Statistics For High Dimensional Data Methods Theory And Applications

The twenty-first century has seen a breathtaking expansion of statistical methodology, both in scope and influence. 'Data science' and 'machine learning' have become familiar terms in the news, as statistical methods are brought to bear upon the enormous data sets of modern science and commerce. How did we get here? And where are we going? How does it all fit together? Now in paperback and fortified with exercises, this book delivers a concentrated course in modern statistical thinking. Beginning with classical inferential theories - Bayesian, frequentist, Fisherian - individual chapters take up a series of influential topics: survival analysis, logistic regression, empirical Bayes, the jackknife and bootstrap, random forests, neural networks, Markov Chain Monte Carlo, inference after model selection, and dozens more. The distinctly modern approach integrates methodology and algorithms with statistical inference. Each chapter ends with class-tested exercises, and the book concludes with speculation on the future direction of statistics and 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

Focuses on a few of the important clustering algorithms in the context of information retrieval.

Motivated by applications to root-cause identification of faults in multistage manufacturing processes which involve a large number of tools or equipments at each stage, we consider multiple testing in regression models whose outputs represent the quality characteristics of a multistage manufacturing process. Because of the large number of input variables that correspond to the tools or equipments used, this falls in the framework of regression modeling in the modern era of big data. On the other hand, with quick fault detection and diagnosis followed by tool rectification, sparsity can be assumed in the regression model. We introduce a new approach to address the multiple testing problem and demonstrate its advantages over existing methods. We also illustrate its performance in an application to semiconductor wafer fabrication that motivated this development. The problem of detection and diagnosis of abrupt changes in a stochastic system on the basis of sequential observations has many applications, some of which are discussed in this thesis. In statistical process control (SPC), the past decade witnessed the emergence of a new direction in quality control because of the availability of big data, making use of contemporaneous developments in the statistics literature on high-dimensional data analysis. It has been noticed that in multivariate and high-dimensional applications, only a sparse subset of quality characteristics or other variables of interest

undergoes abnormal changes that lead to deviations from the state of statistical control. The past decade also witnessed major developments in surveillance over sensor networks, cyber-security and information systems. We give a general theory for sequential fault detection in these stochastic models and also modify and extend it to the much less developed problem of fault diagnosis. This fault diagnosis, or change isolation problem, is to determine upon detection of change in a system which one in a set of possible changes has actually occurred. In this connection, we also develop a parallel theory of sequential multiple hypothesis testing. 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.

This modern approach integrates classical and contemporary methods, fusing theory and practice and bridging the gap to statistical learning.

High-dimensional probability offers insight into the behavior of random vectors, random matrices, random subspaces, and objects used to quantify uncertainty in high dimensions. Drawing on ideas from probability, analysis, and geometry, it lends itself to applications in mathematics, statistics, theoretical computer science, signal processing, optimization, and more. It is the first to integrate theory, key tools, and modern applications of high-dimensional probability. Concentration inequalities form the core, and it covers both classical results such as Hoeffding's and Chernoff's inequalities and modern developments such as the matrix Bernstein's inequality. It then introduces the powerful methods based on stochastic processes, including such tools as Slepian's, Sudakov's, and Dudley's inequalities, as well as generic chaining and bounds based on VC dimension. A broad range of illustrations is embedded throughout, including classical and modern results for covariance estimation, clustering, networks, semidefinite programming, coding, dimension reduction, matrix completion, machine learning, compressed sensing, and sparse regression.

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.

New technologies allow us to handle increasingly large datasets, while monitoring devices are becoming ever more sophisticated. This high-tech progress produces statistical units sampled over finer and finer grids. As the measurement points become closer, the data can be considered as observations varying over a continuum. This intrinsic continuous data (called functional data) can be found in various fields of science, including biomechanics, chemometrics, econometrics, environmetrics, geophysics, medicine, etc. The failure of standard multivariate statistics to analyze such functional data has led the statistical community to develop appropriate statistical methodologies, called Functional Data Analysis (FDA). Today, FDA is certainly one of the most motivating and popular statistical topics due to its impact on crucial societal issues (health, environment, etc). This is why the FDA statistical community is rapidly growing, as are the statistical developments . Therefore, it is

necessary to organize regular meetings in order to provide a state-of-art review of the recent advances in this fascinating area. This book collects selected and extended papers presented at the second International Workshop of Functional and Operatorial Statistics (Santander, Spain, 16-18 June, 2011), in which many outstanding experts on FDA will present the most relevant advances in this pioneering statistical area. Undoubtedly, these proceedings will be an essential resource for academic researchers, master students, engineers, and practitioners not only in statistics but also in numerous related fields of application.

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 studies and applies modern flexible regression models for survival data with a special focus on extensions of the Cox model and alternative models with the aim of describing time-varying effects of explanatory variables. Use of the suggested models and methods is illustrated on real data examples, using the R-package timereg developed by the authors, which is applied throughout the book with worked examples for the data sets.

"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"--

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

This book covers several of the statistical concepts and data analytic skills needed to succeed in data-driven life science research. The authors proceed from relatively basic concepts related to computed p-values to advanced topics related to analyzing highthroughput data. They include the R code that performs this analysis and connect the lines of code to the statistical and mathematical concepts explained.

During the past decade there has been an explosion in computation and information technology. With it have come vast amounts of data in a variety of fields such as medicine, biology, finance, and marketing. The challenge of understanding these data has led to the development of new tools in the field of

statistics, and spawned new areas such as data mining, machine learning, and bioinformatics. Many of these tools have common underpinnings but are often expressed with different terminology. This book describes the important ideas in these areas in a common conceptual framework. While the approach is statistical, the emphasis is on concepts rather than mathematics. Many examples are given, with a liberal use of color graphics. It should be a valuable resource for statisticians and anyone interested in data mining in science or industry. The book's coverage is broad, from supervised learning (prediction) to unsupervised learning. The many topics include neural networks, support vector machines, classification trees and boosting---the first comprehensive treatment of this topic in any book. This major new edition features many topics not covered in the original, including graphical models, random forests, ensemble methods, least angle regression & path algorithms for the lasso, non-negative matrix factorization, and spectral clustering. There is also a chapter on methods for "wide" data (p bigger than n), including multiple testing and false discovery rates. Trevor Hastie, Robert Tibshirani, and Jerome Friedman are professors of statistics at Stanford University. They are prominent researchers in this area: Hastie and Tibshirani developed generalized additive models and wrote a popular book of that title. Hastie co-developed much of the

statistical modeling software and environment in R/S-PLUS and invented principal curves and surfaces. Tibshirani proposed the lasso and is co-author of the very successful An Introduction to the Bootstrap. Friedman is the co-inventor of many data-mining tools including CART, MARS, projection pursuit and gradient boosting.

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.

This ready reference discusses different methods for statistically analyzing and validating data created with high-throughput methods. As opposed to other titles, this book focusses on systems approaches, meaning that no single gene or protein forms the basis of the analysis but rather a more or less complex biological network. From a

methodological point of view, the well balanced contributions describe a variety of modern supervised and unsupervised statistical methods applied to various large-scale datasets from genomics and genetics experiments. Furthermore, since the availability of sufficient computer power in recent years has shifted attention from parametric to nonparametric methods, the methods presented here make use of such computer-intensive approaches as Bootstrap, Markov Chain Monte Carlo or general resampling methods. Finally, due to the large amount of information available in public databases, a chapter on Bayesian methods is included, which also provides a systematic means to integrate this information. A welcome guide for mathematicians and the medical and basic research communities.

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 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.

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.

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.

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. 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.

"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.

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.

An integrated package of powerful probabilistic tools and key applications in modern mathematical data science. Discover New Methods for Dealing with High-Dimensional Data A sparse statistical model has only a small number of nonzero parameters or weights; therefore, it is much easier to estimate and interpret than a dense model. Statistical Learning with Sparsity: The Lasso and Generalizations presents methods that exploit sparsity to help recover the underlying signal in a set of data. Top experts in this rapidly evolving field, the authors describe the lasso for linear regression and a simple coordinate descent algorithm for its computation. They discuss the application of l1 penalties to generalized linear models and support vector machines, cover generalized penalties such as the elastic net and group lasso, and review numerical methods for optimization. They also present statistical inference methods for fitted (lasso) models, including the bootstrap, Bayesian methods, and recently developed approaches. In addition, the book examines matrix decomposition, sparse multivariate analysis, graphical models, and compressed sensing. It concludes with a survey of theoretical results for the lasso. In this age of big data, the number of features measured on a person

or object can be large and might be larger than the number of observations. This book shows how the sparsity assumption allows us to tackle these problems and extract useful and reproducible patterns from big datasets. Data analysts, computer scientists, and theorists will appreciate this thorough and up-to-date treatment of sparse statistical modeling.

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.

In many applications of econometrics and economics, a large proportion of the questions of interest are identification. An economist may be interested in uncovering the true signal when the data could be very noisy, such as time-series spurious regression and weak instruments problems, to name a few. In this book, High-Dimensional Econometrics and Identification, we illustrate the true signal and, hence, identification can be recovered even with noisy data in high-dimensional data, e.g., large panels. High-dimensional data in econometrics is the rule rather than the exception. One

of the tools to analyze large, high-dimensional data is the panel data model.High-Dimensional Econometrics and Identification grew out of research work on the identification and high-dimensional econometrics that we have collaborated on over the years, and it aims to provide an up-todate presentation of the issues of identification and high-dimensional econometrics, as well as insights into the use of these results in empirical studies. This book is designed for high-level graduate courses in econometrics and statistics, as well as used as a reference for researchers.

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

A coherent introductory text from a groundbreaking

researcher, focusing on clarity and motivation to build intuition and understanding.

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.

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.

Praise for the First Edition "…extremely well written…a comprehensiveand up-to-date overview of this important field." –Journal of Environmental Quality Exploration and Analysis of DNA Microarray and OtherHigh-Dimensional Data, Second Edition provides comprehensivecoverage of recent advancements in microarray data analysis. Acutting-edge guide, the Second Edition demonstrates variousmethodologies for analyzing data in biomedical research and offersan overview of the modern techniques used in microarray technologyto study patterns of gene activity. The new edition answers the need for an efficient outline of allphases of this revolutionary analytical technique, frompreprocessing to the analysis stage. Utilizing research andexperience from highly-qualified

authors in fields of dataanalysis, Exploration and Analysis of DNA Microarray and OtherHigh-Dimensional Data, Second Edition features: A new chapter on the interpretation of findings that includes adiscussion of signatures and material on gene set analysis,including network analysis New topics of coverage including ABC clustering, biclustering,partial least squares, penalized methods, ensemble methods, andenriched ensemble methods Updated exercises to deepen knowledge of the presented materialand provide readers with resources for further study The book is an ideal reference for scientists in biomedical andgenomics research fields who analyze DNA microarrays and proteinarray data, as well as statisticians and bioinformaticspractitioners. Exploration and Analysis of DNA Microarray andOther High-Dimensional Data, Second Edition is also a usefultext for graduate-level courses on statistics, computationalbiology, and bioinformatics. 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.

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.

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

This textbook is aimed at computer science undergraduates late in sophomore or early in junior year, supplying a comprehensive background in qualitative and quantitative data analysis, probability, random variables, and statistical methods, including machine learning. With careful treatment of topics that fill the curricular needs for the course, Probability and Statistics for Computer Science features: • A treatment of random variables and expectations dealing primarily with the discrete case. • A practical treatment of simulation, showing how many interesting probabilities and expectations can be extracted, with particular emphasis on Markov chains. • A clear but crisp account of simple point inference strategies (maximum likelihood; Bayesian inference) in simple contexts. This is extended to

cover some confidence intervals, samples and populations for random sampling with replacement, and the simplest hypothesis testing. • A chapter dealing with classification, explaining why it's useful; how to train SVM classifiers with stochastic gradient descent; and how to use implementations of more advanced methods such as random forests and nearest neighbors. • A chapter dealing with regression, explaining how to set up, use and understand linear regression and nearest neighbors regression in practical problems. • A chapter dealing with principal components analysis, developing intuition carefully, and including numerous practical examples. There is a brief description of multivariate scaling via principal coordinate analysis. • A chapter dealing with clustering via agglomerative methods and k-means, showing how to build vector quantized features for complex signals. Illustrated throughout, each main chapter includes many worked examples and other pedagogical elements such as boxed Procedures, Definitions, Useful Facts, and Remember This (short tips). Problems and Programming Exercises are at the end of each chapter, with a summary of what the reader should know. Instructor resources include a full set of model solutions for all problems, and an Instructor's Manual with accompanying presentation slides.