

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 prefatory 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

I	Concepts from modeling, inference, and computing	15
1	Probabilistic modeling and inference	17
1.1	Modeling with data	17
1.1.1	Why do we obtain models from raw data?	17
1.1.2	Why do we formulate models with random variables?	18
1.1.3	Why do our models have parameters?	19
1.2	Working with random variables	20
1.2.1	How to assign probability distributions	20
1.2.2	How to sample from probability distributions	29
1.2.3	Manipulating probability distributions	33
1.3	Data-driven modeling and inference	37
1.4	Exercise problems	40
2	Dynamical systems and Markov processes	45
2.1	Why do we care about stochastic dynamical models?	45
2.2	Forward models of dynamical systems	46
2.3	Systems with discrete state-spaces in continuous time	48
2.3.1	Modeling a system with discrete events	49
2.3.2	Markov jump processes	52
2.3.3	Structured Markov jump processes*	55
2.3.4	Global descriptions of Markov jump processes*	65
2.4	Systems with discrete state-spaces in discrete time	71
2.4.1	Modeling a system at discrete times	71
2.4.2	Modeling kinetic schemes	72
2.4.3	Quantifying state persistence*	75
2.5	Systems with continuous state-spaces in discrete time	76
2.5.1	Modeling under equations of motion	77
2.5.2	Modeling under increments	81
2.5.3	<i>A case study in Langevin dynamics and Brownian motion*</i>	82
2.6	Systems with continuous state-spaces in continuous time	87
2.6.1	Modeling with stochastic differential equations	87
2.6.2	The Fokker-Planck equation	88
2.6.3	Deriving the Fokker-Planck equation*	89
2.6.4	<i>A case study in thermal physics*</i>	90
2.7	Exercise problems	92

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

3	Likelihoods and latent variables	99
3.1	Quantifying measurements with likelihoods	99
3.1.1	Estimating parameters with maximum likelihood	100
3.1.2	Likelihood maximization as an optimization problem*	101
3.2	Observations and associated measurement noise	105
3.2.1	Completed likelihoods	107
3.2.2	The EM algorithm*	108
3.3	Exercise problems	112
4	Bayesian inference	117
4.1	Modeling in Bayesian terms	117
4.1.1	The posterior distribution	117
4.1.2	Bayesian data analysis: the big picture	122
4.2	The logistics of Bayesian formulations: priors	123
4.2.1	Uninformative priors	123
4.2.2	Informative priors	124
4.2.3	Conjugate prior-likelihood pairs	124
4.2.4	Conjugate priors and the exponential family	126
4.2.5	Informative priors for Normal likelihoods	127
4.3	EM for posterior maximization*	129
4.4	Hierarchical Bayesian formulations and graphical representations	130
4.5	Bayesian model selection*	133
4.5.1	The Bayesian information criterion	133
4.5.2	Why the BIC works?	134
4.5.3	<i>A case study in change point detection</i>	134
4.6	Information theory*	137
4.7	Exercise problems	138
5	Computational inference	141
5.1	The fundamentals of statistical computation	141
5.1.1	Monte Carlo methods	141
5.1.2	Markov chain Monte Carlo methods	144
5.1.3	Monte Carlo Markov chain requirements	145
5.2	Basic MCMC samplers	146
5.2.1	Metropolis-Hastings family of samplers	146
5.2.2	Gibbs family of samplers	156
5.3	Processing and interpretation of MCMC	162
5.3.1	Assessing convergence	163
5.3.2	Burn-in removal	163
5.3.3	Thinning	164
5.4	Advanced MCMC samplers*	165
5.4.1	Multiplicative random walk samplers	165
5.4.2	Within Gibbs sampling schemes	167
5.4.3	Auxiliary variable samplers	168
5.4.4	Metropolis-Hastings samplers with deterministic proposals	170
5.4.5	Hamiltonian MCMC samplers	171
5.5	Exercise problems	176

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

II	Statistical models	181
6	Regression models	183
6.1	The regression problem	183
6.2	Nonparametric regression in continuous space: Gaussian process	186
6.2.1	Covariance kernel	187
6.2.2	Sampling the GaussianP	189
6.2.3	GaussianP priors and posteriors	189
6.2.4	Predictive distribution and inducing points	191
6.2.5	GaussianP regression with non-conjugate likelihoods*	192
6.3	Nonparametric regression in discrete space: Beta Process Bernoulli Process*	195
6.3.1	Posterior convergence of the BetaBernP	196
6.4	Exercise problems	197
7	Mixture models	205
7.1	Mixture model formulations with observations	206
7.1.1	Representations of a mixture distribution	206
7.1.2	Likelihoods for MMs	207
7.1.3	State-space labeling and likelihood invariance*	208
7.2	MM in the Bayesian paradigm	209
7.2.1	Classification: Estimating categories	209
7.2.2	Mixture weights and the Dirichlet distribution	210
7.2.3	Estimating weights and other parameters	214
7.3	The infinite MM and the Dirichlet process*	215
7.4	Exercise problems	217
8	Hidden Markov models	221
8.1	Introduction	221
8.2	The Hidden Markov Model	222
8.2.1	Modeling dynamics	222
8.2.2	Modeling overview	223
8.3	The Hidden Markov Model in the frequentist paradigm	224
8.3.1	Evaluation of the likelihood	224
8.3.2	Decoding the state sequence	226
8.3.3	Estimation of the parameters	228
8.3.4	Some computational considerations	232
8.3.5	State-space labeling and likelihood*	234
8.4	The Hidden Markov Model in the Bayesian paradigm	235
8.4.1	Priors for the HMM	236
8.4.2	MCMC inference in the Bayesian HMM	236
8.4.3	Interpretation and label switching*	241
8.5	Dynamical variants of the Bayesian HMM*	242
8.5.1	Modeling time scales	243
8.5.2	Modeling equilibrium	244
8.5.3	Modeling reversible systems	244
8.5.4	Modeling kinetic schemes	245
8.5.5	Modeling factorial dynamics	245
8.6	The infinite Hidden Markov Model*	247
8.7	<i>A case study in fluorescence spectroscopy*</i>	248
8.7.1	Time resolved spectroscopy	248
8.7.2	Discretization of time	249

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

8.7.3	Formulation of the dynamics	250
8.7.4	Formulation of the measurements	250
8.7.5	Modeling overview	251
8.7.6	Reformulation	252
8.7.7	Computational training	254
8.7.8	Bayesian considerations	258
8.8	Exercise problems	259
9	State-space models	263
9.1	State-space models	263
9.2	Gaussian state-space models	265
9.3	Linear Gaussian state-space models	266
9.3.1	Filtering	266
9.3.2	Smoothing	268
9.3.3	Forecasting	269
9.3.4	Simulation	270
9.4	Bayesian state-space models and estimation	270
9.5	Exercise problems	272
10	Continuous time models*	275
10.1	Modeling in continuous time	275
10.2	MJP uniformization and virtual jumps	276
10.2.1	Uniformization sampler	277
10.2.2	Why does uniformization work?	277
10.3	Hidden MJP sampling with uniformization and filtering	278
10.3.1	Embedding uniformization into the MJP trajectory sampler	278
10.3.2	Irregular grid filtering in hidden MJP sampling	279
10.4	Sampling trajectories and model parameters	279
10.4.1	Hidden MJP formulation with transition rates	279
10.4.2	Hidden MJP formulation with transition probabilities	280
10.4.3	Trajectory marginalization	281
10.5	Exercise problems	282
III	Appendix	285
A	Notation and other conventions	287
A.1	Time and other physical quantities	287
A.2	Random variables and other mathematical notions	287
A.3	Collections	287
B	Numerical random variables	289
B.1	Continuous random variables	289
B.1.1	With bounded support	289
B.1.2	With semi-bounded support	290
B.1.3	With unbounded support	292
B.2	Discrete random variables	294
B.2.1	With bounded support	294
B.2.2	With semi-bounded support	295

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

C	The Kronecker and Dirac deltas	299
C.1	Kronecker Δ	299
C.2	Dirac δ	299
C.2.1	Definition	299
C.2.2	Properties	301
D	Memoryless distributions	303
E	Foundational aspects of probabilistic modeling	305
E.1	Outcomes and events	305
E.2	The measure of probability	308
E.3	Random variables	312
E.4	The measurables	313
E.5	A comprehensive modeling overview	315
F	Derivation of key relations	317
F.1	Relations of chapter 2	317
F.2	Relations of chapter 3	318
F.3	Relations of chapter 5	320
F.4	Relations of chapter 6	322
F.5	Relations of chapter 7	325
F.6	Relations of chapter 8	327
F.7	Relations of chapter 9	332
F.8	Relations of chapter 10	337
	Index of terms	339