Bayesian Models for Astrophysical Data

Using R, JAGS, Python, and Stan

This comprehensive guide to Bayesian methods in astronomy enables hands-on work by supplying complete R, JAGS, Python, and Stan code, to use directly or to adapt. It begins by examining the normal model from both frequentist and Bayesian perspectives and then progresses to a full range of Bayesian generalized linear and mixed or hierarchical models, as well as additional types of models such as ABC and INLA. The book provides code that is largely unavailable elsewhere and includes details on interpreting and evaluating Bayesian models. Initial discussions offer models in synthetic form so that readers can easily adapt them to their own data; later the models are applied to real astronomical data. The consistent focus is on hands-on modeling, analysis of data, and interpretations that address scientific questions. A must-have for astronomers, its concrete approach will also be attractive to researchers in the sciences more generally.

Joseph M. Hilbe is Solar System Ambassador with NASA's Jet Propulsion Laboratory, Adjunct Professor of Statistics at Arizona State University, and Professor Emeritus at the University of Hawaii. He is currently President of the International Astrostatistics Association (IAA) and was awarded the IAA's 2016 Outstanding Contributions to Astrostatistics medal, the association's top award. Hilbe is an elected Fellow of both the American Statistical Association and the IAA and is a full member of the American Astronomical Society. He has authored 19 books on statistical modeling, including leading texts on modeling count and binomial data. His book, Modeling Count Data (2014, Cambridge) received the 2015 PROSE honorable mention for books in mathematics.

Rafael S. de Souza is a researcher at Eötvös Loránd University. He is currently Vice-President for development of the International Astrostatistics Association and was awarded the IAA's 2016 Outstanding Publication in Astrostatistics award. He has authored dozens of scientific papers, serving as the leading author for over 20.

Emille E. O. Ishida is a researcher at the Université Clermont-Auvergne (Université Blaise Pascal). She is cochair of the Cosmostatistics Initiative and coordinator of its Python-related projects. She is a specialist in machine learning applications to astronomy with special interests in type Ia supernovae spectral characterization, classification, and cosmology. She has been the lead author of numerous articles in prominent astrophysics journals and currently serves as chair of the IAA public relations committee.

Bayesian Models for Astrophysical Data

Using R, JAGS, Python, and Stan

JOSEPH M. HILBE
Arizona State University and Jet Propulsion Laboratory, California Institute of Technology

RAFAEL S. DE SOUZA
Eötvös Loránd University, Budapest

EMILLE E. O. ISHIDA
Université Clermont-Auvergne (Université Blaise Pascal), France

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

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi – 110002, 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/9781107133082

First published 2017

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

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

Library of Congress Cataloging-in-Publication Data
Names: Hilbe, Joseph M., 1944- | De Souza, Rafael S. | Ishida, Emille E. O.
Title: Bayesian models for astrophysical data using R, JAGS, Python, and Stan /
Joseph M. Hilbe, Jet Propulsion Laboratory and Arizona State University,
Rafael S. de Souza, MTA Eötvös Loránd University, Emille E.O. Ishida,
Université de Clermont-Ferrand II (Université Blaise Pascal), France.
Description: Cambridge : Cambridge University Press, 2017. | Includes
bibliographical references and index.
Identifiers: LCCN 2017009823 | ISBN 9781107133082 (alk. paper)
Subjects: LCSH: Statistical astronomy. | Statistical astronomy–Data
processing. | Astronomy–Data processing.
Classification: LCC QB149 .H55 2017 | DDC 520.1/519542–dc23
LC record available at https://lccn.loc.gov/2017009823

ISBN 978-1-107-13308-2 Hardback

Additional resources for this publication at www.cambridge.org/bayesianmodels

In Memoriam

Joseph M. Hilbe
(30th December, 1944 – 12th March, 2017)

Our mentor, colleague, and friend.

May we all live up to your legacy.

Farewell Joe. It was an honour to meet you.

Contents

Color plates are to be found between pages 198 and 199.

Preface

Bayesian Models for Astrophysical Data provides those who are engaged in the Bayesian modeling of astronomical data with guidelines on how to develop code for modeling such data, as well as on how to evaluate a model as to its fit. One focus in this volume is on developing statistical models of astronomical phenomena from a Bayesian perspective. A second focus of this work is to provide the reader with statistical code that can be used for a variety of Bayesian models.

We provide fully working code, not simply code snippets, in R, JAGS, Python, and Stan for a wide range of Bayesian statistical models. We also employ several of these models in real astrophysical data situations, walking through the analysis and model evaluation. This volume should foremost be thought of as a guidebook for astronomers who wish to understand how to select the model for their data, how to code it, and finally how best to evaluate and interpret it. The codes shown in this volume are freely available online at www.cambridge.org/bayesianmodels. We intend to keep it continuously updated and report any eventual bug fixes and improvements required by the community. We advise the reader to check the online material for practical coding exercises.

This is a volume devoted to applying Bayesian modeling techniques to astrophysical data. Why Bayesian modeling? First, science appears to work in accordance with Bayesian principles. At each stage in the development of a scientific study new information is used to adjust old information. As will be observed when reviewing the examples later in this volume, this is how Bayesian modeling works. A posterior distribution created from the mixing of the model likelihood (derived from the model data) and a prior distribution (outside information we use to adjust the observed data) may itself be used as a prior for yet another enhanced model. New information is continually being used in models over time to advance yet newer models. This is the nature of scientific discovery. Yet, even if we think of a model in isolation from later models, scientists always bring their own perspectives into the creation of a model on the basis of previous studies or from their own experience in dealing with the study data. Models are not built independently of the context, so bringing in outside prior information to the study data is not unusual or overly subjective. Frequentist statisticians choose the data and predictors used to study some variable – most of the time based on their own backgrounds and external studies. Bayesians just make the process more explicit.

A second reason for focusing on Bayesian models is that recently there has been a rapid move by astronomers to Bayesian methodology when analyzing their data. Researchers in many other disciplines are doing the same, e.g., in ecology, environmental science, health outcomes analysis, communications, and so forth. As we discuss later, this is largely the case because computers are now finally able to engage with complex MCMC-based algorithms, which entail thousands of sampling iterations and many millions of calculations in arriving at a single posterior distribution. Moreover, aside from faster computers with much greater memory, statisticians and information scientists have been developing ever more efficient estimation algorithms, which can now be found in many commercial statistical software packages as well as in specially developed Bayesian packages, e.g., JAGS, Stan, OpenBUGS, WinBUGS, and MLwiN. Our foremost use in the book is of JAGS and Stan. The initial release of JAGS by Martyn Plummer was in December 2007. Stan, named after Stanislaw Ulam, co-developer of the original MCMC algorithm in 1949 with Nicholas Metropolis (Metropolis and Ulam, 1949), was first released in late August 2012. However, the first stable release of Stan was not until early December 2015, shortly before we began writing this volume. In fact, Stan was written to overcome certain problems with the convergence of multilevel models experienced with BUGS and JAGS. It is clear that this book could not have been written a decade ago, or even five years ago, as of this writing. The technology of Bayesian modeling is rapidly advancing, indicating that astrostatistics will be advancing with it as well. This book was inspired by the new modeling capabilities being jointly provided by the computer industry and by statisticians, who are developing better methods of analyzing data.

Bayesian Models for Astrophysical Data differs from other books on astrostatistics. The book is foremost aimed to provide the reader with an understanding of the statistical modeling process, and it displays the complete JAGS and, in most cases, Stan code for a wide range of models. Each model is discussed, with advice on when to use it and how best to evaluate it with reference to other models. Following an overview of the meaning and scope of statistical modeling, and of how frequentist and Bayesian models differ, we examine the basic Gaussian or normal model. This sets the stage for us to then present complete modeling code based on synthetic data for what may be termed Bayesian generalized linear models. We then extend these models, discussing two-part models, mixed models, three-parameter models, and hierarchical models. For each model we show the reader how to create synthetic data based on the distributional assumptions of the model being evaluated. The model code is based on the synthetic data but because of that it is generic and can easily be adapted to alternative synthetic data or to real data. We provide full JAGS and Stan code for each model. In the majority of the examples in this volume, JAGS code is run from the R environment and Stan from within the Python environment. In many cases we also display the code for, and run, stand-alone R and Python models.

Following our examination of models, including continuous, binary, proportion, grouped, and count response models we address model diagnostics. Specifically, we discuss information criteria including the Bayesian deviance, the deviance information criterion (DIC), and the pD and model predictor selection methods, e.g., the Kuo and Mallick test and Bayesian LASSO techniques. In Chapter 10 on applications we bring in real astronomical data from previously published studies and analyze them using the models discussed earlier in the book. Examples are the use of time series for sunspot events, lognormal models for the stellar initial mass function, and errors in variables for the analysis of supernova properties. Other models are discussed as well, with the likelihood-free approximate Bayesian computation (ABC) presented alongside a pure Python computation as an alternative for analyzing Sunyaev–Zeldovich surveys.

This book should be thought of as a guidebook to help researchers develop more accurate and useful models for their research studies. In this regard we do not go into deep details of an underlying astrophysical model but use it as a template, so the researcher can understand how to connect a given statistical model with the data at hand and, therefore, how to apply to it to his or her own problem. It is a new approach to learning how to apply statistics to astronomical research, but we believe that the pedagogy is sound.

It is important for astrostatisticians and astroinformaticists to read a variety of books and articles related to the statistical analysis of astronomical data. Doing so enhances the analyst's research acumen. Below we briefly point out the themes or approaches given in three other books that have recently been published on astrostatistics. We do not regard these books as competitors. Rather, they should be thought of as complements to this book.

In our view, Feigelson and Babu (2012a) is generally regarded by astronomical and general astrostatistical communities as the standard text on astrostatistics from the frequentist perspective. The text provides readers with a wide range of both parametric and non-parametric methods for evaluating astronomical data; R is used throughout for examples. We recommend this text to astronomers who wish to have an up-to date volume on astrostatistics from the frequentist tradition. The authors are planning a second edition, which will likely be out in 2018.

The other two books we recommend to accompany the present text are volumes in the *Springer Series on Astrostatistics*, which is co-sponsored by the International Astrostatistics Association (IAA). Andreon and Weaver (2015) provides a Bayesian approach to the physical sciences, with an emphasis on astronomy. It is closest to this volume in approach. JAGS code is displayed in the book for many examples. The emphasis is on Gaussian-based models, although examples using Bayesian Poisson and logistic models are examined. The book provides a number of nicely presented guidelines and theory. Our third recommendation is the general purpose astrostatistics text by Chattopadhyay and Chattopadhyay (2014), which is frequentist oriented but has a very nice component on Monte Carlo simulation. The R code is used for examples. All three books are excellent resources, but they substantially differ from this volume.

The aim of our book is to provide astrostatisticians with the statistical code and information needed to create insightful and useful models of their data. Much of the code presented here cannot be found elsewhere, but it can be used to develop important models of astronomical data. It is central to us to provide readers with better statistical tools and code for advancing our understanding of the Universe. If this book has achieved this goal to any degree, it has been worth the effort.

We assume that readers have a basic background in frequency-based statistical modeling, and maximum likelihood estimation in particular. We do not expect that readers have gone beyond normal or basic linear regression, but having at least some background in logistic regression or Poisson regression will be helpful. It is not necessary, though. It will also be helpful if the reader knows the basics of R and Python programming. We use R as the base language for examples but make every attempt to explain the code as we go along. For most examples complete commented Python scripts are provided, allowing the reader to benefit also from a direct comparison between these programming languages. In addition, we do not expect that readers have a background in Bayesian modeling – the subject of the text – but the more you know already, the better.

Owing to these assumptions we cover frequency-based modeling concepts rather quickly, touching on only major points to be remembered when contrasting frequency-based and Bayesian methodologies. We provide an overview of Bayesian modeling and of how it differs from frequentist-based modeling. However, we do not focus on theory. There are a plethora of books and other publications on this topic. We do attempt, however, to provide sufficient background on Bayesian methodology to allow the reader to understand the logic and purpose of the Bayesian code discussed in the text. Our main emphasis is to provide astrostatisticians with the code and the understanding to employ models on astrophysical data that have previously not been used, or have only seldom been used – but which perhaps should be used more frequently. We provide modeling code and diagnostics using synthetic data as well as using real data from the literature.

We should mention that researchers in disciplines other than astrophysics may also find the book useful. The code and discussion using synthetic data are applicable to nearly all disciplines.

We are grateful to a number of our colleagues for their influence on this work. Foremost we wish to acknowledge Alain F. Zuur of Highlands Statistics in Newburgh, Scotland for his contributions to several of the JAGS models used in the book. The codes for various models are adaptations of code from Zuur, Hilbe, and Ieno (2013), a book on both the frequentist and the Bayesian approaches to generalized linear models and generalized linear mixed models for ecologists. Moreover, we have adopted a fairly uniform style or format for constructing JAGS models on this basis of the work of Dr. Zuur and the first author of this volume, which is reflected in Zuur, Hilbe, and Ieno (2013).

We would like to express our appreciation to Diana Gillooly, statistics editor at Cambridge University Press, for accepting our proposal to write this book. She has been supportive throughout the life of the book's preparation. We thank Esther Miguéliz, our Content Manager, and Susan Parkinson, our freelance copy-editor, for their dedicated work in improving this book in a number of ways. Their professionalism and assistance is greatly appreciated. We also wish to express our gratitude to John Hammersley, CEO of Overleaf (WriteLatex Ld), who provided us with Overleaf Pro so that we could work simultaneously on the manuscript. This new technology makes collaborative authorship endeavors much easier than in the past. Since we live and work in Arizona, Hungary, and France respectively, this was an ideal way to write the book.

Finally, we each have those in our personal lives who have contributed in some way to the creation of this volume.

The third author would like to thank Wolfgang Hillebrandt and Emmanuel Gangler for providing unique working environments which enabled the completion of this project. In addition, she would like to acknowledge all those who supported the Cosmostatistics Initiative and its Residence Programs, where many applications described in this volume were developed. Special thanks to Alberto Krone-Martins, Alan Heavens, Jason McEwen, Bruce Bassett, and Zsolt Frei as well as her fellow co-authors.

The second author thanks all members of the Cosmostatistics Initiative. Particular thanks to Ewan Cameron, Alberto Krone-Martins, Maria Luiza Linhares Dantas, Madhura Killedar, and Ricardo Vilalta, who have demonstrated their incredible support for our endeavor.

The first author wishes to acknowledge Cheryl Hilbe, his wife, who has supported his taking time from other family activities to devote to this project. In addition, he also wishes to expressly thank Eric Feigelson for advice and support over the past seven years as he learned about the astrophysical community and the unique concerns of astrostatistics. Professor Feigelson's experience and insights have helped shape how he views the discipline. He also acknowledges the true friendship which has evolved between the authors of this volume, one which he looks forward to continuing in the coming years. Finally, he dedicates this book to two year old Kimber Lynn Hilbe, his granddaughter, who will likely be witness to a future world that cannot be envisioned by her grandfather.