04 October 2001 **Technometrics: Book Review**

Multivariate Analysis of Quality An Introduction, by Harald Martens and Magni Martens, West Sussex, UK: John Wiley, 2001, ISBN 0-471-97428-5, xx + 445pp., $135.00.

The authors state that this book has been written to meet a need for an entry-level textbook on multivariate data analysis for the practical researcher in academia or industry. The text is presented as a narrative analysis of multivariate data using soft modelling. The authors have little use for classical statistical analysis by stating: (1) " the book makes little or no use of traditional statistical distribution theory, and takes a dim view on hypothesis testing;" (2) "traditional hypothesis testing methods in mathematical statistics focus more on the *noise* than on the interesting *signal* in data;" and (3) "computerised statistical methods for re-sampling have already reduced the need for teaching complicated mathematical statistics to unwilling and terrified non-statisticians".

So how do these chemometricians predict a response matrix **Y** from a matrix of x-predictor variables **X**? First, matrix **X** is represented by the product of a score matrix $\mathbf{T_A}$ (i.e., "A" principal components) and a loading matrix **P** plus an error matrix. Secondly, matrix **Y** is modeled by the product of the $\mathbf{T_A}$ score matrix and a loading matrix **Q** plus an error matrix. The **Y** matrix is also represented by the linear regression of **X** with coefficient matrix **B** plus an error matrix. The authors use cross-validation to determine the optimal number of principal components, to estimate the predictive ability of the final model, to assess the reliability of modeling various X- and Y-variables, and to detect outliers. The coefficients are displayed in terms of a reliability range $\mathbf{b} \pm 2s(\mathbf{b})$ where $s(\mathbf{b})$ is called the estimated standard uncertainty of regression coefficients determined again from cross-validation. Rather than using statistical testing procedures most conclusions about the optimal number (i.e., "A" usually two) of principal components, goodness of fits, and important coefficients are made by viewing plots. About 60% of the book illustrates these points by using two examples, one on viscosity measurements and another on a mixture experiment to optimize the recipe for making cocoa. Of course, there are some variations on this summary, but this is the authors' basic plan for analyzing almost all types of data.

Some of the statistical thinking used to develop models are not followed. For example, the authors' advise, "the choice of what is **X** and what is **Y** does not have to follow unspoken tradition that **X** = cause and **Y** = effect;" and "but the choice is usually not very critical. If in doubt, the user should try to swap **X** and **Y**." The authors also consider the correlation coefficient a unit-free measure of the goodness of fit rather than a measure of linearity. The concept of degrees of freedom is considered too complex for this book. Models that include squared x-variables are called non-linear models rather than second-order models and cross-products representing interaction are given minimal consideration. In the cocoa mixture experiment, these chemometricians ignore that mixture components add to100% and blindly use a full linear model to represent the response.

Even though this is a heuristic approach to data analysis, I was surprised that the authors did not discuss some important topics. In many chemistry experiments, the number of x-variables is larger than the

number of samples in the **X** matrix of dimension $n \times k$ with $k > n$. The authors mention in the appendices that the (**X'X**) matrices can not be inverted for regression analysis and "the inversion is made in a sequence of simple steps." But there is no discussion in the main text about this situation. Cross validation is the main tool for evaluation, but there is no discussion about the number of experimental runs to leave out relative to the total number of experimental runs. Most important, there is no discussion about software to do the recommended soft modeling. This is particularly important because this book provides a minimal amount of mathematics to analyze the data. I am not sure how any student will be able solve any problems using this book. The authors don't even provide any general guidelines or principles for interpreting the plots that are used to evaluate all of their examples.

As I read this book, I wondered who was going to teach this material. I could not envision any statistician using this book as a class text and most applied scientists that I know prefer a more scientific foundation for data interpretation. In some special cases, the authors do allow for statisticians to analyze data: "when working with extremely expensive, dangerous or ethically difficult issues, we recommend that you let a professional statistician take over both the planning and the data analysis."

Charles K. Bayne
Oak Ridge National Laboratory

The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC-05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow other to do so, for U.S. Government purposes.