Financial Analytics with R

Building a Laptop Laboratory for Data Science

Are you innately curious about dynamically inter-operating financial markets? Since the crisis of 2008, there is a need for professionals with more understanding about statistics and data analysis, who can discuss the various risk metrics, particularly those involving extreme events.

By providing a resource for training students and professionals in basic and sophisticated analytics, this book meets that need. It offers both the intuition and basic vocabulary as a step toward the financial, statistical, and algorithmic knowledge required to resolve the industry problems, and it depicts a systematic way of developing analytical programs for finance in the statistical language R. Build a hands-on laboratory and run many simulations. Explore the analytical fringes of investments and risk management.

Bennett and Hugen help profit-seeking investors and data science students sharpen their skills in many areas, including time-series, forecasting, portfolio selection, covariance clustering, prediction, and derivative securities.

Mark J. Bennett is a senior data scientist with a major investment bank and a lecturer in the University of Chicago's Master's program in Analytics. He has held software positions at Argonne National Laboratory, Unisys Corporation, AT&T Bell Laboratories, Northrop Grumman, and XR Trading Securities.

Dirk L. Hugen is a graduate student in the Department of Statistics and Actuarial Science at the University of Iowa. He previously worked as a signal processing engineer.

Financial Analytics with R

Building a Laptop Laboratory for Data Science

MARK J. BENNETT
University of Chicago

DIRK L. HUGEN
University of Iowa

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

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/9781107150751

First published 2016

Printed in the United Kingdom by Clays, St Ives plc

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

Library of Congress Cataloging-in-Publication Data
Names: Bennett, Mark J. (Mark Joseph), 1959– author. | Hugen, Dirk L., author.
Title: Financial analytics with R : building a laptop laboratory for data science / Mark J. Bennett, University of Chicago, Dirk L. Hugen, University of Iowa.
Description: Cambridge, UK : Cambridge University Press, 2016.
Identifiers: LCCN 2016026635 | ISBN 9781107150751
Subjects: LCSH: Finance–Mathematical models–Data processing. | Finance–Databases. | R (Computer program language)
Classification: LCC HG104 .B46 2016 | DDC 332.0285/513--dc23
LC record available at https://lccn.loc.gov/2016026635

ISBN 978-1-107-15075-1 Hardback

To our parents:
Mary and Herb and Patricia and Bernard

and family:
Rachel, Austin, and Cheryl
for all their kindness, love, and support

Contents

Preface

In 1994 the Channel Tunnel opened between England and France, allowing high-speed Eurostar trains to whisk passengers from the continent to the United Kingdom and back on a grand scale. What an amazing engineering feat it was for the time (beyond many people's earlier imaginations), yet we take it for granted today. In 1994, Grumman Aerospace Corporation, the chief contractor on the Apollo Lunar Module, was acquired by Northrop Corporation to form the new aerospace giant, Northrop Grumman. It was the prime contractor of the newly deployed advanced technology B-2 Stealth Bomber. On a much more mundane and personal scale, also in 1994, in a townhouse just outside the City of Chicago, I was performing a tedious daily exercise: looking up daily closing prices each evening in a stack of *Investor's Business Daily* newspapers for the two stock investments that were about to be purchased. This was not only to find out their running rate of return but also to find out their historical volatility relative to other stocks before entering into the positions. Doing this manual calculation was slow and tedious. The World Wide Web was introduced in the form of the Mosaic browser the next year. It was not long before Yahoo! was posting stock quotes and historical price charts, as well as technical indicators on the charts, available on demand for free in just a few seconds via the new web browsers.

The advent of spreadsheet software took analysts to a new level of analytical thinking. No longer were live, human-operated calculations limited to a single dimension. Each row or column could present a time dimension, a production category, a business scenario. And the automated dependency feature made revisions quite easy. Now spreadsheets can be used for a prototype for a more sophisticated and permanent analytical product: the large-scale, analytical computer program.

With modern programming languages like R and Python®, a skilled analyst can now design their analytic logic with significantly less effort than before, using resources such as Yahoo! or other free services for historical quotes. It has been said that Python's terse syntax allows for programs with the same functionality as their Java equivalents, yet four times smaller, and we suspect that R is similar. A small financial laboratory can be built on a laptop costing less than $200 in a matter of weeks, simulating multiple market variables as required. Or, by obtaining a higher-end laptop model with more drive space, the entire market can be loaded with 10 to 20 years' worth of historical data on a scale never before possible.

Once that laboratory is built, one can start to gain insights. Knowledge discovery was once a term for a human process. Now we're talking about computer automation.

Knowledge discovery seems like a bold term, a little too ambitious for anything a computer program could create. For example, the computer science professional society, the Association for Computing Machinery (ACM), has a special interest group called Knowledge Discovery and Data Mining (KDD). Hardly anyone would challenge the "data mining" part of that. After all, statisticians and computer scientists having at it with data is what they do. But discovering knowledge with a machine? Really? Automatically? Now that seems a little too exaggerated to be true. Then again, experiencing firsthand the algorithms that will be described in this book, we soon realized that the programs, using data science techniques, can not only automate very tedious calculations but then very positively yield insights into the human thinking level: insights that would otherwise not be found.

Perhaps one can view the experience with a sports analogy. In many sports we have defense to protect our current position and prevent the opponents from scoring further points. Offense is the ability to piece together athletic feats sequentially to put more points on the scoreboard. The data mining portion of KDD can be thought of as defense: the more disciplined, regimented side of the sport. Single achievements can be effective: thrusting up one's hand to block a pass, throwing a curve ball to prevent a hitter from connecting on the pitch. On the other hand, knowledge discovery is the offensive skill set, going beyond the required and expected data analysis and into the creative side. On offense, an entire series of events needs to be successful to yield progress: a full-field soccer scoring drive or a series of three successive baseball base hits before three outs to score a run. The likelihood of a success on offense is less.

So in the KDD model and sports analogy, data mining is the defense and knowledge discovery is the offense. Achieving knowledge discovery is a rare event with amazing impact. The discovery can be as powerful as human-made ideas, and can certainly enhance them. For example, we may discover that there is a publicly traded stock with uniquely desirable properties. The KDD domain touches the limits of what these machines can do with all the advancements in computer science.

In 1968, a Hollywood movie and novel by author Arthur C. Clarke, *2001: A Space Odyssey*, predicted automated reasoning, natural language speech recognition, video calls, and face recognition. The HAL9000 computer controls the flight to Jupiter while conversing and playing chess with astronaut Dr. Frank Poole and monitoring life conditions of over 300 astronauts. Since then, computer science, specifically simulation, has greatly impacted the research and discovery process in many fields and effectively achieved many of these science fiction goals now. Among others, there are fields of computational biology, computational cosmology, and computational linguistics. Images from these fields are shown in Figure 1.

Throughout this book we are concerned with computer simulation. Computer simulation has become so successful that it is now widely accepted that, after theory and physical experimentation, it is a third scientific method. As the subtitle says, this book can be used to build a simulation laboratory for finance. The book was developed as study material for the graduate Financial Analytics course in the Graham School at the University of Chicago Master of Science in Analytics program and from the undergraduate Investments course in the Department of Finance in the Tippie College of Business

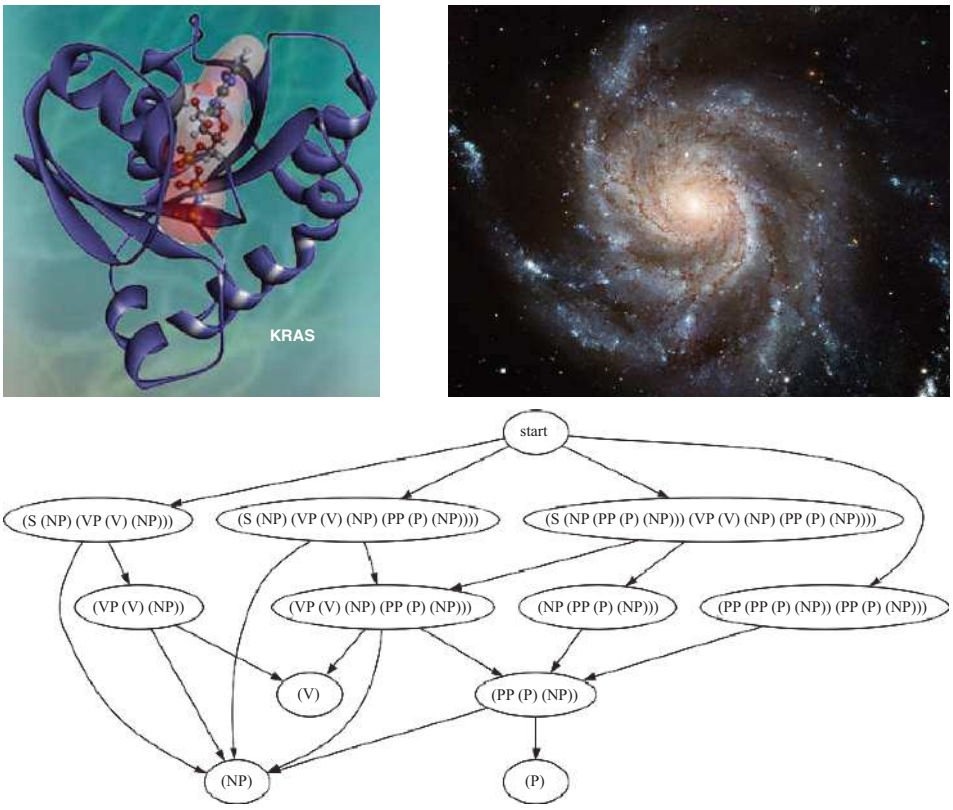

Figure 1 Sample images from computational biology, computational cosmology, and computational linguistics.

at the University of Iowa. It is recommended as a graduate textbook when used at a college or university. With the proper mathematical and computer science background, it could also be used at the advanced undergraduate level.

It is best to have taken a course in statistical analysis, probability and statistics, or, ideally, mathematical statistics for the material in the book, but much of the required material is introduced within the main text and the Appendix. It is best to have an undergraduate-level background in calculus, linear algebra, and enough computer science to be familiar with array manipulation with one or more scientifically based programming languages such as C, C++, Java, C#, Python, or Matlab®. A finance background is not necessary. Any experience with R is, of course, useful.

Financial computer simulation in the R language can be more intricate and challenging than building a spreadsheet. A quantitative optimizer can be better controlled and tailored when its logic is immediately apparent from the surrounding program code. More computer science knowledge is required by our reader to build more robust and sophisticated platforms, and more goes into the compiler and run-time system behind the scenes. But as the pieces are completed, the builder, or operator, or student of financial analytics begins to realize the benefits of simulation performed in a language designed

for statistical simulation. The insights that can be gained from building simulators and from observing the simulations will help deepen understanding for upcoming professional venues. Just as it is now for machines, for people it has always been about learning.

Regarding the exercises at the end of each chapter, data science involves the study of statistical and computational models. In this book, that means that we are unlocking the economic value which exists in the financial markets. Data engineering is the process of implementing models on computers as applied to large datasets, using files, program logic, testing, and continuous improvement. With these exercises, we take advantage of the data science principles of the prior chapters to build and engineer our financial laboratory.

As the reader performs these exercises, they may need to install various R packages from time to time. Various pages found by internet searching will steer the reader to proper instructions for loading R packages and troubleshooting any failed attempts. There are too many packages, conditions, and cases to repeat those instructions here.

The exercises focus individually on the various components so that we obtain an understanding of the logic and data. Each new component builds upon prior components in order to provide the level of sophistication required to answer our financial analytics inquiries.

Acknowledgments

Through our careers we are influenced greatly by people who are special to us. We can never pay them back for the gifts they have bestowed upon us: conversation, insight, opportunity, knowledge, and companionship.

Mark would like to thank Ron Krupp and Barry Finkel at Argonne; Phil Ridinger, Pat Baldwin, Howard Seckler, David Rouse, Bill Neidfelt, and Tom Bishop at Bell Labs; Christopher Marlin and Adrienne Critcher at the University of Iowa; Per Brinch-Hansen and Orna Berry at the University of Southern California; Dave Martin, Milos Ercegovac, Stott Parker, and Dan Berry at the University of California, Los Angeles; Bob MacGregor at System Development; Mark Christensen, Paul Schmitz, Tim Bancroft, Clare Morgan, Margaret Lakins, and Joe Dvorak at Northrop; Shelly Reis, Greg Brim, Brian Ostrow, Ron Netzel, David Joffe, Mack Amin, Thayer Allison, Dilip Nair, Gregg Berger, Haider Sajjad, Yuri Salnikov, Jim Bohmbach, Dante Lomibao, Paul Lee, Li Chen, Steve Zhu, Mario Konrad, Brian Philpott, Nancy Goldberg, Laura Lang, Raja Afandi, Ashish Batra, Krishna Bhamidi, and Chris Leakeas at Nationsbanc-CRT; Harry Georgakopoulos and YeeMan Bergstrom at XR Trading; and Adam Ginensky, Yuri Balasanov, and Sema Barlas at the University of Chicago and Marc Tempkin at the Chicago ACM for being tremendous mentors and colleagues over the years. Inspiration for this effort was provided at these wonderful locales: the bohemian Wolverine Farm book store in Fort Collins, the stately Pentacrest at the University of Iowa campus, and the Crerar Library of the gothic University of Chicago campus.

Dirk would like to thank his advisor Joseph Lang for all his help as well as Luke Tierney, Kate Cowles, Joyee Ghosh, Dale Zimmerman, Rhonda DeCook, Kung-Sik Chan, Erik Lie, David Bates, Ashish Tiwari, Wei Li, and Paul Weller for the excellent statistics and finance coursework at the University of Iowa. Thanks also to Kung-Sik Chan, Brian Ripley, Dirk Eddelbuettel, Hadley Wickham, David A. James, Seth Falcon, Winston Chang, Romain Francois, J.J. Allaire, Kevin Ushey, Qiang Kou, Douglas Bates, Jeffrey A. Ryan, Joshua M. Ulrich, Wouter Thielen, and John Chambers for development of the R packages used in the book.

The authors would particularly like to thank managing editor David Tranah for insightful comments based upon his vast experience, Clare Dennison for the timely content decisions, and Austin Bennett for the creative data scientist caricature on the cover.