
$$AA^T u_j = \sigma_j^2 u_j \quad \textit{and} \quad A^T A v_j = \sigma_j^2 v_j.$$

The values $\sigma_1, \ldots, \sigma_r$ are called the singular values of A. The vectors $\{u_1, \ldots, u_r\}$ and $\{v_1, \ldots, v_r\}$ are said to be left-singular vectors and right-singular vectors, respectively.

Proof. Let us prove that such a decomposition exists. Since AA^T is positive semidefinite, we have a spectral decomposition

$$AA^T = \sum_{j=1}^{r} \lambda_j u_j u_j^T,$$

with $\lambda_1 \geq \ldots \geq \lambda_r > 0$ and $\{u_1, \ldots, u_r\}$ an orthonormal family of $\mathbb{R}^n$. Let us define $v_1, \ldots, v_r$ by $v_j = \lambda_j^{-1/2} A^T u_j$ for $j = 1, \ldots, r$. We have

$$\|v_j\|^2 = \lambda_j^{-1} u_j^T AA^T u_j = u_j^T u_j = 1,$$

and

$$A^T A v_j = \lambda_j^{-1/2} A^T (AA^T) u_j = \lambda_j^{1/2} A^T u_j = \lambda_j v_j,$$

so $\{v_1, \dots, v_r\}$ is an orthonormal family of eigenvectors of $A^T A$. Setting $\sigma_j = \lambda_j{}^{1/2}$, we obtain

$$\begin{aligned}\sum_{j=1}^{r} \sigma_j u_j v_j^T &= \sum_{j=1}^{r} \lambda_j^{1/2}\, \lambda_j^{-1/2} u_j u_j^T A \\ &= \left(\sum_{j=1}^{r} u_j u_j^T \right) A.\end{aligned}$$

We notice that $\sum_{j=1}^{r} u_j u_j^T$ is the projection onto the range of AA^T. To conclude, we recall that $\mathbb{R}^p = \ker(A) \bigoplus \text{range}(A^T)$ is the orthogonal sum of $\ker(A)$ and $\text{range}(A^T)$, so the range of A and the range of AA^T coincide and

$$\sum_{j=1}^{r} \sigma_j u_j v_j^T = \left(\sum_{j=1}^{r} u_j u_j^T \right) A = \text{Proj}_{\text{range}(A)}\, A = A.$$

The proof of Lemma C.1 is complete. □

C.2 Moore–Penrose Pseudo-Inverse

The Moore–Penrose pseudo-inverse A^+ of a matrix A generalizes the notion of inverse for singular matrices. It is a matrix such that $AA^+ y = y$ for all y in the range of A and $A^+ A x = x$ for all x in the range of A^+. Furthermore, the matrices AA^+ and $A^+ A$ are symmetric. When A is nonsingular, we have the identity $A^+ = A^{-1}$. We first describe A^+ for diagonal matrices, then for normal matrices, and finally for arbitrary matrices.

Diagonal matrices

The Moore–Penrose pseudo-inverse of a diagonal matrix D is a diagonal matrix D^+, with diagonal entries $[D^+]_{jj} = 1/D_{jj}$ when $D_{jj} \neq 0$ and $[D^+]_{jj} = 0$ otherwise.

Normal matrices[1]

Write $A = UDU^T$ for a spectral decomposition of A with D diagonal and U unitary[2]. The Moore–Penrose pseudo-inverse of A is given by $A^+ = UD^+U^T$.

Arbitrary matrices

Write $A = \sum_{j=1}^{r} \sigma_j(A) u_j v_j^T$ for a singular value decomposition of A with $r = \text{rank}(A)$. The Moore–Penrose pseudo-inverse of A is given by

$$A^+ = \sum_{j=1}^{r} \sigma_j(A)^{-1} v_j u_j^T.$$

[1] A is normal if $A^T A = AA^T$. In particular, symmetric matrices are normal.

[2] U unitary if$U^T U = UU^T = I$.

We notice that

$$A^+A = \sum_{j=1}^{r} v_j v_j^T = \text{Proj}_{\text{range}(A^T)} \quad \text{and} \quad AA^+ = \sum_{j=1}^{r} u_j u_j^T = \text{Proj}_{\text{range}(A)}.$$

In particular, $AA^+ = A^+A = I$ when A is nonsingular.

C.3 Matrix Norms

In the following, we denote by $\sigma_1(A) \geq \sigma_2(A) \geq \ldots$ the singular values of A. Several interesting norms are related to the singular values.

Frobenius norm

The standard scalar product on matrices is $\langle A, B\rangle_F = \sum_{i,j} A_{i,j} B_{i,j}$. It induces the Frobenius norm

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 = \text{Tr}(A^T A) = \sum_k \sigma_k(A)^2.$$

The last equality follows from the fact that the $\sigma_k(A)^2$ are the eigenvalues of $A^T A$.

Operator norm

The $\ell^2 \to \ell^2$ operator norm is defined by

$$|A|_{\text{op}} = \sup_{\|x\| \leq 1} \|Ax\| = \sigma_1(A).$$

Let us prove this last equality. We have $Ax = \sum_k \sigma_k(A) u_k v_k^T x$, so $\|Ax\|^2 = \sum_k \sigma_k^2(A) \langle v_k, x\rangle^2 \leq \sigma_1(A)^2 \|x\|^2$, with equality for $x = v_1$.

Nuclear norm

The nuclear norm is defined by

$$|A|_* = \sum_{k=1}^{r} \sigma_k(A).$$

The three following inequalities are very useful.

Lemma C.2 *We have*

1. $|A|_* \leq \sqrt{\text{rank}(A)}\, \|A\|_F$,
2. $\langle A, B\rangle_F \leq |A|_* |B|_{\text{op}}$,
3. $\|AB\|_F \leq |A|_{\text{op}} \|B\|_F$.

Proof. The first inequality is simply Cauchy–Schwartz inequality. For the second inequality, we start from

$$\langle A, B\rangle_F = \sum_k \sigma_k(A) \langle u_k v_k^T, B\rangle_F = \sum_k \sigma_k(A) \langle u_k, B v_k\rangle$$

and notice that $\langle u_k, Bv_k \rangle_F \leq \|Bv_k\| \leq |B|_{\text{op}}$ since $\|u_k\| = \|v_k\| = 1$. The inequality

$$\langle A, B \rangle_F \leq \sum_k \sigma_k(A) |B|_{\text{op}}$$

then follows. Let us turn to the third inequality. We denote by B_j the j-th column of B. We observe that $\|B\|_F^2 = \sum_j \|B_j\|^2$, so

$$\|AB\|_F^2 = \sum_j \|(AB)_j\|^2 = \sum_j \|AB_j\|^2 \leq \sum_j |A|_{\text{op}}^2 \|B_j\|^2 = |A|_{\text{op}}^2 \|B\|_F^2.$$

The proof of Lemma C.2 is complete. □

C.4 Matrix Analysis

We present in this section some useful results on singular values. The first result is a geometric characterization of the singular values.

Theorem C.3 Max–Min formula
For an $n \times p$ matrix A and $k \leq \min(n,p)$, we have

$$\sigma_k(A) = \max_{S:\dim(S)=k} \min_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|}, \tag{C.2}$$

where the maximum is taken over all the linear spans $S \subset \mathbb{R}^p$ with dimension k.

Proof. We start from the singular value decomposition $A = \sum_{j=1}^r \sigma_j(A) u_j v_j^T$ and we consider $\{v_{r+1}, \ldots, v_p\}$, such that $\{v_1, \ldots, v_p\}$ is an orthonormal basis of $\mathbb{R}^p$. We define $S_k = \text{span}\{v_1, \ldots, v_k\}$ and $W_k = \text{span}\{v_k, \ldots, v_p\}$. For any linear span $S \subset \mathbb{R}^p$ with dimension k, we have $\dim(S) + \dim(W_k) = p + 1$, so $S \cap W_k \neq \{0\}$. For any nonzero $x \in S \cap W_k$ we have

$$\frac{\|Ax\|^2}{\|x\|^2} = \frac{\sum_{j=k}^r \sigma_j(A)^2 \langle v_j, x \rangle^2}{\sum_{j=k}^p \langle v_j, x \rangle^2} \leq \sigma_k(A)^2,$$

so

$$\max_{S:\dim(S)=k} \min_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|} \leq \sigma_k(A).$$

Conversely, for all $x \in S_k \setminus \{0\}$, we have

$$\frac{\|Ax\|^2}{\|x\|^2} = \frac{\sum_{j=1}^k \sigma_j(A)^2 \langle v_j, x \rangle^2}{\sum_{j=1}^k \langle v_j, x \rangle^2} \geq \sigma_k(A)^2,$$

with equality for $x = v_k$. As a consequence,

$$\max_{S:\dim(S)=k} \min_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \sigma_k(A),$$

with equality for $S = S_k$. □

Corollary C.4 *For an $n \times p$ matrix A and $k \leq \min(n, p)$, we have for any orthogonal projector P in $\mathbb{R}^n$*

$$\sigma_k(PA) \leq \sigma_k(A) \quad \textit{and} \quad \|PA\|_F \leq \|A\|_F. \tag{C.3}$$

Similarly, we have for any orthogonal projector P in $\mathbb{R}^p$

$$\sigma_k(AP) \leq \sigma_k(A) \quad \textit{and} \quad \|AP\|_F \leq \|A\|_F. \tag{C.4}$$

Proof. A projector is 1-Lipschitz so $\|PAx\| \leq \|Ax\|$. The inequalities (C.3) then follows from (C.2). Furthermore, we have $\sigma_k(AP) = \sigma_k(PA^T) \leq \sigma_k(A^T) = \sigma_k(A)$, which gives (C.4). □

The second result characterizes the "projection" on the set of matrices of rank r. It also provides an improvement of the Cauchy–Schwartz inequality $\langle A, B \rangle_F \leq \|A\|_F \|B\|_F$ in terms of the Ky–Fan $(2, q)$-norm

$$\|A\|_{(2,q)}^2 = \sum_{k=1}^{q} \sigma_k(A)^2, \tag{C.5}$$

with $q = \text{rank}(A) \wedge \text{rank}(B)$. We observe that $\|A\|_{(2,q)} \leq \|A\|_F$, with strict inequality if $q < \text{rank}(A)$.

Theorem C.5 *For any matrices $A, B \in \mathbb{R}^{n \times p}$, we set $q = \text{rank}(A) \wedge \text{rank}(B)$. We then have*

$$\langle A, B \rangle_F \leq \|A\|_{(2,q)} \, \|B\|_{(2,q)},$$

where the Ky–Fan $(2, q)$-norm $\|A\|_{(2,q)}$ is defined in (C.5).

As a consequence, for $A = \sum_{k=1}^{r} \sigma_k(A) u_k v_k^T$ and $q < r$, we have

$$\min_{B:\text{rank}(B) \leq q} \|A - B\|_F^2 = \sum_{k=q+1}^{r} \sigma_k(A)^2.$$

In addition, the minimum is achieved for

$$B = \sum_{k=1}^{q} \sigma_k(A) u_k v_k^T.$$

Proof. We can assume, e.g., that the rank of B is not larger than the rank of A. Let us denote by q the rank of B and P_B the projection on the range of B. We have

$$\langle A, B \rangle_F = \langle P_B A, B \rangle_F \leq \|P_B A\|_F \|B\|_F.$$

The rank of $P_B A$ is at most q and previous corollary ensures that $\sigma_k(P_B A) \leq \sigma_k(A)$, so

$$\|P_B A\|_F^2 = \sum_{k=1}^{q} \sigma_k(P_B A)^2 \leq \sum_{k=1}^{q} \sigma_k(A)^2 = \|A\|_{(2,q)}^2.$$

Since $q = \text{rank}(B)$, we have $\|B\|_F = \|B\|_{(2,q)}$, and the first part of the theorem is proved.

According to the first part of the theorem, for any matrix B of rank q, we have

$$\|A-B\|_F^2 = \|A\|_F^2 - 2\langle A,B\rangle_F + \|B\|_F^2 \geq \|A\|_F^2 - 2\|A\|_{(2,q)}\|B\|_F + \|B\|_F^2.$$

The right-hand side is minimum for $\|B\|_F = \|A\|_{(2,q)}$, so

$$\|A-B\|_F^2 \geq \|A\|_F^2 - \|A\|_{(2,q)}^2 = \sum_{k=q+1}^{r} \sigma_k(A)^2.$$

Finally, we observe that this lower bound is achieved for $B = \sum_{k=1}^{q} \sigma_k(A) u_k v_k^T$. □

The last result states that the singular values are 1-Lipschitz with respect to the operator norm.

Theorem C.6 Weyl inequality
For two $n \times p$ matrices A and B, we have for any $k \leq \min(n,p)$

$$|\sigma_k(A) - \sigma_k(B)| \leq \sigma_1(A-B) = |A-B|_{\text{op}}.$$

Proof. For any $x \in \mathbb{R}^p \setminus \{0\}$, we have

$$\frac{\|Ax\|}{\|x\|} \leq \frac{\|Bx\|}{\|x\|} + \frac{\|(A-B)x\|}{\|x\|} \leq \frac{\|Bx\|}{\|x\|} + \sigma_1(A-B).$$

The inequality follows by applying the Max–Min formula (C.3). □

We refer to Horn and Johnson [74] for more involved results on matrix analysis.

Appendix D

Subdifferentials of Convex Functions

D.1 Subdifferentials and Subgradients

A function $F:\mathbb{R}^n\to\mathbb{R}$ is convex if $F(\lambda x+(1-\lambda)y)\le\lambda F(x)+(1-\lambda)F(y)$ for all $x,y\in\mathbb{R}^n$ and $\lambda\in[0,1]$. An equivalent definition is that the epigraph $\{(x,y),\ x\in\mathbb{R}^n,\ y\in[F(x),+\infty[\}$ is a convex subset of $\mathbb{R}^{n+1}$.

Lemma D.1 *When the function $F:\mathbb{R}^n\to\mathbb{R}$ is convex and differentiable, we have*

$$F(y)\ge F(x)+\langle\nabla F(x),y-x\rangle\,,\quad \textit{for all } x,y\in\mathbb{R}^n.$$

Proof. Let $x,h\in\mathbb{R}^n$, and define $f:\mathbb{R}\to\mathbb{R}$ by $f(t)=F(x+th)$. Since F is differentiable, so is f and $f'(t)=\langle\nabla F(x+th),h\rangle$. By Taylor's expansion, we have for some $t^*\in[0,1]$

$$F(x+h)-F(x)=\langle\nabla F(x+t^*h),h\rangle=f'(t^*).$$

Since

$$f(\lambda t+(1-\lambda)s)=F(\lambda(x+th)+(1-\lambda)(x+sh))\le\lambda f(t)+(1-\lambda)f(s)\,,$$

the function f is convex, so

$$F(x+h)-F(x)=f'(t^*)\ge f'(0)=\langle\nabla F(x),h\rangle.$$

We conclude by setting $h=y-x$. □

We define the subdifferential ∂F of a convex function $F:\mathbb{R}^n\to\mathbb{R}$ by

$$\partial F(x)=\{w\in\mathbb{R}^n:\ F(y)\ge F(x)+\langle w,y-x\rangle\ \text{ for all } y\in\mathbb{R}^n\}\,. \tag{D.1}$$

A vector $w\in\partial F(x)$ is called a subgradient of F in x.

Lemma D.2

1. *$F:\mathbb{R}^n\to\mathbb{R}$ is convex if and only if the set $\partial F(x)$ is nonempty for all $x\in\mathbb{R}^n$.*
2. *When F is convex and differentiable in x, $\partial F(x)=\{\nabla F(x)\}$.*

Proof.
1. Assume that $\partial F(x)$ is nonempty for all $x \in \mathbb{R}^n$. For any $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$, there exists $w \in \partial F(\lambda x + (1-\lambda)y)$. By definition of the subdifferential, we have $F(y) \geq F(\lambda x + (1-\lambda)y) + \langle w, \lambda(y-x)\rangle$ and $F(x) \geq F(\lambda x + (1-\lambda)y) + \langle w, (1-\lambda)(x-y)\rangle$. Multiplying the first inequality by $(1-\lambda)$ and the second by λ, we obtain by summing the results

$$(1-\lambda)F(y) + \lambda F(x) \geq F(\lambda x + (1-\lambda)y),$$

so F is convex.

Conversely, if F is convex, then its epigraph is convex in $\mathbb{R}^{n+1}$, so there exists a supporting hyperplane H_x separating the epigraph from any (x,z) with $z < F(x)$. Since F is finite on $\mathbb{R}^n$, this hyperplane is not vertical, so there exists $u \in \mathbb{R}^n$ and $a \in \mathbb{R}$, such that $H_x = \{(\alpha, \beta) : \langle u, \alpha\rangle + \beta = a\}$ and such that any (α, β) in the epigraph of F fulfills $\langle u, \alpha\rangle + \beta \geq a$. Since $(x, F(x)) \in H_x$, we have $a = \langle u, x\rangle + F(x)$. For any $y \in \mathbb{R}^n$, the couple $(y, F(y))$ belongs to the epigraph of F, and therefore

$$\langle u, y\rangle + F(y) \geq a = \langle u, x\rangle + F(x).$$

This ensures that $-u \in \partial F(x)$.

2. Let w be a subgradient of F. When F is differentiable, Taylor's formula gives for any $x, h \in \mathbb{R}^n$ and $t > 0$

$$F(x \pm th) - F(x) = \pm t\langle \nabla F(x), h\rangle + o(t) \geq \pm t\langle w, h\rangle.$$

Letting t go to zero, this enforces $\langle \nabla F(x), h\rangle = \langle w, h\rangle$ for all $h \in \mathbb{R}^n$, so $w = \nabla F(x)$.
□

More generally, when $F : \mathscr{D} \to \mathbb{R}$ is convex on a convex domain $\mathscr{D}$ of $\mathbb{R}^n$, the subdifferential $\partial F(x)$ is nonempty for all x in the interior of $\mathscr{D}$.

It is well-known that the derivative f' of a smooth convex function $f : \mathbb{R} \to \mathbb{R}$ is increasing. The next lemma shows that this result remains valid for subdifferentials.

Lemma D.3 Monotonicity
The subdifferential of a convex function F is monotone increasing:

$$\langle w_x - w_y, x - y\rangle \geq 0, \quad \textit{for all } w_x \in \partial F(x) \textit{ and } w_y \in \partial F(y). \qquad \text{(D.2)}$$

Proof. By definition, we have $F(y) \geq F(x) + \langle w_x, y - x\rangle$ and $F(x) \geq F(y) + \langle w_y, x - y\rangle$. Summing these two inequalities gives $\langle w_x - w_y, x - y\rangle \geq 0$. □

Finally, the minimum of a convex function can be easily characterized in terms of its subdifferential.

Lemma D.4 First order optimality condition
For any convex function $F:\mathbb{R}^n\to\mathbb{R}$, we have

$$x_*\in\operatorname*{argmin}_{x\in\mathbb{R}^n}F(x)\quad\Longleftrightarrow\quad 0\in\partial F(x_*).\tag{D.3}$$

Proof. Both conditions are equivalent to $F(y)\geq F(x_*)+\langle 0,y-x_*\rangle$ for all $y\in\mathbb{R}^n$. □

D.2 Examples of Subdifferentials

As examples, we compute the subdifferential of several common norms. For $x\in\mathbb{R}$, we set $\operatorname{sign}(x)=\mathbf{1}_{x>0}-\mathbf{1}_{x\leq 0}$.

Lemma D.5 Subdifferential of ℓ^1 and ℓ^∞ norms

1. *For $x\in\mathbb{R}^n$, let us set $J(x)=\{j:x_j\neq 0\}$. We have*

$$\partial|x|_1=\left\{w\in\mathbb{R}^n:\ w_j=\operatorname{sign}(x_j)\text{ for }j\in J(x),\ w_j\in[-1,1]\text{ for }j\notin J(x)\right\}.$$

2. *Let us set $J_*=\{j:\ |x_j|=|x|_\infty\}$ and write $\mathscr{P}(J_*)$ for the set of probabilities on J_*. We have for $x\neq 0$*

$$\partial|x|_\infty=\left\{w\in\mathbb{R}^n:\ \begin{array}{l}w_j=0\ \text{ for }j\notin J_*\\ w_j=\lambda_j\,sign(x_j)\ \ for\ j\in J_*\ with\ \lambda\in\mathscr{P}(J_*)\end{array}\right\}.$$

Proof. For $p\in[0,+\infty]$ and q, such that $1/p+1/q=1$, Hölder's inequality ensures that $|x|_p=\sup\{\langle\phi,x\rangle:\ |\phi|_q\leq 1\}$. To prove Lemma D.5, all we need is to check that

$$\partial|x|_p=\left\{\phi\in\mathbb{R}^n:\langle\phi,x\rangle=|x|_p\text{ and }|\phi|_q\leq 1\right\}.$$

i) Consider ϕ_x, such that $\langle\phi_x,x\rangle=|x|_p$ and $|\phi_x|_q\leq 1$. Then, we have for any $y\in\mathbb{R}^n$

$$|y|_p\geq\langle\phi_x,y\rangle=|x|_p+\langle\phi_x,y-x\rangle,$$

and therefore $\phi_x\in\partial|x|_p$.
ii) Conversely, let us consider $w\in\partial|x|_p$. For $y=0$ and $y=2x$, Equation (D.1) give

$$0\geq|x|_p-\langle w,x\rangle\quad\text{and}\quad 2|x|_p\geq|x|_p+\langle w,x\rangle,$$

from which we get $|x|_p=\langle w,x\rangle$. Furthermore, we have $|w|_q=\langle w,\phi_w\rangle$ for some $\phi_w\in\mathbb{R}^n$ fulfilling $|\phi_w|_p\leq 1$. The triangular inequality and (D.1) give

$$|x|_p+|\phi_w|_p\geq|x+\phi_w|_p\geq|x|_p+\langle w,\phi_w\rangle,$$

which finally ensures that $|w|_q=\langle w,\phi_w\rangle\leq|\phi_w|_p\leq 1$. The proof of Lemma D.5 is complete. □

The next lemma characterizes the subdifferential of the nuclear norm. We refer to Appendix C for the definition of the nuclear norm $|\cdot|_*$, the operator norm $|\cdot|_{\text{op}}$, and the singular value decomposition (SVD).

Lemma D.6 Subdifferential of the nuclear norm

Let us consider a matrix B with rank r and singular value decomposition $B = \sum_{k=1}^r \sigma_k u_k v_k^T$. We write P_u for the orthogonal projector onto $\text{span}\{u_1,\ldots,u_r\}$ *and P_v for the orthogonal projector onto* $\text{span}\{v_1,\ldots,v_r\}$. *We also set $P_u^\perp = I - P_u$ and $P_v^\perp = I - P_v$. Then, we have*

$$\partial|B|_* = \left\{ \sum_{k=1}^r u_k v_k^T + P_u^\perp W P_v^\perp : |W|_{\text{op}} \leq 1 \right\}. \tag{D.4}$$

Proof. With the same reasoning as in the proof of Lemma D.5, we have

$$\partial|B|_* = \left\{ Z : \langle Z, B\rangle_F = |B|_* \text{ and } |Z|_{\text{op}} \leq 1 \right\}.$$

All we need is then to check that this set coincides with (D.4).

i) First, we observe that any matrix $Z = \sum_{k=1}^r u_k v_k^T + P_u^\perp W P_v^\perp$ with $|W|_{\text{op}} \leq 1$ fulfills $|Z|_{\text{op}} \leq 1$ and $\langle Z, B\rangle_F = |B|_*$. Therefore, such a matrix Z is in the subdifferential of $|B|_*$.

ii) Conversely, let Z be a matrix fulfilling $\langle Z, B\rangle_F = |B|_*$ and $|Z|_{\text{op}} \leq 1$. Since

$$\sum_{k=1}^r \sigma_k = \langle Z, B\rangle_F = \sum_{k=1}^r \sigma_k \langle Z, u_k v_k^T\rangle_F = \sum_{k=1}^r \sigma_k \langle Z v_k, u_k\rangle_F$$

and $\langle Z v_k, u_k\rangle_F \leq \|Z v_k\| \leq |Z|_{\text{op}} \leq 1$, we then have $\langle Z v_k, u_k\rangle_F = \|Z v_k\| = 1$. Since $\|u_k\| = 1$, this enforces $Z v_k = u_k$. Since $\langle Z v_k, u_k\rangle_F = \langle v_k, Z^T u_k\rangle_F$, we have for the same reasons $Z^T u_k = v_k$. In particular, $u_1,\ldots,u_r$ are eigenvectors of ZZ^T associated to the eigenvalue 1, and $v_1,\ldots,v_r$ are eigenvectors of $Z^T Z$ also associated to the eigenvalue 1. As a consequence, a SVD of Z is given by

$$Z = \sum_{k=1}^r u_k v_k^T + \sum_{k=r+1}^{\text{rank}(Z)} \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^T,$$

where $\widetilde{u}_k$ is orthogonal to $u_1,\ldots,u_r$ and $\widetilde{v}_k$ is orthogonal to $v_1,\ldots,v_r$. Furthermore, we have $\widetilde{\sigma}_k \leq 1$ since $|Z|_{\text{op}} \leq 1$. In particular, we can write

$$\sum_{k=r+1}^{\text{rank}(Z)} \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^T = P_u^\perp W P_v^\perp$$

for some matrix W, fulfilling $|W|_{\text{op}} \leq 1$. The derivation of (D.4) is complete. □

Appendix E

Reproducing Kernel Hilbert Spaces

Reproducing Kernel Hilbert Spaces (RKHS) are some functional Hilbert spaces, where the smoothness of a function is driven by its norm. RKHS also fulfill a special "reproducing property" that is crucial in practice, since it allows efficient numerical computations as in Proposition 9.9 in Chapter 9.

A function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is said to be a positive definite kernel if it is symmetric ($k(x,y) = k(y,x)$ for all $x,y \in \mathscr{X}$), and if for any $N \in \mathbb{N}$, $x_1,\ldots,x_N \in \mathscr{X}$ and $a_1,\ldots,a_N \in \mathbb{R}$ we have

$$\sum_{i,j=1}^{N} a_i a_j k(x_i,x_j) \geq 0. \tag{E.1}$$

Examples of positive definite kernels in $\mathscr{X} = \mathbb{R}^d$:

- linear kernel: $k(x,y) = \langle x,y \rangle$
- Gaussian kernel: $k(x,y) = e^{-\|x-y\|^2/2\sigma^2}$
- histogram kernel ($d = 1$): $k(x,y) = \min(x,y)$
- exponential kernel: $k(x,y) = e^{-\|x-y\|/\sigma}$.

We can associate to a positive definite kernel k a special Hilbert space $\mathscr{F} \subset \mathbb{R}^{\mathscr{X}}$ called Reproducing Kernel Hilbert Space associated to k. In the following, the notation $k(x,.)$ refers to the map $y \to k(x,y)$.

Proposition E.1 Reproducing Kernel Hilbert Space (RKHS)

To any positive definite kernel k on $\mathscr{X}$, we can associate a (unique) Hilbert space $\mathscr{F} \subset \mathbb{R}^{\mathscr{X}}$ fulfilling

1. *$k(x,.) \in \mathscr{F}$ for all $x \in \mathscr{X}$*
2. **reproducing property:**

$$f(x) = \langle f, k(x,.) \rangle_{\mathscr{F}} \quad \text{for all } x \in \mathscr{X} \text{ and } f \in \mathscr{F}. \tag{E.2}$$

The space $\mathscr{F}$ is called the Reproducing Kernel Hilbert Space associated to k.

Proof. From the first property, if the Hilbert space $\mathscr{F}$ exists, it must include the linear space $\mathscr{F}_0$ spanned by the family $\{k(x,.) : x \in \mathscr{X}\}$

$$\mathscr{F}_0 = \left\{ f : \mathscr{X} \to \mathbb{R} : f(x) = \sum_{i=1}^{N} a_i k(x_i,x),\ N \in \mathbb{N},\ x_1,\dots,x_N \in \mathscr{X},\ a_1,\dots,a_N \in \mathbb{R} \right\}.$$

Furthermore, from the reproducing property, if $\mathscr{F}$ exists, we must have $\langle k(x,.), k(y,.)\rangle_{\mathscr{F}} = k(x,y)$ and

$$\left\langle \sum_{i=1}^{N} a_i k(x_i,.), \sum_{j=1}^{M} b_j k(y_j,.) \right\rangle_{\mathscr{F}} = \sum_{i=1}^{N}\sum_{j=1}^{M} a_i b_j k(x_i,y_j).$$

Accordingly, we define for any $f = \sum_{i=1}^{N} a_i k(x_i,.)$ and $g = \sum_{j=1}^{M} b_j k(y_j,.)$ in $\mathscr{F}_0$

$$\langle f,g\rangle_{\mathscr{F}_0} := \sum_{i=1}^{N}\sum_{j=1}^{M} a_i b_j k(x_i,y_j) = \sum_{i=1}^{N} a_i g(x_i) = \sum_{j=1}^{M} b_j f(y_j),$$

where the last two equalities ensures that $\langle f,g\rangle_{\mathscr{F}_0}$ does not depend on the choice of the expansion of f and g, so $\langle f,g\rangle_{\mathscr{F}_0}$ is well-defined. The application $(f,g) \to \langle f,g\rangle_{\mathscr{F}_0}$ is bilinear, symmetric, positive (according to (E.1)), and we have the reproducing property

$$f(x) = \langle f, k(x,.)\rangle_{\mathscr{F}_0} \quad \text{for all } x \in \mathscr{X} \quad \text{and } f \in \mathscr{F}_0. \tag{E.3}$$

The Cauchy–Schwartz inequality $\langle f,g\rangle_{\mathscr{F}_0} \leq \|f\|_{\mathscr{F}_0}\|g\|_{\mathscr{F}_0}$ and the reproducing formula (E.3) give

$$|f(x)| \leq \sqrt{k(x,x)}\ \|f\|_{\mathscr{F}_0}. \tag{E.4}$$

As a consequence $\|f\|_{\mathscr{F}_0} = 0$ implies $f = 0$ so $\langle f,g\rangle_{\mathscr{F}_0}$ is a scalar product on $\mathscr{F}_0$. Therefore $\mathscr{F}_0$ is a pre-Hilbert space fulfilling the reproducing property. We obtain $\mathscr{F}$ by completing $\mathscr{F}_0$. □

Remark 1. Let us consider two sequences $(x_i) \in \mathscr{X}^{\mathbb{N}}$ and $(a_i) \in \mathbb{R}^{\mathbb{N}}$ fulfilling $\sum_{i,j\geq 1} a_i a_j k(x_i,x_j) < +\infty$. According to (E.4), for any $M < N$ and $x \in \mathscr{X}$, we have

$$\left| \sum_{i=M+1}^{N} a_i k(x_i,x) \right| \leq \sqrt{k(x,x)} \sum_{i,j=M+1}^{N} a_i a_j k(x_i,x_j).$$

When $\sum_{i,j\geq 1} a_i a_j k(x_i,x_j)$ is finite, the right-hand side goes to 0 when M,N goes to infinity, so the partial series $\sum_{i=1}^{N} a_i k(x_i,x)$ is Cauchy and it converges when $N \to \infty$. We can therefore define the space

$$\mathscr{F}_0' = \left\{ f : \mathscr{X} \to \mathbb{R} : f(x) = \sum_{i=1}^{\infty} a_i k(x_i,x), \right.$$
$$\left. (x_i) \in \mathscr{X}^{\mathbb{N}},\ (a_i) \in \mathbb{R}^{\mathbb{N}},\ \sum_{i,j\geq 1} a_i a_j k(x_i,x_j) < +\infty \right\}$$

and the bilinear form

$$\langle f,g\rangle_{\mathscr{F}_0'} := \sum_{i,j=1}^{\infty} a_i b_j k(x_i,y_j) = \sum_{i=1}^{\infty} a_i g(x_i) = \sum_{j=1}^{\infty} b_j f(y_j)$$

for $f=\sum_{i=1}^{\infty} a_i k(x_i,.)$ and $g=\sum_{j=1}^{\infty} b_j k(y_j,.)$ in $\mathscr{F}_0'$. Exactly as above, the application $(f,g)\to\langle f,g\rangle_{\mathscr{F}_0'}$ is a scalar product fulfilling the reproduction property

$$f(x)=\langle f,k(x,.)\rangle_{\mathscr{F}_0'} \quad \text{for all } x\in\mathscr{X} \text{ and } f\in\mathscr{F}_0'.$$

In addition, the partial sums $f_N=\sum_{i=1}^{N} a_i k(x_i,.)$ relative to a function $f=\sum_{i=1}^{\infty} a_i k(x_i,.)\in\mathscr{F}_0'$ are Cauchy, since

$$\|f_M-f_N\|_{\mathscr{F}_0}^2 = \sum_{i,j=N+1}^{M} a_i a_j k(x_i,x_j) \overset{N,M\to\infty}{\to} 0,$$

and they converge to f. As a consequence, $\mathscr{F}_0'$ is included in the completion $\mathscr{F}$ of $\mathscr{F}_0$ and the scalar product $\langle .,.\rangle_{\mathscr{F}}$ restricted to $\mathscr{F}_0'$ coincides with $\langle .,.\rangle_{\mathscr{F}_0'}$.

Remark 2. The norm of a function f in an RKHS $\mathscr{F}$ is strongly linked to its smoothness. This appears clearly in the inequality

$$|f(x)-f(x')| = |\langle f,k(x,.)-k(x',.)\rangle_{\mathscr{F}}| \le \|f\|_{\mathscr{F}}\ \|k(x,.)-k(x',.)\|_{\mathscr{F}}. \qquad \text{(E.5)}$$

Let us illustrate this point by describing the RKHS associated to the histogram and Gaussian kernels.

Example 1: RKHS associated to the histogram kernel.
The Sobolev space

$$\mathscr{F}=\left\{f\in C([0,1],\mathbb{R}) : f \text{ is a.e. differentiable, with } f'\in L^2([0,1]) \text{ and } f(0)=0\right\}$$

endowed with the scalar product $\langle f,g\rangle_{\mathscr{F}}=\int_0^1 f'g'$ is an RKHS with reproducing kernel $k(x,y)=\min(x,y)$ on $[0,1]$. Actually, $k(x,.)\in\mathscr{F}$ for all $x\in[0,1]$ and

$$f(x)=\int_0^1 f'(y)\mathbf{1}_{y\le x}\,dy = \langle f,k(x,.)\rangle_{\mathscr{F}}, \quad \text{for all } f\in\mathscr{F} \text{ and } x\in[0,1].$$

In this case the norm $\|f\|_{\mathscr{F}}$ corresponds simply to the L^2-norm of the derivative of f. The smaller is this norm, the smoother is f.

Example 2: RKHS associated to the Gaussian kernel.
Let us write $\mathbf{F}[f]$ for the Fourier transform in $\mathbb{R}^d$ with normalization

$$\mathbf{F}[f](\omega)=\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} f(t)e^{-\mathrm{i}\langle\omega,t\rangle}, \quad \text{for } f\in L^1(\mathbb{R}^d)\cap L^2(\mathbb{R}^d) \text{ and } \omega\in\mathbb{R}^d.$$

For any $\sigma > 0$, the functional space

$$\mathscr{F}_\sigma = \left\{ f \in C_0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) \text{ such that } \int_{\mathbb{R}^d} \left|\mathbf{F}[f](\omega)\right|^2 e^{\sigma|\omega|^2/2} d\omega < +\infty \right\},$$

endowed with the scalar product

$$\langle f, g \rangle_{\mathscr{F}_\sigma} = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \overline{\mathbf{F}[f](\omega)} \mathbf{F}[g](\omega) e^{\sigma|\omega|^2/2} d\omega,$$

is an RKHS associated with the Gaussian kernel $k(x,y) = \exp(-\|y-x\|^2/2\sigma^2)$. Actually, for all $x \in \mathbb{R}^d$ the function $k(x,.)$ belongs to $\mathscr{F}_\sigma$, and straightforward computations give

$$\langle k(x,.), f \rangle_{\mathscr{F}_\sigma} = \mathbf{F}^{-1}\left[\mathbf{F}[f]\right](x) = f(x) \quad \text{for all } f \in \mathscr{F} \text{ and all } x \in \mathbb{R}^d.$$

The space $\mathscr{F}_\sigma$ gathers very regular functions, and the norm $\|f\|_{\mathscr{F}_\sigma}$ directly controls the smoothness of f. We note that when σ increases, the space $\mathscr{F}_\sigma$ shrinks and contains smoother and smoother functions.

We refer to Aronszajn [9] and Schölkopf and Smola [109] for more details on RKHS.

Notations

$$
\begin{aligned}
(x)_+ &= \max(0,x)\\
x\vee y &= \max(x,y)\\
x\wedge y &= \min(x,y)\\
\|\beta\| &= \sqrt{\sum_j \beta_j^2}\\
\langle x,y\rangle &= x^T y \;=\; \sum_j x_j y_j\\
|\beta|_1 &= \sum_j |\beta_j|\\
|\beta|_\infty &= \max_j |\beta_j| \quad \text{or } \max_{i,j} |\beta_{ij}| \text{ if } \beta \text{ is a matrix}\\
\operatorname{supp}(\beta) &= \{j:\ \beta_j \neq 0\}\\
|\beta|_0 &= \operatorname{card}(\operatorname{supp}(\beta))\\
\beta_S &= [\beta_j]_{j\in S}\\
|S| &= \operatorname{card}(S)\\
I_n &= \text{identity matrix on } \mathbb{R}^n\\
\partial_i F &= \text{partial derivative of } F \text{ according to variable } i\\
\operatorname{div}(F) &= \sum_i \partial_i F \quad \text{(divergence of } F\text{)}\\
\nabla F &= \text{gradient of } F\\
\partial F(x) &= \text{subdifferential of } F \text{ at point } x\\
\sigma_j(A) &= j\text{-th largest singular value of } A\\
|A|_{\mathrm{op}} &= \sigma_1(A) \;=\; \sup_{\|x\|\leq 1} \|Ax\|\\
|A|_* &= \sum_j \sigma_j(A)\\
\|A\|_F &= \sqrt{\sum_{ij} A_{ij}^2} \;=\; \sqrt{\sum_j \sigma_j(A)^2}\\
\|A\|_{(2,q)} &= \sqrt{\sum_{k=1}^{q} \sigma_k(A)^2}
\end{aligned}
$$

$\lvert A\rvert_{1,\infty}$	$=$	$\max_j \sum_i \lvert A_{ij}\rvert$
$A_{j:}$	$=$	j-th row of matrix A
A_j	$=$	j-th column of matrix A
Proj_S	$=$	orthogonal projector onto the linear span S
C_p^d	$=$	$\dfrac{p!}{d!(n-d)!}$
$\mathrm{sign}(x)$	$=$	$\mathbf{1}_{x>0} - \mathbf{1}_{x\leq 0}$
x^T	$=$	transpose of vector or matrix x
$\underset{\beta\in\mathscr{C}}{\mathrm{argmin}}\, F(\beta)$	$=$	set of the minimizers in $\mathscr{C}$ of F
$\mathscr{P}(E)$	$=$	set gathering all the subsets of E
$E = F \bigcirc\!\!\!\!\!\perp G$	$=$	decomposition $E = F + G$ with F orthogonal to G
$]a,b]$	$=$	$\{x \in \mathbb{R} : a < x \leq b\}$
i	$=$	imaginary unit
$\mathrm{var}(X)$	$=$	$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$
$\mathrm{sdev}(X)$	$=$	$\sqrt{\mathrm{var}(X)}$

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[2] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, NY, second edition, 1984.

[3] A. Antoniadis. Comments on: ℓ_1-penalization for mixture regression models [mr2677722]. *TEST*, 19(2):257–258, 2010.

[4] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[5] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54. 2009.

[6] S. Arlot and A. Célisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[7] S. Arlot and A. Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.*, 21(4):613–632, 2011.

[8] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279, 2010.

[9] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[10] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

[11] F. Bach. Consistency of trace norm minimization. *J. Mach. Learn. Res.*, 9:1019–1048, 2008.

[12] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 11 2012.

[13] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

[14] Y. Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401, 2011.

[15] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model

selection: ρ-estimation. arXiv:1403.6057, 2014.

[16] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[17] Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the gaussian setting. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(3):1092–1119, 08 2014.

[18] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[19] A.R. Barron. *Are Bayes rules consistent in information? in Open problems in communication and computation. Cover, T.M. and Gopinath, B.* Springer-Verlag, 1987.

[20] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: Overview and implementation. *Statist. Comput.*, 22(2):455–470, 2012.

[21] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, 2009.

[22] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[23] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *JRSS B*, 57(1):289–300, 1995.

[24] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

[25] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 08 2013.

[26] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[27] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory*, 51:1611–1615, 2005.

[28] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[29] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[30] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[31] G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992, 2008.

[32] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: some recent advances. *ESAIM Probability & Statistics*, 9:323–375, 2005.

[33] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013.

[34] S. Bubeck. Theory of convex optimization for machine learning. *arXiv:1405.4980*, 2014.

[35] F. Bunea, J. Lederer, and Y. She. The square root group lasso: theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 30(2):1313–1325, 2014.

[36] F. Bunea, Y. She, and M. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, 40(5):2359–2388, 2012.

[37] F. Bunea, Y. She, and M.H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Stat.*, 39(2):1282–1309, 2011.

[38] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[39] T. Cai, W. Liu, and X. Luo. A constrained ℓ^1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[40] E.J. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[41] R. Castelo and A. Roverato. A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.*, 7:2621–2650, 2006.

[42] O. Catoni. A mixture approach to universal model selection. Technical report, Ecole Normale Superieure, 1997.

[43] O. Catoni. Statistical learning theory and stochastic optimization. In *Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI — 2001*, volume 1851 of *Lecture Notes in Mathematics*, pages 1–269, New York, 2004. Springer.

[44] G. Chagny. Penalization versus goldenshluger–lepski strategies in warped bases regression. *ESAIM: Probability and Statistics*, 17:328–358, 1 2013.

[45] V. Chandrasekaran, P.A. Parrilo, and A.S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 08 2012.

[46] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[47] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of u-statistics. *The Annals of Statistics*, 36:844–874, 2008.

[48] A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning*, 72(1-2):39– 61, 2008.

[49] K.R. Davidson and S.J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages

317–366. North-Holland, Amsterdam, 2001.

[50] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, Berlin, 1996.

[51] L.P. Devroye and T.J. Wagner. The L_1 convergence of kernel density estimates. *Ann. Statist.*, 7(5):1136–1139, 1979.

[52] T. Dickhaus. *Simultaneous Statistical Inference*. Springer, Berlin, 2014.

[53] D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality, 2000. American Math. Society "Math Challenges of the 21st Century.".

[54] D.L. Donoho, I. Johnstone, and A. Montanari. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *Information Theory, IEEE Transactions on*, 59(6):3396–3433, June 2013.

[55] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA*, 97(457):77–87, 2002.

[56] S. Dudoit and M. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics, 2008.

[57] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

[58] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Statist.*, (1):302–332, 2007.

[59] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

[60] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso, 2010. arXiv:1001.0736.

[61] S. Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[62] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *JMLR Workshop and Conference Proceedings*, 19 (COLT 2011 Proceedings):377–396, 2011.

[63] C. Giraud. Estimation of Gaussian graphs by model selection. *Electron. J. Stat.*, 2:542–563, 2008.

[64] C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008.

[65] C. Giraud. Low rank multivariate regression. *Electron. J. Stat.*, 5:775–799, 2011.

[66] C. Giraud, S. Huet, and N. Verzelen. Graph selection with GGMselect. *Stat. Appl. Genet. Mol. Biol.*, 11(3):1–50, 2012.

[67] C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with un-

known variance. *Stat. Sci.*, 27(4):500–518, 2012.

[68] J.J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, 2014.

[69] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

[70] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, and C.D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[71] W. Guo and M.B. Rao. On control of the false discovery rate under no assumption of dependency. *J. Statist. Plann. Inference*, 138(10):3176–3188, 2008.

[72] J.M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Technical report, 1971.

[73] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[74] R.A. Horn and C.R. Johnson. *Matrix analysis. 2nd ed.* Cambridge: Cambridge University Press. xviii, 643 p., 2013.

[75] I.A. Ibragimov, V.N. Sudakov, and B.S. Tsirel'son. *Norms of Gaussian sample functions*, volume 550 of *Lecture Notes in Mathematics*. Springer, 1976. Proceedings of the third Japan-USSR Symposium on Probability Theory.

[76] J. Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864, June 2009.

[77] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, 8:613–636, May 2007.

[78] M. Kalisch and P. Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. *J. Comput. Graph. Statist.*, 17(4):773–789, 2008.

[79] V. Koltchinski, K. Lounici, and A. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

[80] J. Lafferty, H. Liu, and L. Wasserman. Sparse nonparametric graphical models. *Statistical Science*, 27:519–537, 2012.

[81] S.L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[82] E. Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85:717–736, 2005.

[83] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[84] M. Ledoux. *Concentration of measure phenomenon*. AMS, 2001.

[85] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[86] O. Lepski. Asymptotically minimax adaptive estimation. i. upper bounds. optimally adaptive estimates. *Teor. Veroyatnost i Primenen*, 36:645–659, 1991.

[87] M. Lerasle and S. Arlot. Why v=5 is enough in v-fold cross-validation. arXiv:1210.5830, 2012.

[88] G. Leung and A. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

[89] K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987.

[90] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 08 2012.

[91] C.L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[92] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, October 2007.

[93] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[94] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188, 1989.

[95] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[96] N. Meinshausen, G. Rocha, and B. Yu. Discussion: a tale of three cousins: Lasso, L2boosting and Dantzig. *Annals of Statistics*, 35(6):2373–2384, 2007.

[97] F. Mosteller and J.W. Tukey. Data analysis, including statistics. In G. Lindsey and E. Aronson, editors, *Handbook of Social Psychology, Vol. 2*. Addison-wesley, 1968.

[98] Y. Nesterov. A method of solving a convex programming problem with convergence rate O(1/sqr(k)). *Soviet Mathematics Doklady*, 27:372–376, 1983.

[99] J.R. Norris. *Markov chains*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998.

[100] P. Ravikumar, M.J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing, ℓ^1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[101] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 3rd edition, 1999.

[102] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Annals of Statistics*, 39(2):731–771, 2011.

[103] P. Rigollet and A.B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 11 2012.

[104] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 2nd edition, 2004.

[105] C. Robert and G. Casella. *Méthodes de Monte-Carlo avec R*. Springer Paris, 2011.

[106] E. Roquain. Type I error rate control for testing many hypotheses: a survey with proofs. *J. SFdS*, 152(2):3–38, 2011.

[107] V. Sabarly, O. Bouvet, J. Glodt, O. Clermont, D. Skurnik, L. Diancourt, D. de Vienne, E. Denamur, and C. Dillmann. The decoupling between genetic structure and metabolic phenotypes in escherichia coli leads to continuous phenotypic diversity. *Journal of Evolutionary Biology*, 24:1559–1571, 2011.

[108] A. Sanchez-Perez. Time series prediction via aggregation : an oracle bound including numerical cost. In *Modeling and Stochastic Learning for Forecasting in High Dimension*, Lecture Notes in Statistics. Springer, 2014.

[109] B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

[110] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[111] J. Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494, 1993.

[112] D. Slepian. The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

[113] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

[114] N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.

[115] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[116] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[117] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[118] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*,

67(1):91–108, 2005.

[119] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.*, 81:82–86, 1986.

[120] S.A. Van de Geer. High-dimensional generalized linear models and the lasso. *The annals of statistics*, 36(2):614–645, 2008.

[121] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[122] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

[123] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027, 2011.

[124] N. Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Stat.*, 6:38–90, 2012.

[125] F. Villers, B. Schaeffer, C. Bertin, and S. Huet. Assessing the Validity Domains of Graphical Gaussian Models in Order to Infer Relationships among Components of Complex Biological Systems. *Statistical Applications in Genetics and Molecular Biology*, 7, 2008.

[126] M. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009.

[127] A. Wille and P. Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.*, 5:Art. 1, 34 pp. (electronic), 2006.

[128] L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 10 2012.

[129] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

[130] T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inform. Theory*, 57(7):4689–4708, 2011.

[131] Y. Zhang, M. Wainwright, and M. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *JMLR: Workshop and Conference Proceedings*, 35:1–28, 2014.

[132] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[133] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

Statistics

"There is a real need for this book. It can quickly make someone new to the field familiar with modern topics in high-dimensional statistics and machine learning, and it is great as a textbook for an advanced graduate course."
—Marten H. Wegkamp, Cornell University, Ithaca, New York, USA

Ever-greater computing technologies have given rise to an exponentially growing volume of data. Today massive data sets (with potentially thousands of variables) play an important role in almost every branch of modern human activity, including networks, finance, and genetics. However, analyzing such data has presented a challenge for statisticians and data analysts and has required the development of new statistical methods capable of separating the *signal* from the *noise*.

Introduction to High-Dimensional Statistics is a concise guide to state-of-the-art models, techniques, and approaches for handling high-dimensional data. The book is intended to expose the reader to the key concepts and ideas in the most simple settings possible while avoiding unnecessary technicalities.

Offering a succinct presentation of the mathematical foundations of high-dimensional statistics, this highly accessible text:

- Describes the challenges related to the analysis of high-dimensional data
- Covers cutting-edge statistical methods including model selection, sparsity and the lasso, aggregation, and learning theory
- Provides detailed exercises at the end of every chapter with collaborative solutions on a wikisite
- Illustrates concepts with simple but clear practical examples

Introduction to High-Dimensional Statistics is suitable for graduate students and researchers interested in discovering modern statistics for massive data. It can be used as a graduate text or for self-study.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K23198
ISBN: 978-1-4822-3794-8
90000
9 781482 237948
www.crcpress.com