Received: 29 June 2021 | Accepted: 23 July 2021

DOI: 10.1111/biom.13534

BOOK REVIEWS

Statistics for making decisions

Nicholas T. Longford

Boca Raton, FL: Chapman and Hall/CRC, 2021. Hard cover.
pp. 307. $96

Chapman and Hall/CRC has published a practical and unique perspective textbook for statisticians, analysts, and consultants, "*Statistics for Making Decisions*" by Nicholas T. Longford. Making decisions is an important action in everyone's lives, and statistics can help us to provide the options for making a better decision depending on the limited information. However, the general statistical methods omit the client's perspective and value such as greatest benefits and least losses. In this book, the author incorporates the client's concern into the statistical methods to make the proper verdicts. Many examples in different subject areas help the reader to understand the statistical concepts.

The book is organized into 12 chapters covering a wide range of statistical for making decision concepts. Chapter 1 emphasizes making decisions from the client's perspective rather than just the analyst's perspective, and the statistics are meaningful after a comprehensive assessment from each viewpoint. Chapter 2 broadly introduces the familiar approaches for statistical inference and making decisions, including estimation and hypothesis testing. The concept involves the frequentist, Bayesian, and fiducial paradigms. Chapter 3 uses the simple case, one cutpoint with two options for a parameter from the normal distribution, to introduce the notations, terminologies, and theories. Three straightforward loss functions, including constant, linear, and quadratic forms, are explored. The relationship between various loss functions and corresponding verdicts, equilibrium, cutpoint, and errors are explained. The unequivocal and impasse decision-making depends on the setting for the plausible range of loss ratio. The concepts are connected to the commonly used estimation and significant level for hypothesis testing. Two parameters case with plausible rectangles is also extended. Chapter 4 goes on the parameters from the nonnormal distributions, including t, F, binomial, beta, Poisson, and gamma distributions. Chapter 5 starts with the special case that considers two thresholds to separate closer and far from the specific parameter value zero. It uses to explain the complex loss functions, such as piecewise linear, piecewise quadratic, and several combination losses. The multiple options with discrete and continuum cases are extended. Chapter 6 brings in the important issues in study designs and considers the sample size, cost of data collection, and the variance of outputs. Chapter 7 discusses the decision-making in particular medical concerns, such as discriminating the positive and negative disease status, finding the threshold cutpoints specific to subpopulations, and setting the scheme to save the expensive resources. Chapter 8 regards the topics and strategies for multiple decisions, inclusive of several types of error for multiple testing, actions in a sequence, extreme selections, setting the gray zone. Chapter 9 addresses the particular parameter that displayed the institution's performances. The outliers, rare events, and the best with tie situation that both in continuous and dichotomies performance assessment are concerned. Chapter 10 concentrates on the material in clinical trials. The comparison of treatment effects for different study designs is the main concern. Chapter 11 ends with the subjects of model selection, and Chapter 12 provides a short postscript. The author gives many further reading papers and exercises with solutions or R programs in all chapters.

In summary, this book provides a higher and broader perspective of the statistics for decision making. I sincerely recommend the book for the users who possess or will have experience in consulting and want to extensively consider the statistical approach in the client's view.

Yu-Chung Wei

Graduate Institute of Statistics and Information Science, National Changhua University of Education, Chunghua, Taiwan

Correspondence
Yu-Chung Wei, Graduate Institute of Statistics and Information Science, National Changhua University of Education, Chunghua, Taiwan.
Email: weiyuchung@cc.ncue.edu.tw

ORCID
Yu-Chung Wei https://orcid.org/0000-0003-3747-2215

DOI: 10.1111/biom.13533

Structural equation modeling with partial least squares using Stata and R

Mehmet Mehmetoglu | Sergio Venturini

Boca Raton, FL, USA: Chapman and Hall/CRC, 2021. Hard cover. pp. 382. £ 99.99

Partial least squares (PLS) algorithm was first proposed by Swedish statistician Herman Wold in 1960s. In later years, it has two distinct routes of developments: as a dimensional reduction technique, it has gained wide applications first in chemometrics and later in bioinformatics; it has also been promoted, mainly in information technology and business research, as an alternative to the popular covariance-based structural equation models. This book is about the application of PLS in the latter, known as PLS path modeling or PLS structural equation modeling. The basic ideas of PLS structural equation modeling are similar to those of covariance-based structural equation modeling; both incorporate latent variables derived from observed (manifest) variables into their modeling of causal relations among groups of variables. Their differences are both conceptual and technical; latent variables in PLS structural equation modeling are weighted composites of manifest variables (here we only consider reflective models), while latent variables in covariance-based structural equation models are the shared components of manifest variables, just like factors in factor analysis. PLS structural equation modeling does not make assumptions about the joint distribution of manifest variables nor does it have well-defined target functions to minimize, such as the maximum likelihood functions in covariance-based structural equation models. Consequently, it lacks some statistical properties, such as consistency of estimates, and objective criteria for evaluating the model goodness of fit. These have drawn some severe criticisms in the past and have been addressed, if not entirely satisfactorily, in this book.

PLS structural equation modeling has become popular in some research fields owing to a few freely available software packages, such as Smart-PLS, PLS-GUI, and PLS-graph. These packages provide graphical user interface for users to visualize their models by drawing boxes for manifest variables, circles for latent variables, and arrows for specifying the relations between variables. Since version 3.0, Smart-PLS has incorporated many new features but it is no longer free. The authors of this book are also authors of a package plssem for the statistical software Stata. This is certainly a welcome addition to the capability of Stata, and moreover, this book also includes sections on the use of packages in R software for PLS structural equation modeling.

This book consists of eight chapters and an appendix. The first chapter, Framing Structural Equation Modeling, provides a brief discussion of the intertwining histories of PLS and covariance-based structural equation modeling. I may add one bit of history here: Herman Wold was the PhD advisor of Karl Jöreskog who made important contributions to the development of covariance-based structural equation modeling. The second chapter, Multivariates Statistics Requisites, is the longest chapter, providing the basic statistical knowledge for those who are not familiar with bootstrapping, principal component analysis, mixture models and path analysis. Reading this chapter will not make you an expert of multivariate statistics but is essential for those who wish to gain a deeper understanding of PLS structural equation modeling and its recent progress. The third chapter, PLS Structural Equation Modeling: Specification and Estimation, is the core of this book, explaining the fundamental theory of PLS algorithms in terms of how it works and why it works. This chapter also discusses some more advanced applications, such as higher order constructs and consistent PLS structural equation modeling. Examples are used to demonstrate how to use Stata to undertake the analyses, and detailed explanations for the commands and the interpretation of the results are provided. Mathematical details are left in the appendix at the end of the chapter. Chapter 4 discusses how to evaluate the performance of a PLS structural equation model. A covariance-based structural equation models is evaluated by the differences between its estimated and observed variance–covariance matrices of manifest variables in the model, but no similar criterion has been proposed to evaluate PLS structural equation models. Instead, the evaluation focuses on how much of endogenous latent variable variance is explained by exogenous latent variables in the model, such as R^2. It is similar to using model R^2 to evaluate the performance of a linear regression model. Some critics consider this the Achilles Heel of PLS structural equation modeling. A correctly specified regression model may have a low R^2, while an erroneously specified one may have a high R^2. R^2 alone cannot be used to evaluate the verisimilitude of the specified causal relations between variables within a structural equation model.

Next two chapters demonstrate how to undertake mediation analysis and estimating moderation/interaction effects, respectively. The approaches implemented by PLS structural equation modeling are very similar to those by covariance-based structural equation modeling. Therefore, readers who are familiar with mediation analysis and interaction effects in covariance-based structural equation modeling will have an easy read. Both, nevertheless, share the same limitations. For instance, endogenous latent variables and mediators are assumed to be continuous in PLS mediation analysis, and this severely limits its applications to health data, where many variables are categorical. Chapter 7, Detecting Unobserved Heterogeneity in PLS-SEM, demonstrates new methodological developments in PLS structural equation modeling by incorporating latent classes into the model, aiming to identify potential subgroups within the data. The final chapter is very short and uses a real example to show how to write up an article of PLS structural equation modeling. The final Appendix provides some very basics of regression analysis, and its aim is to serve as an introduction to regression analysis using Stata software.

Overall, this book takes a balanced approach to presenting the statistical theory of PLS structural equation modeling and its practical applications. The mathematical level is not high but is higher than most books on PLS structural equation modeling. Although Stata and R software packages are used to analyze the examples, this book will be useful for users of other software with an interest in getting a grasp of statistical theory behind PLS structural equation modeling. It is comprehensive and also accessible. Anyone who is seriously thinking of using PLS structural equation modeling for their research should carefully read through this book before embarking on their first analysis.

Yu-Kang Tu

Institute of Epidemiology & Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

Correspondence
Yu-Kang Tu, Institute of Epidemiology & Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan.
Email: yukangtu@ntu.edu.tw

ORCID
Yu-Kang Tu https://orcid.org/0000-0002-2461-474X

DOI: 10.1111/biom.13521

Introduction to Data Science: Data Analysis and Algorithms with R

Rafael Irrizarry

Boca Raton, FL: CRC Press, 2020. Hard cover. pp. 711.

This book is a very comprehensive book and much needed in this area. It is 711 pages long and divided into 38 chapters in six main parts. The parts are:

1. R: Chapters 1–5
2. Data Visualization: Chapters 6–11
3. Statistics with R: Chapters 12–19
4. Data Wrangling: Chapters 20–26
5. Machine Learning: Chapters 21–34
6. Productivity Tools: Chapters 35–38

In this review, I will provide synopsis of each section and then finally talk about the general applicability of this book.

The first section is a good synopsis of R software. It starts with the basics of what the consoles look like and quickly moves on to *logicals*, to programming tools and spends a good portion of time on **tidyverse**. While the pace is acceptable for people who have prior R experience, it seemed a bit too hurried for a novice. Also, a lot of the examples used in this section are from existing built-in libraries, which while easy to demonstrate, might not give the readers a hands-on experience. In Chapter 5, the authors do talk about importing data and the “read” commands.

The second part of the book deals with data visualization. The section starts with library **ggplot2** and discusses some details of how to change labels, aesthetic mappings, and colors. This section seemed a little out of place, as it seems like there would be a discussion about data types, exploratory data analysis, the issues of stats, before visualization. There are a fair number of details in this section, especially in the chapter “Data visualization principles,” including common misrepresentations, visual cues, and inclusion of 0 in a plot. This is a great discussion on some pertinent issues. However, this section would be better if it came after some discussion on data and statistical problems.

The third section is on using statistics with R. This section seems to lack some flow, as it jumps into Monte Carlo simulation without really discussing random variables and distributions. This section would be okay to follow if you had already had a class on probability distributions and inference. In an effort to incorporate R in the methods, some of the background methods were glossed over. This would potentially better serve audience if the statistics part was done first, then a section like this and data visualization were included.

The Data wrangling section is best part of this book. It goes into some detail and would be useful to statisticians and would-be data scientist. The chapters of *Join* and *Webscraping* are useful and thorough. The section on *Text mining*, is a bit short and felt like an afterthought but was a good addition. This section could definitely be longer. But in general, this section of the book is definitely useful.

The Machine Learning section is detailed but with some missing background material. It covers a lot of ground talking about smoothing, receiver operating characteristic curve. The focus seemed like the caret **package**. Again the flow seemed a bit strange with regression and generalized linear models getting a small mention. However, in general it does an adequate job of presenting how to do machine learning (ML) in R. I really appreciated the fact that the author tried to make a distinction between artificial intelligence and ML in the introduction to this section.

The last section is something a lot of statisticians would not be exposed to: The author called it Productivity Tools. This is where it seemed to have a more computing flavor. It talks about *Git* and *GitHub* and *Rmarkdown*. This is definitely a very useful section for statisticians. This would also be a great set of topics to include for teaching Data Science.

In general, the book covers a vast number of topics related to data science in R. I would have preferred the tile to be "Data Science using R," as the focus of the book is definitely "how to" in R. However, it would be a good book for reading about some terms used in data science and some basic ideas of methods. The best feature of the book is: it is a comprehensive text on using R in Data Science. It would be a good text for Statisticians trying to understand the jargon of data science. However, as a stand-alone book in data science, it is a bit thin on explanation of the methods and the advantages and disadvantages of the methods. Overall, I think the author did a laudable job in summarizing a lot of material.

Nairanjana Dasgupta,
Director of Data Analytics,
Professor of Statistics

Washington State University

Correspondence
Nairanjana Dasgupta, Director of Data Analytics,
Professor of Statistics, Washington State University.
Email: dasgupta@wsu.edu

ORCID
Nairanjana Dasgupta https://orcid.org/0000-0003-4009-6281

DOI: 10.1111/biom.13522

Statistical Foundations of Data Science

Jianqing Fan | Runze Li | Cun-Hui Zhang | Hui Zou

Boca Raton, FL: Chapman and Hall/CRC, 2021. Hard cover. pp. 774.
USD $130.00

Data science has become popular and has been developed rapidly in many fields such as data mining, machine learning, bioinformatics, and computational biology among others. In recent years, many monographs based on computational perspective or algorithms have been published, but little published books have introduced data science from statistical methodologies and theoretical perspective. This paper aims to review *Statistical Foundations of Data Science*, which was published in 2021. As shown in the title, this book aims to introduce data science via statistical tools. Regarding the content in this book, the authors introduce estimation methods to construct statistical models, topics in machine learning including (un)supervised learning and graphical models, and neural networks in deep learning. In addition to standard presentation of methodologies, the authors also provide theoretical properties and their rigorous proofs. Moreover, each chapter has exercise problems, which are crucial issues in the authors' research work and, at the same time, help readers review the materials in each chapter.

This book contains 14 chapters. In Chapter 1, the authors first present rise and ubiquitousness of big data and high dimensionality from different resources, such as gene expression profiles of microarray in biological sciences, internet or social network in computer and information sciences, economical and financial data from stocks, bonds to foreign exchange rates, and earth and astronomy sciences. After that, the authors discuss impacts of big data and some popular developments such as artificial intelligence or machine learning in recent years. On the other hand, big data are usually accompanied by high dimensionality. To more understand this concern, the authors summarize some impacts of high dimensionality, including computation, noise accumulation, and spurious correlation, and then present the insight of high-dimensional statistical learning.

The topic in Chapter 2 is about the fundamental concepts of multiple and nonparametric regression models. The authors first review standard setup of linear regression and summarize important properties and proofs. After that, some extensions are discussed accordingly, including weighted least squares to deal with errors that are heteroscedastic or correlated and the Box-Cox transformation to define transformed variables for model fitting. Motivated by multiple regression, the authors further present spline regression based on several types of spline functions to handle nonparametric regression. To deal with collinearity caused by highly correlated covariates, ridge regression, which is equivalent to L_2 penalized regression, is introduced. In the last two sections, the authors consider a reproducing kernel Hilbert space (RKHS) which focuses on the functional form in nonparametric regression and generalizes previous settings that are under Euclidean space. Finally, since ridge regression and nonparametric regression involve a tuning parameter, to determine its value and achieve the "best" estimation/prediction performance, generalized cross-validation (GCV) is outlined.

Chapters 3 and 4 introduce penalized least squares (PLS) methods that aim to do variable selection and estimation. The main difference between these two chapters is that Chapter 3 discusses key concepts and methodologies, while Chapter 4 focuses on theoretical justification and proofs. Specifically, Chapter 3 starts by several classical variable selection criteria: subset selection, Akaike information criterion (AIC), Mallow's C_p criterion, (generalized) cross-validation criterion, and some extensions of AIC, such as Bayesian information criterion (BIC), risk inflation, and ϕ-criteria. After that, the authors next introduce PLS, which is a combination of the least squares function and the penalty function with a tuning parameter. Some commonly used penalty functions and strategies are outlined, including ridge regression, smoothly clipped absolute deviation (SCAD), minimax concave penalty (MCP), (adaptive) lasso, elastic net, and Dantzig selector. To efficiently solve optimization problems, the authors provide some useful algorithms, such as least angle regression (LARS), iterative shrinkage-thresholding algorithm, and alternating direction method of multiplier. Moreover, similar strategy can be naturally extended to nonparametric regression. The authors take the generalized additive model as an example and introduce group penalty to address variable selection.

Chapter 4 describes some properties of PLS. The authors first present performance benchmarks to assess the performance of the estimator and define several loss functions to assess variable selection. Next, the authors revisit the PLS estimator but now theoretical results and rigorous proofs are provided, including selection consistency, sign consistency, examination of prediction and coefficient estimation errors, and oracle properties. Finally, smaller and sorted penalties that are used to control false discovery rate (FDR) are discussed.

Chapter 5 considers the generalized linear model (GLM) that is a natural extension of linear regression. The authors first introduce exponential family and some concepts of GLM, such as model construction and likelihood estimation. Parallel with the developments in Chapter 3, the authors assume sparse parameters and propose the penalized likelihood method to do variable selection for GLM. Meanwhile, some computational algorithms and selection of tuning parameters in Chapter 3 are applied. Finally, some important results with associated conditions, such as oracle properties, are established.

Chapter 6 discusses M-estimation with penalization, which is regarded as the generalization of least squares estimations and likelihood estimations in Chapters 4 and 5, and M-estimation enables us to deal with a large body of different types of data structures. In this chapter, the authors focus their attentions on some important and challenging models, including (composite) quantile regression, robust regression based on least absolute deviation and Huber loss functions, rank regression, and the proportional hazards (PH) model for survival data. By the spirit of M-estimation and techniques in previous chapters, one is able to construct penalized estimating functions and then handle variable selection and estimation. Similar to preceding chapters, theoretical results are established and numerical algorithms are provided.

Chapter 7 continues the issue of high-dimensional analysis but focuses on the relevant statistical analysis, such as interval estimation and hypothesis testing of low-dimensional features (e.g., different effect) based on high-dimensional data. The main purpose of this study is that little information of low-dimensional features is available, even though optimality properties of the estimators have been established in previous chapters. The authors first introduce two concepts: the first one is semi-low-dimensional (semi-LD) approach where low-dimensional components are main interests and high-dimensional components are regarded as noise; and the second one is the semi-parametric approach where the parametric term is the main target but nonparametric part is the noise term. To present these ideas in detail, several regression models are explored. For example, debiased techniques are proposed to deal with "specific" estimator of main interest for linear regression. GLM is considered and desparsified lasso, decorrelated score estimator, and partial penalized likelihood ratio methods are introduced to test linear hypothesis. Moreover, based on derived asymptotic normality, the authors further discuss asymptotic efficiency for low-dimensional penalized estimator for random design and the partial linear model. On the other hand, based on the concept of PLS, the authors further discuss the estimation of the partial correlation of two univariate random variables, which can be regarded as the strength of the edge in graphical models. Finally, in the last section of this chapter, the authors describe data swap and gradient approximation to address asymptotic normal estimation and efficient inference in the semi-LD approach.

Chapter 8 discusses feature screening, whose purpose aims to retain informative covariates and screen out noninformative ones. Different from regularization methods that are based on estimating functions and penalty functions, the strategy of feature screening is to treat the correlation of the response and the covariate as a signal, and retain variables by choosing large values of signals. To demonstrate this idea, the authors first consider linear regression and treat the sample correlation as a signal. Under this pioneering idea, several extensions are presented in other sections, such as the generalized correlation and rank correlation that treat the marginal maximum likelihood estimator as a signal, nonparametric independent screening for generalized additive models, and model-free approaches that do not impose (non)parametric structures. All existing methods are shown to enjoy the sure screening property, which ensures that all truly informative variables are retained under certain conditions. In addition, to increase the accuracy of feature screening and detect falsely excluded variables, the authors further propose iterated algorithm, feature screening via forward regression, and partial correlation. Finally, to estimate the error variance in linear models and nonparametric regression, the authors propose refitted cross-validation.

Chapter 9 presents the estimation of precision matrices and covariance regularization, which have been widely used in Fisher's discriminant, Gaussian graphical models (GGM), Hotelling T^2-test, and FDR results. In the beginning of this chapter, the authors first review some basic facts and operations for matrices. After that, the authors focus on multivariate normal distributions and propose the penalized likelihood method with different types of thresholding approaches. In reality, however, the sub-Gaussian assumption is too strong. To relax this restriction and provide general approaches, an element-wise robust covariance estimator and the adaptive Huber estimator are developed. The authors next discuss analysis of graphical models, which refer to the detection of network structures or estimation of precision matrices. Different from the PLS in Chapter 7, the authors clearly introduce the structure of GGM and summarize several useful methods, such as penalized likelihood, a.k.a. graphical lasso, and the constrained ℓ_1-minimization for inverse matrix estimation. To relax the normality assumption, some advanced and complex settings are explored, such as applying Kendall's τ to deal with latent GGM and nonparametric graphical models based on binary or the mix of continuous and binary types of data. Finally, the derivations of all theorems in this chapter are placed in the last section.

Chapter 10 continues to discuss the estimation of high-dimensional covariance matrix, but the strategy is under factor models, which are not only used to model the dependence among high-dimensional measurements but also to induce a low-rank plus sparse covariance structure among the measured variables. To see the motivation, the authors start reviewing principal component analysis (PCA) as well as its relevant techniques, such as singular value decomposition and the power method. Next, the authors focus on factor models and show that the covariance structure admits a low-rank and sparse structure. Moreover, the principal orthogonal complement thresholding (POET) is proposed to extract latent factors and the adjusted eigenvalues thresholding (ACT) and eigenvalue ratio estimators are adopted to determine number of factors. On the other hand, in the case that factors are partially known, the strict factor model and robust initial estimator are developed; when side information is available, the projected PCA based on the additive model and a random-effect model are introduced. Finally, in the last two sections of this chapter, the authors summarize theoretical properties of estimated loading matrix, covariance matrix, and idiosyncratic components as well as their proofs.

Different from Chapter 10 that focuses on methodologies and theory of PCA and factor models, Chapter 11 presents applications of factor models and PCA: (1) factor-adjusted regularized model selection, (2) robust multiple testing, (3) factor augmented regression methods, and (4) statistical machine learning. The purpose of the first application is to deal with the case that covariates admit a factor structure, which breaks down the irrepresentable condition and fails model selection consistency due to high correlation among covariates. A factor-adjusted regularized model selector is proposed to address this problem. For the second application, the authors focus on FDR with dependence measurements, and the factor-adjusted robust multiple test is introduced, whose key idea is to apply the t-test to the factor-adjusted data obtained by PCA, and then control FDR by the Benjamini–Hochberg procedure. Regarding the third application, the authors discuss principal component regression, which is widely used to compress high-dimensional variables into a couple of principal components. Finally, the last application aims to adopt PCA to a class of Wigner matrices, whose relevant applications include community detection, topic model, matrix completion, item ranking, and Gaussian mixture models.

Chapter 12 discusses classification problem, which is one of important topics in supervised learning. While this topic has been studied in a huge amount of monographs, the authors comprehensively summarize many methods with their own insights and interpretations. Specifically, the authors first present model-based classifiers, such as linear/quadratic discriminant analysis (LDA/QDA) and logistic regression. Regarding LDA, the authors further relax parametric setting and apply kernel estimations to develop nonparametric classifiers. Next, the authors introduce nearest neighbor, classification tree, support vector machine (SVM), and ensemble learning (e.g., bagging, random forest, and boosting). Moreover, some important extensions are also covered. For example, to address high-dimensional classification problem, the authors develop sparse SVM, sparse large margin classifiers, and sparse discriminant analysis. To relax parametric assumptions and deal with nonparametric and semi-parametric settings, the authors propose feature augmentation via nonparametrics selection, penalized additive logistic regression, and semiparametric sparse discriminant analysis.

Chapter 13 discusses unsupervised learning, which involves no response variables. Typical topics include clustering and PCA. In the first three sections, the authors introduce several methods for cluster analysis (K-means, hierarchical, model based/spectral clustering) and choices of the number of clusters. In the presence of high-dimensional data, several techniques of variable selection in clustering are discussed. The remaining sections in this chapter focus on high-dimensional PCA, in which the approaches are quite different from those in Chapter 10. The authors first outline the inconsistency of the regular PCA under high-dimensional data. To solve this issue, the authors next introduce sparse PCA, iterative SVD thresholding approach, a penalized matrix decomposition method, a semidefinite programming approach, and a generalized power method.

Chapter 14 introduces deep learning, which is one of popular topics and an important application of artificial intelligence (AI) in recent years. To give readers motivation and intuition, the authors first outline the rise of deep learning and emphasize methods based on neural networks. For example, deep neural networks enable us to use composition of nonlinear functions to model nonlinearity; convolutional neural networks (CNN) are used to analyze data with salient spatial structures or image processing; recurrent neural networks (RNN) are adopted to process time series data or other sequence data. In addition, ideas of deep neural networks functions can be extended to unsupervised learning. For example, autoencoders and generative adversarial networks methods are introduced. Moreover, to understand the procedure of training deep neural nets, stochastic gradient descent (SGD) is adopted to deal with the empirical risk minimization. Regarding the optimization problem, some valid strategies are provided. To name a few, ReLU activation function, skip connections, and batch normalization are able to ease numerical instability; weight decay and data augmentation are used to improve the generalization power of trained neural nets. Finally, in the last section, some examples and R commands are presented to demonstrate the methods in this chapter.

In summary, different from other monographs for data science, this book provides clear introduction to statistical inference and machine/deep learning. Comprehensive discussions make people understand the developments in relevant topics. In addition, rigorous derivations enable readers to get insights and key steps to derive theoretical results, and then further develop new theory. It is a definitely useful reference for researchers and is suitable for graduated-level courses as well.

Li-Pang Chen

Department of Statistics, National Chengchi University

Correspondence
Li-Pang Chen, Department of Statistics,
National Chengchi University.
Email: lchen723@nccu.edu.tw

ORCID

Li-Pang Chen https://orcid.org/0000-0001-5440-5036

DOI: 10.1111/biom.13529

Handbook of neuroimaging data analysis

Hernando Ombao | Martin Lindquist |
Wesley Thompson | John Aston

Boca Raton, FL: CRC Press, 2017. Hard cover. pp. 702. 210.00 USD

In the introduction to *Handbook of Neuroimaging Data Analysis*, Dr. John Aston writes, “understanding the brain has become arguably one of the most complex, important and challenging issues in science, and imaging has become an invaluable tool in this endeavor.” The development of imaging methods has provided applied researchers with a set of minimally invasive techniques to study the brain *in vivo*, and heavily influences contemporary medical and psychological research. Since the origin of neuroimaging, the field has always been rich with statistical and computational challenges. From image reconstruction, to multi-image alignment, to modeling and analysis, a complete synthesis of neuroimage data depends on a series of computational techniques. Moreover, raw neuroimage data can exhibit complex spatio-temporal correlation structure, and applied work is often replete with difficult-to-remember acronyms, a host of possible data formats, and at times unintuitive measurements of quantities like “white matter integrity,” or “cortical volume.” For reasons like these, the activation energy required for quantitative researchers to enter into the world of neuroimaging can be quite high. In *Handbook of Neuroimaging Data Analysis* (hereafter, *Handbook*), editors (and contributors) Ombao, Lindquist, Thompson, and Aston work to reduce this barrier by providing readers with a single volume introduction to quantitative neuroimaging analysis.

Reflecting the relative popularity of magnetic resonance imaging (MRI) based applied research, the majority of *Handbook* is dedicated to summarizing developments in statistical analysis of different MRI data types. Much more could be (and has been) written about, for example, image Co-registration algorithms, or the mathematics behind solutions to other preprocessing problems. Writing as a biostatistician, most datasets we encounter will have already been heavily preprocessed by collaborators or the data collection team. After an initial introductory chapter, the text is divided into two main sections: first, an overview of the many different data types that neuroimaging entails, and second, a comprehensive survey of areas of active research into statistical methods to work with each modality. In my review to follow, I will generally summarize *Handbook*’s content by section and topic first. Individual chapters may be listed out of order.

As noted, Chapters 2 through 7 orient readers to specific data types, including positron emission tomography (PET, Chapter 2); MRI (structural MRI—Chapters 3 and 5, diffusion weighted MRI—Chapter 4, and functional MRI—Chapter 6); and electroencephalography (EEG, Chapter 7). Functional near-infrared spectroscopy and magnetoencephalography are notably absent from the survey of imaging modalities, though in practice analyses of these data may have much in common with analysis of functional MRI and EEG. For PET, MRI, and EEG, however, this section of the text very successfully synthesizes the broad domain-specific knowledge of each data type into introductory chapters intended for quantitative researchers.

In particular, though aimed specifically at structural MRI with a clinical flavor, Chapter 5 is written as a part-tutorial to basic features of image preprocessing like magnetic field inhomogeneity correction and affine transformation. The chapter also includes brief code snippets to introduce image manipulation with the R programming language and preprocessing with existing software. This chapter would make an excellent introduction to the basics of MR image analysis in general, and is explained clearly enough to be accessible to a beginning graduate student or advanced undergraduate. Chapter 4 also provides a quite thorough overview of diffusion MRI including tractography. Here, the authors detail what the data represent mathematically while retaining a high-level feel and noting several important open questions for research, such as how to estimate uncertainty in tractographic maps. Similarly, the authors of Chapter 6 summarize the myriad techniques related to functional MRI analysis. Authors Lindquist and Wager review task-related activity localization and discuss notions of “functional connectivity” as a way for researchers to study systems of functionally correlated brain regions. Finally, Chapter 7 introduces EEG data and the many challenges it may present including artifact removal and modeling of stationary and nonstationary time series.

The final section of *Handbook* functions as a thorough overview of modern statistical research into methods for modeling MRI and EEG data. Methods for EEG are covered at the end in a two-chapter (Chapters 20 and 21) *tour de force* surveying time and frequency domain analysis, dynamic linear and vector autoregressive modeling, frequency coherence analysis, and change point detection. Most other topics in this section are covered in relation to MR imaging. For example, Chapter 8 introduces the MR image reconstruction process that underlies absolutely all research with the resultant images. Chapter 10 contains an interesting review of MRI and EEG preprocessing pipelines and a variety of criteria that have been employed in the literature to evaluate the reliability and reproducibility of the output of these pipelines.

For structural MRI, Chapter 9 reviews advanced techniques for direct modeling of structural features or morphology. Chapter 18 meanwhile discusses general methods for modeling longitudinal series of images as outcomes. Outcome images of course need not be structural in nature, but much of the discussion in Chapter 18 is expressed in terms of intuitive, easily visualizable cortical thickness data. Matters of functional MR image analysis are presented starting with estimation of the hemodynamic response function in Chapter 11, and continuing through contemporary approaches to classical group-level modeling in Chapter 12. Chapter 14 begins a series of chapters devoted to functional connectivity methods with an explanation of the many different connectivity metrics used in the literature. The overall discussion of connectivity is deepened by a review of notions of effective connectivity for estimating causal relationships in Chapter 16, and formalized by Ginestet, Kramer, and Kolaczyk in their chapter on graph theory (Chapter 17). Rounding out the set of chapters on functional connectivity, Bowman, Simpson, and Drake present cutting-edge work on joint modeling of functional connectivity and anatomical tractographic imaging (Chapter 19). Other general topics are discussed as well: methods for control of family-wise or false-discovery error rates—always a concern with high-dimensional data—are reviewed in Chapter 13. Last but not least, in Chapter 15, Eloyan, Zipunnikov, Yang, and Caffo have written a particularly clear introduction to matrix decomposition methods and their application to neuroimaging data. This chapter is augmented by a review of literature on computation of independent component analysis (ICA) or blind source separation in high dimensions. ICA in particular is a cornerstone technique for much applied functional connectivity work, and so many researchers may find this section useful.

Overall, *Handbook* functions more as a reference than a text, but very much succeeds in that regard. A researcher looking to embark on, say, research in structural MRI methodology would be very well served by successively reading through Chapters 5, 3, and 10 for an introduction to the data and general familiarity with image preprocessing. Following that or similar sequence, Chapters 18 and 9 might provide a survey of more advanced modeling of structural data types, and importantly supply our researcher with an armful of vetted references for further study. I certainly noted quite a few useful references while reading through *Handbook*, and will be returning to reread chapters for years to come.

Andrew Whiteman

Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA

Correspondence
Andrew Whiteman, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.
Email: awhitem@umich.edu

ORCID

Andrew Whiteman https://orcid.org/0000-0002-5107-0506