Journal of Statistical Software

February 2009, Volume 29, Book Review 11. *http://www.jstatsoft.org/*

Reviewer: John Maindonald
Australian National University

Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning

> Alan Julian Izenman
> Springer-Verlag, New York, 2008.
> ISBN 978-0-387-78188-4. 732+xxvi pp. USD 89.95.
> `http://astro.ocis.temple.edu/~alan/MMST/`

This is a near encyclopedic account of topics that come generally under the headings of regression, classification, cluster analysis and low-dimensional representation, albeit with a strong statistical learning perspective. Graphical displays are used to excellent effect. It is roughly comparable in style and content to Hastie, Tibshirani, and Friedman (2001), due out shortly in a second edition.

Theoretical demands are more severe than for Hastie *et al.* (2001). As with that text, it stays strictly within an independent observations theoretical framework. I could not find any discussion of issues that arise from the common attempt to generalize beyond the population that have generated the data.

A high level of theoretical preparation is expected:

> As prerequisites, readers are expected to have had previous knowledge of probability, statistical theory and methods, multivariate calculus, and linear/matrix algebra. [...] Along with a background in classical statistical theory, it would also be helpful if the reader had some exposure to Bayesian ideas in statistics.

Data sets, from a wide range of disciplines, are available from the book's web site. Many are also available in one or other R package, or from elsewhere on the net. At the end of each chapter are information on software packages, bibliographic notes, and exercises. The bibliographic notes are a helpful starting guide to the literature. They give useful indications of important directions in current research. The exercises are a mix of theoretical derivations, practical data analysis, and coding exercises. R and S-PLUS codes for many of the exercises are promised, in due course, from the book's website.

Chapter 1 comments broadly on data mining and machine learning, then giving an overview of later chapters. Chapter 2 comments broadly on database technology, then moving on to discuss data quality problems.

Chapter 3, on random vectors and matrices, has basic theory that "builds the foundation for the statistical analysis of multivariate data". Here, the reader encounters the singular value decomposition, generalized inverses, multivariate normal theory including the Wishart distribution, and maximum likelihood estimation based on sample vectors from the multivariate normal. Theoretical rigor is not however an over-riding concern. The preface has the quote:

> When new ideas are being developed, the most fruitful approach is often to let rigor rest for a while, and let intuition reign – at least in the beginning. (Inge S. Helland, 2000)

As with Berk (2008) (see also the review by Maindonald 2009), this book is broadly in the tradition of "statistical learning". Izenman comments that: "This book mixes new algorithmic techniques for analysing large multivariate data sets with some of the more classical multivariate techniques". Regression methods (including regression with multiple outcome variables) occupy the greater part of the book. There are chapters on: model assessment and selection in multiple regression, multivariate regression, linear discriminant analysis, recursive partitioning and tree-based methods, artificial neural networks, support vector machines, and committee machines. Committee machines encompass, largely, boosting methods and random forests.

Methods that are not suitably described as regression, or that are more descriptive in character, include: nonparametric density estimation, multidimensional scaling and distance geometry, latent variable models (including factor analysis), nonlinear dimensionality reduction and manifold learning, and correspondence analysis.

This is a comprehensive and challenging overview. A criticism is that, like most books in this area, it assumes too narrow a theoretical framework. The practical data analyst does not always have the luxury of exploiting a deep and comprehensive training in models that assume independent observations. There are too many contexts that demand some broader modeling framework.

References

Berk RA (2008). *Statistical Learning from a Regression Perspective.* Springer-Verlag, New York.

Hastie T, Tibshirani R, Friedman J (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, New York.

Maindonald J (2009). "Review of 'Statistical Learning from a Regression Perspective'." *Journal of Statistical Software, Book Reviews*, **29**(12), 1–4. URL `http://www.jstatsoft.org/v29/b12/`.

Reviewer:

John Maindonald
Australian National University
Centre for Mathematics and Its Applications
Canberra, ACT 0200, Australia
E-mail: john.maindonald@anu.edu.au
URL: http://www.maths.anu.edu.au/~johnm/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 29, Book Review 11 *Published:* 2009-02-09
February 2009