Practical Guide to Chemometrics, Second Edition (Book Review)

Author:
Hibbert, D. Brynn

Publication details:
Chemometrics and Intelligent Laboratory Systems
v. 87
Chapter No. 2
pp. 319-320
0169-7439 (ISSN)

Publication Date:
2007

Publisher DOI:
http://dx.doi.org/10.1016/j.chemolab.2007.03.005

Book review

Practical Guide to Chemometrics, Second edition
Editor, Paul Gemperline
CRC Press, Taylor and Francis Group,
Boca Raton, USA
2006, pp541
ISBN 1-57444-783-1
US$ 179.95

Review by:
D Brynn Hibbert
School of Chemistry
University of New South Wales
Sydney
NSW 2052
Australia

tel: +61 2 9385 4713
fax: +61 2 9385 6141
e-mail: b.hibbert@unsw.edu.au

According to the jacket blurb, the text " … offers an accessible introduction to application oriented-multivariate methods of data analysis and procedures that are highly beneficial to solving a variety of problems using analytical chemistry and statistics." The book "builds on fundamental concepts to reach advanced topics in chemometrics". This is all true; the text, certainly in the earlier chapters is accessible and there is a progression throughout the book from a very basic level of statistics through to some reasonably advanced material.

Herein lies a problem with a book of this kind. Following a short and general introduction by the editor, the second chapter on statistical evaluation of data starts, as these tend to do, with definitions of mean, standard deviation and so on, and makes it to ANOVA some thirty pages later. A student who was starting with the basics would probably need one of the dedicated data analysis books, and certainly would be unlikely to be able to tackle the matrix algebra of some of the later chapters armed only with this material. The accessibility also comes out in the rather loose terminology. Accuracy and precision have interesting definitions here, even if they do not quite accord with the Vocabulary of Basic and General Terms in Metrology. In particular, accuracy as a concept is now used as a description of the state of having adequate trueness with acceptable precision, rather than simply 'getting the answer right'. The combination of strung out measurement results that just happen to average to the 'target' value (Figure 2.1 b) is not fit-for-purpose result. In a real measurement the target value is, of course, not known, and so if the uncertainty is very great, then the result is useless. This might be an example of the different viewpoints of metrologists working in chemistry and chemometricians. Perhaps also the statisticians may have some concerns with this chapter. For example they would argue that the correct explanation of a 95% confidence interval of a mean of *n* observations is that

95% of the 95% confidence intervals of the means of repeated experiments would encompass the population mean.
There follows an interesting chapter on the multivariate normal distribution, which starts with the normal distribution and the central limit theorem, topics which could have graced chapter 2. Example calculations in Matlab are a good addition here.
Throughout the book there are Matlab or Excel examples (or neither); authors have been left to use their preferred software. A disk is included with the book with Matlab 'm' files for examples in chapters 3, 4, 7 and 12. As these only occupy just over 1 MB perhaps some other useful material could have been included on the disk. Matlab and other code for examples in chapter 6 can be obtained from web sites, cited in the chapter. If a disk is to be included with the text then it should offer more comprehensive and structured material.
The editor then writes a chapter on principal components analysis. This is nicely done with a standard development of the theory, backed up with examples from spectroscopy and chromatography. Preprocessing, choosing numbers of factors and residual variance analysis are covered and this chapter provides the background for others in the book (e.g. Chapter 9 on pattern recognition).
Two chapters on calibration follow; titled 'calibration' and 'robust calibration'. With this juxtaposition, the unfortunate inference that the former is 'non-robust calibration' cannot help but be contemplated. In fact chapter 5 is a good introduction to univariate and multivariate inverse methods ('inverse' as the concentration is modelled as a function of observations). The statistics of calibration models are developed and the chapter reaches PCR, PLS and mentions ridge regression. Variable selection is covered well. The chapter on robust calibration is really quite hard and could have been better integrated with the preceding chapter. I think I followed 'affine equivalent covariance estimators' and the difference between affine equivariant and orthogonal equivariant, but am not sure. Perhaps what is missing is the 'practical' (of the book's title) aspects that could guide readers when one or another method should be employed. The example on melons at the end of the chapter is interesting, but does not tell us whether the method employed was the only approach, or how this problem fits into the scope of 'robust' calibrations.
Maeder and Neuhold author a chapter on kinetic modelling of multivariate measurements with non-linear regression. This is a somewhat niche subject, but the treatment is well done and the exposition can be followed easily. Both Matlab and Excel code are given. The use of Solver (part of the Data Analysis Toolbox for Excel) is detailed and described as being little known in chemistry. I am not sure this entirely correct. Certainly Solver is well used in my circle of chemical acquaintances and the popularity of De Levie's book (referenced by the authors) I think has caused its greater use, particularly as De Levie's "Solver Aid" program can give standard errors on the fitted parameters. Tellinghuisen has also published a series of papers showing how uncertainties in the '*x*' value of an unknown can be obtained by a non-linear fit of the calibration points and the instrument response to the unknown. The oscillating Belusov-Zhabotinski reaction is given as an example of the solution of a complex, non-linear system, but then the authors do not give details or the code used to solve the problem saying that the system is complex and coding is prone to error. They recommend the automatic generation of code but only offer literature references to suggest how this might be accomplished. It is not clear why they bother with the example.
Experimental design is treated in chapter 8. This is a very good chapter, with the approach to response surface methodology treated in the context of the design of

experiments. Mixture models are first discussed in detail, then factorial designs of different kinds, followed by non-symmetric optimal designs. The description of software assumes that the reader has access to Matlab's Statistical toolbox, or to a number of specialised software packages, which limits the use of the information. An explanation of the use of Taguchi designs in analytical method validation would have provided a good example. The one used, that of multi-component agrichemicals was spoilt by the need to code all the variables for commercial reasons (although this did not affect the math, people like to understand real examples in context), and the fact that although this was a 'calibration' exactly what was calibrated was not said in the description of the problem.

In chapter 9, pattern recognition catches up with the PCA discussed in chapter 4. To this is added cluster analysis, K-nearest neighbours and SIMCA. Nearly half the chapter is given over to examples that illustrate the methods used, which does mesh well with the 'practical' title. Steven Brown then covers signal processing, discussing the major methods used, such as Fourier transform, wavelets, smoothing algorithms and so on. Multivariate curve resolution then follows (chapter 11) with the different kinds of factor analysis and alternating least squares grouped under non-iterative and iterative methods. Self-modeling curve resolution is also described with a short mention of errors and uncertainty. There are some interesting examples including the resolution of protein structures. The last methods chapter is a rather short discussion of three-way calibration. Multi-way, multivariate methods (PARAFAC, PARAFAC2, Tucker etc) are achieving greater importance as chemical methods delve into multi-dimensions (LC-MS-MS, GC-GC, 2D NMR and so on). In future editions this should be expanded to the general *n*-way case, and could take in both calibration and classification aspects.

To round off the text, the editor writes on the future trends in chemometrics (always a brave thing to do). His concluding remarks deserve repeating "The future of chemometrics lies in the development of innovative solutions to interesting problems. The problem-oriented approach is required because relatively few advances can be expected in the area of new mathematical and numerical methods" (page 516). Perhaps this means that the next edition of this book will be completely changed to discuss problems seeking solutions rather than solutions (methods) seeking problems? As the reader will understand, we have come a long way from the early statistics chapter. Somewhere in the 520 pages of text is an answer to your chemometrics problem. As a single book this does well to cover so much ground. I certainly recommend it for any library or laboratory bookshelf.