Use R!

Radhakrishnan Nagarajan
Marco Scutari
Sophie Lèbre

Bayesian Networks in R

with Applications in Systems Biology

Springer

Use R!

Radhakrishnan Nagarajan
Marco Scutari
Sophie Lèbre

Bayesian Networks in R

with Applications in Systems Biology

Springer

Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

For further volumes:
http://www.springer.com/series/6991

Use R!

Albert: Bayesian Computation with R
Bivand/Pebesma/Gómez-Rubio: Applied Spatial Data Analysis with R
Cook/Swayne: Interactive and Dynamic Graphics for Data Analysis:
With R and GGobi
Hahne/Huber/Gentleman/Falcon: Bioconductor Case Studies
Paradis: Analysis of Phylogenetics and Evolution with R
Pfaff: Analysis of Integrated and Cointegrated Time Series with R
Sarkar: Lattice: Multivariate Data Visualization with R
Spector: Data Manipulation with R

Radhakrishnan Nagarajan • Marco Scutari
Sophie Lèbre

Bayesian Networks in R

with Applications in Systems Biology

Springer

Radhakrishnan Nagarajan
Division of Biomedical Informatics
Department of Biostatistics
University of Kentucky
Lexington, Kentucky, USA

Marco Scutari
Genetics Institute
University College London
London, United Kingdom

Sophie Lèbre
ICube
Université de Strasbourg
France

ISBN 978-1-4614-6445-7 ISBN 978-1-4614-6446-4 (eBook)
DOI 10.1007/978-1-4614-6446-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013935127

© Springer Science+Business Media New York 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Adriana Brogini and Fortunato Pesarin,
who showed me what an academic should be.

Preface

Real world entities work in concert as a system and not in isolation. Understanding the associations between these entities from their digital signatures can provide novel system-level insights and is an important step prior to developing meaningful interventions.

While there have been significant advances in capturing data from the entities across complex real-world systems, their associations and relationships are largely unknown. Associations between the entities may reveal interesting system-level properties that may not be apparent otherwise. Often these associations are hypothesized by superimposing knowledge across distinct reductionist representations of these entities obtained from disparate sources. Such representations, while useful, may provide only an incomplete picture of the associations. This can be attributed to their dependence on prior knowledge and failure of the principle of superposition in general. Such representations may also be unhelpful in discovering novel undocumented associations. A more rigorous approach would be to identify associations from data measured simultaneously across the entities of interest from a given system. These data sets or digital signatures are quantized in time and amplitude and in turn may (dynamic) or may not (static) contain explicit temporal information. Symmetric measures such as correlation have been helpful in modeling direct associations as undirected graphs. However, it is well appreciated that the association between a given pair of entities may be indirect and often mediated through others. Symmetric measures are also immune to the direction of association by their very definition. Graphical models such as Bayesian networks have especially proven to be useful in this regard. The vertices (nodes) represent the entities of interest, the arcs (edges) represent their associations, and the entire Bayesian network represents the joint probability distribution between the entities of interest. Bayesian networks

may also reveal possible causal relationships between these entities under certain implicit assumptions. More specifically, their ability to model associations from observational data sets where no active perturbation is possible has drawn attention across a wide spectrum of disciplines including biology, medicine, and health care.

There have been several noteworthy contributions to Bayesian network modeling and inference along with open-source implementations of the related algorithms. However, many of these prior contributions are extremely involved and demand a high level of sophistication from the reader. This book is unique as it introduces the reader to the essential concepts in conjunction with examples in the open-source statistical environment `R`. The level of sophistication is gradually increased across the chapters. Each chapter is accompanied by examples and exercises with solutions for enhanced understanding and experimentation. Thus this book may appeal to multidisciplinary audience and can potentially assist in teaching graduate-level courses in Bayesian networks and inference that permit hands-on experimentation of the concepts and approaches. The data sets considered essentially consist of publicly available molecular expression profiles. The emphasis on molecular data can be attributed to the growing need in life sciences for discovering novel associations across biological paradigms with minimal precedence and increasing emphasis on data-driven approaches. Classical studies in life sciences have focused on understanding the changes in the expression of a given set of molecules, such as genes and proteins, across distinct phenotypes and disease states. However, with recent advances in high-throughput assays that enable simultaneous screening of a large number of genes, there has been growing interest in understanding the associations between these molecules that may provide system-level insights. Such system-level insights have been argued to be critical prior to developing meaningful interventions. These efforts together fall under the emerging discipline called systems biology. Bayesian networks have especially proven to be useful abstractions of the underlying biological pathways and signaling mechanisms. Their usefulness is also exemplified by their ability to discover new associations in addition to validating known associations between the entities of interest.

While a list of popular open-source `R` packages pertinent to Bayesian networks is listed under Table 2.1 (Chap. 2), the discussion focuses on the packages **bnlearn**, **G1DBN**, and **ARTIVA**.

http://cran.r-project.org/web/packages/bnlearn
http://cran.r-project.org/web/packages/G1DBN
http://cran.r-project.org/web/packages/ARTIVA

We believe that these packages are comprehensive and accommodate the necessary functionalities required across the chapters. We also believe that concentrating on these packages keeps the book more focused with minimal demand on the audience time in learning the functionalities across the various open-source `R` packages.

This book is organized as follows. Chapter 1 introduces the reader to the essentials of graph theory and `R` programming. Chapter 2 discusses the essential definitions and properties of Bayesian networks with an emphasis on static Bayesian

networks. It introduces the reader to structure and parameter learning from multiple independent realizations of data sets without explicit temporal information. Such data sets are quite common and represent a snapshot of the process. The impact of discretization on the network inference with application to molecular expression data is also discussed. The lack of temporal information implicitly excludes the presence of feedback or cycles, resulting in a directed acyclic graphical representation of the associations between the entities. These limitations are overcome by learning networks from data sets with explicit temporal signatures. In Chap. 3, we discuss the usefulness of dynamic Bayesian networks for learning the network structure in the presence of explicit temporal information such as multivariate time series. Homogeneous and nonhomogeneous dynamic Bayesian networks are discussed. In Chap. 4, static and dynamic Bayesian network inference methods are discussed. Some of the network learning algorithms discussed in the earlier chapters are computationally intensive limiting their usefulness across large and high-dimensional data sets. Parallelization options for some of the algorithms discussed in the earlier chapters are discussed in Chap. 5 to overcome some of these limitations.

Lexington, KY — Radhakrishnan Nagarajan
London, UK — Marco Scutari
Strasbourg, France — Sophie Lèbre

Contents

1 Introduction 1
1.1 A Brief Introduction to Graph Theory 1
1.1.1 Graphs, Nodes, and Arcs 1
1.1.2 The Structure of a Graph 2
1.1.3 Further Reading 4
1.2 The R Environment for Statistical Computing 4
1.2.1 Base Distribution and Contributed Packages 4
1.2.2 A Quick Introduction to R 5
1.2.3 Further Reading 10
Exercises 11

2 Bayesian Networks in the Absence of Temporal Information 13
2.1 Bayesian Networks: Essential Definitions and Properties 13
2.1.1 Graph Structure and Probability Factorization 13
2.1.2 Fundamental Connections 15
2.1.3 Equivalent Structures 15
2.1.4 Markov Blankets 16
2.2 Static Bayesian Networks Modeling 17
2.2.1 Constraint-Based Structure Learning Algorithms 17
2.2.2 Score-Based Structure Learning Algorithms 19
2.2.3 Hybrid Structure Learning Algorithms 20
2.2.4 Choosing Distributions, Conditional Independence Tests, and Network Scores 20
2.2.5 Parameter Learning 23
2.2.6 Discretization 23
2.3 Static Bayesian Networks Modeling with R 24
2.3.1 Popular R Packages for Bayesian Network Modeling 24
2.3.2 Creating and Manipulating Network Structures 26
2.3.3 Plotting Network Structures 34
2.3.4 Structure Learning 35

2.3.5 Parameter Learning . . . 40
2.3.6 Discretization . . . 42
2.4 Pearl's Causality . . . 44
2.5 Applications to Gene Expression Profiles . . . 46
2.5.1 Model Averaging . . . 47
2.5.2 Choosing the Significance Threshold . . . 51
2.5.3 Handling Interventional Data . . . 53
Exercises . . . 56

3 Bayesian Networks in the Presence of Temporal Information . . . 59
3.1 Time Series and Vector Auto-Regressive Processes . . . 59
3.1.1 Univariate Time Series . . . 59
3.1.2 Multivariate Time Series . . . 60
3.2 Dynamic Bayesian Networks: Essential Definitions and Properties . 63
3.2.1 Definitions . . . 63
3.2.2 Dynamic Bayesian Network Representation of a VAR Process . . . 66
3.3 Dynamic Bayesian Network Learning Algorithms . . . 67
3.3.1 Least Absolute Shrinkage and Selection Operator . . . 67
3.3.2 James–Stein Shrinkage . . . 68
3.3.3 First-Order Conditional Dependencies Approximation . . . 68
3.3.4 Modular Networks . . . 69
3.4 Non-homogeneous Dynamic Bayesian Network Learning . . . 69
3.5 Dynamic Bayesian Network Learning with `R` . . . 72
3.5.1 Multivariate Time Series Analysis . . . 72
3.5.2 LASSO Learning: **lars** and **simone** . . . 74
3.5.3 Other Shrinkage Approaches: **GeneNet**, **G1DBN** . . . 78
3.5.4 Non-homogeneous Dynamic Bayesian Network Learning: **ARTIVA** . . . 80
Exercises . . . 81

4 Bayesian Network Inference Algorithms . . . 85
4.1 Reasoning Under Uncertainty . . . 85
4.1.1 Probabilistic Reasoning and Evidence . . . 85
4.1.2 Algorithms for Belief Updating: Exact and Approximate Inference . . . 87
4.1.3 Causal Inference . . . 90
4.2 Inference in Static Bayesian Networks . . . 91
4.2.1 Exact Inference . . . 91
4.2.2 Approximate Inference . . . 93
4.3 Inference in Dynamic Bayesian Networks . . . 94
Exercises . . . 100

5 Parallel Computing for Bayesian Networks 103
5.1 Foundations of Parallel Computing 103
5.2 Parallel Programming in R 105
5.3 Applications to Structure and Parameter Learning 108
5.3.1 Constraint-Based Structure Learning Algorithms 109
5.3.2 Score-Based Structure Learning Algorithms 112
5.3.3 Hybrid Structure Learning Algorithms 114
5.3.4 Parameter Learning 115
5.4 Applications to Inference Procedures 115
5.4.1 Bootstrap 115
5.4.2 Cross-Validation 117
5.4.3 Conditional Probability Queries 120
Exercises 123

Solutions 125

References 149

Index 155