DATA SCIENCE SERIES

STATISTICAL FOUNDATIONS OF DATA SCIENCE

JIANQING FAN
RUNZE LI
CUN-HUI ZHANG
HUI ZOU

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Copyrighted material

First edition published 2020
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

© 2020 Taylor & Francis Group, LLC

CRC Press is an imprint of Taylor & Francis Group

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.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access www.copyright.com or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact mpkbookspermissions@tandf.co.uk

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

ISBN: 978-1-466-51084-5 (hbk)

Visit the eResources: https://www.routledge.com/Statistical-Foundations-of-Data-Science/Fan-Li-Zhang-Zou/p/book/9781466510845

Copyrighted material

Contents

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Preface

Big data are ubiquitous. They come in varying volume, velocity, and variety. They have a deep impact on systems such as storages, communications and computing architectures and analysis such as statistics, computation, optimization, and privacy. Engulfed by a multitude of applications, data science aims to address the large-scale challenges of data analysis, turning big data into smart data for decision making and knowledge discoveries. Data science integrates theories and methods from statistics, optimization, mathematical science, computer science, and information science to extract knowledge, make decisions, discover new insights, and reveal new phenomena from data. The concept of data science has appeared in the literature for several decades and has been interpreted differently by different researchers. It has nowadays become a multi-disciplinary field that distills knowledge in various disciplines to develop new methods, processes, algorithms and systems for knowledge discovery from various kinds of data, which can be either low or high dimensional, and either structured, unstructured or semi-structured. Statistical modeling plays critical roles in the analysis of complex and heterogeneous data and quantifies uncertainties of scientific hypotheses and statistical results.

This book introduces commonly-used statistical models, contemporary statistical machine learning techniques and algorithms, along with their mathematical insights and statistical theories. It aims to serve as a graduate-level textbook on the statistical foundations of data science as well as a research monograph on sparsity, covariance learning, machine learning and statistical inference. For a one-semester graduate level course, it may cover Chapters 2, 3, 9, 10, 12, 13 and some topics selected from the remaining chapters. This gives a comprehensive view on statistical machine learning models, theories and methods. Alternatively, a one-semester graduate course may cover Chapters 2, 3, 5, 7, 8 and selected topics from the remaining chapters. This track focuses more on high-dimensional statistics, model selection and inferences but both paths strongly emphasize sparsity and variable selections.

Frontiers of scientific research rely on the collection and processing of massive complex data. Information and technology allow us to collect big data of unprecedented size and complexity. Accompanying big data is the rise of dimensionality, and high dimensionality characterizes many contemporary statistical problems, from sciences and engineering to social science and humanities. Many traditional statistical procedures for finite or low-dimensional data are still useful in data science, but they become infeasible or ineffective for

Copyrighted material

dealing with high-dimensional data. Hence, new statistical methods are indispensable. The authors have worked on high-dimensional statistics for two decades, and started to write the book on the topics of high-dimensional data analysis over a decade ago. Over the last decide, there have been surges in interest and exciting developments in high-dimensional and big data. This led us to concentrate mainly on statistical aspects of data science.

We aim to introduce commonly-used statistical models, methods and procedures in data science and provide readers with sufficient and sound theoretical justifications. It has been a challenge for us to balance statistical theories and methods and to choose the topics and works to cover since the number of publications in this emerging area is enormous. Thus, we focus on the foundational aspects that are related to sparsity, covariance learning, machine learning, and statistical inference.

Sparsity is a common assumption in the analysis of high-dimensional data. By sparsity, we mean that only a handful of features embedded in a huge pool suffice for certain scientific questions or predictions. This book introduces various regularization methods to deal with sparsity, including how to determine penalties and how to choose tuning parameters in regularization methods and numerical optimization algorithms for various statistical models. They can be found in Chapters 3–6 and 8.

High-dimensional measurements are frequently dependent, since these variables often measure similar things, such as aspects of economics or personal health. Many of these variables have heavy tails due to a large number of collected variables. To model the dependence, factor models are frequently employed, which exhibit low-rank plus sparse structures in data matrices and can be solved by robust principal component analysis from high-dimensional covariance. Robust covariance learning, principal component analysis, as well as their applications to community detection, topic modeling, recommender systems, etc. are also a feature of this book. They can be found in Chapters 9–11. Note that factor learning or more generally latent structure learning can also be regarded as unsupervised statistical machine learning.

Machine learning is critical in analyzing high-dimensional and complex data. This book also provides readers with a comprehensive account on statistical machine learning methods and algorithms in data science. We introduce statistical procedures for supervised learning in which the response variable (often categorical) is available and the goal is to predict the response based on input variables. This book also provides readers with statistical procedures for unsupervised learning, in which the responsible variable is missing and the goal concentrates on learning the association and patterns among a set of input variables. Feature creations and sparsity learning also arise in these problems. See Chapters 2, 12–14 for details.

Statistical inferences on high-dimensional data are another focus of this book. Statistical inferences require one to characterize the uncertainty, estimate the standard errors of the estimated parameters of primary interest and

Copyrighted material

derive the asymptotic distributions of the resulting estimates. This is very challenging under the high-dimensional regime. See Chapter 7.

Fueled by the surging demands on processing high-dimensional and big data, there have been rapid and vast developments in high-dimensional statistics and machine learning over the last decade, contributed by data scientists from various fields such as statistics, computer science, information theory, applied and computational mathematics, and others. Even though we have narrowed the scope of the book to the statistical aspects of data science, the field is still too broad for us to cover. Many important contributions that do not fit our presentation have been omitted. Conscientious effort was made in the composition of the reference list and bibliographical notes, but they merely reflect our immediate interests. Omissions and discrepancies are inevitable. We apologize for their occurrence.

Although we all contribute to various chapters and share the responsibility for the whole book, Jianqing Fan was the lead author for Chapters 1, 3 and 9–11, 14 and some sections in other chapters, Runze Li for Chapters 5, and 8 and part of Chapters 6–7, Cun-Hui Zhang for Chapters 4 and 7, and Hui Zou for Chapters 2, 6, 11 and 12 and part of Chapter 5.

Many people have contributed importantly to the completion of this book. In particular, we would like to thank the editor, John Kimmel, who has been extremely helpful and patient with us for over 10 years! We greatly appreciate a set of around 10 anonymous reviewers for valuable comments that have led to the improvement of the book. We are particularly grateful to Cong Ma and Yiqiao Zhong for preparing a draft of Chapter 14, to Zhao Chen for helping us with putting our unsorted and non-uniform references into the present form, to Tracy Ke, Bryan Kelly, Dacheng Xiu and Jia Wang for helping us with constructing Figure 1.3, to Krishna Balasubramanian, Cong Ma, Lingzhou Xue, Boxiang Wang, Kaizheng Wang, Yi Yang, and Ziwei Zhu for producing some figures. Various people have carefully proof-read certain chapters of the book and made useful suggestions. They include Krishna Balasubramanian, Pierre Bayle, Alexander Chen, Elynn Chen, Wenyan Gong, Yongyi Guo, Bai Jiang, Cong Ma, Igor Silin, Qiang Sun, Francesca Tang, Bingyan Wang, Kaizheng Wang, Weichen Wang, Yuling Yan, Zhuoran Yang, Mengxin Yu, Wen-Xin Zhou, Yifeng Zhou, and Ziwei Zhu. We owe them many thanks.

We used a draft of this book as a textbook for a first-year graduate course at Princeton University in 2019 and 2020 and a senior graduate topic course at Pennsylvania State University in 2019. We would like to thank the graduate students in the classes for their careful readings. In particular, we are indebted to Cong Ma, Kaizheng Wang and Zongjun Tan for assisting in preparing the homework problems used for the Princeton course, most of which are now a part of our exercise at the end of each chapter. At Princeton, we covered chapters 2-3, 5, 8.1, 8.3, 9-12, 13.1-13.3, 14.

We are grateful to our teachers who educate us and to all of our collaborators for many enjoyable and stimulating collaborations. Finally, we would like to thank our families for their love and support.

Copyrighted material

Jianqing Fan
Runze Li
Cun-Hui Zhang
Hui Zou

January 2020.

Copyrighted material

Book websites:

http://personal.psu.edu/ril4/DataScience/

https://orfe.princeton.edu/~jqfan/DataScience/

Copyrighted material

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Copyrighted material

Chapter 1

Introduction

The first two decades of this century have witnessed the explosion of data collection in a blossoming age of information and technology. The recent technological revolution has made information acquisition easy and inexpensive through automated data collection processes. The frontiers of scientific research and technological developments have collected huge amounts of data that are widely available to statisticians and data scientists via internet dissemination. Modern computing power and massive storage allow us to process this data of unprecedented size and complexity. This provides mathematical sciences great opportunities with significant challenges. Innovative reasoning and processing of massive data are now required; novel statistical and computational methods are needed; insightful statistical modeling and theoretical understandings of the methods are essential.

1.1 Rise of Big Data and Dimensionality

Information and technology have revolutionized data collection. Millions of surveillance video cameras, billions of internet searches and social media chats and tweets produce massive data that contain vital information about security, public health, consumer preference, business sentiments, economic health, among others; billions of prescriptions, and an enormous amount of genetics and genomics information provide critical data on health and precision medicine; numerous experiments and observations in astrophysics and geosciences give rise to big data in science.

Nowadays, *Big Data* are ubiquitous: from the internet, engineering, science, biology and medicine to government, business, economy, finance, legal, and digital humanities. "There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days", according to Eric Schmidt, the CEO of Google, in 2010; "Data are becoming the new raw material of business", according to Craig Mundie, Senior Advisor to the CEO at Microsoft; "Big data is not about the data", according to Gary King of Harvard University. The first quote is on the volume, velocity, variety, and variability of big data nowadays, the second is about the value of big data and its impact on society, and the third quote is on the importance of the smart analysis of big data.

Copyrighted material

Accompanying *Big Data* is rising dimensionality. Frontiers of scientific research depend heavily on the collection and processing of massive complex data. Big data collection and high dimensionality characterize many contemporary statistical problems, from sciences and engineering to social science and humanities. For example, in disease classification using microarray or proteomics data, tens of thousands of expressions of molecules or proteins are potential predictors; in genome-wide association studies, hundreds of thousands of single-nucleotide polymorphisms (SNPs) are potential covariates; in machine learning, millions or even billions of features are extracted from documents, images and other objects; in spatial-temporal problems in economics and earth sciences, time series of hundreds or thousands of regions are collected. When interactions are considered, the dimensionality grows much more quickly. Yet, interaction terms are needed for understanding the synergy of two genes, proteins or SNPs or the meanings of words. Other examples of massive data include high-resolution images, high-frequency financial data, e-commerce data, warehouse data, functional and longitudinal data, among others. See also Donoho (2000), Fan and Li (2006), Hastie, Tibshirani and Friedman (2009), Bühlmann and van de Geer (2011), Hastie, Tibshirani and Wainwright (2015), and Wainwright (2019) for other examples.

1.1.1 Biological sciences

Bioimaging technology allows us to simultaneously monitor tens of thousands of genes or proteins as they are expressed differently in the tissues or cells under different experimental conditions. Microarray measures expression profiles of genes, typically in the order of tens of thousands, in a single hybridization experiment, depending on the microarray technology being used. For customized microarrays, the number of genes printed on the chip can be much smaller, giving more accurate measurements on the genes of focused interest. Figure 1.1 shows two microarrays using the Agilent microarray technology and cDNA micorarray technology. The intensity of each spot represents the level of expression of a particular gene. Depending on the nature of the studies, the sample sizes range from a couple to tens or hundreds. For cell lines, the individual variations are relatively small and the sample size can be very small, whereas for tissues from different human subjects, the individual variations are far larger and the sample sizes can be a few hundred.

RNA-seq (Nagalakshmi, et al., 2008), a methodology for RNA profiling based on next-generation sequencing (NGS, Shendure and Ji, 2008), has replaced microarrays for the study of gene expression. Next-generation sequencing is a term used to describe a number of different modern sequencing technologies that allow us to sequence DNA and RNA much more quickly and cheaply. RNA-seq technologies, based on assembling short reads 30∼400 base pairs, offer advantages such as a wider range of expression levels, less noise, higher throughput, in addition to more information to detect allele-specific expression, novel promoters, and isoforms. There are a number of papers on

Copyrighted material

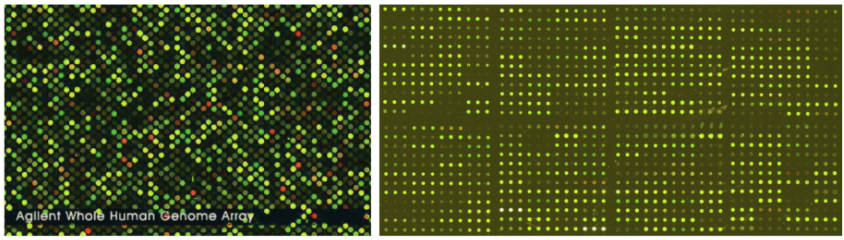

Figure 1.1: Gene expression profiles of microarrays. The intensity at each spot represents the gene expression profile (e.g. Agilent microarray, left panel) or relative profile (e.g. cDNA-microarray, right panel).

statistical methods for detecting differentially expressed genes across treatments/conditions; see Kvam, Liu and Si (2012) for an overview.

After the gene/RNA expression measurements have been properly normalized through RNA-seq or microarray technology, one can then select genes with different expressions under different experimental conditions (e.g. treated with cytokines) or tissues (e.g. normal versus tumor) and genes that express differently over time after treatments (time course experiments). See Speed (2003). This results in a lot of various literature on statistical analysis of controlling the *false discovery rate* in large scale hypothesis testing. See, for example, Benjamini and Hochberg (1995), Storey (2002), Storey and Tibshirani (2003), Efron (2007, 2010b), Fan, Han and Gu (2012), Barber and Candés (2015), Candés, Fan, Janson and Lv (2018), Fan, Ke, Sun and Zhou (2018), among others. The monograph by Efron (2010a) contains a comprehensive account on the subject.

Other aspects of analysis of gene/RNA expression data include association of gene/RNA expression profiles with clinical outcomes such as disease stages or survival time. In this case, the gene expressions are taken as the covariates and the number of variables is usually large even after preprocessing and screening. This results in high-dimensional regression and classification (corresponding to categorical responses, such as tumor types). It is widely believed that only a small group of genes are responsible for a particular clinical outcome. In other words, most of the regression coefficients are zero. This results in high-dimensional sparse regression and classification problems.

There are many other high throughput measurements in biomedical studies. In proteomics, thousands of proteins expression profiles, which are directly related to biological functionality, are simultaneously measured. Similar to genomics studies, the interest is to associate the protein expressions with clinical outcomes and biological functionality. In genomewide association studies, many common genetic variants (typically single-nucleotide polymorphisms or *SNPs*) in different individuals are examined to study if any variant is associated with a trait (heights, weights, eye colors, yields, etc.) or a disease. These

Copyrighted material

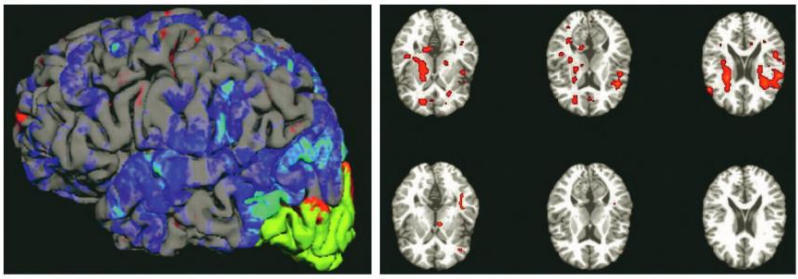

Figure 1.2: Schematic illustration of a brain response to a cognitive task and several slices of its associated fMRI measurements .

genetic variants are referred to as the *quantitative trait loci* (QTL) and hundreds of thousands or millions of SNPs are available for examination. The need for understanding pathophysiology has also led to investigating the so-called *eQTL* studies, the association between SNPs and the expressions of nearby genes. In this case, the gene expressions are regarded as the responses whereas the individual SNPs are taken as the covariates. This again results in high-dimensional regression problems.

High throughput measurements are also commonly used in neuroscience, astronomy, and agriculture and resource surveys using satellite and other imaging technology. In neuroscience, for example, *functional magnetic resonance imaging* (fMRI) technology is frequently applied to measure Blood Oxygenation Level-Dependent (*BOLD*) response to stimuli. This allows investigators to determine which areas of the brain are involved in a cognitive task, or more generally, the functionality of brains. Figure 1.2 gives a schematic illustration. fMRI data contain time-course measurements over tens or hundreds of thousand voxels, resulting in high-dimensional statistical problems.

1.1.2 Health sciences

Health scientists employ many advanced bioinformatic tools to understand molecular mechanisms of disease initiation and progression, and the impact of genetic variations on clinical outcomes. Many health studies also collect a number of risk factors as well as clinical responses over a period of time: many covariates and responses of each subject are collected at different time points. These kinds of longitudinal studies can give rise to high-dimensional big data.

A famous example is the *Framingham Heart Study*, initiated in 1948 and sponsored by the National Heart, Lung and Blood Institute. Documentation of its first 55 years can be found at the website

http://www.framinghamheartstudy.org/.

More details on this study can be found from the website of the American Heart Association. Briefly, the study follows a representative sample of 5,209

Copyrighted material

adult residents and their offspring aged 28-62 years in Framingham, Massachusetts. These subjects have been tracked using standardized biennial cardiovascular examination, daily surveillance of hospital admissions, death information and information from physicians and other sources outside the clinic. In 1971, the study enrolled a second-generation group, consisting of 5,124 of the original participants' adult children and their spouses, to participate in similar examinations.

The aim of the Framingham Heart Study is to identify risk factors associated with heart disease, stroke and other diseases, and to understand the circumstances under which cardiovascular diseases arise, evolve and end fatally in the general population. In this study, there are more than 25,000 samples, each consisting of more than 100 variables. Because of the nature of this longitudinal study, some participants cannot be followed up due to their migrations. Thus, the collected data contain many missing values. During the study, cardiovascular diseases may develop for some participants, while other participants may never experience cardiovascular diseases. This implies that some data are censored because the event of particular interest never occurs. Furthermore, data between individuals may not be independent because data for individuals in a family are clustered and likely positively correlated. Missing, censoring and clustering are common features in health studies. These three issues make the data structure complicated and identification of important risk factors more challenging.

High-dimensionality is frequently seen in many other biomedical studies. It also arises in the studies of health costs, health care and health records.

1.1.3 Computer and information sciences

The development of information and technology itself collects massive amounts of data. For example, there are billions of web pages on the internet, and an internet search engine needs to statistically learn the most likely outcomes of a query and fast algorithms need to evolve with empirical data. The input dimensionality of queries can be huge. In Google, Facebook and other social networks, algorithms are designed to predict the potential interests of individuals in certain services or products. A familiar example of this kind is amazon.com in which related books are recommended online based on user inputs. This kind of recommendation system applies to other types of services such as music and movies. These are just a few examples of statistical learning in which the data sets are huge and highly complex, and the number of variables is ultrahigh.

Machine learning algorithms have been widely applied to pattern recognition, search engines, computer vision, document and image classification, bioinformatics, medical diagnosis, natural language processing, knowledge graphs, automatic driving machines, internet doctors, among others. The development of these algorithms is based on high-dimensional statistical regres-

Copyrighted material

Figure 1.3: Some illustrations of machine learning. Top panel: the word clouds of sentiments of a company (Left: Negative Words; Right: Positive Words). The plots were constructed by using data used in Ke, Kelly and Xiu (2019). Bottom left: It is challenging for a computer to recognize the pavilion from the background in computer vision. Bottom right: Visualization of the friendship connections in Facebook.

sion and classification with a large number of predictors and a large amount of empirical data. For example, in text and document classification, the data of documents are summarized by word-document information matrices: the frequencies of the words and phrases x in document y are computed. This step of *feature extraction* is very important for the accuracy of classification. A specific example of document classification is E-mail spam in which there are only two classes of E-mails, junk or non-junk. Clearly, the number of features should be very large in order to find important features for accurate document classifications. This results in high-dimensional classification problems.

Similar problems arise for image or object classifications. Feature extractions play critical roles. One approach for such a feature extrapolation is the classical *vector quantization* technique, in which images are represented by many small subimages or *wavelet* coefficients, which are further reduced by summary statistics. Again, this results in high-dimensional predictive variables. Figure 1.3 illustrates a few problems that arise in machine learning.

Copyrighted material

1.1.4 Economics and finance

Thanks to the revolution in information and technology, high-frequency financial data have been collected for a host of financial assets, from stocks, bonds, and commodity prices to foreign exchange rates and financial derivatives. The asset correlations among 500 stocks in the S&P500 Index already involve over a hundred thousand parameters. This poses challenges in accurately measuring the financial risks of the portfolios, systemic risks in the financial systems, bubble migrations, and risk contagions, in additional to portfolio allocation and management (Fan, Zhang and Yu, 2012; Brownlees and Engle, 2017). For an overview of high-dimensional economics and finance, see, for example, Fan, Lv and Qi (2012).

To understand the dynamics of financial assets, large panels of financial time series are widely available within asset classes (e.g. components of Russell 3000 stocks) and across asset classes (e.g. stocks, bonds, options, commodities, and other financial derivatives). This is important for understanding the dynamics of price co-movements, time-dependent large volatility matrices of asset returns, systemic risks, and bubble migrations.

Large panel data also arise frequently in economic studies. To analyze the joint evolution of macroeconomic time series, hundreds of macroeconomic variables are compiled to better understand the impact of government policies and to gain better statistical accuracy via, for example, the vector autoregressive model (Sims, 1980). The number of parameters is very large since it grows quadratically with the number of predictors. To enrich the model information, Bernanke et al. (2005) propose to augment standard VAR models with estimated factors (FAVAR) to measure the effects of monetary policy. Factor analysis also plays an important role in prediction using large dimensional data sets (for reviews, see Stock and Watson (2006), Bai and Ng (2008)). A comprehensive collection of 131 macroeconomics time series (McCracken and Ng, 2015) with monthly updates can be found in the website

https://research.stlouisfed.org/econ/mccracken/fred-databases/ .

Spatial-temporal data also give rise to big data in economics. Unemployment rates, housing price indices and sale data are frequently collected in many regions, detailed up to zip code level, over a period of time. The use of spatial correlation enables us to better model the joint dynamics of the data and forecast future outcomes. In addition, exploring homogeneity enables us to aggregate a number of homogeneous regions to reduce the dimensionality, and hence statistical uncertainties, and to better understand heterogeneity across spatial locations. An example of this in prediction of housing appreciation was illustrated in the paper by Fan, Lv, and Qi (2012). See Figure 1.4 and Section 3.9.

Copyrighted material

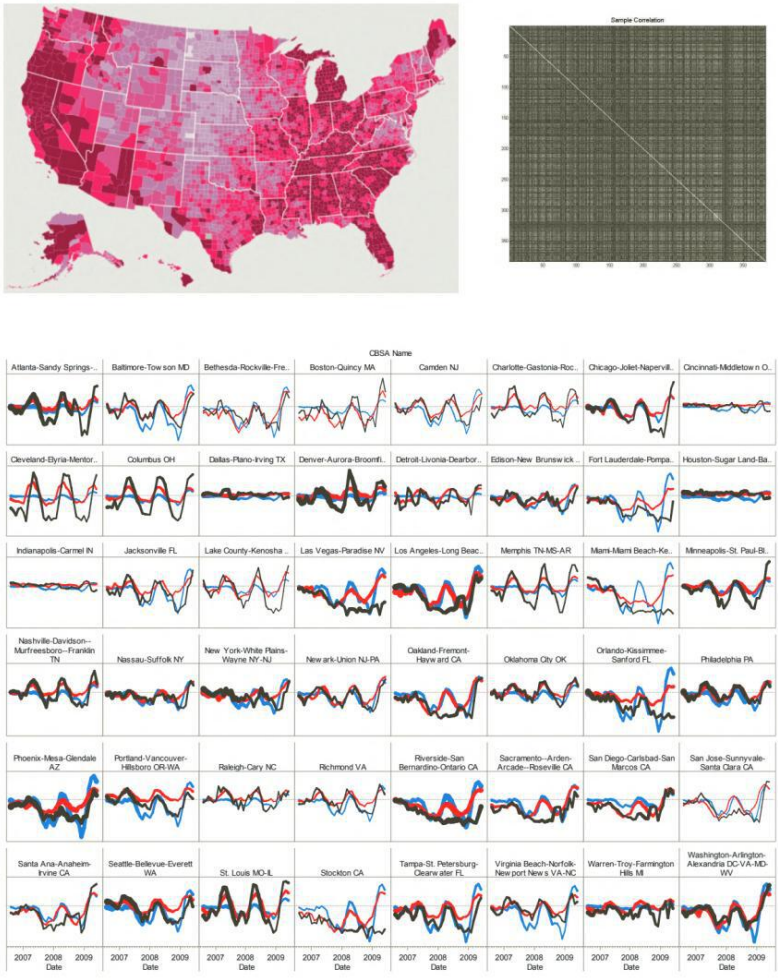

Figure 1.4: Prediction of monthly housing appreciation. Top panel-left: Choropleth map for the 2009 U.S. unemployment rate by county. Top panel-right: Spatial correlation of monthly housing price appreciation among 352 largest counties in the United States from January 2000 to December 2009 (from Fan, Lv, and Qi, 2012). Bottom panel: Prediction of monthly housing pricing appreciation in 48 regions from January 2006 to December 2009 using a large sparse econometrics model with 352 monthly time series from January 2000 to December 2005. Blue: OLS. Red: PLS. Black: Actual. Thickness: Proportion to repeated sales. Adapted from Fan, Lv, and Qi (2012).

Copyrighted material

1.1.5 Business and program evaluation

Big data arises frequently in marketing and program evaluation. Multi-channel strategies are frequently used to market products, such as drugs and medical devices. Data from hundreds of thousands of doctors are collected with different marketing strategies over a period of time, resulting in big data. The design of marketing strategies and the evaluation of a program's effectiveness are important to corporate revenues and cost savings. This also applies to online advertisements and AB-tests.

Similarly, to evaluate government programs and policies, large numbers of confounders are collected, along with many individual responses to the treatment. This results in big and high-dimensional data.

1.1.6 Earth sciences and astronomy

Spatial-temporal data have been widely available in the earth sciences. In meteorology and climatology studies, measurements such as temperatures and precipitations are widely available across many regions over a long period of time. They are critical for understanding climate changes, local and global warming, and weather forecasts, and provide an important basis for energy storage and pricing weather based financial derivatives.

In astronomy, sky surveys collect a huge amount of high-resolution imaging data. They are fundamental to new astronomical discoveries and to understanding the origin and dynamics of the universe.

1.2 Impact of Big Data

The arrival of *Big Data* has had deep impact on data systems and analysis. It poses great challenges in terms of storage, communication and analysis. It has forever changed many aspects of computer science, statistics, and computational and applied mathematics: from hardware to software; from storage to super-computing; from data base to data security; from data communication to parallel computing; from data analysis to statistical inference and modeling; from scientific computing to optimization. The efforts to provide solutions to these challenges gave birth to a new disciplinary science, data science. Engulfed by the applications in various disciplines, *data science* consists of studies on data acquisition, storage and communication, data analysis and modeling, and scalable algorithms for data analysis and artificial intelligence. For an overview, see Fan, Han, and Liu (2014).

Big Data powers the success of statistical prediction and artificial intelligence. Deep *artificial neural network* models have been very successfully applied to many *machine learning* and prediction problems, resulting in a discipline called *deep learning* (LeCun, Bengio and Hinton, 2015; Goodfellow, Bengio and Courville, 2016). Deep learning uses a family of over parameterized models, defined through deep neural networks, that have small modeling biases. Such an over-parameterized family of models typically has large vari-

Copyrighted material

ances, too big to be useful. It is the big amount of data that reduces the variance to an acceptable level, achieving bias and variance trade-offs in prediction. Similarly, such an over-parameterized family of models typically is too hard to find reasonable local minima, and it is modern computing power and cheap GPUs that make the implementation possible. It is fair to say that today's success of deep learning is powered by the arrivals of big data and modern computing power. These successes will be further carried into the future, as we collect even bigger data and become even better computing architecture.

As Big Data are typically collected by automated process and by different generations of technologies, the quality of data is low and measurement errors are inevitable. Since data are collected from various sources and populations, the problem of *heterogeneity* of big data arises. In addition, since the number of variables is typically large, many variables have high kurtosis (much higher than the normal distribution). Moreover, *endogeneity* occurs incidentally due to high-dimensionality that has huge impacts on model selection and statistical inference (Fan and Liao, 2014). These intrinsic features of Big Data have significant impacts on the future developments of big data analysis techniques, from heterogeneity and heavy tailedness to endogeneity and measurement errors. See Fan, Han, and Liu (2014).

Big data are often collected at multiple locations and owned by different parties. They are often too big and unsafe to be stored in one single machine. In addition, the processing power required to manipulate big data is not satisfied by standard computers. For these reasons, big data are often distributed in multiple locations. This creates the issues of communications, privacy and owner issues.

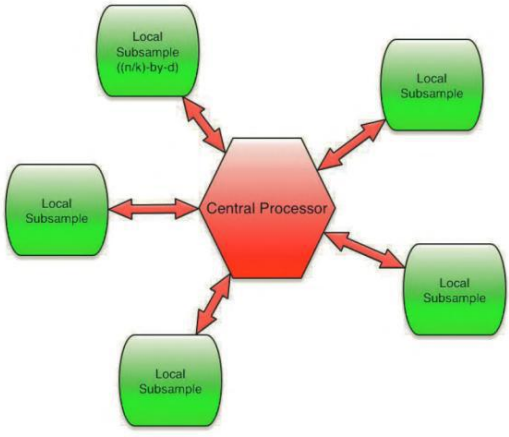

Figure 1.5: Schematic illustration of the distributed data analysis and computing architecture.

Copyrighted material

A simple architecture that tackles simultaneously the storage, communication, privacy and ownership issues is the *distributed data analysis* in Figure 1.5. Here, each node analyzes the local data and communicates only the results to the central machine. The central machine then aggregates the results and reports the final results (one-shot analysis) or communicates the results back to each node machine for further analysis (multi-shot analysis). For recent developments on this subject, see Shamir, Srebro and Zhang (2014), Zhang, Duchi and Wainwright (2015), Jordan, Lee and Yang (2018) for low-dimensional regression; Chen and Xie (2014), Lee, Liu, Sun and Taylor (2017), Battey, Fan, Liu, Lu and Zhu (2018) for high-dimensional sparse regression and inference, and El Karoui and d'Aspremont (2010), Liang, et al. (2014), Bertrand and Moonen (2014), Schizas and Aduroja (2015), Garber, Shamir and Srebro (2017), and Fan, Wang, Wang and Zhu (2019) for *principal component analysis*.

As mentioned before, big data are frequently accompanied by high-dimensionality. We now highlight the impacts of dimensionality on data analysis.

1.3 Impact of Dimensionality

What makes high-dimensional statistical inference different from traditional statistics? High-dimensionality has a significant impact on computation, spurious correlation, noise accumulation, and theoretical studies. We now briefly touch these topics.

1.3.1 Computation

Statistical inferences frequently involve numerical optimization. Optimizations in millions and billions of dimensional spaces are not unheard of and arise easily when interactions are considered. High-dimensional optimization is not only expensive in computation, but also slow in convergence. It also creates numerical instability. Algorithms can easily get trapped at local minima. In addition, algorithms frequently use iteratively the inversions of large matrices, which causes many instability issues in addition to large computational costs and memory storages. Scalable and stable implementations of high-dimensional statistical procedures are very important to statistical learning.

Intensive computation comes also from the large number of observations, which can be in the order of millions or even billions as in marketing and machine learning studies. In these cases, computation of summary statistics such as correlations among all variables is expensive, yet statistical methods often involve repeated evaluations of summation of loss functions. In addition, when new cases are added, it is ideal to only update some of the summary statistics, rather than to use the entire updated data set to redo the computation. This also saves considerable data storage and computation. Therefore,

Copyrighted material

scalability of statistical techniques to both dimensionality and the number of cases is paramountly important.

The high dimensionality and the availability of big data have reshaped statistical thinking and data analysis. Dimensionality reduction and feature extraction play pivotal roles in all high-dimensional statistical problems. This helps reduce computation costs as well as improve statistical accuracy and scientific interpretability. The intensive computation inherent in these problems has altered the course of methodological developments. Simplified methods have been developed to address the large-scale computational problems. Data scientists are willing to trade statistical efficiencies with computational expediency and robust implementations. Fast and stable implementations of optimization techniques are frequently used.

1.3.2 Noise accumulation

High-dimensionality has significant impact on statistical inference in at least two important aspects: *noise accumulation* and *spurious correlation*. Noise accumulation refers to the fact that when a statistical rule depends on many parameters, each estimated with stochastic errors, the estimation errors in the rule can accumulate. For high-dimensional statistics, noise accumulation is more severe, and can even dominate the underlying signals. Consider, for example, a linear classification rule which classifies a new data point $\mathbf{x}$ to class 1 if $\mathbf{x}^T\boldsymbol{\beta} > 0$. This rule can have high discrimination power when $\boldsymbol{\beta}$ is known. However, when an estimator $\widehat{\boldsymbol{\beta}}$ is used instead, due to accumulation of errors in estimating the high-dimensional vector $\widehat{\boldsymbol{\beta}}$, the classification rule can be as bad as random guessing.

To illustrate the above point, let us assume that we have random samples $\{\mathbf{X}_i\}_{i=1}^n$ and $\{\mathbf{Y}_i\}_{i=1}^n$ from class 0 and class 1 with the population distributions $N(\boldsymbol{\mu}_0, \mathbf{I}_p)$ and $N(\boldsymbol{\mu}_1, \mathbf{I}_p)$, respectively. To mimic the gene expression data, we take $p = 4500$, $\boldsymbol{\mu}_0 = 0$ without loss of generality, and $\boldsymbol{\mu}_1$ from a realization of $0.98\delta_0 + 0.02 * \text{DE}$, a mixture of point mass 0 with probability 0.98 and the standard double exponential distribution with probability 0.02. The realized $\boldsymbol{\mu}_1$ is shown in Figure 1.6, which should have about 90 non-vanishing components and is taken as true $\boldsymbol{\mu}_1$. The components that are considerably different from zero are numbered far less than 90, around 20 to 30 or so.

Unlike high-dimensional regression problems, high-dimensional classification does not have implementation issues if the Euclidian distance based classifier is used; see Figure 1.6. It classifies $\mathbf{x}$ to class 1 if

$$\|\mathbf{x} - \boldsymbol{\mu}_1\|^2 \le \|\mathbf{x} - \boldsymbol{\mu}_0\|^2 \qquad \text{or} \qquad \boldsymbol{\beta}^T(\mathbf{x} - \boldsymbol{\mu}) \ge 0, \tag{1.1}$$

where $\boldsymbol{\beta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$. For the particular setting in the last paragraph, the distance-based classifier is the Fisher classifier and is the optimal Bayes classifier if prior probability of class 0 is 0.5. The misclassification probability for $\mathbf{x}$ from class 1 into class 0 is $\Phi(-\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|/2)$. This reveals the fact that components with large differences contribute more to differentiating

Copyrighted material

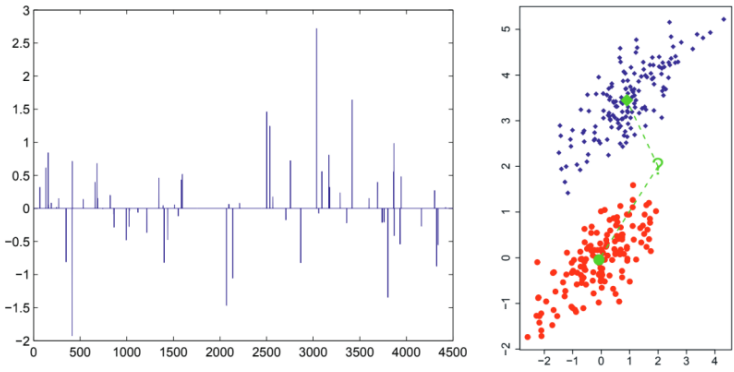

Figure 1.6: Illustration of Classification. Left panel: a realization of $\{\mu_j\}_{j=1}^{4500}$ from the mixture distribution $0.98\delta_0 + 0.02 * \text{DE}$, where DE stands for the standard Double Exponential distribution. Right panel: Illustration of the Euclidian distance based classifier, which classifies the query to a class according to its distances to the centroids.

the two classes, and the more components the smaller the discrimination error. In other words, $\Delta_p = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|$ is a nondecreasing function of p. Let $\Delta_{(m)}$ be the distance computed based on the m largest components of the difference vector $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. For our particular specification in the last paragraph, the misclassification rate is around $\Phi(-\sqrt{2^2 + 2.5^2}/2) = 0.054$ when the two most powerful components are used ($m = 2$). In addition, $\Delta_{(m)}$ stops increasing noticeably when m reaches 30 and will be constant when $m \geq 100$.

The practical implementation requires estimates of the parameters such as $\widehat{\boldsymbol{\beta}}$. The actual performance of the classifiers can differ from our expectation due to the noise accumulation. To illustrate the noise accumulation phenomenon, let us assume that the rank of the importance of the p features is known to us. In this case, if we use only two features, the classification power is very high. This is shown in Figure 1.7(a). Since the dimensionality is low, the noise in the estimated parameters is negligible. Now, if we take $m = 100$, the signal strength Δ_m increases. On the other hand, we need to estimate 100 coefficients $\boldsymbol{\beta}$, which accumulate stochastic noises in the classifier. To visualize this, we project the observed data onto the first two principal components of these 100-dimensional selected features. From Figure 1.7(b), it is clear that signal and noise effect cancel. We still have classification power to differentiate the two classes. When $m = 500$ and 4500, there is no further increase of signals and noise accumulation effect dominates. The performance is as poorly as random guessing. Indeed, Fan and Fan (2008) show that almost all high-dimensional classifiers can perform as poorly as random guessing unless the signal is excessively strong. See Figure 1.7(c) and (d).

Copyrighted material

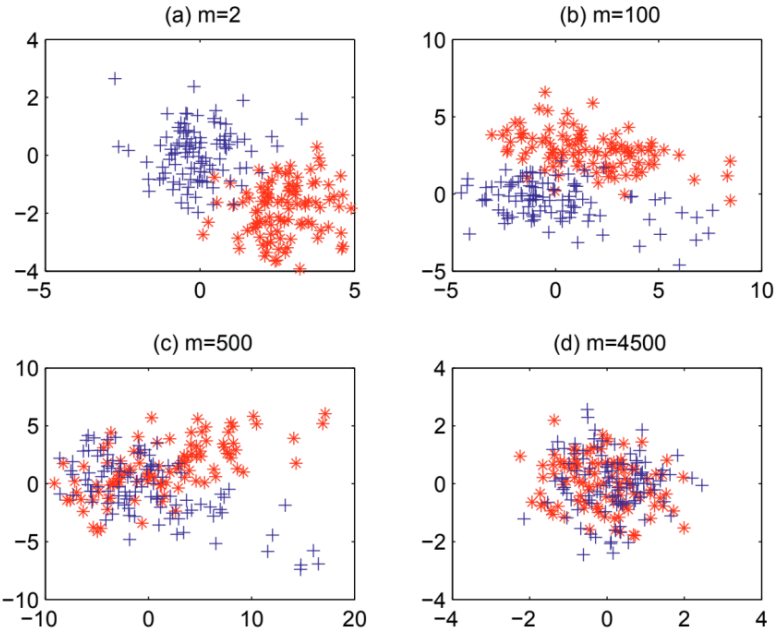

Figure 1.7: Illustration of noise accumulation. Left panel: Projection of observed data ($n = 100$ from each class) onto the first two principal components of m-dimensional selected feature space. The m most important features are extracted before applying the principal component analysis.

Fan and Fan (2008) quantify explicitly the price paid with use of more features. They demonstrate that the classification error rate depends on $\Delta_m/\sqrt{m}$. The numerator shows the benefit of the dimensionality through the increase of signals Δ_m, whereas the denominator represents the noise accumulation effect due to estimation of the unknown parameters. In particular, when $\Delta_p/\sqrt{p} \to \infty$ as $p \to \infty$, Hall, Pittelkow and Ghosh (2008) show that the problem is perfectly classifiable (error rate converges to zero).

The above illustration of the noise accumulation phenomenon reveals the pivotal role of feature selection in high dimensional statistical endeavors. Not only does it reduce the prediction error, but also improves the interpretability of the classification rule. In other words, the use of sparse $\boldsymbol{\beta}$ is preferable.

1.3.3 Spurious correlation

Spurious correlation refers to the observation that two variables which have no population correlation have a high sample correlation. The analogy is that two persons look alike but have no genetic relation. In a small village,

Copyrighted material

spurious correlation rarely occurs. This explains why spurious correlation is not an issue in traditional low-dimensional statistics. In a moderate sized city, however, spurious correlations start to occur. One can find two similar looking persons with no genetic relation. In a large city, one can easily find two persons with similar appearances who have no genetic relation. In the same vein, high dimensionality easily creates issues of spurious correlation.

To illustrate the above concept, let us generate a random sample of size $n = 50$ of $p+1$ independent standard normal random variables $Z_1, \cdots, Z_{p+1} \sim_{i.i.d.} N(0,1)$. Theoretically, the sample correlation between any of two random variables is small. When p is small, say $p = 10$, this is indeed the case and the issue of spurious correlation is not severe. However, when p is large, the spurious correlation starts to be noticeable. To illustrate this, let us compute

$$\widehat{r} = \max_{j \geq 2} \widehat{\text{cor}}(Z_1, Z_j) \tag{1.2}$$

where $\widehat{\text{cor}}(Z_1, Z_j)$ is the sample correlation between the variables Z_1 and Z_j. Similarly, let us compute

$$\widehat{R} = \max_{|\mathcal{S}|=5} \widehat{\text{cor}}(Z_1, \mathbf{Z}_{\mathcal{S}}) \tag{1.3}$$

where $\widehat{\text{cor}}(Z_1, \mathbf{Z}_{\mathcal{S}})$ is the multiple correlation between Z_1 and $\mathbf{Z}_{\mathcal{S}}$, namely, the correlation between Z_1 and its best linear predictor using $\mathbf{Z}_{\mathcal{S}}$. To avoid computing all $\binom{p}{5}$ multiple R^2 in (1.3), we use the forward selection algorithm to compute $\widehat{R}$. The actual value of $\widehat{R}$ is larger than what we present here. We repeat this experiment 200 times and present the distributions of $\widehat{r}$ and $\widehat{R}$ in Figure 1.8.

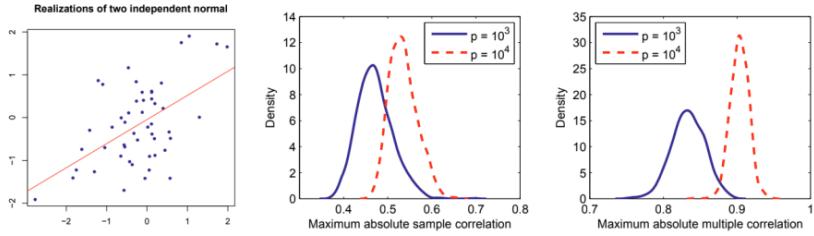

Figure 1.8: Illustration of spurious correlation. Left panel: a typical realization of Z_1 with its mostly spuriously correlated variable ($p = 1000$); middle and left panels: distributions of $\widehat{r}$ and $\widehat{R}$ for $p = 1,000$ and $p = 10,000$, respectively. The sample size is $n = 50$.

The maximum spurious correlation $\widehat{r}$ is around 0.45 for $p = 1000$ and 0.55 for $p = 10,000$. They become 0.85 and 0.91 respectively when multiple correlation $\widehat{R}$ in (1.3) is considered. Theoretical results on the order of these spurious correlations can be found in Cai and Jiang (2012) and Fan, Guo and

Copyrighted material

Hao (2012), and more comprehensively in Fan, Shao, and Zhou (2018) and Fan and Zhou (2016).

The impact of *spurious correlation* includes false scientific discoveries and false statistical inferences. Since the correlation between Z_1 and $\mathbf{Z}_{\widehat{S}}$ is around 0.9 for a set $\widehat{S}$ with $|\widehat{S}| = 5$ (Figure 1.8), Z_1 and $\mathbf{Z}_{\widehat{S}}$ are practically indistinguishable given $n = 50$. If Z_1 represents the gene expression of a gene that is responsible for a disease, we will also discover 5 genes $\widehat{S}$ that have a similar predictive power although they have no relation to the disease.

To further appreciate the concept of spurious correlation, let us consider the neuroblastoma data used in Oberthuer et al. (2006). The study consists of 251 patients, aged from 0 to 296 months at diagnosis with a median age of 15 months, of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004. Neuroblastoma is a common pediatric solid cancer, accounting for around 15% of pediatric cancers. 251 neuroblastoma specimens were analyzed using a customized oligonucleotide microarray with $p = 10,707$ gene expressions available after preprocessing. The clinical outcome is taken as the indicator of whether a neuroblastoma child has a 3 year event-free survival. 125 cases are taken at random as the training sample (with 25 positives) and the remaining data are taken as the testing sample. To illustrate the spurious correlation, we now replace the gene expressions by artificially simulated Gaussian data. Using only $p = 1000$ artificial variables along with the traditional forward selection, we can easily find 10 of those artificial variables that perfectly classify the clinical outcomes. Of course, these 10 artificial variables have no relation with the clinical outcomes. When the classification rule is applied to the test samples, the classification result is the same as random guessing.

To see the impact of spurious correlation on statistical inference, let us consider a linear model

$$Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon, \qquad \sigma^2 = \mathrm{Var}(\varepsilon). \tag{1.4}$$

Let $\widehat{\mathcal{S}}$ be a selected subset and we compute the residual variances based on the selected variables $\widehat{\mathcal{S}}$:

$$\widehat{\sigma}^2 = \mathbf{Y}^T(I_n - \mathbf{P}_{\widehat{\mathcal{S}}})\mathbf{Y}/(n - |\widehat{\mathcal{S}}|), \qquad \mathbf{P}_{\widehat{\mathcal{S}}} = \mathbf{X}_{\widehat{\mathcal{S}}}(\mathbf{X}_{\widehat{\mathcal{S}}}^T\mathbf{X}_{\widehat{\mathcal{S}}})^{-1}\mathbf{X}_{\widehat{\mathcal{S}}}^T. \tag{1.5}$$

In particular, when $\boldsymbol{\beta} = 0$, all selected variables are spurious. In this case, $\mathbf{Y} = \boldsymbol{\varepsilon}$ and

$$\widehat{\sigma}^2 \approx (1 - \gamma_n^2)\|\boldsymbol{\varepsilon}\|^2/n \approx (1 - \gamma_n^2)\sigma^2, \tag{1.6}$$

when $|\widehat{\mathcal{S}}|/n \to 0$, where $\gamma_n^2 = \boldsymbol{\varepsilon}^T\mathbf{P}_{\widehat{\mathcal{S}}}\boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|^2$. Therefore, σ^2 is underestimated by a factor of γ_n^2

Suppose that we select only one spurious variable, then that variable must be mostly correlated with $\mathbf{Y}$. Since the spurious correlation is high, the bias is large. The two left panels of Figure 1.9 depicts the distribution of γ_n along with the associated estimates of $\widehat{\sigma}^2$ for different choices of p. Clearly, the bias increases with the dimensionality p.

Copyrighted material

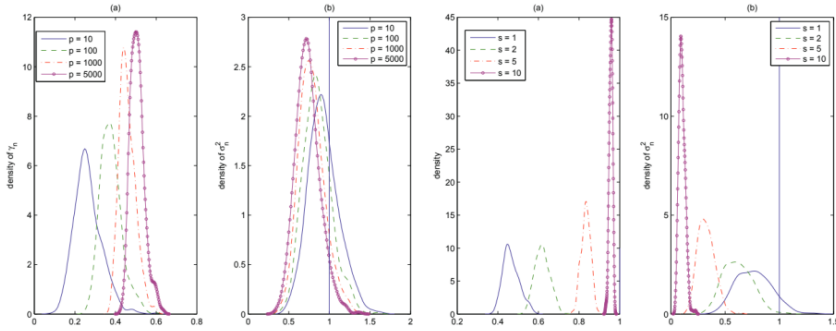

Figure 1.9: Distributions of spurious correlations. Left panel: Distributions of γ_n for the null model when $|\widehat{\mathcal{S}}| = 1$ and their associated estimates of $\sigma^2 = 1$ for various choices of p. Right panel: Distributions of γ_n for the model $Y = 2X_1 + 0.3X_2 + \varepsilon$ and their associated estimates of $\sigma^2 = 1$ for various choices of $|\widehat{\mathcal{S}}|$ but fixed $p = 1000$. The sample size $n = 50$. Adapted from Fan, Guo, and Hao (2012).

Spurious correlation gets larger when more than one spurious variables are selected, as seen in Figure 1.8. To see this, let us consider the linear model $Y = 2X_1 + 0.3X_2 + \varepsilon$ and use forward selection methods to recruit variables. Again, the spurious variables are selected mainly due to their spurious correlation with ε, the unobservable but realized random noises. As shown in the right panel of Figure 1.9, the spurious correlation is very large and $\widehat{\sigma}^2$ gets notably more biased when $|\widehat{\mathcal{S}}|$ gets larger.

Underestimate of residual variance leads to further wrong statistical inferences. More variables will be called statistically significant and that further leads to wrong scientific conclusions. There is active literature on selective inference for dealing with such kinds of issues, starting from Lockhart, Taylor, Tibshirani and Tibshirani (2014); see also Taylor and Tibshirani (2015) and Tibshirani, Taylor, Lockhart and Tibshirani (2016).

1.3.4 Statistical theory

High dimensionality has a strong impact on statistical theory. The traditional asymptotic theory assumes that sample size n tends to infinity while keeping p fixed. This does not reflect the reality of the high dimensionality and cannot explain the observed phenomena such as noise accumulation and spurious correlation. A more reasonable framework is to assume p grows with n and investigate how high the dimensionality p_n a given procedure can handle given the sample size n. This new paradigm is now popularly used in the literature.

Copyrighted material

High dimensionality gives rise to new statistical theory. Many new insights have been unveiled and many new phenomena have been discovered. Subsequent chapters will unveil some of these.

1.4 Aim of High-dimensional Statistical Learning

As shown in Section 1.1, high-dimensional statistical learning arises from various different scientific contexts and has very different disciplinary goals. Nevertheless, its statistical endeavor can be abstracted as follows. The main goals of high dimensional inferences, according to Bickel (2008), are

(a) to construct a method as effective as possible to predict future observations and

(b) to gain insight into the relationship between features and responses for scientific purposes, as well as, hopefully, to construct an improved prediction method.

This view is also shared by Fan and Li (2006). The former appears in problems such as text and document classifications or portfolio optimizations, in which the performance of the procedure is more important than understanding the features that select spam e-mail or stocks that are chosen for portfolio construction. The latter appears naturally in many genomic studies and other scientific endeavors. In these cases, scientists would like to know which genes are responsible for diseases or other biological functions, to understand the molecular mechanisms and biological processes, and predict future outcomes. Clearly, the second goal of high dimensional inferences is more challenging.

The above two objectives are closely related. However, they are not necessarily the same and can be decisively different. A procedure that has a good mean squared error or, more generally risk properties, might not have model selection consistency. For example, if an important variable is missing in a model selection process, the method might find 10 other variables, whose linear combination acts like the missing important variable, to proxy it. As a result, the procedure can still have good prediction power. Yet, the absence of that important variable can lead to false scientific discoveries for objective (b).

As will be seen in Sec. 3.3.2, Lasso (Tibshirani, 1996) has very good risk properties under mild conditions. Yet, its model selection consistency requires the restricted *irrepresentable condition* (Zhao and Yu, 2006; Zou, 2006; Meinshausen and Bühlmann, 2006). In other words, one can get optimal rates in mean squared errors, and yet the selected variables can still differ substantially from the underlying true model. In addition, the estimated coefficients are biased. In this view, Lasso aims more at objective (a). In an effort to resolve the problems caused by the L_1-penalty, a class of *folded-concave* penalized least-squares or likelihood procedures, including SCAD, was introduced by Fan and Li (2001), which aims more at objective (b).

Copyrighted material

1.5 What Big Data Can Do

Big Data hold great promise for the discovery of heterogeneity and search for personalized treatments and precision marketing. An important aim for big data analysis is to understand heterogeneity for personalized medicine or services from large pools of variables, factors, genes, environments and their interactions as well as latent factors. Such a kind of understanding is only possible when sample size is very large, particularly for rare diseases.

Another important aim of big data is to discover the commonality and weak patterns, such as the impact of drinking teas and wines on health, in the presence of large variations. Big data allow us to reduce large variances of complexity models such as deep neural network models, as discussed in Section 1.2. The successes of *deep learning* technologies rest to quite an extent on the variance reduction due to big data so that a stable model can be constructed.

1.6 Scope of the Book

This book will provide a comprehensive and systematic account of theories and methods in high-dimensional data analysis. The statistical problems range from high-dimensional sparse regression, compressed sensing, sparse likelihood-based models, supervised and unsupervised learning, large covariance matrix estimation and graphical models, high-dimensional survival analysis, robust and quantile regression, among others. The modeling techniques can either be parametric, semi-parametric or nonparametric. In addition, variable selection via regularization methods and sure independent feature screening methods will be introduced.

Copyrighted material

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Copyrighted material

Chapter 2

Multiple and Nonparametric Regression

2.1 Introduction

In this chapter we discuss some popular linear methods for regression analysis with continuous response variable. We call them linear regression models in general, but our discussion is not limited to the classical multiple linear regression. They are extended to multivariate nonparametric regression via the kernel trick. We first give a brief introduction to multiple linear regression and least-squares, presenting the basic and important ideas such as inferential results, Box-Cox transformation and basis expansion. We then discuss linear methods based on regularized least-squares with ridge regression as the first example. We then touch on the topic of nonparametric regression in a reproducing kernel Hilbert space (RKHS) via the kernel trick and kernel ridge regression. Some basic elements of the RKHS theory are presented, including the famous representer theorem. Lastly, we discuss the leave-one-out analysis and generalized cross-validation for tuning parameter selection in regularized linear models.

2.2 Multiple Linear Regression

Consider a *multiple linear regression* model:

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{2.1}$$

where Y represents the *response* or *dependent variable* and the X variables are often called *explanatory variables* or *covariates* or *independent variables*. The intercept term can be included in the model by including 1 as one of the covariates, say $X_1 = 1$. Note that the term "random error" ε in (2.1) is a generic name used in statistics. In general, the "random error" here corresponds to the part of the response variable that cannot be explained or predicted by the covariates. It is often assumed that "random error" ε has zero mean, uncorrelated with covariates X, which is referred to as *exogenous* variables. Our goal is to estimate these β's, called *regression coefficients*, based on a random sample generated from model (2.1).

Suppose that $\{(X_{i1}, \cdots, X_{ip}, Y_i)\}, i = 1, \cdots, n$ is a random sample from

Copyrighted material

model (2.1). Then, we can write

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i. \tag{2.2}$$

The method of least-squares is a standard and popular technique for data fitting. It was advanced early in the nineteenth century by Gauss and Legendre. In (2.2) we have the residuals (r_i's)

$$r_i = Y_i - \sum_{j=1}^{p} X_{ij}\beta_j.$$

Assume that random errors ε_i's are *homoscedastic*, i.e., they are uncorrelated random variables with mean 0 and common variance σ^2. The *least-squares method* is to minimize the residual sum-of-squares (RSS):

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2. \tag{2.3}$$

with respect to $\boldsymbol{\beta}$. Since (2.3) is a nice quadratic function of $\boldsymbol{\beta}$, there is a closed-form solution. Denote by

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X}_j = \begin{pmatrix} X_{1j} \\ \vdots \\ X_{nj} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ldots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then (2.2) can be written in the matrix form

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

The matrix $\mathbf{X}$ is known as the *design matrix* and is of crucial importance to the whole theory of linear regression analysis. The RSS($\boldsymbol{\beta}$) can be written as

$$\text{RSS}(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Differentiating RSS($\boldsymbol{\beta}$) with respect to $\boldsymbol{\beta}$ and setting the gradient vector to zero, we obtain the *normal equations*

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}.$$

Here we assume that $p < n$ and $\mathbf{X}$ has rank p. Hence $\mathbf{X}^T\mathbf{X}$ is invertible and the normal equations yield the least-squares estimator of $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.4}$$

In this chapter $\mathbf{X}^T\mathbf{X}$ is assumed to be invertible unless specifically mentioned otherwise.

Copyrighted material

The fitted Y value is

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

and the regression residual is

$$\widehat{\mathbf{r}} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}.$$

Theorem 2.1 *Define* $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. *Then we have*

$$\mathbf{P}\mathbf{X}_j = \mathbf{X}_j, \quad j = 1, 2, \cdots, p;$$

$$\mathbf{P}^2 = \mathbf{P} \quad \text{or} \quad \mathbf{P}(\mathbf{I}_n - \mathbf{P}) = \mathbf{0},$$

namely $\mathbf{P}$ *is a projection matrix onto the space spanned by the columns of* $\mathbf{X}$.

Proof. It follows from the direct calculation that

$$\mathbf{P}\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X}.$$

Taking the j column of the above equality, we obtain the first results. Similarly,

$$\mathbf{P}\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P}.$$

This completes the proof. ∎

By Theorem 2.1 we can write

$$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}, \quad \widehat{\mathbf{r}} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} \tag{2.5}$$

and we see two simple identities:

$$\mathbf{P}\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}, \quad \widehat{\mathbf{Y}}^T\widehat{\mathbf{r}} = 0.$$

This reveals an interesting geometric interpretation of the method of least-squares: the least-squares fit amounts to projecting the response vector onto the linear space spanned by the covariates. See Figure 2.1 for an illustration with two covariates.

2.2.1 The Gauss-Markov theorem

We assume the linear regression model (2.1) with

- *exogeneity*: $\mathrm{E}(\varepsilon|X) = 0$;
- *homoscedasticity*: $\mathrm{Var}(\varepsilon|X) = \sigma^2$.

Theorem 2.2 *Under model (2.1) with exogenous and homoscedastic error, it follows that*

(i) (unbiasedness) $\mathrm{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$.

Copyrighted material

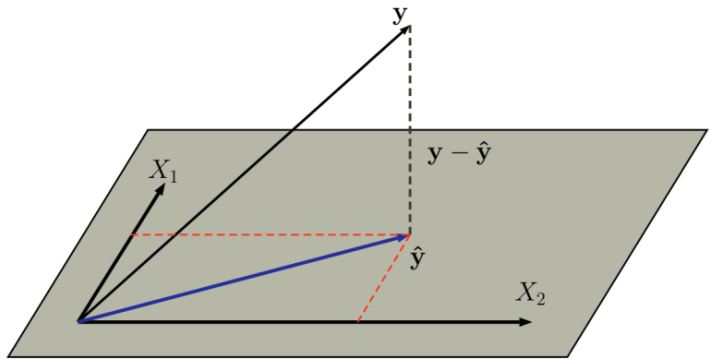

Figure 2.1: Geometric view of least-squares. The fitted value is the blue arrow, which is the projection of $\mathbf{Y}$ on the plane spanned by X_1 and X_2.

(ii) (conditional standard errors) $\mathrm{Var}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

(iii) (BLUE) *The least-squares estimator $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE). That is, for any given vector $\mathbf{a}$, $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$ is a linear unbiased estimator of the parameter $\theta = \mathbf{a}^T\boldsymbol{\beta}$. Further, for any linear unbiased estimator $\mathbf{b}^T\mathbf{Y}$ of θ, its variance is at least as large as that of $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$.*

Proof. The first property follows directly from $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}.$$

To prove the second property, note that for any linear combination $\mathbf{AY}$, its variance-covariance matrix is given by

$$\mathrm{Var}(\mathbf{AY}|\mathbf{X}) = \mathbf{A}\,\mathrm{Var}(\mathbf{Y}|\mathbf{X})\mathbf{A}^T = \sigma^2\mathbf{A}\mathbf{A}^T. \tag{2.6}$$

Applying this formula to the least-squares estimator with $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, we obtain the property (ii).

To prove property (iii), we first notice that $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of the parameter $\theta = \mathbf{a}^T\boldsymbol{\beta}$, with the variance

$$\mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{a}^T\,\mathrm{Var}(\widehat{\boldsymbol{\beta}}|\mathbf{X})\mathbf{a} = \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}.$$

Now, consider any linear unbiased estimator, $\mathbf{b}^T\mathbf{Y}$, of the parameter θ. The unbiasedness requires that

$$\mathbf{b}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta},$$

Copyrighted material

namely $\mathbf{X}^T\mathbf{b} = \mathbf{a}$. The variance of this linear estimator is

$$\sigma^2\mathbf{b}^T\mathbf{b}.$$

To prove (iii), we need only to show that

$$\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \leq \mathbf{b}^T\mathbf{b}.$$

Note that

$$\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \mathbf{b}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{b} = \mathbf{b}^T\mathbf{P}\mathbf{b}.$$

$\mathbf{P} = \mathbf{P}^2$ means that the eigenvalues of $\mathbf{P}$ are either 1 or 0 and hence $\mathbf{I}_n - \mathbf{P}$ is a semi-positive matrix. Thus,

$$\mathbf{b}^T(\mathbf{I}_n - \mathbf{P})\mathbf{b} \geq 0,$$

or equivalently $\mathbf{b}^T\mathbf{b} \geq \mathbf{b}^T\mathbf{P}\mathbf{b}$. ∎

Property (ii) of Theorem 2.2 gives the variance-covariance matrix of the least-squares estimate. In particular, the conditional standard error of $\widehat{\beta}_i$ is simply $\sigma a_{ii}^{1/2}$ and the covariance between $\widehat{\beta}_i$ and $\widehat{\beta}_j$ is $\sigma^2 a_{ij}$, where a_{ij} is the (i,j)-th element of matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.

In many applications σ^2 is often an unknown parameter of the model in addition to the regression coefficient vector $\boldsymbol{\beta}$. In order to use the variance-covariance formula, we first need to find a good estimate of σ^2. Given the least-squares estimate of $\boldsymbol{\beta}$, RSS can be written as

$$\text{RSS} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = (\mathbf{Y} - \widehat{\mathbf{Y}})^T(\mathbf{Y} - \widehat{\mathbf{Y}}). \tag{2.7}$$

Define

$$\widehat{\sigma}^2 = \text{RSS}\,/(n-p).$$

We will show in Theorem 2.3 that $\widehat{\sigma}^2$ is an unbiased estimator of σ^2.

Theorem 2.3 *Under the linear model (2.1) with homoscedastic error, it follows that*

$$\text{E}(\widehat{\sigma}^2|\mathbf{X}) = \sigma^2.$$

Proof. First by Theorem 2.1 we have

$$\text{RSS} = \|(\mathbf{I}_n - \mathbf{P})\mathbf{Y}\|^2 = \|(\mathbf{I}_n - \mathbf{P})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|^2 = \boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}.$$

Let $\text{tr}(\mathbf{A})$ be the trace of the matrix $\mathbf{A}$. Using the property that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, we have

$$\text{RSS} = \text{tr}\{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\}.$$

Hence,

$$\text{E}(\text{RSS}\,|\mathbf{X}) = \sigma^2\,\text{tr}(\mathbf{I}_n - \mathbf{P}).$$

Copyrighted material

Because the eigenvalues of $\mathbf{P}$ are either 1 or 0, its trace is equal to its rank which is p under the assumption that $\mathbf{X}^T\mathbf{X}$ is invertible. Thus,

$$\mathrm{E}(\widehat{\sigma}^2|\mathbf{X}) = \sigma^2(n-p)/(n-p) = \sigma^2.$$

This completes the proof. ∎

2.2.2 Statistical tests

After fitting the regression model, we often need to perform some tests on the model parameters. For example, we may be interested in testing whether a particular regression coefficient should be zero, or whether several regression coefficients should be zero at the same time, which is equivalent to asking whether these variables are important in the presence of other covariates. To facilitate the discussion, we focus on the fixed design case where $\mathbf{X}$ is fixed. This is essentially the same as the random design case but conditioned upon the given realization $\mathbf{X}$.

We assume a homoscedastic model (2.1) with normal error. That is, ε is a Gaussian random variable with zero mean and variance σ^2, written as $\varepsilon \sim N(0, \sigma^2)$. Note that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}. \tag{2.8}$$

Then it is easy to see that

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2). \tag{2.9}$$

If we look at each $\widehat{\beta}_j$ marginally, then $\widehat{\beta}_j \sim N(\beta_j, v_j\sigma^2)$ where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. In addition,

$$(n-p)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n-p} \tag{2.10}$$

and $\widehat{\sigma}^2$ is independent of $\widehat{\boldsymbol{\beta}}$. The latter can easily be shown as follow. By (2.7), $\widehat{\sigma}^2$ depends on $\mathbf{Y}$ through $\mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}$ whereas $\widehat{\boldsymbol{\beta}}$ depends on $\mathbf{Y}$ through (2.8) or $\mathbf{X}^T\boldsymbol{\varepsilon}$. Note that both $(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}$ and $\mathbf{X}^T\boldsymbol{\varepsilon}$ are jointly normal because they are linear transforms of normally distributed random variables, and therefore their independence is equivalent to their uncorrelatedness. This can easily be checked by computing their covariance

$$\mathrm{E}(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}(\mathbf{X}^T\boldsymbol{\varepsilon})^T = \mathrm{E}(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X} = \sigma^2(\mathbf{I}_n - \mathbf{P})\mathbf{X} = 0.$$

If we want to test the hypothesis that $\beta_j = 0$, we can use the following t test statistic

$$t_j = \frac{\widehat{\beta}_j}{\sqrt{v_j}\widehat{\sigma}} \tag{2.11}$$

which follows a t-distribution with $n - p$ degrees of freedom under the null

Copyrighted material

hypothesis $H_0 : \beta_j = 0$. A level α test rejects the null hypothesis if $|t_j| > t_{n-p,1-\alpha/2}$, where $t_{n-p,1-\alpha/2}$ denotes the $100(1-\alpha/2)$ percentile of the t-distribution with $n-p$ degrees of freedom.

In many applications the null hypothesis is that a subset of the covariates have zero regression coefficients. That is, this subset of covariates can be deleted from the regression model: they are unrelated to the response variable given the remaining variables. Under such a null hypothesis, we can reduce the model to a smaller model. Suppose that the reduced model has p_0 many regression coefficients. Let RSS and RSS_0 be the residual sum-of-squares based on the least-squares fit of the full model and the reduced smaller model, respectively. If the null hypothesis is true, then these two quantities should be similar: The RSS reduction by using the full model is small, in relative terms. This leads to the *F-statistic*:

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}\,/(n-p)}. \tag{2.12}$$

Under the null hypothesis that the reduced model is correct, $F \sim F_{p-p_0,n-p}$.

The normal error assumption can be relaxed if the sample size n is large. First, we know that $(\mathbf{X}^T\mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma$ always has zero mean and an identity variance-covariance matrix. On the other hand, (2.8) gives us

$$(\mathbf{X}^T\mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma = (\mathbf{X}^T\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^T\boldsymbol{\varepsilon}/\sigma.$$

Observe that $(\mathbf{X}^T\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^T\boldsymbol{\varepsilon}/\sigma$ is a linear combination of n i.i.d. random variables $\{\varepsilon_i\}_{i=1}^n$ with zero mean and variance 1. Then the central limit theorem implies that under some regularity conditions,

$$\widehat{\boldsymbol{\beta}} \xrightarrow{D} N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2). \tag{2.13}$$

Consequently, when n is large, the distribution of the t test statistic in (2.11) is approximately $N(0,1)$, and the distribution of the F test statistic in (2.12) is approximately $\chi^2_{p-p_0}/(p-p_0)$.

2.3 Weighted Least-Squares

The method of least-squares can be further generalized to handle the situations where errors are *heteroscedastic* or correlated. In the linear regression model (2.2), we would like to keep the assumption $\text{E}(\varepsilon|\mathbf{X}) = 0$ which means there is no structure information left in the error term. However, the constant variance assumption $\text{Var}(\varepsilon_i|\mathbf{X}_i) = \sigma^2$ may not likely hold in many applications. For example, if y_i is the average response value of the ith subject in a study in which k_i many repeated measurements have been taken, then it would be more reasonable to assume $\text{Var}(\varepsilon_i|\mathbf{X}_i) = \sigma^2/k_i$.

Let us consider a modification of model (2.1) as follows

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i; \quad \text{Var}(\varepsilon_i|\mathbf{X}_i) = \sigma^2 v_i \tag{2.14}$$

Copyrighted material

where v_is are known positive constants but σ^2 remains unknown. One can still use the ordinary least-squares (OLS) estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. It is easy to show that the OLS estimator is unbiased but no longer BLUE. In fact, the OLS estimator can be improved by using the *weighted least-squares* method.

Let $Y_i^* = v_i^{-1/2}Y_i$, $X_{ij}^* = v_i^{-1/2}X_{ij}$, $\varepsilon_i^* = v_i^{-1/2}\varepsilon_i$. Then the new model (2.14) can be written as

$$Y_i^* = \sum_{j=1}^{p} X_{ij}^*\beta_j + \varepsilon_i^* \tag{2.15}$$

with $\text{Var}(\varepsilon_i^*|\mathbf{X}_i^*) = \sigma^2$. Therefore, the working data $\{(X_{i1}^*, \cdots, X_{ip}^*, Y_i^*)\}_{i=1}^n$ obey the standard *homoscedastic* linear regression model. Applying the standard least-squares method to the working data, we have

$$\widehat{\boldsymbol{\beta}}^{wls} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n}\left(Y_i^* - \sum_{j=1}^{p} X_{ij}^*\beta_j\right)^2 = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} v_i^{-1}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2.$$

It follows easily from Theorem 2.2 that the weighted least-squares estimator is the BLUE for $\boldsymbol{\beta}$.

In model (2.14) the errors are assumed to be uncorrelated. In general, the method of least-squares can be extended to handle heteroscedastic and correlated errors.

Assume that

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

and the variance-covariance matrix of $\boldsymbol{\varepsilon}$ is given

$$\text{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{W}, \tag{2.16}$$

in which $\mathbf{W}$ is a known positive definite matrix. Let $\mathbf{W}^{-1/2}$ be the square root of $\mathbf{W}^{-1}$, i.e.,

$$(\mathbf{W}^{-1/2})^T\mathbf{W}^{-1/2} = \mathbf{W}^{-1}.$$

Then

$$\text{Var}(\mathbf{W}^{-1/2}\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I},$$

which are homoscedastic and uncorrelated.

Define the working data as follows:

$$\mathbf{Y}^* = \mathbf{W}^{-1/2}\mathbf{Y}, \quad \mathbf{X}^* = \mathbf{W}^{-1/2}\mathbf{X}, \quad \boldsymbol{\varepsilon}^* = \mathbf{W}^{-1/2}\boldsymbol{\varepsilon}.$$

Then we have

$$\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*. \tag{2.17}$$

Thus, we can apply the standard least-squares to the working data. First, the residual sum-of-squares (RSS) is

$$\text{RSS}(\boldsymbol{\beta}) = ||\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}||^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.18}$$

Copyrighted material

Then the *general least-squares* estimator is defined by

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \text{argmin}_{\boldsymbol{\beta}} \, \text{RSS}(\boldsymbol{\beta}) \\ &= (\mathbf{X}^{*T}\mathbf{X}^{*})^{-1}\mathbf{X}^{*T}\mathbf{Y}^{*} \\ &= (\mathbf{X}^{T}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{W}^{-1}\mathbf{Y}. \end{aligned} \tag{2.19}$$

Again, $\widehat{\boldsymbol{\beta}}$ is the BLUE according to Theorem 2.2.

In practice, it is difficult to know precisely the $n \times n$ covariance matrix $\mathbf{W}$; the misspecification of $\mathbf{W}$ in the general least-squares seems hard to avoid. Let us examine the robustness of the general least-squares estimate. Assume that $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W}_0$, where $\mathbf{W}_0$ is unknown to us, but we employ the general least-squares method (2.19) with the wrong covariance matrix $\mathbf{W}$. We can see that the general least-squares estimator is still unbiased:

$$\text{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}^{T}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{W}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Furthermore, the variance-covariance matrix is given by

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{T}\mathbf{W}^{-1}\mathbf{X})^{-1}(\mathbf{X}^{T}\mathbf{W}^{-1}\mathbf{W}_0\mathbf{W}^{-1}\mathbf{X})(\mathbf{X}^{T}\mathbf{W}^{-1}\mathbf{X})^{-1},$$

which is of order $O(n^{-1})$ under some mild conditions. In other words, using the wrong covariance matrix would still give us a root-n consistent estimate. So even when errors are heteroscedastic and correlated, the ordinary least-squares estimate with $\mathbf{W} = \mathbf{I}$ and the weighted least-squares estimate with $\mathbf{W} = \text{diag}(\mathbf{W}_0)$ still give us an unbiased and $n^{-1/2}$ consistent estimator. Of course, we still prefer using a working $\mathbf{W}$ matrix that is identical or close to the true $\mathbf{W}_0$.

2.4 Box-Cox Transformation

In practice we often take a transformation of the response variable before fitting a linear regression model. The idea is that the transformed response variable can be modeled by the set of covariates via the classical multiple linear regression model. For example, in many engineering problems we expect $Y \propto X_1^{\beta_1}X_2^{\beta_2} \cdot X_p^{\beta_p}$ where all variables are positive. Then a linear model seems proper by taking logarithms: $\log(Y) = \sum_{j=1}^{p} \beta_j X_j + \varepsilon$. If we assume $\varepsilon \sim N(0, \sigma^2)$, then in the original scale the model is $Y = (\prod_j^p X_j^{\beta_j})\varepsilon^*$ where ε^* is a log-normal random variable: $\log \varepsilon^* \sim N(0, \sigma^2)$.

Box and Cox (1964) advocated the variable transformation idea in linear regression and also proposed a systematic way to estimate the transformation function from data. Their method is now known as the *Box-Cox transform* in the literature. Box and Cox (1964) suggested a parametric family for the transformation function. Let $Y^{(\lambda)}$ denote the transformed response where λ parameterizes the transformation function:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(Y) & \text{if } \lambda = 0 \end{cases}.$$

Copyrighted material

The Box-Cox model assumes that

$$Y^{(\lambda)} = \sum_{j=1}^{p} X_j \beta_j + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$.

The likelihood function of the Box-Cox model is given by

$$L(\lambda, \boldsymbol{\beta}, \sigma^2) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2\sigma^2}\|\mathbf{Y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta}\|^2} \cdot J(\lambda, \mathbf{Y})$$

where $J(\lambda, \mathbf{Y}) = \prod_{i=1}^{n} |\frac{dy_i^{(\lambda)}}{dy_i}| = (\prod_{i=1}^{n} |y_i|)^{\lambda - 1}$. Given λ, the maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$ and σ^2 are obtained by the ordinary least-squares:

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^{(\lambda)}, \quad \widehat{\sigma}^2(\lambda) = \frac{1}{n}\|\mathbf{Y}^{(\lambda)} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^{(\lambda)}\|^2.$$

Plugging $\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}^2(\lambda)$ into $L(\lambda, \boldsymbol{\beta}, \sigma^2)$ yields a likelihood function of λ

$$\log L(\lambda) = (\lambda - 1)\sum_{i=1}^{n} \log(|y_i|) - \frac{n}{2}\log \widehat{\sigma}^2(\lambda) - \frac{n}{2}.$$

Then the MLE of λ is

$$\widehat{\lambda}_{mle} = \text{argmax}_\lambda \log L(\lambda),$$

and the MLE of $\boldsymbol{\beta}$ and σ^2 are $\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{mle})$ and $\widehat{\sigma}^2(\widehat{\lambda}_{mle})$, respectively.

2.5 Model Building and Basis Expansions

Multiple linear regression can be used to produce nonlinear regression and other very complicated models. The key idea is to create new covariates from the original ones by adopting some transformations. We then fit a multiple linear regression model using augmented covariates.

For simplicity, we first illustrate some useful transformations in the case of $p = 1$, which is closely related to the curve fitting problem in *nonparametric regression*. In a nonparametric regression model

$$Y = f(X) + \varepsilon,$$

we do not assume a specific form of the regression function $f(x)$, but assume only some qualitative aspects of the regression function. Examples include that $f(\cdot)$ is continuous with a certain number of derivatives or that $f(\cdot)$ is convex. The aim is to estimate the function $f(x)$ and its derivatives, without a specific parametric form of $f(\cdot)$. See, for example Fan and Gijbels (1996), Li and Racine (2007), Hastie, Tibshirani and Friedman (2009), among others.

Copyrighted material

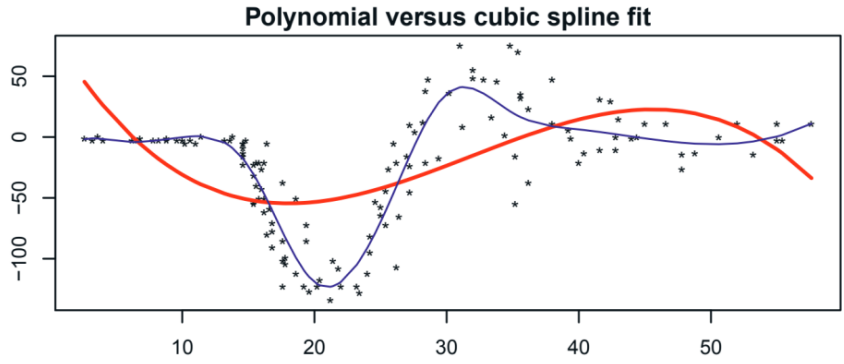

Figure 2.2: Scatter plot of time (in milliseconds) after a simulated impact on motorcycles against the head acceleration of a test object. Red = cubic polynomial fit, blue = cubic spline fit.

2.5.1 Polynomial regression

Without loss of generality, assume X is bounded on $[0, 1]$ for simplicity. The Weierstrass approximation theorem states that any continuous $f(x)$ can be uniformly approximated by a polynomial function up to any precision factor. Let us approximate the model by

$$Y = \underbrace{\beta_0 + \beta_1 X + \cdots + \beta_d X^d}_{\approx f(X)} + \varepsilon$$

This *polynomial regression* is a multiple regression problem by setting $X_0 = 1, X_1 = X, \cdots, X_d = X^d$. The design matrix now becomes

$$\mathbf{B}_1 = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & \cdots & x_n^d \end{pmatrix}.$$

We estimate $f(x)$ by

$$\widehat{f}(x) = \widehat{\beta}_0 + \sum_{m=1}^{d} \widehat{\beta}_m x^m,$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{B}_1^T \mathbf{B}_1)^{-1} \mathbf{B}_1^T \mathbf{Y}$ is the least-squares estimate.

Polynomial functions have derivatives everywhere and are global functions. They are not very flexible in approximating functions with local features such as functions with various degrees of smoothness at different locations. Figure 2.2 shows the cubic polynomial fit to a motorcycle data. Clearly, it does not fit the data very well. Increasing the order of the polynomial fits will

Copyrighted material

help reduce the bias issue, but will not solve the lack of fit issue. This is because the underlying function cannot be economically approximated by a polynomial function. It requires high-order polynomials to reduce approximation biases, but this increases both variances and instability of the fits. This leads to the introduction of spline functions that allow for more flexibility in function approximation.

2.5.2 Spline regression

Let $\tau_0 < \tau_1 < \cdots < \tau_{K+1}$. A *spline function* of degree d on $[\tau_0, \tau_{K+1}]$ is a piecewise polynomial function of degree d on intervals $[\tau_j, \tau_{j+1})$ $(j = 0, \cdots, K)$, with continuous first $d-1$ derivatives. The points where the spline function might not have continuous d^{th} derivatives are $\{\tau_j\}_{j=1}^{K}$, which are called *knots*. Thus, a cubic spline function is a piecewise polynomial function with continuous first two derivatives and the points where the third derivative might not exist are called knots of the cubic spline. An example of a cubic fit is given by Figure 2.2.

All spline functions of degree d form a linear space. Let us determine its basis functions.

Linear Splines: A continuous function on $[0, 1]$ can also be approximated by a piecewise constant or linear function. We wish to use a continuous function to approximate $f(x)$. Since a piecewise constant function is not continuous unless the function is a constant in the entire interval, we use a continuous piecewise linear function to fit $f(x)$. Suppose that we split the interval $[0, 1]$ into three regions: $[0, \tau_1], [\tau_1, \tau_2], [\tau_2, 1]$ with given knots τ_1, τ_2. Denote by $l(x)$ the continuous piecewise linear function. In the first interval $[0, \tau_1]$ we write

$$l(x) = \beta_0 + \beta_1 x, \ x \in [0, \tau_1],$$

as it is linear. Since $l(x)$ must be continuous at τ_1, the newly added linear function must have an intercept 0 at point τ_1. Thus, in $[\tau_1, \tau_2]$ we must have

$$l(x) = \beta_0 + \beta_1 x + \beta_2 (x - \tau_1)_+, \ x \in [\tau_1, \tau_2],$$

where z_+ equals z if $z > 0$ and zero otherwise. The function is linear in $[\tau_1, \tau_2]$ with slope $\beta_1 + \beta_2$. Likewise, in $[\tau_2, 1]$ we write

$$l(x) = \beta_0 + \beta_1 x + \beta_2 (x - \tau_1)_+ + \beta_3 (x - \tau_2)_+, \ x \in [\tau_2, 1].$$

The function is now clearly a piecewise linear function with possible different slopes on different intervals. Therefore, the basis functions are

$$B_0(x) = 1, B_1(x) = x, B_2(x) = (x - \tau_1)_+, B_3(x) = (x - \tau_2)_+; \qquad (2.20)$$

which are called a *linear spline* basis. We then approximate the nonparametric regression model as

$$Y = \underbrace{\beta_0 B_0(X) + \beta_1 B_1(X) + \beta_2 B_2(X) + \beta_3 B_3(X)}_{\approx f(X)} + \varepsilon.$$

Copyrighted material

This is again a multiple regression problem where we set $X_0 = B_0(X), X_1 = B_1(X), X_2 = B_2(X), X_3 = B_3(X)$. The corresponding design matrix becomes

$$\mathbf{B}_2 = \begin{pmatrix} 1 & x_1 & (x_1-\tau_1)_+ & (x_1-\tau_2)_+ \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n-\tau_1)_+ & (x_n-\tau_2)_+ \end{pmatrix},$$

and we estimate $f(x)$ by

$$\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2(x-\tau_1)_+ + \widehat{\beta}_3(x-\tau_2)_+,$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{B}_2^T\mathbf{B}_2)^{-1}\mathbf{B}_2^T\mathbf{Y}$. The above method applies more generally to a multiple knot setting for the data on any intervals.

Cubic Splines: We can further consider fitting piecewise polynomials whose derivatives are also continuous. A popular choice is the so-called cubic spline that is a piecewise cubic polynomial function with continuous first and second derivatives. Again, we consider two knots and three regions: $[0,\tau_1], [\tau_1,\tau_2], [\tau_2,1]$. Let $c(x)$ be a cubic spline. In $[0,\tau_1]$ we write

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,\ x \le \tau_1.$$

And $c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \delta(x)$ in $[\tau_1,\tau_2]$. By definition, $\delta(x)$ is a cubic function in $[\tau_1,\tau_2]$ and its first and second derivatives equal zero at $x = \tau_1$. Then we must have

$$\delta(x) = \beta_4(x-\tau_1)_+^3,\ x \in [\tau_1,\tau_2]$$

which means

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x-\tau_1)_+^3,\ x \in [\tau_1,\tau_2].$$

Likewise, in $[\tau_2,1]$ we must have

$$c(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x-\tau_1)_+^3 + \beta_5(x-\tau_2)_+^3,\ x > \tau_2.$$

Therefore, the basis functions are

$$B_0(x) = 1, B_1(x) = x, B_2(x) = x^2, B_3(x) = x^3$$
$$B_4(x) = (x-\tau_1)_+^3, B_5(x) = (x-\tau_2)_+^3.$$

The corresponding transformed design matrix becomes

$$\mathbf{B}_3 = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & (x_1-\tau_1)_+^3 & (x_1-\tau_2)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & (x_n-\tau_1)_+^3 & (x_n-\tau_2)_+^3 \end{pmatrix},$$

Copyrighted material

and we estimate $f(x)$ by

$$\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x^3 + \widehat{\beta}_4 (x - \tau_1)_+^3 + \widehat{\beta}_5 (x - \tau_2)_+^3,$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{B}_3^T \mathbf{B}_3)^{-1} \mathbf{B}_3^T \mathbf{Y}$ is the least-squares estimate of the coefficients.

In general, if there are K knots $\{\tau_1, \cdots, \tau_K\}$, then the *basis functions* of *cubic splines* are

$$B_0(x) = 1, B_1(x) = x, B_2(x) = x^2, B_3(x) = x^3$$
$$B_4(x) = (x - \tau_1)_+^3, \cdots, B_{K+3}(x) = (x - \tau_K)_+^3.$$

By approximating the nonparametric function $f(X)$ by the spline function with knots $\{\tau_j\}_{j=1}^K$, we have

$$Y = \underbrace{\beta_0 B_0(X) + \beta_1 B_1(X) + \cdots + \beta_{K+3} B_{K+3}(X)}_{\approx f(X)} + \varepsilon \tag{2.21}$$

This *spline regression* is again a multiple regression problem.

Natural Cubic Splines: Extrapolation is always a serious issue in regression. It is not wise to fit a cubic function to a region where the observations are scarce. If we must, extrapolation with a linear function is preferred. A *natural cubic spline* is a special cubic spline with additional constraints: the cubic spline must be linear beyond two end knots. Consider a natural cubic spline, $\mathrm{NC}(x)$, with knots at $\{\tau_1, \cdots, \tau_K\}$. By its cubic spline representation, we can write

$$\mathrm{NC}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^K \beta_{3+j} (x - \tau_j)_+^3.$$

First, $\mathrm{NC}(x)$ is linear for $x < \tau_1$, which implies that

$$\beta_2 = \beta_3 = 0.$$

Second, $\mathrm{NC}(x)$ is linear for $x > \tau_K$, which means that

$$\sum_{j=1}^K \beta_{3+j} = 0, \qquad \sum_{j=1}^K \tau_j \beta_{3+j} = 0,$$

corresponding to the coefficients for the cubic and quadratic term of the polynomial $\sum_{j=1}^K \beta_{3+j} (x - \tau_j)^3$ for $x > \tau_K$. We solve for β_{K+2}, β_{K+3} from the above equations and then write $\mathrm{NC}(x)$ as

$$\mathrm{NC}(x) \quad = \quad \sum_{j=0}^{K-1} \beta_j B_j(x),$$

Copyrighted material

where the *natural cubic spline* basis functions are given by

$$B_0(x) = 1, B_1(x) = x,$$
$$B_{j+1}(x) = \frac{(x-\tau_j)_+^3 - (x-\tau_K)_+^3}{\tau_j - \tau_K} - \frac{(x-\tau_{K-1})_+^3 - (x-\tau_K)_+^3}{\tau_{K-1} - \tau_K}$$
for $j = 1, \cdots, K-2$.

Again, by approximating the nonparametric function with the natural cubic spline, we have

$$Y = \sum_{j=0}^{K-1} \beta_j B_j(X) + \varepsilon. \tag{2.22}$$

which can be solved by using multiple regression techniques.

2.5.3 Multiple covariates

The concept of polynomial regression extends to multivariate covariates. The simplest example is the bivariate regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 X_2 + \beta_5 X_2^2 + \varepsilon.$$

The term $X_1 X_2$ is called the *interaction*, which quantifies how X_1 and X_2 work together to contribute to the response. Often, one introduces interactions without using the quadratic term, leading to a slightly simplified model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon.$$

More generally, the multivariate quadratic regression is of the form

$$Y = \sum_{j=1}^{p} \beta_j X_j + \sum_{j \le k} \beta_{jk} X_j X_k + \varepsilon \tag{2.23}$$

and the multivariate regression with main effects (the linear terms) and interactions is of the form

$$Y = \sum_{j=1}^{p} \beta_j X_j + \sum_{j < k} \beta_{jk} X_j X_k + \varepsilon. \tag{2.24}$$

This concept can also be extended to the multivariate spline case. The basis function can be the tensor of the univariate spline basis function for not only unstructured $f(\mathbf{x})$, but also other basis functions for structured $f(\mathbf{x})$. Unstructured nonparametric functions are not very useful: If each variable uses 100 basis functions, then there are 100^p basis functions in the tensor products, which is prohibitively large for say, $p = 10$. Such an issue is termed the "curse-of-dimensionality" in literature. See Hastie and Tibshirani (1990) and Fan and Gijbels (1996). On the other hand, for the structured multivariate

Copyrighted material

model, such as the following additive model (Stone, 1985, 1994; Hastie and Tibshirani, 1990),

$$Y = f_1(X_1) + \cdots + f_p(X_p) + \varepsilon \tag{2.25}$$

the basis functions are simply the collection of all univariate basis functions for approximating $f_1, \cdots, f_p$. The total number grows only linearly with p.

In general, let $B_m(\mathbf{x})$ be the basis functions $m = 1, \cdots, M$. Then, we approximate the multivariate nonparametric regression model $Y = f(\mathbf{X}) + \varepsilon$ by

$$Y = \sum_{m=1}^{M} \beta_j B_j(\mathbf{X}) + \varepsilon. \tag{2.26}$$

This can be fit using a multiple regression technique. The new design matrix is

$$\mathbf{B} = \begin{pmatrix} B_1(\mathbf{X}_1) & \cdots & B_M(\mathbf{X}_1) \\ \vdots & \cdots & \vdots \\ B_1(\mathbf{X}_n) & \cdots & B_M(\mathbf{X}_n) \end{pmatrix}$$

and the least-squares estimate is given by

$$\widehat{f}(\mathbf{x}) = \sum_{m=1}^{M} \widehat{\beta}_m B_m(\mathbf{x}),$$

where

$$\widehat{\boldsymbol{\beta}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y}.$$

The above fitting implicitly assumes that $M \ll n$. This condition in fact can easily be violated in unstructured multivariate nonparametric regression. For the additive model (2.25), in which we assume $f(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j)$ where each $f_j(x_j)$ is a smooth univariate function of x_j, the univariate basis expansion ideas can be readily applied to approximation of each $f_j(x_j)$:

$$f_j(x_j) \approx \sum_{m=1}^{M_j} B_{jm}(x_j)\beta_{jm}$$

which implies that the fitted regression function is

$$f(\mathbf{x}) \approx \sum_{j=1}^{p} \sum_{m=1}^{M_j} B_{jm}(x_j)\beta_{jm}.$$

In Section 2.6.5 and Section 2.7 we introduce a fully nonparametric multiple regression technique which can be regarded as a basis expansion method where the basis functions are given by kernel functions.

Copyrighted material

2.6 Ridge Regression

2.6.1 Bias-variance tradeoff

Recall that the ordinary least squares estimate is defined by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ when $\mathbf{X}$ is of full rank. In practice, we often encounter highly correlated covariates, which is known as the *collinearity* issue. As a result, although $\mathbf{X}^T\mathbf{X}$ is still invertible, its smallest eigenvalue can be very small. Under the homoscedastic error model, the variance-covariance matrix of the OLS estimate is $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. Thus, the collinearity issue makes $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ large.

Hoerl and Kennard (1970) introduced the *ridge regression* estimator as follows:

$$\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2.27}$$

where $\lambda > 0$ is a regularization parameter. In the usual case ($\mathbf{X}^T\mathbf{X}$ is invertible), ridge regression reduces to OLS by setting $\lambda = 0$. However, ridge regression is always well defined even when $\mathbf{X}$ is not full rank.

Under the assumption $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, it is easy to show that

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\sigma^2. \tag{2.28}$$

We always have $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda) < (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. The ridge regression estimator reduces the estimation variance by paying a price in estimation bias:

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_\lambda) - \boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta} = -\lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}. \tag{2.29}$$

The overall estimation accuracy is gauged by the mean squared error (MSE). For $\widehat{\boldsymbol{\beta}}_\lambda$ its MSE is given by

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_\lambda) = \mathrm{E}(\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|^2). \tag{2.30}$$

By (2.28) and (2.29) we have

$$\begin{aligned}
\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_\lambda) &= \mathrm{tr}\left((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\sigma^2\right) \\
&\quad + \lambda^2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-2}\boldsymbol{\beta} \\
&= \mathrm{tr}\left((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-2}[\lambda^2\boldsymbol{\beta}\boldsymbol{\beta}^T + \sigma^2\mathbf{X}^T\mathbf{X}]\right).
\end{aligned} \tag{2.31}$$

It can be shown that $\frac{d\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_\lambda)}{d\lambda}|_{\lambda=0} < 0$, which implies that there are some proper λ values by which ridge regression improves OLS.

Copyrighted material

2.6.2 ℓ_2 penalized least squares

Define a penalized residual sum-of-squares (PRSS) as follows:

$$\text{PRSS}(\boldsymbol{\beta}|\lambda) = \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{2.32}$$

Then let

$$\widehat{\boldsymbol{\beta}}_\lambda = \text{argmin}_{\boldsymbol{\beta}}\, \text{PRSS}(\boldsymbol{\beta}|\lambda). \tag{2.33}$$

Note that we can write it in a matrix form

$$\text{PRSS}(\boldsymbol{\beta}|\lambda) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2.$$

The term $\lambda\|\boldsymbol{\beta}\|^2$ is called the ℓ_2-penalty of $\boldsymbol{\beta}$. Taking derivatives with respect to $\boldsymbol{\beta}$ and setting it to zero, we solve the root of the following equation

$$-\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta} = 0,$$

which yields

$$\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The above discussion shows that ridge regression is equivalent to the ℓ_2 penalized least-squares.

We have seen that ridge regression can achieve a smaller MSE than OLS. In other words, the ℓ_2 penalty term helps regularize (reduce) estimation variance and produces a better estimator when the reduction in variance exceeds the induced extra bias. From this perspective, one can also consider a more general ℓ_q penalized least-squares estimate

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \tag{2.34}$$

where q is a positive constant. This is referred to as the Bridge estimator (Frank and Friedman, 1993). The ℓ_q penalty is strictly concave when $0 < q < 1$, and strictly convex when $q > 1$. For $q = 1$, the resulting ℓ_1 penalized least-squares is also known as the Lasso (Tibshirani, 1996). Chapter 3 covers the Lasso in great detail. Among all Bridge estimators only the ridge regression has a nice closed-form solution with a general design matrix.

2.6.3 Bayesian interpretation

Ridge regression has a neat Bayesian interpretation in the sense that it can be a formal Bayes estimator. We begin with the homoscedastic Gaussian error model:

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i$$

Copyrighted material

and $\varepsilon_i|\mathbf{X}_i \sim N(0,\sigma^2)$. Now suppose that β_j's are also independent $N(0,\tau^2)$ variables, which represent our knowledge about the regression coefficients before seeing the data. In Bayesian statistics, $N(0,\tau^2)$ is called the prior distribution of β_j. The model and the prior together give us the posterior distribution of $\boldsymbol{\beta}$ given the data (the conditional distribution of $\boldsymbol{\beta}$ given $\mathbf{Y},\mathbf{X}$). Straightforward calculations yield

$$P(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) \propto \exp(-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2)\exp(-\frac{1}{2\tau^2}\|\boldsymbol{\beta}\|^2). \tag{2.35}$$

A maximum posteriori probability (MAP) estimate is defined as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{MAP}} &= \text{argmax}_{\boldsymbol{\beta}} P(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) \\ &= \text{argmax}_{\boldsymbol{\beta}} \left\{ -\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 - \frac{1}{2\tau^2}\|\boldsymbol{\beta}\|^2 \right\}. \end{aligned} \tag{2.36}$$

It is easy to see that $\widehat{\boldsymbol{\beta}}^{\text{MAP}}$ is ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$. Another popular Bayesian estimate is the posterior mean. In this model, the posterior mean and posterior mode are the same.

From the Bayesian perspective, it is easy to construct a generalized ridge regression estimator. Suppose that the prior distribution for the entire $\boldsymbol{\beta}$ vector is $N(0,\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a general positive definite matrix. Then the posterior distribution is computed as

$$P(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) \propto \exp(-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2)\exp(-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}). \tag{2.37}$$

The corresponding MAP estimate is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{MAP}} &= \text{argmax}_{\boldsymbol{\beta}} P(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) \\ &= \text{argmax}_{\boldsymbol{\beta}} \left\{ -\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 - \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} \right\}. \end{aligned} \tag{2.38}$$

It is easy to see that

$$\widehat{\boldsymbol{\beta}}^{\text{MAP}} = (\mathbf{X}^T\mathbf{X}+\sigma^2\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.39}$$

This generalized ridge regression can take into account different scales of covariates, by an appropriate choice of $\boldsymbol{\Sigma}$.

2.6.4 Ridge regression solution path

The performance of ridge regression heavily depends on the choice of λ. In practice we only need to compute ridge regression estimates at a fine grid of λ values and then select the best from these candidate solutions. Although ridge regression is easy to compute for a λ owing to its nice closed-form solution expression, the total cost could be high if the process is repeated many times.

Copyrighted material

Through a more careful analysis, one can see that the solutions of ridge regression at a fine grid of λ values can be computed very efficiently via singular value decomposition.

Assume $n > p$ and $\mathbf{X}$ is full rank. The singular value decomposition (SVD) of $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where $\mathbf{U}$ is a $n \times p$ orthogonal matrix, $\mathbf{V}$ is a $p \times p$ orthogonal matrix and $\mathbf{D}$ is a $p \times p$ diagonal matrix whose diagonal elements are the ordered (from large to small) singular values of $\mathbf{X}$. Then

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T,$$

$$\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{I} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})\mathbf{V}^T,$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^T.$$

The ridge regression estimator $\widehat{\boldsymbol{\beta}}_\lambda$ can now be written as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_\lambda &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{Y} \\ &= \sum_{j=1}^{p} \frac{d_j}{d_j^2 + \lambda}\langle \mathbf{U}_j, \mathbf{Y}\rangle \mathbf{V}_j, \end{aligned} \tag{2.40}$$

where d_j is the j^{th} diagonal element of $\mathbf{D}$ and $\langle \mathbf{U}_j, \mathbf{Y}\rangle$ is the inner product between $\mathbf{U}_j$ and $\mathbf{Y}$ and $\mathbf{U}_j$ ($\mathbf{V}_j$ are respectively the j^{th} column of $\mathbf{U}$ and $\mathbf{V}$). In particular, when $\lambda = 0$, ridge regression reduces to OLS and we have

$$\widehat{\boldsymbol{\beta}}^{\text{OLS}} = \mathbf{V}D^{-1}\mathbf{U}^T\mathbf{Y} = \sum_{j=1}^{p} \frac{1}{d_j}\langle \mathbf{U}_j, \mathbf{Y}\rangle \mathbf{V}_j. \tag{2.41}$$

Based on (2.40) we suggest the following procedure to compute ridge regression at a fine grid $\lambda_1, \cdots, \lambda_M$:

1. Compute the SVD of $\mathbf{X}$ and save $\mathbf{U}, \mathbf{D}, \mathbf{V}$.
2. Compute $\mathbf{w}_j = \frac{1}{d_j}\langle \mathbf{U}_j \cdot \mathbf{Y}\rangle \mathbf{V}_j$ for $j = 1, \cdots, p$ and save $\mathbf{w}_j$s.
3. For $m = 1, 2 \cdots, M$,

   (i). compute $\gamma_j = \frac{d_j^2}{d_j^2 + \lambda_m}$

   (ii). compute $\widehat{\boldsymbol{\beta}}_{\lambda_m} = \sum_{j=1}^{p} \gamma_j \mathbf{w}_j$.

The essence of the above algorithm is to compute the common vectors $\{\mathbf{w}_j\}_{j=1}^p$ first and then utilize (2.40).

Copyrighted material

2.6.5 Kernel ridge regression

In this section we introduce a nonparametric generalization of ridge regression. Our discussion begins with the following theorem.

Theorem 2.4 *Ridge regression estimator is equal to*

$$\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{Y} \tag{2.42}$$

and the fitted value of Y at $\mathbf{x}$ is

$$\widehat{y} = \mathbf{x}^T\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{x}^T\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{Y} \tag{2.43}$$

Proof. Observe the following identity

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{X}^T = \mathbf{X}^T\mathbf{X}\mathbf{X}^T + \lambda\mathbf{X}^T = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}).$$

Thus, we have

$$\mathbf{X}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})$$

and

$$\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1} \quad = \quad (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T.$$

Then by using (2.27) we obtain (2.42) and hence (2.43). ∎

It is important to see that $\mathbf{X}\mathbf{X}^T$ and not $\mathbf{X}^T\mathbf{X}$ appears in the expression for $\widehat{\boldsymbol{\beta}}_\lambda$. Note that $\mathbf{X}\mathbf{X}^T$ is a $n \times n$ matrix and its ij elements are $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Similarly, $\mathbf{x}^T\mathbf{X}^T$ is an n-dimensional vector with the ith element being $\langle \mathbf{x}, \mathbf{x}_i \rangle$ $i = 1, \cdots, n$. Therefore, the prediction by ridge regression boils down to computing the inner product between p-dimensional covariate vectors. This is the foundation of the so-called "*kernel trick*".

Suppose that we use another "inner product" to replace the usual inner product in Theorem 2.4; then we may end up with a new ridge regression estimator. To be more specific, let us replace $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ with $K(\mathbf{x}_i, \mathbf{x}_j)$ where $K(\cdot, \cdot)$ is a known function:

$$\mathbf{x}^T\mathbf{X}^T \to (K(\mathbf{x}, \mathbf{X}_1), \cdots, K(\mathbf{x}, \mathbf{X}_n)),$$

$$\mathbf{X}\mathbf{X}^T \to \mathbf{K} = (K(\mathbf{X}_i, \mathbf{X}_j))_{1 \le i,j \le n}.$$

By doing so, we turn (2.43) into

$$\widehat{y} = (K(\mathbf{x}, \mathbf{X}_1), \cdots, K(\mathbf{x}, \mathbf{X}_n))(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{Y} = \sum_{i=1}^{n} \widehat{\alpha}_i K(\mathbf{x}, \mathbf{X}_i), \tag{2.44}$$

where $\widehat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{Y}$. In particular, the fitted $\mathbf{Y}$ vector is

$$\widehat{\mathbf{Y}} = \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{Y}. \tag{2.45}$$

The above formula gives the so-called kernel ridge regression. Because $\mathbf{X}\mathbf{X}^T$ is at least positive semi-definite, it is required that $\mathbf{K}$ is also positive semi-definite. Some widely used kernel functions (Hastie, Tibshirani and Friedman, 2009) include

Copyrighted material

- *linear kernel*: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$,
- *polynomial kernel*: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^d$, $d = 2, 3, \cdots$,
- *radial basis kernel*: $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$, $\gamma > 0$, which is the *Gaussian kernel*, and $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|}$, $\gamma > 0$, which is the *Laplacian kernel*.

To show how we get (2.45) more formally, let us consider to approximate the multivariate regression by using the kernel basis functions $\{K(\cdot, \mathbf{x}_j)\}_{j=1}^n$ so that our observed data are now modeled as

$$Y_i = \sum_{j=1}^{n} \alpha_j K(\mathbf{X}_i, \mathbf{X}_j) + \varepsilon_i$$

or in matrix form $\mathbf{Y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$. If we apply the ridge regression

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{K}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}\mathbf{K}\boldsymbol{\alpha},$$

the minimizer of the above problem is

$$\widehat{\boldsymbol{\alpha}} = (\mathbf{K}^T\mathbf{K} + \lambda\mathbf{K})^{-1}\mathbf{K}^T\mathbf{Y} = \{\mathbf{K}(\mathbf{K} + \lambda\mathbf{I})\}^{-1}\mathbf{K}\mathbf{Y},$$

where we use the fact that $\mathbf{K}$ is symmetric. Assuming $\mathbf{K}$ is invertible, we easily get (2.45).

So far we have only derived the kernel ridge regression based on heuristics and the kernel trick. In Sec. 2.7 we show that the kernel ridge regression can be formally derived based on the theory of function estimation in a reproducing kernel Hilbert space.

2.7 Regression in Reproducing Kernel Hilbert Space

A *Hilbert space* is an abstract vector space endowed by the structure of an inner product. Let $\mathcal{X}$ be an arbitrary set and $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$, endowed by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The evaluation functional over the Hilbert space of functions $\mathcal{H}$ is a linear functional that evaluates each function at a point x:

$$L_x : f \to f(x), \forall f \in \mathcal{H}.$$

A Hilbert space $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS) if, for all $x \in \mathcal{X}$, the map L_x is continuous at any $f \in \mathcal{H}$, namely, there exists some $C > 0$ such that

$$|L_x(f)| = |f(x)| \leq C\|f\|_{\mathcal{H}}, \qquad \forall f \in \mathcal{H}.$$

By the Riesz representation theorem, for all $x \in \mathcal{X}$, there exists a unique element $K_x \in \mathcal{H}$ with the reproducing property

$$f(x) = L_x(f) = \langle f, K_x \rangle_{\mathcal{H}}, \qquad \forall f \in \mathcal{H}.$$

Copyrighted material


image
not
available

This proves part (i). Now we apply part (i) to get part (ii) by letting $g(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_2)$.

For part (iii) we observe that

$$\begin{aligned} \|g\|^2_{\mathcal{H}_K} &= \langle \sum_{i=1}^n \alpha_i K(\mathbf{x}, \mathbf{x}_i), \sum_{j=1}^n \alpha_j K(\mathbf{x}, \mathbf{x}_j) \rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle K(\mathbf{x}, \mathbf{x}_i), K(\mathbf{x}, \mathbf{x}_j) \rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \end{aligned}$$

where we have used part (ii) in the final step.

■

Consider a general regression model

$$Y = f(\mathbf{X}) + \varepsilon \tag{2.49}$$

where ε is independent of $\mathbf{X}$ and has zero mean and variance σ^2. Given a realization $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the above model, we wish to fit the regression function in $\mathcal{H}_K$ via the following penalized least-squares:

$$\widehat{f} = \text{argmin}_{f \in \mathcal{H}_K} \sum_{i=1}^n [Y_i - f(\mathbf{X}_i)]^2 + \lambda \|f\|^2_{\mathcal{H}_K}, \quad \lambda > 0. \tag{2.50}$$

Note that without the $\|f\|^2_{\mathcal{H}_K}$ term there are infinitely many functions in $\mathcal{H}_K$ that can fit the observations perfectly, i.e., $Y_i = f(\mathbf{X}_i)$ for $i = 1, \cdots, n$. By using the eigen-function expansion of f

$$f(\mathbf{x}) = \sum_{j=1}^\infty \beta_j \psi_j(\mathbf{x}), \tag{2.51}$$

an equivalent formulation of (2.50) is

$$\min_{\{\beta_j\}_{j=1}^\infty} \sum_{i=1}^n [Y_i - \sum_{j=1}^\infty \beta_j \psi_j(\mathbf{X}_i)]^2 + \lambda \sum_{j=1}^\infty \frac{1}{\gamma_j} \beta_j^2. \tag{2.52}$$

Define $\beta_j^* = \frac{\beta_j}{\sqrt{\gamma_j}}$ and $\psi_j^* = \sqrt{\gamma_j} \psi_j$ for $j = 1, 2, \cdots$. Then (2.52) can be rewritten as

$$\min_{\{\beta_j^*\}_{j=1}^\infty} \sum_{i=1}^n [Y_i - \sum_{j=1}^\infty \beta_j^* \psi_j^*(\mathbf{X}_i)]^2 + \lambda \sum_{j=1}^\infty (\beta_j^*)^2. \tag{2.53}$$

The above can be seen as a ridge regression estimate in an infinite dimensional

Copyrighted material


image
not
available


image
not
available


image
not
available

Table 2.2: A list of commonly used regression methods and their $\mathbf{S}$ matrices. d_js are the singular values of $\mathbf{X}$ and γ_is are the eigenvalues of $\mathbf{K}$.

Method	$\mathbf{S}$	$\operatorname{tr}\mathbf{S}$
Multiple Linear Regression	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$	p
Ridge Regression	$\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T$	$\sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda}$
Kernel Regression in RKHS	$\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}$	$\sum_{i=1}^{n}\frac{\gamma_i}{\gamma_i+\lambda}$

fitting process n times to compute the leave-one-out CV. Fortunately, we can avoid much computation for many popular regression methods.

A fitting method is called a *linear smoother* if we can write

$$\widehat{\mathbf{Y}} = \mathbf{S}\mathbf{Y} \tag{2.63}$$

for any dataset $\{(\mathbf{X}_i, Y_i\}_1^n$ where $\mathbf{S}$ is a $n \times n$ matrix that only depends on $\mathbf{X}$. Many regression methods are linear smoothers with different $\mathbf{S}$ matrices. See Table 2.2.

Assume that a linear smoother is fitted on $\{\mathbf{X}_i, Y_i\}_{i=1}^n$. Let $\mathbf{x}$ be a new covariate vector and $\widehat{f}(\mathbf{x})$ be its the predicted value by using the linear smoother. We then augment the dataset by including $(\mathbf{x}, \widehat{f}(\mathbf{x}))$ and refit the linear smoother on this augmented dataset. The linear smoother is said to be *self-stable* if the fit based on the augmented dataset is identical to the fit based on the original data regardless of $\mathbf{x}$.

It is easy to check that the three linear smoothers in Table 2.2 all have the self-stable property.

Theorem 2.7 *For a linear smoother* $\widehat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$ *with the self-stable property, we have*

$$Y_i - \widehat{f}^{(-i)}(\mathbf{X}_i) = \frac{Y_i - \widehat{Y}_i}{1 - S_{ii}}, \tag{2.64}$$

and its leave-one-out CV error is equal to $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i-\widehat{Y}_i}{1-S_{ii}}\right)^2$.

Proof. We first apply the linear smoother to all the data except the ith to compute $\widehat{f}^{(-i)}(\mathbf{X}_i)$. Write $\widetilde{y}_j = y_j$ for $j \neq i$ and $\widetilde{y}_i = \widehat{f}^{(-i)}(\mathbf{X}_i)$. Then we apply the linear smoother to the following working dataset:

$$\{(\mathbf{X}_j, Y_j), j \neq i, (\mathbf{X}_i, \widetilde{Y}_i)\}$$

The self-stable property implies that the fit stays the same. In particular,

$$\widetilde{Y}_i = \widehat{f}^{(-i)}(\mathbf{X}_i) = (\mathbf{S}\widetilde{\mathbf{Y}})_i = S_{ii}\widetilde{Y}_i + \sum_{j\neq i} S_{ij}Y_j \tag{2.65}$$

Copyrighted material

(c) If $\mathbf{\Sigma}$ is the equi-correlation matrix with unknown correlation ρ, what is the solution to part (a)?

2.5 Suppose that $Y_1, \cdots, Y_n$ are random variables with common mean μ and covariance matrix $\sigma^2\mathbf{V}$, where $\mathbf{V}$ is of the form $v_{ii} = 1$ and $v_{ij} = \rho$ $(0 < \rho < 1)$ for $i \neq j$.

(a) Find the generalized least squares estimate of μ.

(b) Show that it is the same as the ordinary least squares estimate.

2.6 Suppose that data $\{X_{i1}, \cdots, X_{ip}, Y_i\}, i = 1, \cdots, n$, are an independent and identically distributed sample from the model

$$Y = f(X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon),$$

where $\varepsilon \sim N(0, \sigma^2)$ with unknown σ^2, and $f(\cdot)$ is a known, differentiable, strictly increasing, non-linear function.

(a) Consider transform $Y_i^* = h(Y_i)$, where $h(\cdot)$ is a differentiable function yet to be determined. Show that $\mathrm{Var}(Y_i^*)$ =constant for all i leads to the equation: $[h'\{f(u)\}]^2\{f'(u)\}^2$ =constant for all u.

(b) Let $f(x) = x^p$ $(p > 1)$. Find the corresponding $h(\cdot)$ using the equation in (a).

(c) Let $f(x) = \exp(x)$. Find the corresponding h transform.

2.7 The data set 'hkepd.txt' consists of daily measurements of levels of air pollutants and the number of total hospital admissions for circulatory and respiratory problems from January 1, 1994 to December 31, 1995 in Hong Kong. This data set can be downloaded from this book website. Of interest is to investigate the association between the number of total hospital admissions and the levels of air pollutants.

We set the Y variable to be the number of total hospital admissions and the X variables the levels of air pollutants. Define

X_1 = the level of sulfur dioxide $(\mu g/m^3)$;

X_2 = the level of nitrogen dioxide $(\mu g/m^3)$;

X_3 = the level of dust $(\mu g/m^3)$.

(a) Fit the data to the following linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \tag{2.67}$$

and test whether the level of each air pollutant has significant impact on the number of total hospital admissions.

(b) Construct residual plots and examine whether the random error approximately follows a normal distribution.

(c) Take $Z = \log(Y)$ and fit the data to the following linear regression model

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \tag{2.68}$$

Copyrighted material

Author Index

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material

Copyrighted material