Data-Driven Computational Methods

Parameter and Operator Estimations

Modern scientific computational methods are undergoing a transformative change; big data and statistical learning methods now have the potential to outperform the classical first-principles modeling paradigm. This book bridges this transition, connecting the theory of probability, stochastic processes, functional analysis, numerical analysis, and differential geometry. It describes two classes of computational methods to leverage data for modeling dynamical systems. The first is concerned with data fitting algorithms to estimate parameters in parametric models that are postulated on the basis of physical or dynamical laws. The second is on operator estimation, which uses the data to nonparametrically approximate the operator generated by the transition function of the underlying dynamical systems.

This self-contained book is suitable for graduate studies in applied mathematics, statistics, and engineering. Carefully chosen elementary examples with supplementary MATLAB® codes and appendices covering the relevant prerequisite materials are provided, making it suitable for self-study.

John Harlim is a Professor of Mathematics and Meteorology at the Pennsylvania State University. His research interests include data assimilation and stochastic computational methods. In 2012, he received the Frontiers in Computational Physics award from the *Journal of Computational Physics* for his research contributions on computational methods for modeling Earth systems. He has previously co-authored another book, *Filtering Complex Turbulent Systems* (Cambridge, 2012).

Data-Driven Computational Methods

Parameter and Operator Estimations

JOHN HARLIM
The Pennsylvania State University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, 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

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108472470
DOI: 10.1017/9781108562461

First published 2018

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

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

ISBN 978-1-108-47247-0 Hardback

This book is dedicated to the joys of my life,
my wife, Leonie, and my son, Kelvin.

I also dedicate this book to my parents, Gemiati and Siang Jong,
who have made my childhood dream to become a student forever come true.

Contents

Preface

Stochastic modeling of dynamical systems has been instrumental in various applied fields, including material sciences, atmospheric and ocean sciences, biology, chemistry, etc. With the rapid advancement in data collection, an important emerging scientific discipline is to leverage this new information to improve modeling and prediction of complex dynamical systems. While the theory of stochastic processes is a well-established field, the development of the data-driven computational tools for their practical implementation has emerged to become an important new discipline in applied mathematics and engineering science. The aim of this book is to provide a survey of such computational tools. In particular, the book covers computational methods for stochastic modeling of dynamical systems.

In general, there are two classes of mathematical/statistical modeling: parametric and nonparametric paradigms. Arguably, one can merge these two classes and invent a semi-parametric paradigm. Parametric modeling of complex dynamical systems usually involves proposing a model based on some physical laws (such as Newtonian, conservation laws, etc.), inferring the model parameters from the available observed data, and verifying the results against the observables. In contrast, one can also use the data to build nonparametric models with minimal assumptions on the underlying dynamics. In such an approach, the notion of nonparametric modeling follows from the standard statistical literature which makes no assumption either about how the dynamics should behave or about the distribution of the underlying dynamics. Instead, we let the data determine the dynamics.

In this book, we discuss computational methods both for the parametric approach and for the non-parametric approach. For the parametric approach, our choice will be to employ parameter estimation methods with Bayesian inference, which have been widely used in many applications. For this topic, we cover two basic approaches. The first one is the Markov-Chain Monte Carlo (MCMC) method in Chapter 2, which aims to estimate the distribution of the hidden parameters of dynamical systems from the noisy data. The second method is the ensemble Kalman filter (EnKF) in Chapter 3, which has been successfully used in numerical weather forecasting applications. EnKF is complementary to MCMC in the sense that it estimates only the first- and second-order moments of

the hidden parameters of dynamical systems from the noisy data. In this book, we will neglect the non-Bayesian parameter estimation techniques.

For the nonparametric modeling of dynamical systems, we will provide a rigorous treatment of a recently developed computational method, the so-called diffusion forecasting model, which allows one to approximate the solution operator corresponding to the Fokker–Planck equation of the Itô diffusion completely from the data. One of the main emphases of this book is the intention to show readers that this nonparametric approach is a natural generalization of the central idea in uncertainty quantification (UQ), namely the representation of random variables with a linear superposition of polynomial basis functions of appropriate Hilbert space. In the traditional UQ approach, one usually chooses the polynomial basis functions by assuming that the random variables that are to be represented belong to a certain class of known distributions on some Euclidean domain. The diffusion forecast generalizes this idea by representing the semigroup operator generated by the transition function of the Itô drifted diffusion processes with basis functions that are purely constructed from the data that lies on a (possibly non-Euclidean) manifold. We shall see that the diffusion forecasting approach is indeed a spectral Galerkin method that uses the data-driven basis functions to represent the Fokker–Planck equation nonparametrically. To facilitate this generalization viewpoint, we give a brief review of a basic non-data-driven UQ approach, namely the stochastic spectral method with polynomial chaos expansion in Chapter 4. Since the construction of the data-driven basis functions relies on a kernel-based manifold learning method, namely the so-called diffusion maps algorithm, we provide a review of the classical Karhunen–Loève expansion in Chapter 5. The key point is to show that proper orthogonal decomposition (POD), which is a popular linear manifold learning algorithm, is an application of the Karhunen–Loève expansion that exploits Mercer's theorem. The theoretical discussion in this chapter, which ties together the eigenfunctions of kernel-based integral operators and the orthonormal basis functions of a Hilbert space, will become handy in understanding the construction of the diffusion maps algorithm. These two chapters are included to give more solid understanding of the operator estimation technique discussed in Chapter 6.

This book is designed for applied mathematicians and engineers ranging from first-year graduate students to senior researchers interested in leveraging data to model stochastic dynamics. Selected elementary examples, together with the MATLAB® scripts (in the supplementary material), are provided to help readers' self-study. While we expect readers to be familiar with basic probability theory, stochastic processes, and differential geometry language, they are not essential. In fact, we provide three appendices reviewing these basic materials.

Acknowledgments

Parts of this book are summaries based on the author's joint works with Andrew Majda, Tyrus Berry, Dimitris Giannakis, Xiantao Li, Adam Mahdi, Haizhao

Yang, Shixiao Jiang, and Yicun Zhen. The author thanks these colleagues for their explicit and implicit contributions to this material. The author also thanks Tyrus Berry, Nan Chen, Wen Shen, Xin Tong, He Zhang and anonymous reviewers for reading through the manuscript and their suggestions.

Special thanks are due to Juliani and Guy Vachon for their hospitality. A large part of this book was written during my stay at their home in Asheville, North Carolina.

The author gratefully acknowledges the generous support from the Office of Naval Research through Reza Malek-Madani and Scott Harper and from the National Science Foundation through Leland Jameson. These research funds made this book a reality. Special thanks are due to the students in the graduate course MA597E Uncertainty Quantification Methods in Fall 2015 at Pennsylvania State University, who motivated the author to assemble this text. The author also thanks the undergraduate students in the REU program in Summer 2016 who worked on a few of the examples in this book.

John Harlim
University Park, PA