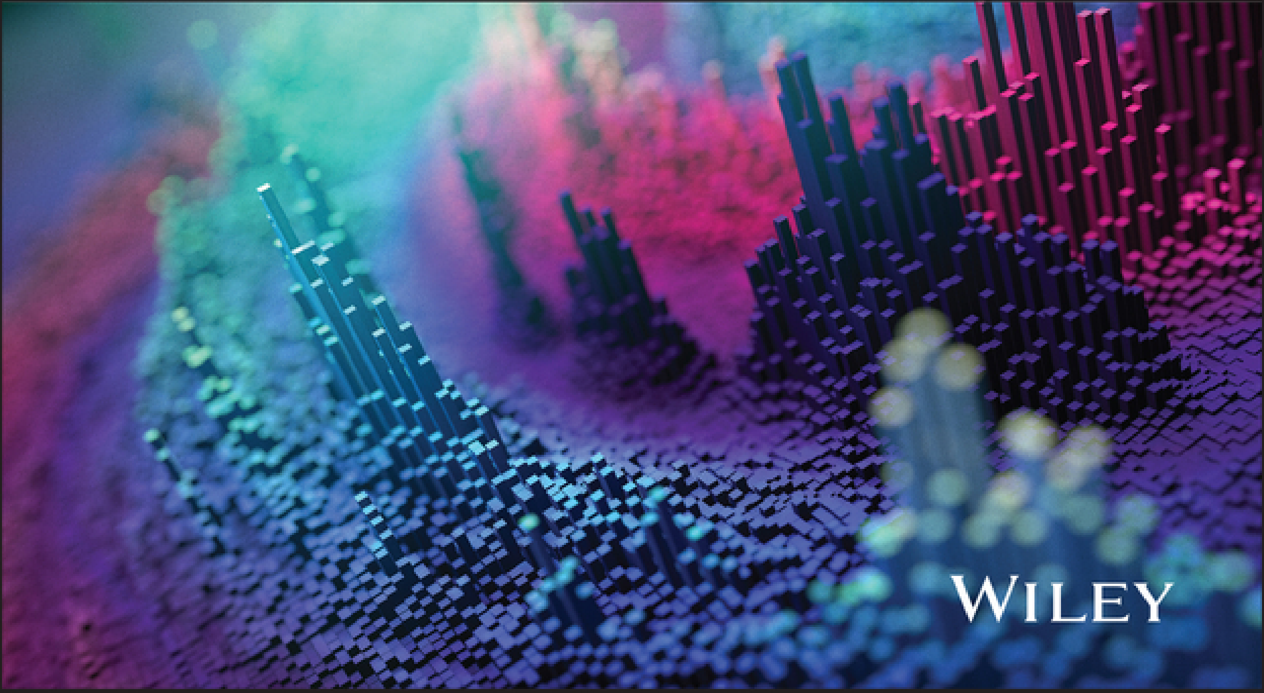
Wiley Finance Series

FINANCIAL DATA ANALYTICS

with Machine Learning, Optimization and Statistics

SAM CHEN
KA CHUN CHEUNG
PHILLIP YAM



WILEY

Praise for *Financial Data Analytics*

Really interesting, and an impressive masterpiece! *Financial Data Analytics* contains a rich amount of material, with original research findings in almost every chapter; many parts of the book will even be directly helpful for my own teaching in business school. In view of its dedication towards data-driven analytical tools genuinely needed in financial problems, I believe that it is the very book that defines the scope of financial data analytics.

—Alain Bensoussan, Fellow of AMS, IEEE, and SIAM; President of *INRIA* (1984–1996); President of *CNES (Centre National d'Etudes Spatiales)* (1996–2003); Chairman of *ESA Council (European Space Agency)* (1999–2002); Former Member of Advisory Board, *Mathematical Finance*; Lars Magnus Ericsson Chair Professor of Management, *Naveen Jindal School of Management, University of Texas at Dallas*

Financial Data Analytics provides a timely and thorough exploration of crucial topics in contemporary data science, specifically tailored for a quantitative finance audience. It skillfully balances introductory and advanced concepts, seamlessly integrating mathematical foundations with detailed coding examples. Designed to appeal to a broad audience, the content accommodates varying levels of familiarity with the mathematical and computational aspects of quantitative finance. The authors' adept presentation of complex ideas, coupled with practical applications, renders *Financial Data Analytics* an invaluable resource for both novice and seasoned professionals alike.

—KC Gary Chan, Fellow of ASA and IMS; President (Western North American Region), *International Biometrics Society* (2022); Professor, Department of Statistics, *University of Washington*

This book presents a wide coverage of state-of-the-art topics in data analytics, which are crucial in our current era of big data. Its organic blend of mathematical derivations of the theory and practical applications in FinTech and InsurTech via tailor-made implementable Python and **R** codes is exceptional. To given but a single example, based on my own research interests, the novel CIBer is a very interesting and original new tool. It is a joy to read *Financial Data Analytics*, a book that cannot be missed on the bookshelf of any researcher or student interested in this topic.

—Jan Dhaene, Full Professor, Director of Master of Science in Financial and Actuarial Engineering, and Head of Actuarial Research Group, Department of Accountancy, Finance and Insurance, Faculty of Business and Economics, *KU Leuven*; Head of Division "Actuariële Toepassingen voor Verzekerings-ondernemingen en Pensioenfondsbeheer", *KU Leuven Research and Development*; Member of *Institute of Actuaries of Belgium*, Member and Vice-chair of Actuarial Education Network, *International Actuarial Association*

Financial Data Analytics is a fantastic book that offers rare gifts to industry practitioners. The important theories are brought to life through **R** and Python program codes, developed by the authors for the book, and easily adaptable for industry use. The book has comprehensive coverage of state-of-the-art techniques for every need. I like reading the practical applications, which help develop intuition for the more complicated methodologies, and surely someone can make a handsome profit implementing them. A super read, and a must-have for professionals if numbers rule your world.

—Kaiser Fung, Bestselling author, *Numbers Rule Your World* and *Numbersense*; Founding Director, MSc programme in Applied Analytics, *Columbia University*; Founder, *Principal Analytics Prep*

Financial Data Analytics is an exceptional book that integrates mathematics, practical examples, and real-life scenarios. With its focus on real datasets and practical programming codes in Python and **R**, the book offers a comprehensive exploration of various topics. It presents novel research findings and provides valuable insights for researchers, practitioners, and actuarial students. The book strikes a balance between foundational concepts and advanced techniques, making it an invaluable reference for professionals in the field. Additionally, its relevance extends to actuarial students preparing for their professional examinations. By redefining the landscape of financial data analytics in FinTech and InsurTech, this book establishes itself as a trusted guide in the industry.

—Simon Lam, Fellow of SOA, CFA, and FRM; President of *The Actuarial Society of Hong Kong* (2018, 2023); Deputy CEO & General Manager, *Munich Re (Hong Kong)*

The book will certainly play an impactful role in the advancement of financial analytics and should be on the bookshelf of every serious student of the topic.

—Wai Keung Li, Fellow of ASA and IMS; Emeritus Professor, *The University of Hong Kong*; Dean, *Faculty of Liberal Arts and Social Sciences, The Education University of Hong Kong*

Financial Data Analytics is a masterfully written book that encompasses a wide spectrum of statistical models and algorithms, with a special emphasis on financial and insurance applications. Drawing upon their multidisciplinary background and extensive research experience, as well as their close connection with the industry, the authors skillfully explain the theoretical underpinnings of both conventional and contemporary statistical methods that are truly relevant to the industry (including but not limited to regression learning, classification trees, neural networks, as well as the specification and assessment of these models), and amply illustrate the practical applications of these methods in various disciplines, by an abundance of real financial and insurance data, and using both Python and **R**. The dual focus on theory and applications, together with the discussion on recent advancements of the fields, makes the book one of a kind, even field-defining, among books on similar topics, and an ideal resource for anyone interested in understanding and implementing statistical models in this era of big data, as well as for students preparing for professional examinations on data analytics, such as the SRM, PA and ATPA exams of the Society of Actuaries.

—Ambrose Lo, Fellow of SOA, Chartered Enterprise Risk Analyst; Author of *ACTEX Study Manual for SOA Exam SRM*, *ACTEX Study Manual for SOA Exam PA*, and *ACTEX Study Manual for SOA Exam ATPA*

It is a tome!

—Suresh P. Sethi, Fellow of INFORMS, IEEE, POMS, and SIAM; Eugene McDermott Chair Professor of Operations Management, *Naveen Jindal School of Management, University of Texas at Dallas*

Financial Data Analytics is an encyclopedic documentation of in-depth and extensive statistical analysis in finance and beyond. It provides an end-to-end systematic approach to academics and practitioners with theories, tremendous examples and data, and algorithms with coding that are readily applicable in real life. The book encompasses four dimensions of coverage—theoretical framework to application and coding, distributional characteristics to data diagnosis and simulation, learning, and lastly coverage of both qualitative and quantitative data. The book consolidates classical knowledge with the most contemporary research in all subjects. *Financial Data Analytics* is one comprehensive biblical handbook for academic researchers, financial practitioners, and graduate students for both methodologies and applications. The book also lays a systematic framework for future extension and enrichment for financial data analytics.

—Nai-pan Tang, Former Chief Risk Officer and Member of Executive Committee, Hang Seng Bank; Former Deputy CEO and Chief Risk Officer, Shanghai Commercial Bank Ltd.; Former Director of the Board, Deputy CEO, Alternative CEO, Chief Risk Officer, and Vice Chairman of Asset Management, China CITIC Bank International; Director, The Hong Kong Institute of Bankers (2019–2021); Professor of Practice, Department of Finance, Chinese University of Hong Kong

Financial Data Analytics is a very impressive work with extensive coverage and many interesting topics. It stands out among similar books by the unique blend of detailed mathematical derivations, practical examples, real-life datasets, and readily available programme codes in Python and **R**. It also contains many novel results from the authors' recent research. Whether you are researchers on related fields, practitioners in the financial industry, or students preparing for a few exams at The Society of Actuaries or The Institute and Faculty of Actuaries, *Financial Data Analytics* will certainly be a valuable reference book.

—Hailiang Yang, Associate of SOA and Honorary Fellow of IFoA; Editor of *Insurance: Mathematics and Economics*; Professor, *Department of Financial and Actuarial Mathematics, Xi'an Jiaotong–Liverpool University*

Throughout my career at JP Morgan Chase and then CITIC Securities, I have participated in and witnessed how various data analytics tools revolutionized financial industry. *Financial Data Analytics* is a perfect example echoing this trend, by discussing a wide spectrum of modern tools in data analytics, from both a theoretical viewpoint and a practical aspect, with readily implementable programme codes in both Python and **R** for real-life examples. This combination is so unique amongst the few books on data analytics; it not only reflects the broad range of theoretical knowledge of the authors, but also demonstrates their close ties with the finance and insurance industries. I actually had the opportunity to testify some profit-making strategies mentioned in the book, and their performance was genuinely impressive. This book is far more than an academic monograph for scholars; it is certainly an illuminating guide for practitioners to explore their own alchemy of finance.

—Wei Zhou, Executive Director of Equity Derivatives Quantitative Research, *JP Morgan Chase* (2016–2021); Executive Director and Head of Quantitative Modelling, *CITIC Securities*

Financial Data Analytics

Founded in 1807, John Wiley & Sons is the oldest independent publishing company in the United States. With offices in North America, Europe, Australia and Asia, Wiley is globally committed to developing and marketing print and electronic products and services for our customers' professional and personal knowledge and understanding.

The Wiley Finance series contains books written specifically for finance and investment professionals as well as sophisticated individual investors and their financial advisors. Book topics range from portfolio management to e-commerce, risk management, financial engineering, valuation and financial instrument analysis, as well as much more.

For a list of available titles, visit our Web site at www.WileyFinance.com.

Financial Data Analytics

with Machine Learning, Optimization and Statistics

SAM CHEN
Hang Seng University of Hong Kong

KA CHUN CHEUNG
University of Hong Kong

PHILLIP YAM
Chinese University of Hong Kong

with programme codes by Kaiser Fan

WILEY

This edition first published 2025

Registered Office(s)
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.
Wiley also publishes its books in a variety of electronic formats and by print-on- demand. Some content that appears in standard print versions of this book may not be available in other formats. Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data Is Available:

ISBN 9781119863373 (Cloth)
ISBN 9781119863380 (ePDF)
ISBN 9781119863397 (ePub)
ISBN 9781119863403 (oBook)

Cover Design: Wiley
Cover Image: © da-kuk/Getty Images

To our parents and families

Contents

PART ONE
Data Cleansing and Analytical Models

CHAPTER 8
Risk Measures, Extreme Values, and Copulae 345

PART TWO
Linear Models

CHAPTER 9
Principal Component Analysis and Recommender Systems 409

CHAPTER 10
Regression Learning 467

About the Authors

Yongzhao Chen (**Sam**) received his BSc in Actuarial Science with first class honours and PhD in Actuarial Science from The University of Hong Kong. He is currently an Assistant Professor at the Department of Mathematics, Statistics and Insurance of the Hang Seng University of Hong Kong. His research interests include actuarial science, especially credibility theory, and data analytics.

Ka Chun Cheung received his BSc in Actuarial Science with first class honours and PhD from The University of Hong Kong. He was the Director of the Actuarial Science Programme, and is currently Head and full Professor at the Department of Statistics and Actuarial Science in School of Computing and Data Science, The University of Hong Kong. He is an Associate of the Society of Actuaries and an elected member of the International Statistical Institute. He is serving on the editorial boards of Insurance, Mathematics and Economics and Journal of Industrial and Management Optimization. His current research interests include various topics in actuarial science, including optimal reinsurance, stochastic orders, dependence structures, and extreme value theory.

Phillip Yam received his BSc in Actuarial Science with first class honours and MPhil from The University of Hong Kong. Supported by the two scholarships awarded by the Croucher Foundation (Hong Kong), he obtained an MASt (Master of Advanced Study) degree, Part III of the Mathematical Tripos, with Distinction in Mathematics from University of Cambridge and a DPhil in Mathematics from University of Oxford. During his postgraduate studies, he was awarded with the E. M. Burnett Prize in Mathematics from University of Cambridge, and the junior research fellowship from The Erwin Schrödinger International Institute for Mathematics and Physics of University of Vienna.

Phillip is currently the Co-Director of the Interdisciplinary Major Programme in Quantitative Finance and Risk Management Science, and a full Professor at the Department of Statistics of The Chinese University of Hong Kong (CUHK). He is also Assistant Dean (Education) of CUHK Faculty of Science, and Fellow of the Centre for Promoting Science Education in the Faculty. He has been appointed as a research fellow in the Hausdorff Research Institute for Mathematics at the University of Bonn and a Visiting Professor in both the Department of Statistics at Columbia University in the City of New York and Naveen Jindal of Management at University of Texas at Dallas. He has published about a hundred journal articles in actuarial science, applied mathematics, data analytics, engineering, financial mathematics, operations management, and statistics, and has also been serving in editorial boards of several journals in these fields. Together with Alain Bensoussan and Jens Frehse, he wrote the first monograph on mean field games and mean field type control theory. His research project with the title “Comonotone-independence Bayes Classifier (CIBer)” was awarded a Silver Medal in the 48th International Exhibition of Inventions Geneva in 2023. Besides academia, he has provided consulting services for various financial institutions and insurance companies, and established close connections in these industries; many of his students also work in international investment banking and insurance companies.

Kaiser Fan received his BSc in Risk Management Science with first class honours and MPhil from The Chinese University of Hong Kong under the guidance of Professor Phillip Yam. As a data scientist, his research interests include data analytics, and machine learning especially in deep learning. He contributes to the programming and many examples and illustrations in the book.

Foreword

To the memory of

Tze Leung Lai (1945–2023)
Late Ray Lyman Wilbur Professor of Statistics,
Stanford University

We were saddened to hear about the sudden passing away of Professor Tze Leung Lai. We would like to thank him for his care for the younger generation, including ourselves, as well as all of his valuable guidance in the past two decades, since Ka Chun's and Phillip's senior-year undergraduate and master studies at The University of Hong Kong; we all learned a lot from him, both indirectly or directly. He was certainly a renowned scholar. Due to the pandemic, we could not make the trip to visit him in person; we were hoping to meet him again last summer after the pandemic eventually came to an end, only to learn that he departed too soon. During the book writing process, we sent a draft version of the book to him. He was glad of what we had achieved and also graciously offered to write a foreword for this book, which can no longer become a reality now. However, this foreword is always reserved for him. We thank you again for your generous offer, Professor Lai; thank you, our mentor, may you rest in peace.

Winter, 2023

Preface

In the field of finance, nothing is more important than gaining profits, and any innovation that draws people's attention must lead to at least the same level of profit as the existing methods; indeed, this has always been the main driving force of updates to relevant curricula over time. For instance, with the development of option pricing and portfolio selection in 1970s, the financial training from mid-1980s to 2000s heavily involved Itô's stochastic calculus and partial differential equations. On the other hand, volatility models such as GARCH were proposed by Robert Engle in the early 1980s for a better estimation of parameters facilitating derivative valuation and portfolio management, and various academic curricula quickly followed suit by placing more emphasis on financial econometrics. Quantitative analysis of game theory, particularly the numerical algorithm for discovering equilibrium points, gained more importance and attention in academia after John Nash won the Nobel Memorial Prize in Economic Sciences in 1994; the trend continues today with further sophistication and generalization towards the context of mean field games in the last dozen years. In the 2000s, as behavioural finance was gaining increasing attention in society, people wanted to learn more methods in the realm of experimental behavioural economics and finance, especially on how the market makes use of statistical methods to understand the impact of different human behaviour and devise advertising strategies accordingly, which explains why case studies and primitive statistics have been prevalent in the financial classes in recent decades.

Recently, attention has been diverted towards AI. The revolutionary developments in machine learning and deep learning have brought new elements of data analytics into finance, particularly including the heated areas of *InsurTech*, *FinTech* and *RegTech*. To catch up with the trend, curriculum designs should be revised to cover financial or business data analytics in a comprehensive manner, and statistics is certainly at the core of them; this is precisely why we wanted to write this book. Among the few books in this field, involving the use of standard statistics in financial analysis with either Python or **R**, and statistical applications in financial engineering, focus is usually put on the possible financial applications of conventional statistical tools, yet some practical problems may require tools beyond traditional statistics, and we aim to address a few relevant issues in this book. Another important motivation for us is certainly the positive feedback from students regarding our teaching materials in the past decade, which we have consolidated as the foundation of this book.

This book investigates contemporary practical techniques of financial data analytics that are specific for real-life scenarios and leave room for a high profit-making potential, with 15 chapters in total covering a wide range of important and frontier topics in this field. For example, we shall explore data analytics in investment strategy, financial forensics, and the immediate use of deep learning in finance. We also critically discuss the pros and cons of machine learning tools. While we raise caution against potential pitfalls of new approaches like deep learning, we also propose a novel feature engineering scheme as part of CIBer (see Chapter 12) to overcome limitations of existing methods regarding input features, which also achieves a promising classification performance. Examples are provided throughout the whole book, in which we focus on a few typical datasets from real-life financial markets to facilitate intuitive comparisons among models, allowing readers to form their own judgements on their pros and cons, and hence apply suitable data analysis methods to their own datasets. Executable detailed programme codes in Python and **R** are also readily available with corresponding examples. Practitioners including quants and fund managers can gain insights into the latest developments of data analytics from the book, and help to formulate effective investment strategies or to facilitate better product designs. This up-to-date knowledge may further help them conduct novel applied research in different business disciplines. It is also our hope that senior-year students and postgraduates can deepen their understanding on this field and find the book useful for their future academic research. Meanwhile, the contents of this book also cover a large part of syllabi of modules from different public professional examinations on predictive analytics, including but not limited to the Statistics for Risk Modeling (SRM) Exam and Predictive Analytics (PA) Exam of the *Society of Actuaries*, making it a suitable main or supplementary reading for these professional examinations.

To benefit the most from this book, it is advisable that readers have a solid background in probability and statistics, linear algebra, and advanced calculus, preferably at the sophomore level. For some parts of the book, some basic knowledge in real and complex analysis would enhance a full understanding of them. Especially, some sections marked by asterisks necessitate a higher level of mathematical understanding and may be omitted during the initial reading. Anyhow, for the convenience of the readers, some of the relevant basic knowledge is reviewed in Chapter 1. Acquaintance with programming languages such as Python or **R** is also instrumental. To make the book more self-contained, a quick overview of these two programming languages is provided in Chapter 2. In addition, readers interested in more sophisticated investment strategies and derivative pricing also need a rudimentary background in Itô's stochastic calculus and continuous martingale theory, which is unavoidable given the technical nature of the subject matter.

In writing this book, our multidisciplinary background proved helpful; we all had diverse training in actuarial science, economics and finance, mathematics, probability and statistics, and our research also involves the application of data analytics in diverse applied areas. We have established long-term research collaborations with industry practitioners, and many of our undergraduate and postgraduate students, friends, and colleagues are also working in world-leading companies in finance and insurance sectors. Growing up in the traditional global financial hub of Hong Kong has also equipped us with practical financial knowledge that benefits our pragmatic research, while we are not bounded by the routine methods in solving both research and practical problems. In addition, our graduate student Kaiser Fan also made a unique contribution by implementing most programme codes in the examples in a tailor-made and illuminating fashion.

With the publication of this book, we welcome valuable feedback and comments from readers. Due to limitations in both scope and time, we could not delve into all topics in detail, and we apologize for any missing information. While we drew inspiration from a wide range of literature, and we tried to cite all of them, some may still have unintentionally slipped our mind over time, and we sincerely apologize for any oversights. We also benefited from courses on financial data analytics or equivalents, including teaching materials and course design, in renowned universities in Asia, Australia, Europe and North America.

While we believe our book offers a valuable collection of tools in financial data analytics, we deliberately left out blockchains, as they have shifted towards being an internet and network security concern rather than an analytical tool for financial information. Meanwhile, we have an upcoming book on deep learning with some applications in finance, to which we briefly hint in the closing chapter of this book. Hopefully readers will enjoy the current book and stay tuned for the upcoming release.

Sam Chen, Ka Chun Cheung, and Phillip Yam
Hong Kong, December 2023

Acknowledgements

First and foremost, we would like to thank Wiley for their kind consideration of our book, as well as their assistance and patience in different stages of preparation, so that we could have more time to polish this highly original book.

This book would also be impossible without the valuable teaching materials and datasets generously provided by Pui Lam Leung, and we would like to express our sincere gratitude to him. Moreover, as the soul figure of the whole book, Phillip Yam took the initiative to design its framework, and led the writing of all chapters using his solid mathematical knowledge and deep financial insights; the other authors express their sincere gratitude for his dedicated effort and unique role played.

Throughout the writing of this book over the last five years (mostly during the COVID-19 pandemic), we received much positive feedback and constructive suggestions from experts in both academia and industry after their careful reading of our book, including but not limited to Alain Bensoussan, Jan Dhaene, Tze Leung Lai, Wai Keung Li, Hailiang Yang; as well as Kaiser Fung, Wah Tung Lau, Kam Pui Wat, John Alexander Wright, and Wei Zhou. Our special thanks go to all of them, particularly Professor Lai; he will be remembered for long.

The assistance on typesetting and proofreading by other members in our research team, namely Pok Him Cheng, Adrian Patrick Kennedy, Junyou Li, Kenneth Tsz Hin Ng, Zhengyao Sun, and Jing Zhang, is also critically important to us. We truly appreciate their efforts, without which our book cannot be completed within a limited period of time at the last stage.

We also acknowledge funding support from various research and teaching grants. Yongzhao Chen (Sam) was partially supported by UGC/FDS14/P03/23 with the project title "*InsurTech: Risk Classification and Premium Calibration with Data Analytics*"; Ka Chun Cheung was partially supported by HKGRF-17303721 with the project title "*A contemporary study of optimal reinsurance arrangements*"; Phillip Yam was partially supported by HKGRF-14300319 with the project title "*Shape-constrained Inference: Testing for Monotonicity*", HKGRF-14301321 with the project title "*General Theory for Infinite Dimensional Stochastic Control: Mean Field and Some Classical Problems*", HKGRF-14300123 with the project title "*Well-posedness of Some Poisson-driven Mean Field Learning Models and their Applications*", CUHK Teaching Development and Language Enhancement Grant (TDLEG) for the 2022–25 Triennium with the project title "*Computational Thinking (CT) as a Problem-solving Skill – A Multidisciplinary Virtual Learning Package*", CUHK Teaching Development and Language Enhancement Grant (TDLEG) for the 2019–22 Triennium: Funding Scheme for Engaging Postgraduate Students in Teaching and Teaching Development with the project title "*Supporting Statistics Research Postgraduates to Teach Quantitative Data Analysis to Postgraduate Students without*

Statistics Background – Phase I", and CUHK Teaching Development and Language Enhancement Grant (TDLEG) for the 2022–25 Triennium: Funding Scheme for Engaging Postgraduate Students in Teaching and Teaching Development with project title "*Supporting Statistics Research Postgraduates to Teach Quantitative Data Analysis to Postgraduate Students without Statistics Background – Phase II*". The work described in this book was supported by a grant from the Germany/Hong Kong Joint Research Scheme sponsored by the Research Grants Council of Hong Kong and the German Academic Exchange Service of Germany (Reference No. G-CUHK411/23).

Last but certainly not least, we give heartful thanks to our families for their unwavering support. The writing of this book occupied much of our spare time that we could have spent with our families. We are immensely grateful for their understanding and support throughout. We can never thank them enough, and we love them from the bottom of our hearts.

Introduction

"We know the past but cannot control it. We control the future but cannot know it."

—Claude Shannon [18]

DEVELOPMENT OF FINANCIAL DATA ANALYTICS

Financial data analytics can be described as the application of statistical models and algorithms to learn from the data in financial markets and investment portfolios in a timely manner, and then using the knowledge gained to control their future performance trends and make predictions accordingly. This recently emerging term highlights the importance of incorporating real-world information, especially those from extremal events, in the era of big data. The use of computational intelligence, statistical rules and algorithms allows businesses to make informed decisions and predictions. Furthermore, under an increasing emphasis on the concept of risk management, market practitioners can better measure the econometric patterns and trends using the large datasets available.

While this term sounds quite new, the practice itself has actually been present throughout mankind's history, although in different forms. Even in ancient times, our ancestors already engaged in commercial trading and recorded data related to their daily lives; indeed, economic transactions in the form of exchange among goods, services, and commodities such as crops, livestock, and textiles, were common among ancient civilizations. They also developed some methods to systematically record data as well as to track and document their economic activities, though they are certainly less mature or complex than modern ones. Take the *Babylonians* for example: clay tablets have been discovered from *Mesopotamia*, which contain records of various business transactions, contracts, and trade agreements. While we cannot assert that the tablets were specifically dedicated to financial data, these archaeological findings do provide insights into early business activities and trade. In fact, the Babylonians are also well-known for their advanced mathematical and astronomical knowledge, and they might be the first to develop accounting and record-keeping to track trade, debts, and other business transactions. These represent the initial attempts of humans to systematically record and organize economic information; see more discussions in [11].

As the notions of contracts and stocks emerged, financial activities saw further developments in the medieval era. For example, the first forward contracts were traded on the *Dojima Rice Exchange* in 1697 in Japan, as the *Samurai* were paid in rice and needed a stable method of converting rice to currency; such contracts were beneficial to agricultural producers, who could use them to hedge against future price changes. While people paid more attention to the patterns behind the data, their investment decisions depended mainly on such market sentiments; this also contributed to the financial bubbles during this period. The *Tulip Mania* in the Netherlands during the 1630s is often regarded as one of the most famous examples of a speculative bubble in financial history. During this crisis, the prices of tulip bulbs skyrocketed to extraordinarily high levels before eventually collapsing. Since then, similar bubbles have appeared from time to time, including the well-known *South Sea Company Bubble* in 1720, when many investors were lured to invest in the stock but ended up with a huge loss. The most famous name among these unfortunate stockholders is arguably *Sir Isaac Newton*; despite being an excellent mathematician, his losses in this investment exceeded £20,000, which is comparable with around $20 million nowadays. Allegedly, Newton commented on this incident that he could "*calculate the motions of the heavenly bodies, but not the madness of people*"; one can find many more interesting stories in [3].

In the same era, attempts at data analytics were also made in the field of insurance; in particular, the renowned English astronomer and mathematician, *Edmond Halley*, studied the demographic records in *Breslau* (now *Wrocław* in Poland) and created life tables accordingly to estimate life expectancies and survival probabilities for different age groups, which now play a crucial role in pricing annuities and insurance policies [7]. Compared with the *Ulpian*'s life table, developed in ancient *Rome* without reference to valid data sources, Halley's data-driven studies in the late 17th Century truly laid the foundation for the development of actuarial science as a quantitative subject; also see [22].

Moving on to the mid 19th century, the modern version of commodity and futures markets developed from forward contracts began to take shape in metropolitan areas like *New York* and *London*. The formation of organized exchanges brought several benefits to market participants. It enhanced market liquidity by bringing together a larger number of buyers and sellers, facilitating efficient price discovery. Standardized contracts and rules were introduced, ensuring uniformity and transparency in trading practices. However, during this period, these markets relied mainly on raw business data rather than a large-scale systematic study or model fitting for prediction purposes.

The next wave of advance in finance, particularly financial modelling and option pricing, was initiated at the turn of the 20th century by the French mathematician *Louis Bachelier*. His groundbreaking thesis, "*Théorie de la spéculation*" [1], introduced the use of the *arithmetic Brownian motion* (ABM) to model security prices. Although Bachelier did not have access to reliable market data or sophisticated data analysis techniques, he attempted to replicate real market data as closely as possible. He used a volatile model to capture the movement of prices, yet it was not the best representation of the observed market dynamics. The importance of his work was later acknowledged by the renowned economist *Paul Samuelson*, who further improved the model for asset prices by using geometric Brownian motion (GBM) instead of ABM, as GBM can better capture the long-term growth trend observed in financial markets by the drift term; see [17]. Further along this line of development, in 1973, economists *Fischer Black* and *Myron Scholes* proposed the game-changing