Data Modeling for the Sciences

Applications, Basics, Computations

Steve Pressé and Ioannis Sgouralis

(This draft was last modified on August 12, 2022)

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 already 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 were used as class notes in a course at Arizona State University since 2017 to 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 an exclamation mark; these lengthy derivations are relegated to appendix F at the end of the text.

The text 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. The second half of the text is devoted to specific models starting from basic mixture models followed by Gaussian processes, hidden Markov models, 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 JS Bach (1723) as a prefactory note to his own keyboard exercises (two and three part inventions). That is, we not only wish to inspire a clear way by means 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 provided to the reader 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 include topics that we perceive as easier for students to understand after reading this book, such as specialized cases of topics covered herein.

Note to the instructors– Various iterations of this course have been taught for 5 years at ASU in Physics and Chemistry and 2 years at UTK in Mathematics. In a one semester course, we cover chapters 1 and 6 and the beginning of chapter 8 (as time allows). We also exclude all topics labeled advanced. Taught from start to finish, the text would be more appropriate as a two semester course.

“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 judgements, 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].

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

Contents

*This is an advanced topic and could be skipped on a first reading.

*This is an advanced topic and could be skipped on a first reading.

II Statistical models 181

*This is an advanced topic and could be skipped on a first reading.

*This is an advanced topic and could be skipped on a first reading.