Book review

by **Marjan Tušar** and **Marjana Novič**

Ljubljana, Slovenia

Introduction to Multivariate Statistical Analysis in Chemometrics

Kurt Varmuza and **Peter Filzmoser**

Taylor & Francis - CRC Press, Boca Raton, FL, 2009

Published in Slovenian language in *Acta Chim. Slov.*, **58**, suppl. (2011)

Why to read "yet another book" on statistical methods? This is a question that may arise among chemometricians, chemists, students or anyone using chemometric methods in their research or everyday routine work. However, this book is a very good contribution to a better understanding of statistics because it really captures the point of potential misunderstanding between chemists and mathematicians and fills this gap with clear and exact explanations, mathematically firm and properly equipped with examples. Through the whole book one can observe the spirit of collaboration between a chemist and a mathematician, which is especially valuable because it contributes to a profound understanding of statistical methods, their applicability and limitations. The book provides us also with a historical overview of numerous topics in chemometrics and with an exhaustive list of references at the end of the chapters. Sometimes ingenious explanations are given about the background of terms, for example Bootstrap, Chapter 4.2.6. - "to lever off to great success from a small beginning" with etymological explanation.

The topics covered in the book are

- Definition and processing of multivariate data
- Principal component analysis
- Calibration methods
- Classification methods
- Cluster analysis
- Preprocessing

In addition, there are three appendices with (i) list of abbreviations, (ii) basic principles of matrix algebra and (iii) introduction to ***R***, open source software environment.

The ***R***, open source software environment mainly intended for statistical computing, is referred in every book chapter, offering particular algorithms with the information about the computer libraries and functions accessible to public from this open source software. Many figures in the book are generated by special functions defined in ***R***

and are all well documented in individual chapters, so that then user can easily reproduce them or apply in their own case studies.

In chapter “Introduction” more common topics as data distribution and spread of data are explained in short. Most tests for the data distribution are only listed with short examples showing how to use them with ***R*** software. At the end of this chapter reader can obtain impression of reading ***R*** manual.

But in the following chapters the authors demonstrate their extensive knowledge of statistical methods.

Problems of preprocessing multivariate data are explained in details and upgraded with practical examples. At the end of second chapter “Multivariate data” there is short and concise summary with helpful tips on how to start working with new set of multivariate data.

Principal Component Analysis (PCA) is explained schematically (blocks - matrices -examples) as well as with all explicit equations. Three algorithms for solving the PCA are described in detail. In this chapter also complementary methods (Factor analysis, Cluster analysis, Kohonen mapping, Sammon´s nonlinear mapping, and Multiway PCA) are described. At the end of the chapter a short and concise summary enables reader to use PCA and complementary methods in proper way.

In the chapter “Calibration” first Overfitting, Underfitting and performance criteria of regression models are explained. Then difference between linear regression models, which are calculated with Ordinary Least Squares (OLS) and Robust Regression method, is explained. Several methods for variable selection are presented such as Stepwise selection method, Best-Subset regression, Variable selection based on PCA and PLS models, Genetic algorithms and Cluster analysis of variables. With practical example authors suggest that Genetic Algorithm, Best-Subset Regression and Stepwise Selection Method (start with empty mode) are proper methods for variable selection. Principal Component Regression method (PCR) is briefly explained. Chapter “Calibration” is mainly focused on Partial Least Squares (PLS) method. All PLS mathematical aspects are explained schematically. Several algorithms for solving the PLS are described in detail. It is surprising that popular Variable Importance in the Projection (VIP) coefficients are not mentioned in the text. But nevertheless VIP coefficients can be calculated with ***R*** software. Related methods (Canonical Correlation analysis, Ridge and Lasso Regression) as well as nonlinear regression methods (Basis expansions, Kernel Methods, Regression Trees and Artificial Neural Networks) are described. The importance of the validation of data-driven models is shown. The validation depends on the nature of the modeling methods, so not all validation methods are applicable. For this reason there are several paragraphs for different validation approaches. Since Ridge Regression method has been shown by practical example as the best method, it was of special interest for our research group.

In the chapter “Classification” it is shown that several methods used for calibration are useful tool for classification (OLS, PLS and Artificial Neural Networks). First Linear classification methods are explained. Two different approaches to Linear discriminant analysis are described: Bayes discriminant analysis and Fisher discriminant analysis. From Linear regression methods for discriminant analysis authors present Binary classification, OLS, PLS, and Logistic regression method. SIMCA, Gaussian Mixture Model and k-NN classification are described as Kernel and prototype methods. There are detailed descriptions of Classification Trees, Artificial Neural Networks and Support Vector Machines (SVM). With examples it is shown that different classification methods need different approaches and they can not be directly compared. But on selected data we can notice that k-NN classification, Classification Trees and SVM performed better than other methods.

In the chapter “Cluster Analysis” authors explain that Cluster Analysis methods perform unsupervised learning and Classification methods supervised learning. For unsupervised learning Partition Methods, Hierarchical Clustering Methods, Fuzzy Clustering and Model-based Clustering can be used. To evaluate clusters authors describe also Cluster validity and Clustering tendency methods. With the examples Model-based Clustering was selected as the best method for cluster analysis.

In the last chapter the authors shortly describe Preprocessing methods.

In the book there are three appendices: “Symbols and Abbreviations”, “Matrix Algebra” and “Introduction to ***R***”. With the last appendix we can argue that this is not "yet another book" on statistical methods, but also a very useful handbook for chemometrics and a tutorial with detailed instructions for using ***R*** software.