Chemometrics and Intelligent Laboratory Systems 103 (2010) 19–24

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Chemometric view on "comprehensive chemometrics"

Alexey L. Pomerantsev *, Oxana Ye. Rodionova

Semenov Institute of Chemical Physics RAS, Kosygin 4, 119991 Moscow, Russia

ARTICLE INFO

Article history:
Received 16 April 2010
Accepted 8 May 2010
Available online 15 May 2010

Keywords:
Comprehensive chemometrics
Science metrics
State of the art

ABSTRACT

This paper presents a critical review on "Comprehensive chemometrics. Chemical and biochemical data analysis", Steven D. Brown, Romà Tauler, Beata Walczak (Eds.), Elsevier, ISBN: 978-0-444-52702-8, Hardcover, 2.200pp, March 2009, € 1360, which is a four-volume set written by about 160 authors from 20 countries. Book evaluation is presented in two forms. The first one is a traditional subjective opinion of the reviewers after a careful study of the whole book set. The second approach is an attempt to employ objective multivariate analysis to the presented material and connect it with the state of the art of modern chemometrics in periodical publications. Objective explorative analysis confirms our personal impression that this book gives an adequate presentation of the modern chemometrics and should find a merited place on bookshelves alongside its predecessors.

1. Introduction

The up-to-date state and new applications of chemometrics as an interdisciplinary and continue to growth area needs to be available to a wide range of new and experience practitioners. Many books devoted to the basic and highly tailored aspects of chemometric can be found elsewhere.

Elsevier maintains a special series called "Data handling in science and technology" [1]. The first attempt to present a comprehensive chemometric textbook inside this series should be attributed to "Chemometrics: A textbook" printed in 1988 [2]. This is a one-volume book of 488 pages. The second book of such a kind [3,4] was published ten years later. It consists of two volumes and of about 1600 pages. Now, some ten years later, it is a turn for the modern version of Comprehensive Chemometrics. The presented resource [5] is a four-volume book comprising 90 chapters and about 2400 pages. In contrast to the previous ones the latter does not present a holistic text written by a small team of the authors but a set of reasonably self-contained chapters offered by about 160 researches.

The editors planned a work "that would cover all of the major areas of chemometric research and a wide sample of current applications" and also "a resource that captures the practice of chemometrics in the early twenty-first century." Of course the editors tried to solve a very ambitious problem. The aim of the current review is to present this giant work to the readers trying to answer on the following questions.

1. For whom the first and foremost this book is written? What is the most appropriate way of the employment of this book?
2. Does this book adequately reflect the modern state of chemometrics, both by the depth and breadth of presented materials?
3. Where are the frontiers of chemometrics now?

We tried to evaluate this book from two points of view. The first one is our own opinion after careful study of the whole resource. Of course, such an opinion is subjective. But we also tried to find some objective criteria for the book evaluation using powerful chemometric approach, i.e. multivariate data analysis. This approach was based on the datasets that were extracted from the book and the relevant periodic publication. These data can be found in the supplementary materials.

2. Materials

The Comprehensive Chemometrics set of books consists of four volumes that are organized in $3+3+5+1=12$ sections.

The first volume outlines the fundamental principles, which constitute the basis of chemometrics, underlies all methods employed in the area. The volume consists of three sections. The first part (8 chapters, edited by L. Sarabia) is devoted to Statistics. The presented topics include: Theory of Sampling, various aspects of Quality Assurance, Resampling and Robust techniques, and Bayesian approach. The overall exposition is not even; very good written texts are mixed with the chapters that suffer from the serious mistakes.

The second part of Volume 1 titled Experimental Design (7 chapters, edited by R. Phan-Tan-Luu) surveys different aspects of the technique, including Screening, Factorial, and Mixture Designs, and Response Surface method.

Optimization is considered in the third part (5 chapters, edited by R. Leardi). It is opened with a short review on the relevant techniques. Specific topics include Sequential, Gradient, and Multicriteria

* Corresponding author. Tel./fax: +7 495 9397483.
E-mail address: forecasts@chph.ras.ru (A.L. Pomerantsev).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.05.001

Methods. The last chapter represents an excellent introduction into Genetic Algorithms.

The second volume consists of three parts. The first part (11 chapters, edited by J. Trygg) presents various aspects of data preprocessing as the inherit part of each multivariate data analysis. The selection of the appropriate method depends on the nature of the measurement data and the problem under consideration. It is possible to divide all preprocessing methods on two categories. The first one is a set of unsupervised methods, applied before data modeling for reducing the influence of the random variation without altering the signal and also classical methods for spectral normalization and differentiation. The latter category is the model dependent methods such as OSC, OPLS and O2PLS. Both categories are presented in detailed and rigorous manner with algorithms and very often also with the Matlab codes. Some repetitious, such as twice presentation of Savitzky–Golay algorithm (in chapters 2.02 and 2.03), can be attributed to the fact of its importance and wide-spread application in chemometrics. It was unexpected to find chapter 2.10 "Batch Process Modeling and MSPC" in the preprocessing part of the book.

The title of the second part "Linear Soft-Modeling" (14 chapters edited by A. de Juan) seems to be too general. The input information for different methodologies of this part is presented as a single data with various degree of complexity: two-way, multiway, and multiset data. Chapter 2.13 is devoted to the principal component analysis (PCA) as the basis for all other techniques presented in this part. PCA is presented in a rather traditional manner, unfortunately such an important aspect as different approach for the determination of the number of principal components does not considered at all. This aspect has a special importance as a lot of curve resolution methods presented in the consequent chapters are started with determination of the rank of a bilinear systems based on PCA methods and have reference to chapter 2.13. It is worthy of mentioning that chapters devoted to various aspects of multivariate curve resolution (MCR) methods and also to the multiway data analysis are written in a strict and clear manner. In comparison with this chapters it was strange to read chapter 2.14 "Independent Component Analysis", that is written in a careless mathematical manner. The following expressions can be found in this chapter. Page 228: "The eigenvalues of X are ordered according to value", but when matrix X is not a square matrix such an entity as eigenvalue is not defined for X. Page 229: "$T = W^{-1}$", what does the matrix inverse mean for non-square matrix? These are only several of numerous ungenerous expressions of this chapter. It is also unclear why PCA is described in this chapter right after the whole chapter devoted to PCA. The same can be said about MCR which is described in details in four consequent chapters.

The third part of volume 2 "Unsupervised Data Mining" (6 chapters, edited by D. Coomans) deals with classical pattern recognition methods. Four of six chapters are presented by mathematicians from data mining community. Therefore unsupervised classification tools are presented wider than ordinary used in chemometric applications. The method description is done in a rigorous and clear manner with formulas and algorithms. We consider that this material can serve as valuable guide in chemometric practice.

The third volume is divided into 5 sections. The first part titled Linear Regression Modeling (9 chapters, edited by J. Kalivas) begins with a nice review of this topic. Other chapters discuss various problems related to calibration: Diagnostics, Validation, Preprocessing, Variable Selection, Missing Data, Robust methods, Model Transfer, and Three-Way methodology.

The title of the second section of volume 3 is Non-Linear Regression, (5 chapters, edited by L. Buydens). However, you will not find here the classical theory of the non-linear regression. The first chapter presents an introduction to the kinetics modeling. The next chapter discusses SVM regression and classification. It is rather unusual to find SVM in the regression section instead of the next, Classification, part. Other topics presented in this section outline Locally Weighted Regression, Neural and Fuzzy approach, Classification and Regression Trees, and Projection Pursuit Regression. The last chapter gives a nice introduction into Neural Networks methods.

The third part of volume 3, Classification (5 chapters, edited by B. Lavine), begins with an overview of Basic Concepts. The methods presented here are limited to Statistical Discriminant Analysis, Decision Tree Modeling, and Feed-Forward ANN. The last chapter discusses the problem of validation. We have already mentioned that SVM is presented in the regression section. Another strange fact is that very popular SIMCA approach has a rather sketchy exposition in this section.

The next section is called Feature Selection, (4 chapters, edited by B. Lavine). It includes, besides introduction, the outlines of Uninformative Variable Elimination method, Genetic and wavelet algorithms in application for the variable and feature selection.

The last section of volume 3, Multivariate Robust Techniques, edited by P. van Espen consists of only one chapter Robust Multivariate Methods in Chemometrics that presents excellent overview of this problem.

Volume 4 (15 chapters) is completely devoted to various application areas where chemometric methods are the inherent part of the researches. Chapters are presented by recognized specialists in these areas. Each chapter is a self-contained material and presents the specific subject in an individual manner. Some chapters are organized as a condescend guide which can be used by practitioners in their further studies. Chapter 4.01 presents the theory of sampling and also discusses test set validation problems. Chapter 4.02 deals with all aspects of multivariate statistical process control and fault detection. This chapter partly overlapped with chapter 1.04 (Statistical Control of Measures and Processes) and chapter 2.10 (Batch Process Modeling and MSPC), but due to detailed and consecutive material presentation may be used as self-reliant textbook on MSPC. Chapter 4.10 "Chemometric role within PAT context" is very close to chapter 4.02.

Some chapters present mainly the reviews of modern state of the art in different areas. For example chapter 4.03 presents the review of chemometric applications in environmental studies and underlines specific models and approaches attributed to these studies. Chapter 4.04 is a review of chemometric applications in food sciences. This chapter also contains a set of 21 real data as Excel-files that can be helpful for material understanding and used in future as some kind of "standard data sets". The emphasis of this review is on possible mistakes and wrong conclusions that can be done by poor understanding of chemometric background. In the same manner the material is presented in chapter 4.12 — Chemometric Analysis of Sensory Data.

Chapter 4.06 is devoted to the application of hyperspectral imaging with combination with PLS and ANN that is an important opportunity for rapid monitoring of biological and agricultural products.

Chapters 4.07–4.09 describe various biological applications, and they may, as it is mentioned in the book, "contributing to the building of a bridge across the gap between the two scientific communities", biochemists who are not trained in data analysis and data analysts who lack biological training.

Chapter 4.15 is devoted to application of GRID Computing. Though the authors consider that "The increasing size and resolution of chemometric datasets requires more sophisticated methods and, in turn more sophisticated computing techniques such as Grid computing." it should be mentioned that application of GRID computing is rare even for such huge datasets as omics-data.

3. Methods

The considered four volume book comprises a huge collection of 90 papers. This arises a problem for a reviewer. If each paper is evaluated by 50 words (3–4 sentences), it gives ca 4500 words, or about 16 pages of plain text. Needless to say that such a description will be highly subjective as it is not enough to claim that a chapter is

inappropriate but some arguments must be presented to elucidate this slighting judgment. Therefore, it has been decided that some quantitative method is required for the book evaluation. This, more or less, objective approach is based on the chemometric explorative analysis applied to two datasets acquired from the papers' collection.

3.1. Internal data set

The first dataset is obtained in the following manner. At first, we selected 24 keywords that are believed to represent the modern state of chemometrics. The list includes the simple words like *outlier*, *robust*, popular abbreviations such as *PLS*, *PCA*, and complex logical expressions, e.g. <*SNV* OR *MSC*>. We did not use the keywords with a general meaning, such as *calibration*, *classification*, *regression*, etc., because they give too high number of citations. Each keyword was searched throughout the book and the number of occurrences in each chapter was counted. The procedure resulted in the 90×24 matrix **X**, in which 90 chapters are the samples (i) and 24 descriptors are the variables (j).

Utilized keywords are presented in Fig. 1, where label VS indicates the combination <(*variable* OR *feature*) AND *selection*> and label "way" means the expression <*multi-way* OR *n-way* OR *three-way*>. In this plot the overall keywords' frequencies, i.e. the column-wise sums $\sum_i x_{ij}$, are presented. The most popular in the book is a word *PLS* that was used for 1365 times, and the least popular is *PAT* with 124 occurrences.

Calculating the row-wise sums, i.e. $\sum_j x_{ij}$, we obtain the total usage of these keywords in a chapter. For example, chapter 3.24 "Robust Multivariate Methods in Chemometrics" has 726 occurrences of all keywords, while chapter 2.05 "Denoising and Signal-to-Noise Ratio Enhancement: Splines" has only 4 mentioned.

Internal data set aims at the achievement of two goals. First, it can be used for understanding the modern structure of chemometrics, i.e. the tendencies, the up to day topics, etc. Second, it can be employed for the assessment of internal book composition, correlations, links, and references between the chapters.

3.2. External data set

The external data set was established as a vector of 90 elements (the number of chapters in the book). Vector's elements were calculated as follows. Each chapter was characterized by a descriptor, which is a combination of words selected in a way providing the best presentation of the chapter topic. For example Chapter 3.03 "Validation and Error" was presented by the following logical expression <("standard errors" OR "figures of merit") and "multivariate calibration">. The database for the descriptor search was the collection of the original and review papers published from year 1999 to 2009 in 28 journals, which are the most popular in the chemometric area: Analytical Chemistry, Analytica Chimica Acta, Analytical and Bioanalytical Chemistry, Chemometrics and Intelligent Laboratory Systems, etc. Using the Scopus system [6], 42 215 papers have been explored for the presence of a particular keyword combination. The obtained values (number of papers) form the external data set, **y**, which is presented in Fig. 2.

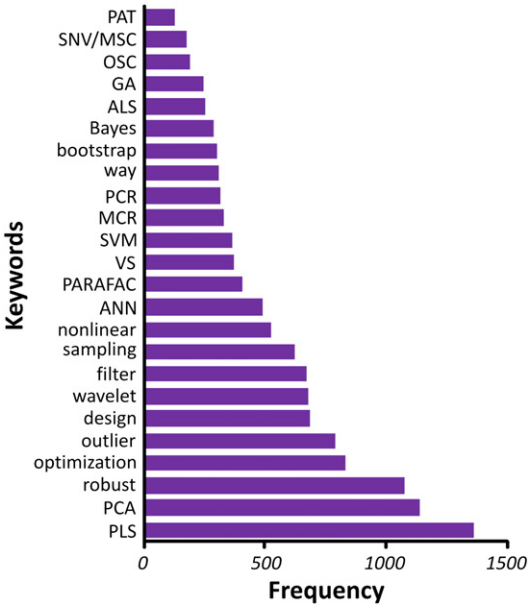

Fig. 1. Keywords frequencies in the book.

The highest value, 2975, was obtained for chapter 2.03 "Denoising and Signal-to-Noise Ratio Enhancement: Wavelet Transform and Fourier Transform" with the key combination <"Wavelet Transform" OR "Fourier Transform">. The lowest value, 0, was found for chapter 4.15 "High-Performance GRID Computing in Chemoinformatics" represented by descriptor <"GRID computing">. Several chapters were not evaluated. Among them there were all introductory texts, such as chapters 1.09 "Experimental Design: Introduction" and 1.15 "Experimental Designs: Conclusions, Terminology, and Symbols". The substantial chapter 2.04 "Denoising and Signal-to-Noise Ratio Enhancement: Derivatives" was not evaluated too. In this case we could not find a relevant keyword combination, since *numerical differentiation* gave only 2 occurrences, but *derivative* was found in 12,968 papers.

The external data set was intended for the assessment of the topics' relevance, i.e. for evaluation of their "popularity" amongst the modern chemometric publications.

4. Results and discussion

4.1. Internal data set

It is natural to subject the internal data set **X** to PCA. Data were centered but not scaled. The preliminary results are presented in Fig. 3.

The first plot (a) demonstrates the Explained Residual Variance in Calibration. It can be concluded that chemometrics, as it is presented in the reviewed book, is a rather complicated science that needs more than 10 PC for the explanation. Plots (b) and (c) are the Bi-plots, in which the scores (dots and squares) are shown together with the loadings values (crosses). It can be seen that the first PC is connected with keywords *robust* and *outlier* presented in chapters 3.24 and 3.07. The second PC is linked to *wavelet* (chapters 2.03 and 3.23) and *filter* Chapter 2.02). The third PC is associated with *PLS* (chapter 3.04) and the forth PC with *sampling* (chapters 1.01 and 4.01).

However, some irregularity in the structure of the PCA scores can be seen. Several samples (chapters) manifest the extreme behavior with respect to the majority of the objects. To reveal these samples a SIMCA like test was performed. The score (h_i, leverages) and orthogonal (v_i, residual variances) distances were calculated and considered together in the influence plot, shown in Fig. 3d. The acceptance area was calculated in line with the method described in [7] for significance 0.05. For better visibility the axes were scaled using the average values h_0 and v_0 and the root square transformed. Nine samples marked with squares (Fig. 3d) have been revealed as outliers. The term outlier has no negative meaning in this context, because each extreme chapter is mainly devoted to a specific topic, which is so far not very popular or wide spread in the modern chemometrics. E.g. chapter 4.01 "Representative Sampling, Data Quality, Validation – A Necessary Trinity in Chemometrics" mentioned the keywords for 244 times including: 1 *SNV*, 1 *optimization*, 3 *PAT*, and 238 *sampling*. It would be better to call such chapters as frontiers.

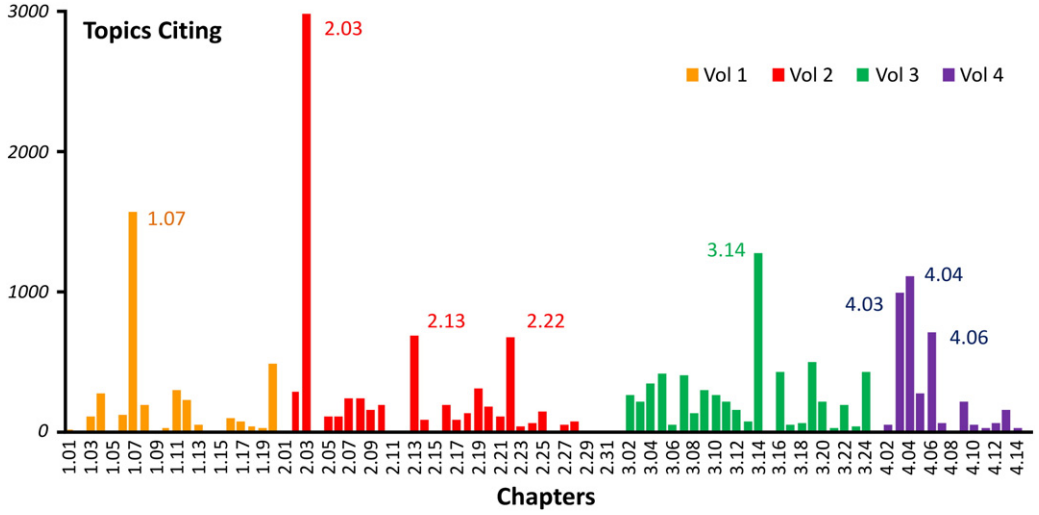

Fig. 2. Topics (chapters) citing.

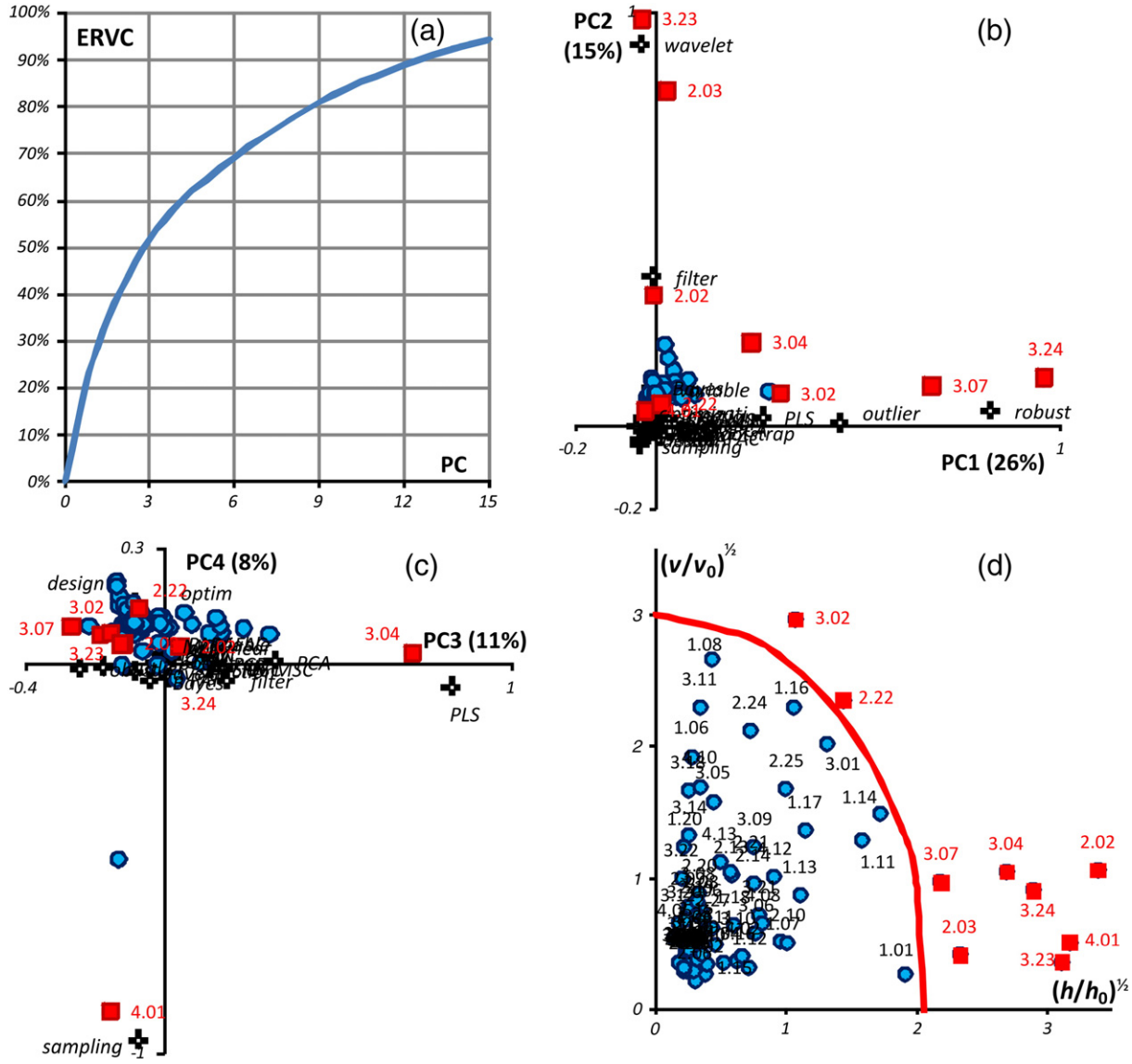

Fig. 3. Preliminary PCA results for internal data set **X**. Outliers are marked with red squares.

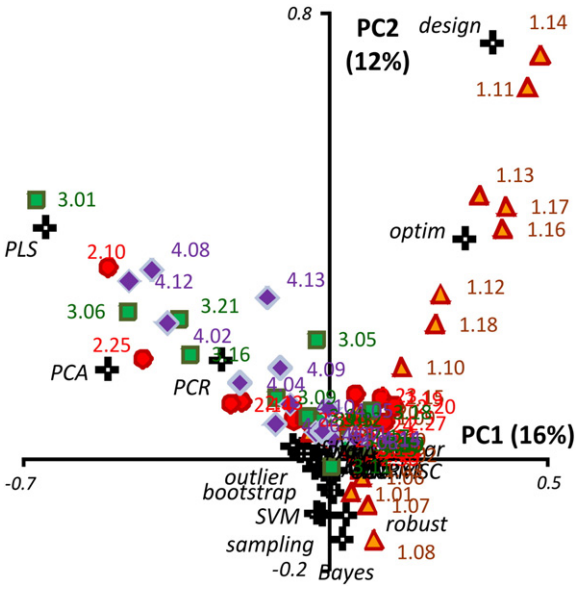

Fig. 4. PCA bi-plot. Reduced internal data set. Triangles mark Volume1, dots mark Volume 2, squares stand for Volume 3, and diamonds mark Volume 4 chapters.

The results of PCA applied to the reduced data set (without 9 frontiers) are shown in Fig. 4. Studding this bi-plot one can conclude that chapters from Volume 1 (triangles) form a separated subset that is completely orthogonal to the rest of the book. In Volume 1 the main keywords are *design* and *optimization*, while in volumes 2–4 they are *PLS*, *PCR* and *PCA*. At the same time both subsets take up a clear diagonal position in the Bi-plot. This means that both Volume 1 and the rest of the book equally contribute to the PC space spanning.

Fig. 5 presents the keywords distribution in the reduced internal dataset with respect to the volume number. To normalize the values, they are divided by the number of chapters in the subset. It may be seen that Volume 1 does not use words *PLS* or *PCA*, while the rest of the book rarely employs *design* and *Bayes*.

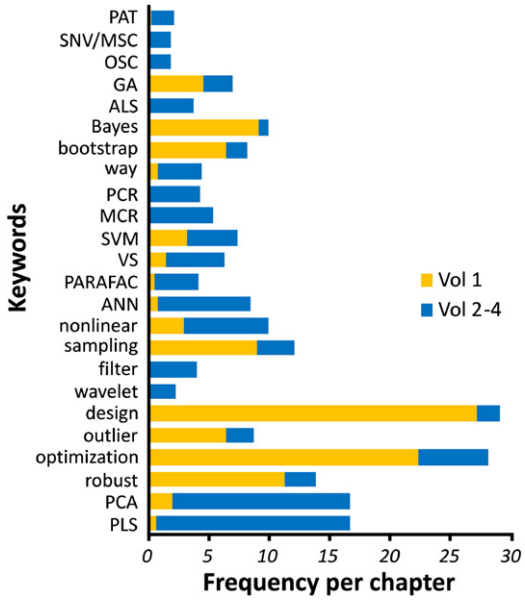

Fig. 5. Keywords' frequencies per chapter with respect to the volume number. Reduced internal data set.

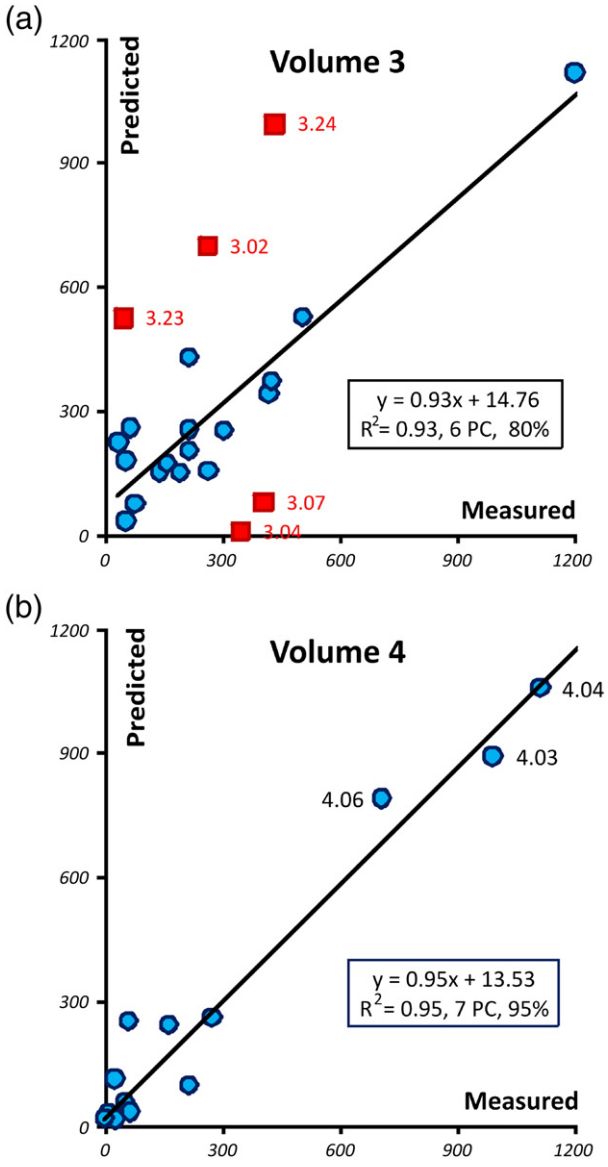

Fig. 6. PLS results for external data set evaluation. Outliers are marked with red squares.

4.2. External data set

Joint evaluation of the internal and external data sets was performed by PLS regression analysis where internal data **X** is considered as a predictors' block and the external data set **y** is a response vector. This regression model can explain the relationship between the book's internal structure and the chemometric world tendencies as they are presented in the periodical literature.

The first attempts to relate **X** and **y** was not successful; the squared correlation coefficient R^2 does not exceed 0.32 up to 24 PCs. Therefore, having in mind the different structure of the volumes, we tried to fit data separately for each volume subset. This approach gave much better results. In Fig. 6 two plots of the "Measured vs. Predicted" are presented. Models for Volumes 1 and 2 are very similar to the Volume 3 regression and are not shown here. All these models have several outliers, which mainly coincide with the PCA extremes. Models complexity is 5–6 PCs that explains about 85% of **y** variation. We did not concern with models' validation since the PLS regressions are not intended for any prediction in the future.

It is interesting that only the Volume 4 model has no outliers. This could be explained by the scope of this part of the book that is similar to the regular research or review papers published in the periodic literature.

5. Conclusions

Comprehensive Chemometrics is not a convention textbook in the sense that its contents are carefully distributed in the chapters. It is more like a very helpful dictionary or an encyclopedia, which chapters are organized in a self-sufficient manner, constituting together a number of good introductions for obtaining comprehensive and essential information in the different areas of modern chemometrics. Every chapter can be studied separately; therefore it is possible that the reader can proceed directly to the chapter of her/his interest without any loss in information. The penalty for such a layout is that some information is repeated in different places in similar or different words. This, however, could help the reader to recall or rethink a topic in a different way following the author's preferences.

Objective explorative analysis confirms our personal impression that this book gives an adequate presentation of the modern chemometrics. To answer the question about the frontiers it is enough to list the chapter descriptors that relate to the outliers found in the analysis of the internal data set. They are: *wavelets*, *n-way*, *outliers*, *homoscedasticity*, *preprocessing*, *robust*, *feature selection*, and *sampling*. The result is not very surprising and many fellow-chemometricians could agree with these conclusions.

It has been about 20 years since the first chemometric handbooks appeared and now everybody has to acknowledge their significance in analytical chemistry. The chemometric concept in these books has been presented to the interested non-chemometrician in a manner that does not suppose a very good background in statistics or matrix algebra. The new collection provides an essential input in presenting and clarifying the state of the art in chemometrics, and it can be highly recommended to anyone working in this field. For this reason this book should find a merited place on bookshelves alongside its predecessors.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemolab.2010.05.001.

References

[1] http://www.elsevier.com/wps/find/bookdescription.cws_home/BS_DHST/description.
[2] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman (Eds.), Chemometrics: A textbook, 1988, 488 pp.
[3] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke (Eds.), Handbook of Chemometrics and Qualimetrics: Part A, 1998, 867 pp.
[4] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke (Eds.), Handbook of Chemometrics and Qualimetrics: Part B, 1998, 713 pp.
[5] Comprehensive Chemometrics. Chemical and Biochemical Data Analysis. Editors-in-Chief: Stephen D. Brown, Romà Tauler, and Beata Walczak, 2009, 4 volumes.
[6] http://www.scopus.com/home.url.
[7] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chem. 22 (2008) 601–609.