Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Big Data in Omics and Imaging: Integrated Analysis and Causal Inference.

Momiao Xiong. Boca Raton, FL: Chapman & Hall/CRC Press, 2018, xxix + 736 pp., $129.95(H), ISBN: 978-0-81-538710-7.

Oliver Y. Chén

To cite this article: Oliver Y. Chén (2020) Big Data in Omics and Imaging: Integrated Analysis and Causal Inference., Journal of the American Statistical Association, 115:529, 487-488, DOI: 10.1080/01621459.2020.1721249

To link to this article: https://doi.org/10.1080/01621459.2020.1721249

Accepted author version posted online: 03 Feb 2020.
Published online: 19 Mar 2020.

three areas mentioned in the book's subtitle: Probability theory (including random variables and vectors but largely neglecting the matrices), stochastic processes (including stochastic integrals and differential equations) and (additionally) statistical inference.

The author bases his book on the lecture notes of his classes at the University of Missouri, dedicating the book to graduate and PhD students and full-time academics in statistics or mathematics. Although this target group typically has some prior knowledge of probability theory, Chapter 1 of the book is a non-rigorous introduction to elementary probability theory under the title "rudimentary models." Chapter 2 deals with statistical inference with a focus on Bayesian approaches, Micheas' area of expertise. Being the only chapter on statistics, it feels like an excursion and is not directly relevant for the remainder of the book. The third chapter on measure theory establishes the basis for the rigorous introduction to probability theory presented in Chapters 4 and 5. The subsequent Chapters 6 and 7 on discrete- and continuous-time stochastic processes cover Markov chains, martingales, the Poisson process, Brownian motion, general Markov processes, and conclude by defining stochastic integrals and stochastic differential equations with respect to Brownian motion. The book finishes with a very short treatise of stochastic partial differential equations at the end of Chapter 7 as promised on the book's back cover. Every chapter in the book ends with a summary that provides useful references for the covered topics.

Clearly, including such a vast array of subdisciplines into 336 pages (excluding appendix and bibliography) comes at a price and here the price is paid in terms of precision and depth of the presented material. With regard to precision, the "rudimentary models" in the first chapter lack accuracy, and slight mathematical inconsistencies appear at several places in the book (e.g., stochastic processes are introduced with general index set T, but properties like independent increments are formulated for $T = [0, \infty)$ without specifying this). With regard to depth, various statements are listed without proof or with the proof left as an exercise for the reader (e.g., the standard central limit theorem for independent, identically distributed random variables is buried in an exercise in Chapter 2). Every chapter comes with a collection of exercises, and a solution manual for these is promised to be found at the book's accompanying webpage. Unfortunately—to this day (January 2020)—the webpage only provides the solutions to the exercises of Chapter 1, an errata sheet, and the Matlab code used to generate the two figures of realized paths of Brownian motion on the book's front cover.

Navigation through the book is complicated as definitions and examples are more prominently highlighted than section titles. Neither the design nor the typesetting (which allows too many "orphans" and "widows," dangling all alone at the bottoms or tops of pages) encourages the reader to use the book as a resource to create lectures or simply as a student textbook.

To conclude, this book does not fulfill the expectations of a textbook since graduate students will have to consult secondary literature for a full understanding of the theory. Academic instructors may find it helpful as a source of possible topics to include in their own classes. Various potential lecture series based on the book's materials are outlined by the author in the preface.

Anita D. Behme
Technische Universität Dresden
Dresden, Germany
anita.behme@tu-dresden.de

Big Data in Omics and Imaging: Integrated Analysis and Causal Inference. Momiao Xiong. Boca Raton, FL: Chapman & Hall/CRC Press, 2018, xxix+736 pp., $129.95(H), ISBN: 978-0-81-538710-7.

Despite our expanding knowledge, we are still very ignorant about many parts of the function and functioning of the human brain and the genes. The challenges and ignorance are equally large for statisticians and data scientists (Fan, Han, and Liu 2014). The difficulties, in part, lie in the large number of features (e.g., neuroimaging data measured from a million voxels, or sequencing data generated from the whole genome consisting of billions of nucleotides) obtained from hundreds of individuals that amount to dozens of Petabytes in storage size, and hundreds of Terabytes after conversion and preprocessing. This is beyond what traditional statistical methods and computer programs can efficiently handle. Large-scale features may vary dynamically in time, such as whole brain imaging data measured by functional magnetic resonance imaging (fMRI); some of them may interact with one another in space, such as whole genome profiling data measured by next generation sequencing technology. These temporally and spatially varying activities are further coupled and intertwined with environmental factors and effects from other agents. All these causes—interlaced—shape our traits, behaviors, and actions throughout our development.

To disentangle, extract, and understand these intricate, intercorrelated, and large-scale features that are dynamic in space and time, neuroscience and genetics need statistics. In his book, Professor Xiong introduces, discusses, and implements a rich variety of statistical tools that can be used to study large-scale features obtained from the human brain and genome, map neural and genetic signatures to behavioral and disease outcomes, and make causal enquiries into their relationships. The scope of the book is comprehensive, the concepts deep, and technicalities oftentimes mathematically heavy. In spite of a focus on omics and (medical) imaging, the book discusses statistical concepts and devices that readers may find useful in studying general problems in human neuroscience and human genetics.

Large-scale brain and genetic data can be illustrated and studied using graphical models, where a node represents a variable (e.g., a brain area or a single nucleotide polymorphism (SNP)) and an edge or the link between two nodes indicates the potential causal association between the nodes (e.g., brain connectivities or pathways of gene expression). Chapter 1 introduces directed graphical models, where there is an order from a node to another node via a directed edge, and undirected graphical models, where the edge has no ordering. A web of associated genotypes and phenotypes and the edges linking them form the topology of a genotype-phenotype network. The latter half of the chapter discusses how to use statistical devices, such as structural equation models, to study such a network, and

how to make causal inference about it. There are, in general, two ways to learn the causal structure of networks. One is to test whether two nodes are (conditionally) independent; if not, then it suggests that there is a potential causal relationship between them. This approach, however, is sensitive to noise. The other approach is to assign a score to each edge (between a pair of nodes) to represent the edge strength, and use the size of the score to evaluate the causal relationship. Chapter 2 examines network biology through causal lenses, exploring the score-based learning to uncover network structures using Bayesian (and a few non-Bayesian) frameworks.

The relationships between signals from different brain regions, various SNPs, and between neural and genetic signatures and behavioral outcomes (such as disease status and severity) may vary in space and in time. Biosensors record data from multiple genome locations (thus tracing the spatial variability) and brain activations over semi-continuous time points (thus tracing the temporal variability). Today, in the pursuit of automated disease diagnosis and real time healthcare monitoring, scientists may gain some inspiration and insights from the rich spatial and temporal information encoded in biosensor data. Chapter 3 introduces three approaches (functional principal component analysis, differential equations, and deep learning, with a focus on convolutional neural networks) to explore the time- and space-varying (causal) relationships on data recorded from biosensors. To investigate how the association between time-varying features and the dynamic outcomes can assist longitudinal (health status and disease severity) prediction, the latter part of the chapter discusses functional regression models.

Chapter 4 sails into the sea of RNA-seq data analysis, discussing statistical tools (including a few discussed in previous chapters) that are suitable for studying single cell, gene co-expression, gene network, and dynamic and longitudinal gene expression. Chapters 5 and 6 discuss statistical and machine-learning devices for analyzing (high-dimensional) methylation data and imaging genomics, respectively. Together, these two methodologically intercorrelated chapters introduce advanced regression models, functional (principal component and structural equation) models, and neural networks. Finally, echoing Chapter 2, Chapter 7 ties the causal knot by introducing statistical frameworks to make enquiries into complex cause-and-effect problems that involve discrete data, multivariate features, multiple and multilevel networks, and confounders. Each chapter closes with simulation studies or real data analyses.

Professor Xiong has written a statistics book for analyzing biological data that explores mathematical concepts and gives only cursory attention to biology at large. In a future edition of this book, I would be delighted to see more biological intuitions and justifications of the statistical models. As a simple example, the author has mentioned a few times that, when the number of features is greater than the sample size (e.g., p. 3 and p. 11), estimation of the inverse covariance matrix via maximum likelihood estimation is not feasible (since the covariance matrix is singular) and thus, as a treatment, one applies a penalty (on the number of nonzero entries of the matrix). This is mathematically (perfectly) sound; but scientists may be wondering whether and why this is biologically suitable, as the problem is scientific in nature. A brief discussion that links statistical analysis with biological insight would be helpful. In this regard, the existence of a small number of network hubs (a class of highly connected nodes) and many poorly connected nodes in cells (Barabási and Oltvai 2004), the brain (van den Heuvel and Sporns 2013), and the genes (Leclerc 2008) could shed some biological light on sparse networks, and hence could provide some justification for penalization. Additionally, the extensive mathematical arguments may be inaccessible to most neuroscientists and geneticists (and some statisticians and biostatisticians) who are mainly interested in learning and implementing statistical frameworks within their knowledge comfort-zone. This, in my view, seems to be a minor oversight.

Statistical, neurobiological, and genetic studies should rejoice that, thanks to the accumulation of neuroimaging and genetics data and multi-site and multi-disciplinary collaborations, they are enriched not only by a wealth of information, but also by an increasing number of powerful statistical analytical devices (many of which are covered in this book), to unravel the intricacies of the human brain and human genetics and how they give rise to behavior, cognition, perception, and how their malfunctioning causes illnesses, which may someday leave profound marks on our everlasting enquiry to understanding who we are.

Oliver Y. Chén
University of Oxford
Oxford, UK
yibing.chen@seh.ox.ac.uk

References

Barabási, A. L., and Oltvai, Z. N. (2004), "Network Biology: Understanding the Cell's Functional Organization," *Nature Reviews Genetics*, 5, 101–113, DOI: 10.1038/nrg1272. [488]

Fan, J., Han, F., and Liu, H. (2014), "Challenges of Big Data Analysis," *National Science Review*, 1, 293–314, DOI: 10.1093/nsr/nwt032. [487]

Leclerc, R. D. (2008), "Survival of the Sparsest: Robust Gene Networks Are Parsimonious," *Molecular Systems Biology*, 4, 213, DOI: 10.1038/msb.2008.52. [488]

van den Heuvel, M. P., and Sporns, O. (2013), "Network Hubs in the Human Brain," *Trends in Cognitive Sciences*, 17, 683–696, DOI: 10.1016/j.tics.2013.09.012. [488]