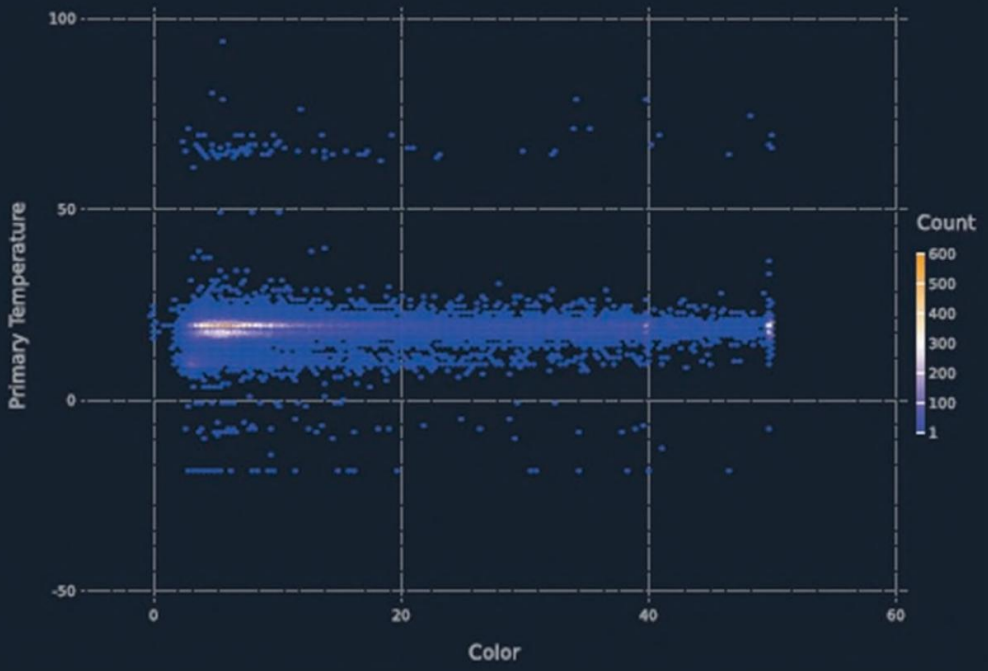

Data Science with Julia

Paul D. McNicholas
Peter A. Tait

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Data Science with Julia

By Paul D. McNicholas and Peter A. Tait

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20191119

International Standard Book Number-13: 978-1-138-49998-0 (Paperback)

Library of Congress Cataloging-in-Publication Data

Names: McNicholas, Paul D., author. | Tait, Peter A., author.
Title: Data science with Julia / Paul D. McNicholas, Peter A. Tait.
Description: Boca Raton : Taylor & Francis, CRC Press, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2018025237 | ISBN 9781138499980 (pbk.)
Subjects: LCSH: Julia (Computer program language) | Data structures (Computer science)
Classification: LCC QA76.73.J85 M37 2018 | DDC 005.7/3--dc23
LC record available at https://lccn.loc.gov/2018025237

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Foreword

The 21st century will probably be the century of the data revolution. Our numerical world is creating masses of data every day and the volume of generated data is increasing more and more (the number of produced numerical data is doubling every two years according to the most recent estimates). In such a context, data science is nowadays an unavoidable field for anyone interested in exploiting data. People may be interested in either understanding a phenomenon or in predicting the future behavior of this phenomenon.

To this end, it is important to have significant knowledge of both the rationale (the theory) behind data science techniques and their practical use on real-world data. Indeed, data science is a mix of data, statistical/machine learning methods and software. Software is actually the link between data and data science techniques. It allows the practitioner to load the data and apply techniques on it for analysis. It is therefore important to master at least one of the data science languages.

The choice of the software language(s) mainly depends on your background and the expected level of analysis. R and Python are probably the two most popular languages for data science. On the one hand, R has been made by statisticians... mostly for statisticians! It is, however, an excellent tool for data science since the most recent statistical learning techniques are provided on the R platform (named CRAN). Using R is probably the best way to be directly connected to current research in statistics and data science through the packages provided by researchers. Python is, on the other hand, an actual computer science language (with all appropriate formal aspects) for which some advanced libraries for data science exist. In this context, the Julia language has the great advantage to permit users to interact with both R and Python (but also C, Fortran, etc.), within a software language designed for efficient and parallel numerical computing while keeping a high level of human readability.

Professor Paul McNicholas and Peter Tait propose in this book to learn both fundamental aspects of data science: theory and application. First, the book will provide you with the significant elements to understand the mathematical aspects behind the most used data science techniques. The book will also allow you to discover advanced recent techniques, such as probabilistic principal components analysis (PPCA), mixtures of PPCAs, and gradient boosting. In addition, the book will ask you to dive into the Julia language such that you directly apply the learned techniques on concrete examples. This is, in my opinion, the most efficient way to learn such an applied science. In addition, the focus made by this book on the Julia language is a great choice because of the numerous qualities of this language regarding data science practice. These include ease of learning for people familiar with R or Python, nice syntax, easy code debugging, the speed of the compiled language, and code reuse.

Both authors have extensive experience in data science. Professor Paul McNicholas is Canada Research Chair in Computational Statistics at McMaster University and Director of the MacDATA Institute of the same university. In his already prolific career, McNicholas has made important contributions to statistical learning. More precisely, his research is mainly focused on model-based learning with high-dimensional and skew-distributed data. He is also a researcher deeply involved in the spreading of research products through his numerous contributions to the R software with packages. Peter Tait is currently a Ph.D. student but, before returning to academia, he had a professional life dedicated to data science in industry. His strong knowledge of the needs of industry regarding data science problems was really an asset for the book.

This book is a great way to both start learning data science through the promising Julia language and to become an efficient data scientist.

Professor Charles Bouveyron
Professor of Statistics
INRIA Chair in Data Science
Université Côte d'Azur
Nice, France

Preface

This is a book for people who want to learn about the Julia language with a view to using it for data science. Some effort has gone into making this book suitable for someone who has familiarity with the R software and wants to learn about Julia. However, prior knowledge of R is not a requirement. While this book is not intended as a textbook for a course, some may find it a useful book to follow for a course that introduces statistics or data science students to Julia. It is our sincere hope that students, researchers and data scientists in general, who wish to learn Julia, will find this book beneficial.

More than twenty years have passed since the term data science was described by Dr. Chikio Hayashi in response to a question at a meeting of the International Federation of Classification Societies (Hayashi, 1998). Indeed, while the term data science has only gained notoriety over the past few years, much of the work it describes has been practiced for far longer. Furthermore, whatever the precise meaning of the term, there is no doubt that data science is important across virtually all areas of endeavour. This book is born out of a mixture of experiences all of which led to the conclusion that the use of Julia, as a language for data science, should be encouraged.

First, part of the motivation to write this book came from experience gained trying to teach material in data science without the benefit of a relatively easily understood base language that is effective for actually writing code. Secondly, there is the practical, and related, matter of writing efficient code while also having access to excellent code written by other researchers. This, of course, is the major advantage of R, where many researchers have contributed packages — sometimes based on code written in another language such as C or Fortran — for a wide variety of statistics and data science tasks. As we illustrate in this book, it is straightforward to call R from Julia and to thereby access whatever R packages are needed. Access to R packages and a growing selection of Julia

packages, together with an accessible, intuitive, and highly efficient base language, makes Julia a formidable platform for data science.

This book is not intended as an exhaustive introduction to data science. In fact, this book is far from an exhaustive introduction to data science. There are many very good books that one can consult to learn about different aspects of data science (e.g., Bishop, 2006; Hastie et al., 2009; Schutt, 2013; White, 2015; Efron and Hastie, 2016), but this book is primarily about Julia. Nevertheless, several important topics in data science are covered. These include data visualization, supervised learning, and unsupervised learning. When discussing supervised learning, we place some focus on gradient boosting — a machine learning technique — because we have found this approach very effective in applications. However, for unsupervised learning, we take a more statistical approach and place some focus on the use of probabilistic principal components analyzers and a mixture thereof.

This monograph is laid out to progress naturally. In Chapter 1, we discuss data science and provide some historical context. Julia is also introduced as well as details of the packages and datasets used herein. Chapters 2 and 3 cover the basics of the Julia language was well as how to work with data in Julia. After that (Chapter 4), a crucially important topic in data science is discussed: visualization. The book continues with selected techniques in supervised (Chapter 5) and unsupervised learning (Chapter 6), before concluding with details of how to call R functions from within Julia (Chapter 7). This last chapter also provides further examples of mixture model-based clustering as well as an example that uses random forests. Some appendices are included to provide readers with some relevant mathematics, Julia performance tips and a list of useful linear algebra functions in Julia.

There is a large volume of Julia code throughout this book, which is intended to help the reader gain familiarity with the language. We strongly encourage readers to run the code for themselves and play around with it. To make the code as easy as possible to work with, we have interlaced it with comments. As readers begin to get to grips with Julia, we encourage them to supplement or replace our comments with their own. For the reader's convenience, all of the code from this book is available on GitHub: `github.com/paTait/dswj`.

We are most grateful to David Grubbs of the Taylor & Francis Group for his support in this endeavour. His geniality and professionalism are always very much appreciated. Special thanks to

Professor Charles Bouveyron for kindly agreeing to lend his expertise in the form of a wonderful Foreword to this book. Thanks also to Dr. Joseph Kang and an anonymous reviewer for their very helpful comments and suggestions. McNicholas is thankful to Eamonn Mullins and Dr. Myra O'Regan for providing him with a solid foundation for data science during his time as an undergraduate student. Dr. Sharon McNicholas read a draft of this book and provided some very helpful feedback for which we are most grateful.

A final word of thanks goes to our respective families; without their patience and support, this book would not have come to fruition.

Paul D. McNicholas and Peter A. Tait
Hamilton, Ontatio

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

About the Authors

Paul D. McNicholas is the Canada Research Chair in Computational Statistics at McMaster University, where he is a Professor and University Scholar in the Department of Mathematics and Statistics as well as Director of the MacDATA Institute. He has published extensively in computational statistics, with the vast majority of his work focusing on mixture model-based clustering. He is one of the leaders in this field and recently published a monograph devoted to the topic (*Mixture Model-Based Classification*; Chapman & Hall/CRC Press, 2016). He is a Senior Member of the IEEE and a Member of the College of the Royal Society of Canada.

Peter A. Tait is a Ph.D. student at the School of Computational Science and Engineering at McMaster University. His research interests span multivariate and computational statistics. Prior to returning to academia, he worked as a data scientist in the software industry, where he gained extensive practical experience.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

CHAPTER 1

Introduction

DATA SCIENCE is discussed and some important connections, and contrasts, are drawn between statistics and data science. A brief discussion of big data is provided, the Julia language is briefly introduced, and all Julia packages used in this monograph are listed together with their respective version numbers. The same is done for the, albeit smaller number of, R packages used herein. Providing such details about the packages used helps ensure that the analyses illustrated herein can be reproduced. The datasets used in this monograph are also listed, along with some descriptive characteristics and their respective sources. Finally, the contents of this monograph are outlined.

1.1 DATA SCIENCE

What is data science? It is an interesting question and one without a widely accepted answer. Herein, we take a broad view that data science encompasses all work related to data. While this includes data analysis, it also takes in a host of other topics such as data cleaning, data curation, data ethics, research data management, etc. This monograph discusses some of those aspects of data science that are commonly handled in Julia, and similar software; hence, its title.

The place of statistics within the pantheon of data science is a topic on which much has been written. While statistics is certainly a very important part of data science, statistics should not be taken as synonymous with data science. Much has been written about the relationship between data science and statistics. On the one extreme, some might view data science — and data analysis, in particular — as a retrogression of statistics; yet, on the other

extreme, some may argue that data science is a manifestation of what statistics was always meant to be. In reality, it is probably an error to try to compare statistics and data science as if they were alternatives. Herein, we consider that statistics plays a crucial role in data analysis, or data analytics, which in turn is a crucial part of the data science mosaic.

Contrasting data analysis and mathematical statistics, Hayashi (1998) writes:

> ... mathematical statistics have been prone to be removed from reality. On the other hand, the method of data analysis has developed in the fields disregarded by mathematical statistics and has given useful results to solve complicated problems based on mathematico-statistical methods (which are not always based on statistical inference but rather are descriptive).

The views expressed by Hayashi (1998) are not altogether different from more recent observations that, insofar as analysis is concerned, data science tends to focus on prediction, while statistics has focused on modelling and inference. That is not to say that prediction is not a part of inference but rather that prediction is a part, and not the goal, of inference. We shall return to this theme, i.e., inference versus prediction, several times within this monograph.

Breiman (2001b) writes incisively about two cultures in statistical modelling, and this work is wonderfully summarized in the first few lines of its abstract:

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.

The viewpoint articulated here leans towards a view of data analysis as, at least partly, arising out of one culture in statistical modelling.

In a very interesting contribution, Cleveland (2001) outlines a blueprint for a university department, with knock-on implications for *curricula*. Interestingly, he casts data science as an "altered

Index