Probabilistic Data-Driven Modeling

This book introduces relevant and established data-driven modeling tools currently in use or in development, that will help readers master the art and science of constructing models from data and dive into different application areas. It presents statistical tools useful to individuate regularities and discover patterns and laws in complex datasets and demonstrates how to apply them to devise models that help to understand these systems and predict their behaviors. By focusing on the estimation of multivariate probabilities, the book shows that the entire domain, from linear regressions to deep learning neural networks, can be formulated in probabilistic terms. This book provides the right balance between accessibility and mathematical rigor for applied data science or operations research students, graduate students in CSE, and machine learning and uncertainty quantification researchers who use statistics in their field. A background in probability theory and undergraduate mathematics is assumed.

Tomaso Aste is Professor of Complexity Science at the Computer Science Department at University College London (UCL). A trained physicist, he has substantially contributed to research in complex systems modeling, from materials to markets. He has authored over 300 research papers and several books and collaborated with over 100 researchers across all continents. Professor Aste is founder and Head of the Financial Computing and Analytics Group at UCL, and he is the founder and editor-in-chief of the *Data-Driven Modelling* journal.

Probabilistic Data-Driven Modeling

Tomaso Aste

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009221856
DOI: 10.1017/9781009221887

When citing this work, please include a reference to the DOI 10.1017/9781009221887

First published 2025

Printed in the United Kingdom by CPI Group Ltd, Croydon CR0 4YY

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Aste, Tomaso, author.
Title: Probabilistic data-driven modeling / Tomaso Aste, University College London.
Description: First edition. | Cambridge, United Kingdom ; New York, NY, USA : Cambridge University
Press, 2024. | Includes bibliographical references and index.
Identifiers: LCCN 2024011634 | ISBN 9781009221856 (hardback)
Subjects: LCSH: Probabilities–Data processing. | Multivariate analysis.
Classification: LCC QA273.19.E4 A88 2024 | DDC 519.2–dc23/eng/20240412
LC record available at https://lccn.loc.gov/2024011634

ISBN 978-1-009-22185-6 Hardback

Dedicated to my daughters

Contents

Preface

What is this book about and why might you want to read it?

This book introduces and guides the reader through a selection of methodologies and approaches to construct models from data. These data-driven approaches were originally developed in different fields, from statistics to complexity science. They are general procedures and tools that apply to any domain where models must be built from observational data. This book is also about probabilities and their identification and estimation from data. I approach the general topic of data-driven modeling from the perspective that the entire domain, from linear regressions to deep learning neural networks, can be formulated in terms of the estimation of multivariate probabilities from data. This perspective provides a unifying frame of reference and an interpretation tool for all the topics and methods discussed in this book.

I provide practical tools to characterize, classify, and model real systems starting from data. The material I present has become my modeling "toolbox" that I have been using and refining for my research over the last 30 years. I included in this book what – I believe – is a relevant, meaningful, and essential selection of tools. In the case when several methods could be applied, I try to avoid listing all of them; rather, I choose to focus on the one that I believe is the most effective or the one I find simplest or – sometimes – that I like the most. I also give details about some of the often-overlooked aspects in this domain. For instance, how to deal with modeling when the number of observations is small or noisy, or non-stationary. I also discuss the interpretation of statistical validation results from a practical perspective and the complexity of many real-world systems.

The book aims to be readable by anyone with a background in mathematics at the bachelor's degree level. It is intended to be used both by students as textbook support and by professionals and academics as a reference. I try to avoid technical jargon and I introduce all new concepts in a way intended to be as self-contained as possible. Experts in some of the subjects might find some of the introductory parts too basic but, for them, there is no need to read everything from cover to cover. Indeed, I have organized the content in a way that makes it easy for an experienced reader to fast-forward through some basic parts of the book, skipping most of the text while retaining the book's perspective on the topics and then focusing on the more advanced parts.

The book presents my perspective on modeling real complex systems. In doing so it introduces some of the most relevant, established, data-driven modeling tools currently in use, and also some approaches that are still in development. I present statistical tools useful to individuate regularities, discover patterns and laws in complex datasets, and apply them to devise models that help to understand these systems and help to predict their behaviors. Specifically, this book provides the mathematical instruments and the knowledge needed to:

- analyze and characterize complex datasets
- compute relevant statistical quantities
- quantify inter-dependency and causality structure between different variables
- quantify the reliability of data
- construct graphical representations of the dataset
- build probabilistic models for description and prediction of real systems
- validate hypothesis and models
- select between alternative models.

I organized this book as a complete guide through the complex, rich, and fascinating field of data-driven modeling. Practical issues on data analysis and statistics are covered using specific examples.

I divided the book into four parts. Part I is about preliminary essential concepts. Part II is about foundations and it provides the theoretical basis for probabilistic data-driven modeling. Part III is about the actual construction of models from data. It gives the practical tools and methodologies. Part IV is the closing. Although unorthodox, a fruitful way to read this book could be starting from Part III and then referring to the definitions and the fundamental concepts in Part I and II when needed. Indeed, I provide cross-references to relevant sections and definitions to help the reader to navigate the book.

I have taught the content of this book across several universities in Australia and the UK for several years. I had thousands of students who learned how to model real complex systems from this material, and I believe I have developed a sense of the content that matters and how to deliver it.

There is a great need to increase data analytics and data-driven modeling capability in the industry. People with data-driven modeling skills are in great and increasing demand. The instruments and tools provided in this book are essential to understanding, modeling, and making practical use of the very large quantity of data that most human activities are currently producing and collecting.

I adopt the perspective that real, complex, systems should be modeled in terms of the multivariate probability of all variables involved in the system. This would classify my approach and perspective in the realm of Bayesian statistics. However, I am a trained physicist and I have not been educated at statistics schools. Therefore, I can qualify myself neither as a Bayesianist nor as a frequentist. I must say that these classifications mean little to me. In some parts of the book, I report the Bayesian perspective and in other parts instead, I adopt what is considered to be the frequentist perspective. I do this following what I consider the most intuitive way to present and understand the problem or, what I believe, is the most powerful instrument for the specific problem. I also avoid using more formal approaches to probability. For instance, I only mention the essentials of Kolmogorov's axiomatic formulation of probability theory in Appendix A to set a sound basis. I do not report proofs of theorems. However, I do report mathematically sound demonstrations when I believe they can be useful for a better understanding of the context. Despite the avoidance of jargon and the absence of some formalizations, I try to be precise and mathematically rigorous.

Exercises

All central chapters, from Chapters 5 to 18, have exercises. They are meant to encourage the reader to challenge themselves on some aspects of the topic presented in each chapter. The exercises tend to be mostly of a theoretical nature. For the numerical challenges, the reader must refer to some of the examples and the supporting material.

Examples

The book includes several examples that provide a practical perspective on the topics discussed in the chapters. These examples give hands-on information on how to use the data-driven tools introduced, thereby completing the chapter discussions. While it is possible to skip these examples, doing so would greatly impoverish the reader's experience. Engaging with these examples enhances understanding and application of the theoretical concepts, making them an essential part of the learning process.

Supporting Material

The book has a large body of supporting material consisting of data, examples, Python and Matlab codes which are provided on the GitHub page:
https://github.com/FinancialComputingUCL/DataDrivenModeling/.

This material is organized by chapter number to help the reader to identify the relevant material while reading the book. However, it sometimes mixes topics and methodologies from other chapters. There is also more general and self-contained extra material that covers topics that span across the entire book and focus on specific applications, certain methodologies, and particular narratives. This supporting material is designed to be dynamic and will continuously evolve, be updated, and be enhanced over time.

Utilizing Colored Boxes for Enhanced Navigation

Throughout this book, colored boxes serve as visual cues to categorize and highlight various segments of content, facilitating content navigation. Specifically, I employ the following color scheme:

- **Definitions** are encapsulated in light yellow boxes, providing a clear and distinct presentation of key terms and concepts.
- **Examples** are encapsulated in light green boxes, offering practical applications and scenarios to elucidate the material discussed.
- **Remarks** are highlighted in light blue boxes.
- **Algorithms** are highlighted in light orange boxes.

Moreover, to accommodate diverse reading preferences and focus areas, light gray boxes are used to encase more mathematically intensive discussions and deeper technical explorations. Readers primarily interested in the foundational aspects of the subject may opt to bypass these gray-boxed sections without losing access to critical information necessary for subsequent chapters.

Acknowledgments

A large number of people have played a vital role in shaping this book, both directly and indirectly. I am grateful to the diverse group of colleagues, collaborators, friends, and family who generously devoted their time to reading the drafts, offering valuable suggestions, and uncovering numerous mistakes.

This book is the result of a long journey involving the effort of colleagues, students, teaching assistants, and many others. I am immensely grateful to each and every person who contributed to its creation. Your involvement has enriched the content and made this scientific endeavor a truly rewarding experience. Thank you all for being part of this journey.

In particular, I owe a tremendous debt of gratitude to the Financial Computing and Analytics group at UCL, a community of brilliant and talented scientists, academicians, and students who have supported me in numerous ways. They have contributed directly by reading the draft, providing insightful feedback, and spotting mistakes, they have also contributed indirectly by inspiring me with their intellect and expertise. Thank you, Silvia Bartolucci, Paolo Barucca, Fabio Caccioli, Geoff Goodell, Denise Gorse, Giacomo Livan, Carolyn Phelan, Jiahua Xu and all other members and affiliates of this remarkable group. A special thanks to Guido Germano who has been particularly patient in spotting mistakes and suggesting changes, especially for what concerns the part of probability theory where he granted me permission to use his lecture notes. I express my deepest gratitude to Antonio Briola, who, coinciding with the time I dedicated to writing this book, has been my Ph.D. student. Antonio's involvement went far beyond mere discussions and suggestions for improvements. In fact, he is the major contributor to the GitHub repository, comprising Python codes and datasets that support this book. I thank my student Marwin Smith who diligently read the draft pointing out several errors. I am very grateful to two other exceptional Ph.D. students, Jeremy Turiel and Raymond Wang, who made substantial contributions to shaping the content of this book through discussions.

While I was writing, I sent the draft version of the book to colleagues who are specialists in their respective fields asking for guidance. I received a great

deal of very valuable feedback, suggestions, and corrections. Let me mention some of the people towards whom I feel most grateful. A special thanks must go to my esteemed colleague and friend Enrico Scalas, who read the whole draft manuscript spotting a multitude of mistakes and suggesting many useful and meaningful changes and several references. Thank you, Enrico, your contribution has been invaluable. Many thanks to Antonio Coniglio, who helped to make some better sense of the multifractality topic. Many thanks to Ginestra Bianconi who greatly helped to better draft the network part.

Finally, to my daughters, my partner, and my family, thank you for your unwavering support and understanding. I love you.

Symbols

$\mathrm{Cov}(X,Y)=\mathbb{E}((X-\mu_X)(Y-\mu_Y))$	covariance between random variables X and Y
$\mathrm{Corr}(X,Y)=\rho_{X,Y}$	correlation coefficient between X and Y
$f(X)$	probability density function (PDF)
$\hat{f}(X)$	estimate of the PDF
$\tilde{f}(X)$	model PDF
$F(X)=P(X \leq x)$	cumulative distribution function (CDF)
$1-F(X)=P(X>x)$	complementary CDF (CCDF)
$\hat{F}(X)$	estimate of the CDF
$g(X)$	function of the random variable X
$\mathcal{G}=(\mathbf{V},\mathbf{E})$	graph with vertex set $\mathbf{V}$ and edge set $\mathbf{E}$
$\mathbb{E}(g(X))$	expected value of $g(X)$
$H(X)$	entropy associated with random variable X
$\hat{H}(X)$	entropy estimate
$H(X\|Y)$	conditional entropy of X given Y
$I(X;Y)$	mutual information between X and Y
$\mathbf{J}=\mathbf{\Sigma}^{-1}$	inverse covariance matrix, or precision matrix
ℓ	log-likelihood
$\mathcal{L}$	likelihood
L_n	n-norm
$\\|\mathbf{M}\\|_n$	n-norm of matrix $\mathbf{M}$
$m_k=\mathbb{E}(X^k)$	kth moment
$\hat{m}_k=1/q\sum_{k=1}^{q}\hat{x}_k^k$	sample kth moment
$\mu=\mathbb{E}(X)$	expected value, mean
$\mu_k=\mathbb{E}((X-\mu)^k)$	kth central moment
$\hat{\mu}=1/q\sum_{k=1}^{q}\hat{x}_k$	sample mean
$\hat{\mu}_k=1/q\sum_{k=1}^{q}(\hat{x}_k-\hat{\mu})^k$	sample kth central moment
$\mathbb{N}$	natural numbers
$\mathcal{O}(\cdot)$	upper bound of an algorithm's complexity
$\sigma_X^2=\mathrm{Var}(X)=\mathbb{E}((X-\mu)^2)$	variance
$\hat{\sigma}_X^2=1/q\sum_{k=1}^{q}(\hat{x}_k-\hat{\mu})^2$	sample variance
σ_X	standard deviation
$\hat{\sigma}_X$	sample standard deviation
$\mathbf{\Sigma}$	covariance matrix
$\hat{\mathbf{\Sigma}}$	sample covariance matrix
$\|\mathbf{\Sigma}\|$	determinant of matrix $\mathbf{\Sigma}$
$\phi(\omega)=\mathbb{E}(e^{i\omega x})$	characteristic function

$P(X)$	probability measure of random variable X
$P(X, Y)$	joint probability of X and Y
$P(\mathbf{X})$	joint probability of the set of variables $\mathbf{X}$
$P(X\|Y)$	conditional probability of X given Y
$Q(\gamma)$	γ-quantile
$\mathbb{R}$	real numbers
$T_{X \to Y}$	transfer entropy of X to Y
$\\|\mathbf{v}\\|_n$	n-norm of vector $\mathbf{v}$
x	value (realization) of random variable X
$\hat{x}$	observed value (outcome) of a random variable X
X	random variable X
$\mathbf{X} = (X_1, ..., X_p)^\top$	column vector of a set of p random variables