Data Modeling for the Sciences

With the increasing prevalence of big data and sparse data, and rapidly growing data-centric approaches to scientific research, students must develop effective data analysis skills at an early stage of their academic careers. This detailed guide to data modeling in the sciences is ideal for students and researchers keen to develop their understanding of probabilistic data modeling beyond the basics of p-values and fitting residuals. The textbook begins with basic probabilistic concepts. Models of dynamical systems and likelihoods are then presented to build the foundation for Bayesian inference, Monte Carlo samplers, and filtering. Modeling paradigms are then seamlessly developed, including mixture models, regression models, hidden Markov models, state-space models and Kalman filtering, continuous time processes, and uniformization. The text is self-contained and includes practical examples and numerous exercises. This would be an excellent resource for courses on data analysis within the natural sciences, or as a reference text for self-study.

Steve Pressé is Professor of Physics and Chemistry at Arizona State University, Tempe. His research lies at the interface of biophysics and chemical physics with an emphasis on inverse methods. He has extensive experience in teaching data analysis and modeling at both undergraduate and graduate level with funding from the National Institutes of Health (NIH) and NSF in data modeling applied to the interpretation of single molecule dynamics and image analysis.

Ioannis Sgouralis is Assistant Professor of Mathematics at the University of Tennessee, Knoxville. His research is focused on computational modeling and applied mathematics, particularly the integration of data acquisition with data analysis across biology, chemistry, and physics.

Data Modeling for the Sciences

Applications, Basics, Computations

STEVE PRESSÉ
Arizona State University

IOANNIS SGOURALIS
University of Tennessee

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/9781009098502

DOI: 10.1017/9781009089555

First published 2023

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

Library of Congress Cataloging-in-Publication Data
Names: Pressé, Steve, author. | Sgouralis, Ioannis, author.
Title: Data modeling for the sciences : applications, basics, computations / Steve Pressé, Ioannis Sgouralis.
Description: Cambridge ; New York, NY : Cambridge University Press, 2023. | Includes bibliographical references and index.
Identifiers: LCCN 2023002178 (print) | LCCN 2023002179 (ebook) | ISBN 9781009098502 (hardback) | ISBN 9781009089555 (epub)
Subjects: LCSH: Research–Statistical methods. | Science–Statistical methods. | Probabilities. | Mathematical statistics.
Classification: LCC Q180.55.S7 P74 2023 (print) | LCC Q180.55.S7 (ebook) | DDC 001.4/22–dc23/eng20230506
LC record available at https://lccn.loc.gov/2023002178
LC ebook record available at https://lccn.loc.gov/2023002179

ISBN 978-1-009-09850-2 Hardback

In memoriam

SP: Adelina "Zia Adi" D'Orta (1912–2000);
IS: Pinelopi Sgourali (1906–2001).

Si l'on considère les méthodes analytiques auxquelles la théorie des probabilités a déjà donné naissance, et celles qu'elle peut faire naître encore, [...], si l'on observe ensuite que dans les choses même qui ne peuvent être soumises au calcul, cette théorie donne les aperçus les plus sûrs qui puissent nous guider dans nos jugemen[t]s, et qu'elle apprend à se garantir des illusions qui souvent nous égarent; on verra qu'il n'est point de science plus digne de nos méditations, et dont les résultats soient plus utiles.

[Considering analytical methods already engendered by the theory of probability, and those that could still arise, [...], and then considering that in those matters that do not lend themselves to [exact] calculation, this theory yields the surest of insights guiding us in our judgments, and teaching us to warrant against those illusions driving us astray; we will see that there exists no science worthier of our inquiry, and whose results are as useful.]

Comte Pierre-Simon de Laplace, *Théorie analytique des probabilités*, 1812

Contents

Preface

Data analysis courses that go beyond teaching elementary topics, such as fitting residuals, are rarely offered to students of the natural sciences. As a result, data analysis, much like programming, remains improvised. Yet, with an explosion of experimental methods generating large quantities of diverse data, we believe that students and researchers alike would benefit from a clear presentation of methods of analysis, many of which have only become feasible due to the practical needs and computational advances of the last decade or two.

The framework for data analysis that we provide here is inspired by new developments in data science, machine learning and statistics in a language accessible to the broader community of natural scientists. As such, this text is ambitiously aimed at making topics such as statistical inference, computational modeling, and simulation both approachable and enjoyable to natural scientists.

It is our goal, if nothing else, to help develop an appreciation for data-driven modeling and what data analysis choices are available alongside what approximations are inherent to the choices explicitly or implicitly made. We do so because theoretical modeling in the natural sciences has traditionally provided limited emphasis on data-driven approaches. Indeed, the prevailing philosophy is to first propose models and then verify or otherwise disprove these by experiments or simulations. But this approach is not data-centric. Nor is it rigorous except for the cleanest of data sets as one's perceived choice in how to compare, say, models and experiments may have dramatic consequences in whether the model is ultimately shown improbable. As we move toward monitoring events on smaller or faster timescales or complex events otherwise sparsely sampled, examples of clean data are increasingly few and far between.

Organization of the Text

We designed the text as a self-contained single semester course in data analysis, statistical modeling, and inference. Earlier versions have been used as class notes in a course at Arizona State University since 2017 for first-year chemistry and physics graduate students as well as upper-level undergraduates across the sciences and engineering. Since 2020, they have also been used in mathematics at the University of Tennessee. While

the text is appropriate for upper-level undergraduates in the sciences, its intended audience is at the master's level. The concepts presented herein are self-contained, though a basic course in computer programming and prior knowledge of undergraduate-level calculus is assumed.

Our text places equal emphasis on explaining the foundations of existing methods and their implementation. It correspondingly places little emphasis on formal proofs and research topics yet to be settled. Along core sections, we have interspersed sections and topics designated by an asterisk. These contain more advanced materials that may be included at the instructor's discretion and are otherwise not necessary upon a first reading. Similarly, we avoid long derivations in the text by marking designated equations with asterisks; these lengthy derivations are relegated to Appendix F.

Part I begins with a survey of modeling concepts to motivate the problem of parameter estimation from data. This leads to a discussion of frequentist and Bayesian inference tools. Along the way, we introduce computational techniques including Monte Carlo methods necessary for a comprehensive exposition of the most recent advances. Part II is devoted to specific models starting from basic mixture models followed by Gaussian processes, hidden Markov models and their adaptations, as well as models appropriate to continuous space and time.

In writing some end-of-chapter exercises, we are reminded of a quote from J. S. Bach (1723) as a prefatory note to his own keyboard exercises (two and three part inventions). That is, we not only wish to inspire a clear way by which to tackle data analysis problems but also create a strong foretaste for the proper independent development of the reader's own analysis tools. Indeed, some end-of-chapter exercises bring together notions intended to broaden the reader's scope of what is possible and spark their interest in developing further inference schemes as complex and realistic as warranted by the application at hand.

Finally, we made clear choices on what topics to include in the book. These were sometimes based on personal interest, though, most often, these choices were based on what we believe is most relevant. To keep our presentation streamlined, however, we have excluded many topics. Some of these we perceive as easier for students to understand after reading this book, such as specialized cases of topics covered herein.

Acknowledgments

First and foremost, we thank our families for their support and for patiently staying by our side as we undertook this effort.

We also thank Weiqing Xu for helping generate some of the figures and Alex Rojewski for her critical help with the solution manual; Weiqing Xu, Sina Jazani, and Zeliha Kilic for suggesting scientific arguments used in select chapters; Banu Ozkan, Julian Lee, and Corey Weistuch for reading over the entirety of earlier drafts of the text; John Fricks for his thoughts on the chapter on foundations; Bob Zigon for pointing us to analysis topics that eventually grew into our interest in nonparametrics; Oliver Beckstein for his thoughts on likelihood maximization; Carlos Bustamante for providing his time to share his contagious passion for (stochastic) single molecule data; Ken Dill for his inspiration in pursuing inverse methods; anonymous referees who provided feedback on the text; and funding agencies for their continued support (Pressé acknowledges the NSF, NIH, and Army Research Office (ARO) and Sgouralis acknowledges his University of Tennessee startup and Eastman Chemical Company).

We also thank the many members of the Pressé lab and students at Arizona State University taking CHM/PHY 598 ("Unraveling the noise") who have proposed and inspired homework problems and identified typos and confusions across earlier drafts of the text. Finally, we thank our many experimental colleagues, often first encountered over the course of Telluride (TSRC) workshops, who have suggested multiple topics of interest and inspired us to learn many of the methods presented herein.

Any remaining typos and omissions are ours alone.

Expanded Note for Instructors

Various iterations of this course have been taught since 2017 at ASU where the course was cross-listed in both physics and chemistry. The course was taken by senior undergraduates and first year graduate students from both of these disciplines as well as by students from various engineering disciplines. The course has also been offered for two years at UTK in mathematics.

In a one-semester course, we cover Chapters 1–7 and the beginning of Chapter 8 (as time allows) in the order in which the material is presented. We exclude all topics labeled advanced and skip all extensive derivations, which are relegated to the appendices. We do otherwise summarize the simpler derivations appearing in the text. Taught from start to finish, without skipping advanced material, the text stands as a complete two-semester course in data modeling.

In presenting the material, we created boxed environments specifically with students and instructors in mind. For instance, the Note environments highlight special issues we hope will not elude the reader. The Example environments are meant to tie the mathematical material back to applications. Most important are the Algorithm environments, which focus on implementation of the ideas presented. The algorithms are intentionally presented in a general manner independent of any coding language. For this reason, students in class have typically used the coding language of their choice in the problem sets.

The end-of-chapter problems appear in two forms: exercises and projects. The exercises are simpler and we provide detailed solutions of sample exercises to verified instructors that we have used in our own offering of the course. The solutions we provide exceed what would be needed to cover weekly or biweekly problem sets for a one-semester course.

The projects, however, are more demanding and are meant to inspire the formulation of broader and deeper questions that may be addressed using the tools presented in each chapter. Indeed, some projects may be worthy of independent publication when completed.

The index is ordered by topic for ease of reference. For instance, "friction coefficient" will be found as a subitem under "physics" Similarly, specific algorithms (that may otherwise introduce confusion by being indexed by long or abbreviated names) simply appear as subitems under "algorithm."

Finally, we appreciate that the material we present is normally not available to students of the natural sciences as it would require advanced prerequisites in probability and stochastic processes that cannot otherwise

fit into their schedules. For this reason, we pay special attention to notation and concepts that would not be familiar to students of the sciences in order to allow them to work from the very basics (Chapter 1) to state-of-the-art modeling (Chapters 6–10). We relegate theoretical topics in probability to an appendix entitled "Foundational Aspects of Probabilistic Modeling."