Statistics For High Dimensional Data Methods Theory And

"Geometric Structure of High-Dimensional Data and Dimensionality Reduction" adopts data geometry as a framework to address various methods of dimensionality reduction. In addition to the introduction to well-known linear methods, the book moreover stresses the recently developed nonlinear methods and introduces the applications of dimensionality reduction in many areas, such as face recognition, image segmentation, data classification, data visualization, and hyperspectral imagery data analysis. Numerous tables and graphs are included to illustrate the ideas, effects, and shortcomings of the methods. MATLAB code of all dimensionality reduction algorithms is provided to aid the readers with the implementations on computers. The book will be useful for mathematicians, statisticians, computer scientists, and data analysts. It is also a valuable handbook for other practitioners who have a basic background in mathematics, statistics and/or computer algorithms, like internet search engine designers, physicists, geologists, electronic engineers, and economists. Jianzhong Wang is a Professor of Mathematics at Sam Houston State University, U.S.A.

Praise for the first edition: "[This book] succeeds singularly at providing a structured introduction to this active field of research. ... it is arguably the most accessible overview yet published of the mathematical ideas and principles that one needs to master to enter the field of high-dimensional statistics. ... recommended to anyone interested in the main results of current research in high-dimensional statistics as well as anyone interested in acquiring the core mathematical skills to enter this area of research." —Journal of the American Statistical Association Introduction to High-Dimensional Statistics, Second Edition preserves the philosophy of the first edition: to be a concise guide for students and researchers discovering the area and interested in the mathematics involved. The main concepts and ideas are presented in simple settings, avoiding thereby unessential technicalities. High-dimensional statistics is a fast-evolving field, and much progress has been made on a large variety of topics, providing new insights and methods. Offering a succinct presentation of the mathematical foundations of high-dimensional statistics, this new edition: Offers revised chapters from the previous edition, with the inclusion of many additional materials on some important topics, including compress sensing, estimation with convex constraints, the slope estimator, simultaneously low-rank and row-sparse linear regression, or aggregation of a continuous set of estimators. Introduces three new chapters on iterative algorithms, clustering, and minimax lower bounds. Provides enhanced appendices, minimax lower-bounds mainly with the addition of the Davis-Kahan perturbation bound and of two simple versions of the Hanson-Wright concentration inequality. Covers cutting-edge statistical methods including model selection, sparsity and the Lasso, iterative hard thresholding, aggregation, support vector machines, and learning theory. Provides detailed exercises at the end of every chapter with collaborative solutions on a wiki site. Illustrates concepts with simple but clear practical examples.

It was undoubtedly a necessary task to collect all the results on the concentration of measure during the past years in a monograph. The author did this very successfully and the book is an important contribution to the topic. It will surely influence further research in this area considerably. The book is very well written, and it was a great pleasure for the reviewer to read it. --Mathematical Reviews The observation of the concentration of measure phenomenon is inspired by isoperimetric inequalities. A familiar example is the way the uniform measure on the standard sphere S^n becomes concentrated around the equator as the dimension gets large. This property may be interpreted in terms of functions on the sphere with small oscillations, an idea going back to Levy. The phenomenon also occurs in probability, as a version of the law of large numbers, due to Emile Borel. This book offers the basic techniques and examples of the concentration of measure phenomenon. The concentration of measure phenomenon was put forward in the early seventies by V. Milman in the asymptotic geometry of Banach spaces. It is of powerful interest in applications in various areas, such as geometry, functional analysis and infinite-dimensional integration, discrete mathematics and complexity theory, and probability theory. Particular emphasis is on geometric, functional, and probabilistic tools to reach and describe measure concentration in a number of settings. The book presents concentration functions and inequalities, isoperimetric and functional examples, spectrum and topological applications, product measures, entropic and transportation methods, as well as aspects of M. Talagrand's deep investigation of concentration in product spaces and its application in discrete mathematics and probability theory, supremum of Gaussian and empirical processes, spin glass, random matrices, etc. Prerequisites are a basic background in measure theory, functional analysis, and probability theory.

Accessible to a variety of readers, this book is of interest to specialists, graduate students and researchers in mathematics, optimization, computer science, operations research, management science, engineering and other applied areas interested in solving optimization problems. Basic principles, potential and boundaries of applicability of stochastic global optimization techniques are examined in this book. A variety of issues that face specialists in global optimization are explored, such as multidimensional spaces which are frequently ignored by researchers. The importance of precise interpretation of the mathematical results in assessments of optimization methods is demonstrated through examples of convergence in probability of random search. Methodological issues concerning construction and applicability of stochastic global optimization methods are discussed, including the one-step optimal average improvement method based on a

statistical model of the objective function. A significant portion of this book is devoted to an analysis of high-dimensional global optimization problems and the so-called 'curse of dimensionality'. An examination of the three different classes of high-dimensional optimization problems, the geometry of high-dimensional balls and cubes, very slow convergence of global random search algorithms in large-dimensional problems , and poor uniformity of the uniformly distributed sequences of points are included in this book.

Big and Complex Data Analysis
High-dimensional Data Analysis
An Introduction with Applications in Data Science
The Grammar of Graphics
Principles, Computation, and Applications
With Applications in R

An integrated package of powerful probabilistic tools and key applications in modern mathematical data science.

This volume conveys some of the surprises, puzzles and success stories in high-dimensional and complex data analysis and related fields. Its peer-reviewed contributions showcase recent advances in variable selection, estimation and prediction strategies for a host of useful models, as well as essential new developments in the field. The continued and rapid advancement of modern technology now allows scientists to collect data of increasingly unprecedented size and complexity. Examples include epigenomic data, genomic data, proteomic data, high-resolution image data, high-frequency financial data, functional and longitudinal data, and network data. Simultaneous variable selection and estimation is one of the key statistical problems involved in analyzing such big and complex data. The purpose of this book is to stimulate research and foster interaction between researchers in the area of high-dimensional data analysis. More concretely, its goals are to: 1) highlight and expand the breadth of existing methods in big data and high-dimensional data analysis and their potential for the advancement of both the mathematical and statistical sciences; 2) identify important directions for future research in the theory of regularization methods, in algorithmic development, and in methodologies for different application areas; and 3) facilitate collaboration between theoretical and subject-specific researchers.

'Big data' poses challenges that require both classical multivariate methods and contemporary techniques from machine learning and engineering. This modern text equips you for the new world - integrating the old and the new, fusing theory and practice and bridging the gap to statistical learning. The theoretical framework includes formal statements that set out clearly the guaranteed 'safe operating zone' for the methods and allow you to assess whether data is in the zone, or near enough. Extensive examples showcase the strengths and limitations of different methods with small classical data, data from medicine, biology, marketing and finance, high-dimensional data from bioinformatics, functional data from proteomics, and simulated data. High-dimension low-sample-size data gets special attention. Several data sets are revisited repeatedly to allow comparison of methods. Generous use of colour, algorithms, Matlab code, and problem sets complete the package. Suitable for master's/graduate students in statistics and researchers in data-rich disciplines.

A comprehensive examination of high-dimensional analysis of multivariate methods and their real-world applications Multivariate Statistics: High-Dimensional and Large-Sample Approximations is the first book of its kind to explore how classical multivariate methods can be revised and used in place of conventional statistical tools. Written by prominent researchers in the field, the book focuses on high-dimensional and large-scale approximations and details the many basic multivariate methods used to achieve high levels of accuracy. The authors begin with a fundamental presentation of the basic tools and exact distributional results of multivariate statistics, and, in addition, the derivations of most distributional results are provided. Statistical methods for high-dimensional data, such as curve data, spectra, images, and DNA microarrays, are discussed. Bootstrap approximations from a methodological point of view, theoretical accuracies in MANOVA tests, and model selection criteria are also presented. Subsequent chapters feature additional topical coverage including: High-dimensional approximations of various statistics High-dimensional statistical methods Approximations with computable error bound Selection of variables based on model selection approach Statistics with error bounds and their appearance in discriminant analysis, growth curve models, generalized linear models, profile analysis, and multiple comparison Each chapter provides real-world applications and thorough analyses of the real data. In addition, approximation formulas found throughout the book are a useful tool for both practical and theoretical statisticians, and basic results on exact distributions in multivariate analysis are included in a comprehensive, yet accessible, format. Multivariate Statistics is an excellent book for courses on probability theory in statistics at the graduate level. It is also an essential reference for both practical and theoretical statisticians who are interested in multivariate analysis and who would benefit from learning the applications of analytical probabilistic methods in statistics.

Mathematical Foundations of Infinite-Dimensional Statistical Models
Foundations of Data Science
The Concentration of Measure Phenomenon
Geometric Structure of High-Dimensional Data and Dimensionality Reduction

Large-dimensional Panel Data Econometrics: Testing, Estimation And Structural Changes

Statistical Analysis for High-Dimensional Data

"Focusing on methodology and computation more than on theorems and proofs, this book provides computationally feasible and statistically efficient methods for estimating sparse and large covariance matrices of high-dimensional data. Extensive in breadth and scope, it features ample applications to a number of applied areas, including business and economics, computer science, engineering, and financial mathematics; recognizes the important and significant contributions of longitudinal and spatial data; and includes various computer codes in R throughout the text and on an author-maintained web site"--

The “Stats in the Château” summer school was held at the CRC château on the campus of HEC Paris, Jouy-en-Josas, France, from August 31 to September 4, 2009. This event was organized jointly by faculty members of three French academic institutions ‒ ENSAE ParisTech, the Ecole Polytechnique ParisTech, and HEC Paris ‒ which cooperate through a scientific foundation devoted to the decision sciences. The scientific content of the summer school was conveyed in two courses, one by Laurent Cavalier (Université Aix-Marseille I) on "Ill-posed Inverse Problems", and one by Victor Chernozhukov (Massachusetts Institute of Technology) on "High-dimensional Estimation with Applications to Economics". Ten invited researchers also presented either reviews of the state of the art in the field or of applications, or original research contributions. This volume contains the lecture notes of the two courses. Original research articles and a survey complement these lecture notes. Applications to economics are discussed in various contributions.

Concentration inequalities have been recognized as fundamental tools in several domains such as geometry of Banach spaces or random combinatorics. They also turn to be essential tools to develop a non asymptotic theory in statistics. This volume provides an overview of a non asymptotic theory for model selection. It also discusses some selected applications to variable selection, change points detection and statistical learning.

Connects fundamental mathematical theory with real-world problems, through efficient and scalable optimization algorithms.

Principles and Methods for Data Science

With Exercises and R Labs

Statistics for High-Dimensional Data

With High-Dimensional Data

Regularization in High-dimensional Statistics

Concentration Inequalities and Model Selection

Statistics for High-Dimensional DataMethods, Theory and ApplicationsSpringer Science & Business Media

This textbook provides a step-by-step introduction to the tools and principles of high-dimensional statistics. Each chapter is complemented by numerous exercises, many of them with detailed solutions, and computer labs in R that convey valuable practical insights. The book covers the theory and practice of high-dimensional linear regression, graphical models, and inference, ensuring readers have a smooth start in the field. It also offers suggestions for further reading. Given its scope, the textbook is intended for beginning graduate and advanced undergraduate students in statistics, biostatistics, and bioinformatics, though it will be equally useful to a broader audience.

Multivariate analysis is a mainstay of statistical tools in the analysis of biomedical data. It concerns with associating data matrices of n rows by p columns, with rows representing samples (or patients) and columns attributes of samples, to some response variables, e.g., patients outcome. Classically, the sample size n is much larger than p, the number of variables. The properties of statistical models have been mostly discussed under the assumption of fixed p and infinite n. The advance of biological sciences and technologies has revolutionized the process of investigations of cancer. The biomedical data collection has become more automatic and more extensive. We are in the era of p as a large fraction of n, and even much larger than n. Take proteomics as an example. Although proteomic techniques have been researched and developed for many decades to identify proteins or peptides uniquely associated with a given disease state, until recently this has been mostly a laborious process, carried out one protein at a time. The advent of high throughput proteome-wide technologies such as liquid chromatography-tandem mass spectroscopy make it possible to generate proteomic signatures that facilitate rapid development of new strategies for proteomics-based detection of disease. This poses new challenges and calls for scalable solutions to the analysis of such high dimensional data. In this volume, we will present the systematic and analytical approaches and strategies from both biostatistics and bioinformatics to the analysis of correlated and high-dimensional data.

Ever-greater computing technologies have given rise to an exponentially growing volume of data. Today massive data sets (with potentially thousands of variables) play an important role in almost every branch of modern human activity, including networks, finance, and genetics. However, analyzing such data has presented a challenge for statisticians

Stats in the Château Summer School, August 31 - September 4, 2009

Large Sample Covariance Matrices and High-Dimensional Data Analysis

A Non-Asymptotic Viewpoint

High-Dimensional Data Analysis in Cancer Research

Independence Screening in High-Dimensional Data

High-Dimensional Probability

High-dimensional data, data in which the number of dimensions exceeds the number of observations, is increasingly common in statistics. The term "ultra-high dimensional" is defined by Fan and Lv (2008) as describing the situation where log(p) is of order O(na) for some a in the interval (0, 1 ⁄ 2). It arises in many contexts such as gene expression data, proteomic data, imaging data, tomography, and finance, as well as others. High-dimensional data present a challenge to traditional statistical techniques. In traditional statistical settings, models have a small number of features, chosen based on an assumption of what features may be relevant to the response of interest. In the high-dimensional setting, many of the techniques of traditional feature selection become computationally intractable, or does not yield unique solutions. Current research in modeling high-dimensional data is heavily focused on methods that screen the features before modeling; that is, methods that eliminate noise-features as a pre-modeling dimension reduction. Typically noise feature are identified by exploiting properties of independent random variables, thus the term "independence screening." There are methods for modeling high-dimensional data without feature screening first (e.g. LASSO or SCAD), but simulation studies show screen-first methods perform better as dimensionality increases. Many proposals for independence screening exist, but in my literature review certain themes recurred: A) The assumption of sparsity: that all the useful information in the data is actually contained in a small fraction of the features (the "active features"), the rest being essentially random noise (the "inactive" features). B) In many newer methods, initial dimension reduction by feature screening reduces the problem from the high-dimensional case to a classical case; feature selection then proceeds by a classical method. C) In the initial screening, removal of features independent of the response is highly desirable, as such features literally give no information about the response. D) For the initial screening, some statistic is applied pairwise to each feature in combination with the response; the specific statistic chosen so that in the case that the two random variables are independent, a specific known value is expected for the statistic. E) Features are ranked by the absolute difference between the calculated statistic and the expected value of that statistic in the independent case, i.e. features that are most different from the independent case are most preferred. F) Proof is typically offered that, asymptotically, the method retains the true active features with probability approaching one. G) Where possible, an iterative version of the process is explored, as iterative versions do much better at identifying features that are active in their interactions, but not active individually.

Over the last few years, significant developments have been taking place in high-dimensional data analysis, driven primarily by a wide range of applications in many fields such as genomics and signal processing. In particular, substantial advances have been made in the areas of feature selection, covariance estimation, classification and regression. This book intends to examine important issues arising from high-dimensional data analysis to explore key ideas for statistical inference and prediction. It is structured around topics on multiple hypothesis testing, feature selection, regression, classification, dimension reduction, as well as applications in survival analysis and biomedical research. The book will appeal to graduate students and new researchers interested in the plethora of opportunities available in high-dimensional data analysis.

Over the last few years, significant developments have been taking place in highdimensional data analysis, driven primarily by a wide range of applications in many fields such as genomics and signal processing. In particular, substantial advances have been made in the areas of feature selection, covariance estimation, classification and regression. This book intends to examine important issues arising from highdimensional data analysis to explore key ideas for statistical inference and prediction. It is structured around topics on multiple hypothesis testing, feature selection, regression, cla.

High-dimensional data appear in many fields, and their analysis has become increasingly important in modern statistics. However, it has long been observed that several well-known methods in multivariate analysis become inefficient, or even misleading, when the data dimension p is larger than, say, several tens. A seminal example is the well-known inefficiency of Hotelling's T2-test in such cases. This example shows that classical large sample limits may no longer hold for high-dimensional data; statisticians must seek new limiting theorems in these instances. Thus, the theory of random matrices (RMT) serves as a much-needed and welcome alternative framework. Based on the authors' own research, this book provides a first-hand introduction to new high-dimensional statistical methods derived from RMT. The book begins with a detailed introduction to useful tools from RMT, and then presents a series of high-dimensional problems with solutions provided by RMT methods.

Statistical Foundations of Data Science
Inverse Problems and High-Dimensional Estimation
High-Dimensional Data Analysis with Low-Dimensional Models
Multivariate Statistics
Analyzing High-Dimensional Data
High-Dimensional Covariance Matrix Estimation

This book provides an introduction to the mathematical and algorithmic foundations of data science, including machine learning, high-dimensional geometry, and analysis of large networks. Topics include the counterintuitive nature of data in high dimensions, important linear algebraic techniques such as singular value decomposition, the theory of random walks and Markov chains, the fundamentals of and important algorithms for machine learning, algorithms and analysis for clustering, probabilistic models for large networks, representation learning including topic modelling and non-negative matrix factorization, wavelets and compressed sensing. Important probabilistic techniques are developed including the law of large numbers, tail inequalities, analysis of random projections, generalization guarantees in machine learning, and moment methods for analysis of phase transitions in large random graphs. Additionally, important structural and complexity measures are discussed such as matrix norms and VC-dimension. This book is suitable for both undergraduate and graduate courses in the design and analysis of algorithms for data.

This book provides a unified exposition of some fundamental theoretical problems in high-dimensional statistics. It specifically considers the canonical problems of detection and support estimation for sparse signals observed with noise. Novel phase-transition results are obtained for the signal support estimation problem under a variety of statistical risks. Based on a surprising connection to a concentration of maxima probabilistic phenomenon, the authors obtain a complete characterization of the exact support recovery problem for thresholding estimators under dependent errors.

This book aims to fill the gap between panel data econometrics textbooks, and the latest development on 'big data', especially large-dimensional panel data econometrics. It introduces important research questions in large panels, including testing for cross-sectional dependence, estimation of factor-augmented panel data models, structural breaks in panels and group patterns in panels. To tackle these high dimensional issues, some techniques used in Machine Learning approaches are also illustrated. Moreover, the Monte Carlo experiments, and empirical examples are also utilised to show how to implement these new inference methods. Large-Dimensional Panel Data Econometrics: Testing, Estimation and Structural Changes also introduces new research questions and results in recent literature in this field.

Modern statistics deals with large and complex data sets, and consequently with models containing a large number of parameters. This book presents a detailed account of recently developed approaches, including the Lasso and versions of it for various models, boosting methods, undirected graphical modeling, and procedures controlling false positive selections. A special characteristic of the book is that it contains comprehensive mathematical theory on high-dimensional statistics combined with methodology, algorithms and illustrations with real data examples. This in-depth approach highlights the methods' great potential and practical applicability in a variety of settings. As such, it is a valuable resource for researchers, graduate students and experts in statistics, applied mathematics and computer science.

Concentration of Maxima and Fundamental Limits in High-dimensional Testing and Inference

Methods, Theory and Applications

Model-Based Clustering and Classification for Data Science

Analysis of Multivariate and High-Dimensional Data

Econophysics, Bioinformatics, and Pattern Recognition

New Directions in Statistical Physics

This book presents the latest research on the statistical analysis of functional, high-dimensional and other complex data, addressing methodological and computational aspects, as well as real-world applications. It covers topics like classification, confidence bands, density estimation, depth, diagnostic tests, dimension reduction, estimation on manifolds, high- and infinite-dimensional statistics, inference on functional data, networks, operatorial statistics, prediction, regression, robustness, sequential learning, small-ball probability, smoothing, spatial data, testing, and topological object data analysis, and includes applications in automobile engineering, criminology, drawing recognition, economics, environmetrics, medicine, mobile phone data, spectrometrics and urban environments. The book gathers selected, refereed contributions presented at the Fifth International Workshop on Functional and Operatorial Statistics (IWFOS) in Brno, Czech Republic. The workshop was originally to be held on June 24-26, 2020, but had to be postponed as a consequence of the COVID-19 pandemic. Initiated by the Working Group on Functional and Operatorial Statistics at the University of Toulouse in 2008, the IWFOS workshops provide a forum to discuss the latest trends and advances in functional statistics and related fields, and foster the exchange of ideas and international collaboration in the field.

This book provides a unique insight into the latest breakthroughs in a consistent manner, at a level accessible to undergraduates, yet with enough attention to the theory and computation to satisfy the professional researcher Statistical physics addresses the study and understanding of systems with many degrees of freedom. As such it has a rich and varied history, with applications to thermodynamics, magnetic phase transitions, and order/disorder transformations, to name just a few. However, the tools of statistical physics can be profitably used to investigate any system with a large number of components. Thus, recent years have seen these methods applied in many unexpected directions, three of which are the main focus of this volume. These applications have been remarkably successful and have enriched the financial, biological, and engineering literature. Although reported in the physics literature, the results tend to be scattered and the underlying unity of the field overlooked.

Principles and Methods for Data Science, Volume 43 in the Handbook of Statistics series, highlights new advances in the field, with this updated volume presenting interesting and timely topics, including Competing risks, aims and methods, Data analysis and mining of microbial community dynamics, Support Vector Machines, a robust prediction method with applications in bioinformatics, Bayesian Model Selection for Data with High Dimension, High dimensional statistical inference: theoretical development to data analytics, Big data challenges in genomics, Analysis of microarray gene expression data using information theory and stochastic algorithm, Hybrid Models, Markov Chain Monte Carlo Methods: Theory and Practice, and more. Provides the authority and expertise of leading contributors from an international board of authors Presents the latest release in the Handbook of Statistics series Updated release includes the latest information on Principles and Methods for Data Science

• Real-world problems can be high-dimensional, complex, and noisy • More data does not imply more information • Different approaches deal with the so-called curse of dimensionality to reduce irrelevant information • A process with multidimensional information is not necessarily easy to interpret nor process • In some real-world applications, the number of elements of a class is clearly lower than the other. The models tend to assume that the importance of the analysis belongs to the majority class and this is not usually the truth • The analysis of complex diseases such as cancer are focused on more-than-one dimensional omic data • The increasing amount of data thanks to the reduction of cost of the high-throughput experiments opens up a new era for integrative data-driven approaches • Entropy-based approaches are of interest to reduce the dimensionality of high-dimensional data

The Abel Symposium 2014

Methodologies and Applications

Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003

High-Dimensional Covariance Estimation

Functional and High-Dimensional Statistics and Related Fields

Proven Methods for Big Data Analysis As big data has become standard in many application areas, challenges have arisen related to methodology and software development, including how to discover meaningful patterns in the vast amounts of data. Addressing these problems, Applied Biclustering Methods for Big and High-Dimensional Data Using R shows how to apply biclustering methods to find local patterns in a big data matrix. The book presents an overview of data analysis using biclustering methods from a practical point of view. Real case studies in drug discovery, genetics, marketing research, biology, toxicity, and sports illustrate the use of several biclustering methods. References to technical details of the methods are provided for readers who wish to investigate the full theoretical background. All the methods are accompanied with R examples that show how to conduct the analyses. The examples, software, and other materials are available on a supplementary website.

Analyzing high-dimensional gene expression and DNA methylation data with R is the first practical book that shows a `\`pipeline" of analytical methods with concrete examples starting from raw gene expression and DNA methylation data at the genome scale. Methods on quality control, data pre-processing, data mining, and further assessments are presented in the book, and R programs based on simulated data and real data are included. Codes with example data are all reproducible. Features: • Provides a sequence of analytical tools for genome-scale gene expression data and DNA methylation data, starting from quality control and pre-processing of raw genome-scale data. • Organized by a parallel presentation with explanation on statistical methods and corresponding R packages/functions in quality control, pre-processing, and data analyses (e.g., clustering and networks). • Includes source codes with simulated and real data to reproduce the results. Readers are expected to gain the ability to independently analyze genome-scaled expression and methylation data and detect potential biomarkers. This book is ideal for students majoring in statistics, biostatistics, and bioinformatics and researchers with an interest in high dimensional genetic and epigenetic studies.

Cluster analysis finds groups in data automatically. Most methods have been heuristic and leave open such central questions as: how many clusters are there? Which method should I use? How should I handle outliers? Classification assigns new observations to groups given previously classified observations, and also has open questions about parameter tuning, robustness and uncertainty assessment. This book frames cluster analysis and classification in terms of statistical models, thus yielding principled estimation, testing and prediction methods, and sound answers to the central questions. It builds the basic ideas in an accessible but rigorous way, with extensive data examples and R code; describes modern approaches to high-dimensional data and networks; and explains such recent advances as Bayesian regularization, non-Gaussian model-based clustering, cluster merging, variable selection, semi-supervised and robust classification, clustering of functional data, text and images, and co-clustering. Written for advanced undergraduates in data science, as well as researchers and practitioners, it assumes basic knowledge of multivariate calculus, linear algebra, probability and statistics.

A coherent introductory text from a groundbreaking researcher, focusing on clarity and motivation to build intuition and understanding.

High-Dimensional Statistics

High-Dimensional and Large-Sample Approximations

Bayesian and High-Dimensional Global Optimization

Fundamentals of High-Dimensional Statistics

Applied Biclustering Methods for Big and High-Dimensional Data Using R

Introduction to High-Dimensional Statistics

This book features research contributions from The Abel Symposium on Statistical Analysis for High Dimensional Data, held in Nyvågar, Lofoten, Norway, in May 2014. The focus of the symposium was on statistical and machine learning methodologies specifically developed for inference in "big data" situations, with particular reference to genomic applications. The contributors, who are among the most prominent researchers on the theory of statistics for high dimensional inference, present new theories and methods, as well as challenging applications and computational solutions. Specific themes include, among others, variable selection and screening, penalised regression, sparsity, thresholding, low dimensional structures, computational challenges, non-convex situations, learning graphical models, sparse covariance and precision matrices, semi- and non-parametric formulations, multiple testing, classification, factor models, clustering, and preselection. Highlighting cutting-edge research and casting light on future research directions, the contributions will benefit graduate students and researchers in computational biology, statistics and the machine learning community.

Modern datasets are growing in terms of samples but even more so in terms of variables. We often encounter datasets where samples consists of time series, images, even movies, so that each sample has thousands, even millions of variables. Classical statistical approaches are inadequate for working with such high-dimensional data because they rely on theoretical and computational tools developed without such data in mind. The work in this thesis seeks to close the apparent gap between the growing size of emerging datasets and the capabilities of existing approaches to statistical estimation, inference, and computing. This thesis focuses on two problems that arise in learning from high-dimensional data (versus black-box approaches that do not yield insights into the underlying data-generation process). They are: 1. model selection and post-selection inference: discover the latent low-dimensional structure in high-dimensional data; 2. scalable statistical computing: design scalable estimators and algorithms that avoid communication and minimize ``passes'' over the data. The work relies crucially on results from convex analysis and geometry. Many of the algorithms and proofs are inspired by results from this beautiful but dusty corner of mathematics.

Methods for estimating sparse and large covariance matrices Covariance and correlation matrices play fundamental roles in every aspect of the analysis of multivariate data collected from a variety of fields including business and economics, health care, engineering, and environmental and physical sciences. High-Dimensional Covariance Estimation provides accessible and comprehensive coverage of the classical and modern approaches for estimating covariance matrices as well as their applications to the rapidly developing areas lying at the intersection of statistics and machine learning. Recently, the classical sample covariance methodologies have been modified and improved upon to meet the needs of statisticians and researchers dealing with large correlated datasets. High-Dimensional Covariance Estimation focuses on the methodologies based on shrinkage, thresholding, and penalized likelihood with applications to Gaussian graphical models, prediction, and mean-variance portfolio management. The book relies heavily on regression-based ideas and interpretations to connect and unify many existing methods and algorithms for the task. High-Dimensional Covariance Estimation features chapters on: Data, Sparsity, and Regularization Regularizing the Eigenstructure Banding, Tapering, and Thresholding Covariance Matrices Sparse Gaussian Graphical Models Multivariate Regression The book is an ideal resource for researchers in statistics, mathematics, business and economics, computer sciences, and engineering, as well as a useful text or supplement for graduate-level courses in multivariate analysis, covariance estimation, statistical learning, and high-dimensional data analysis.

Statistical Foundations of Data Science gives a thorough introduction to 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 and a research monograph on high-dimensional statistics, sparsity and covariance learning, machine learning, and statistical inference. It includes ample exercises that involve both theoretical studies as well as empirical applications. The book begins with an introduction to the stylized features of big data and their impacts on statistical analysis. It then introduces multiple linear regression and expands the techniques of model building via nonparametric regression and kernel tricks. It provides a comprehensive account on sparsity explorations and model selections for multiple regression, generalized linear models, quantile regression, robust regression, hazards regression, among others. High-dimensional inference is also thoroughly addressed and so is feature screening. The book also provides a comprehensive account on high-dimensional covariance estimation, learning latent factors and hidden structures, as well as their applications to statistical estimation, inference, prediction and machine learning problems. It also introduces thoroughly statistical machine learning theory and methods for classification, clustering, and prediction. These include CART, random forests, boosting, support vector machines, clustering algorithms, sparse PCA, and deep learning.

Statistical Diagnostics for Cancer

Statistical Inference from High Dimensional Data

An Introduction to Random Matrix Theory

Analyzing High-Dimensional Gene Expression and DNA Methylation Data with R

Written for statisticians, computer scientists, geographers, research and applied scientists, and others interested in visualizing data, this book presents a unique foundation for producing almost every quantitative graphic found in scientific journals, newspapers, statistical packages, and data visualization systems. It was designed for a distributed computing environment, with special attention given to conserving computer code and system resources. While the tangible result of this work is a Java production graphics library, the text focuses on the deep structures involved in producing quantitative graphics from data. It investigates the rules that underlie pie charts, bar charts, scatterplots, function plots, maps, mosaics, and radar charts. These rules are abstracted from the work of Bertin, Cleveland, Kosslyn, MacEachren, Pinker, Tufte, Tukey, Tobler, and other theorists of quantitative graphics.

Ever-greater computing technologies have given rise to an exponentially growing volume of data. Today massive data sets (with potentially thousands of variables) play an important role in almost every branch of modern human activity, including networks, finance, and genetics. However, analyzing such data has presented a challenge for statisticians and data analysts and has required the development of new statistical methods capable of separating the signal from the noise. Introduction to High-Dimensional Statistics is a concise guide to state-of-the-art models, techniques, and approaches for handling high-dimensional data. The book is intended to expose the reader to the key concepts and ideas in the most simple settings possible while avoiding unnecessary technicalities. Offering a succinct presentation of the mathematical foundations of high-dimensional statistics, this highly accessible text: Describes the challenges related to the analysis of high-dimensional data Covers cutting-edge statistical methods including model selection, sparsity and the lasso, aggregation, and learning theory Provides detailed exercises at the end of every chapter with collaborative solutions on a wikisite Illustrates concepts with simple but clear practical examples Introduction to High-Dimensional Statistics is suitable for graduate students and researchers interested in discovering modern statistics for massive data. It can be used as a graduate text or for self-study.

In nonparametric and high-dimensional statistical models, the classical Gauss–Fisher–Le Cam theory of the optimality of maximum likelihood estimators and Bayesian posterior inference does not apply, and new foundations and ideas have been developed in the past several decades. This book gives a coherent account of the statistical theory in infinite-dimensional parameter spaces. The mathematical foundations include self-contained 'mini-courses'

on the theory of Gaussian and empirical processes, approximation and wavelet theory, and the basic theory of function spaces. The theory of statistical inference in such models - hypothesis testing, estimation and confidence sets - is presented within the minimax paradigm of decision theory. This includes the basic theory of convolution kernel and projection estimation, but also Bayesian nonparametrics and nonparametric maximum likelihood estimation. In a final chapter the theory of adaptive inference in nonparametric models is developed, including Lepski's method, wavelet thresholding, and adaptive inference for self-similar functions. Winner of the 2017 PROSE Award for Mathematics.