Selected Aspects of 40 Years Applied Chemometrics

Kurt Varmuza
Vienna University of Technology, Austria
Email: kurt.varmuza@tuwien.ac.at; www.lcm.tuwien.ac.at/vk/

4th Autumn School of Chemoinformatics, Tokyo, Japan, 26 November 2015
Abstract for lecture

1. Introduction

In year 1975 the American chemist and mathematician Bruce R. Kowalski published a first overview about a chemical discipline called chemometrics (Kowalski 1995). The name itself has been suggested 1972 and 1974 by the Swedish chemist Svante Wold (Wold 1972, 1974). A commonly accepted definition of chemometrics is "Chemometrics is a chemical discipline that uses statistical and mathematical methods to design or select optimum procedures and experiments, and to provide maximum chemical information by analyzing chemical data." This definition is very broad, however, the essential part of chemometrics is still the application of multivariate data analysis. The latter is an important area in statistics, and actually, during the last decades, statistics and chemometrics experienced mutual benefit, although not without frictions. The areas of QSP(A)R - quantitative structure-property (activity) relationships and interpretation of molecular spectra (systematic elucidation of chemical structures) are bridges to chemo- and bio-informatics. Chemometrics provides formal methods and the software tools for transforming chemistry-related data into useful, non-trivial predictions.

Chemometric methods are now routinely used in various fields of analytical chemistry and chemical technology, however, the easy and routine use of (commercial) software packages sometimes hide the necessity of an - at least basic - understanding of the applied methods and their limits. Especially, a critical evaluation of empirical multivariate models is essential, because such models are typically not based on theory (first principles) but are data driven and based on assumptions about useful relationships between the data. Considering the parsimonious principle "make models as simple as possible, but not too simple", is important.

Good sources for the traditional chemometric methods are still the *blue books* by Massart and co-authors (Massart et al. 1988, 1997, Vandeginste et al. 1998). A subjective selection of more recent introductory overview books comprises (Brereton 2006, 2007, 2009), (Mark et. al. 2007), (Otto 2007). The books (Varmuza et al. 2009) and (Wehrens 2011) include codes for the free programming environment R, nowadays often used beside traditional Matlab. For method development the book (Hastie et al. 2008) is a substantial reference to relevant statistics.

In this tutorial a few selected topics of chemometrics are discussed: robust multivariate methods (focus of this abstract), linear latent variables (including random projection), and evaluation of regression and classification models. Some examples use mass spectral data measured on meteorite samples and on comet particles (near a comet, by COSIMA/Rosetta); (ESA 2015, Kissel et al. 2007, Schulz et al. 2015, Varmuza et al. 2015).

2. Overview of robust methods in chemometrics

For details see textbooks on statistics; R codes are, e. g., contained in (Varmuza et al. 2009).

2.1. Chemometrics often deals with experimental data which may contain outliers and may not have a requested distribution. Robust statistical methods are less influenced by outliers or, e. g., by deviations from a normal distribution. They use other criteria (estimators) than classical methods (e. g., the median instead of the arithmetic mean), or give the observations weights depending on their outlying behavior (e. g., in robust regression) - however, avoid a yes/no-elimination of potential outliers. Trimmed estimators (e. g., as performance measures for a classification model) exclude extreme values (e. g., considering only the 5% to 95% range).

2.2. Basic descriptive measures for a set of n numbers (x_1, x_2, ..., x_n) characterize their central value and their spread. The classical estimators arithmetic mean (*mean*) and standard deviation (s) are sensitive to outliers. Simple robust counterparts are the median (*med*) and the interquartile range (*IQR*). In the case of a normal distribution, s_{IQR} = 0.7413 *IQR* provides a robust estimation of s. An alternative to *IQR* is the median absolute deviation (*MAD*), defined as the median of the absolute differences $|med - x_i|$; another robust estimation of s is s_{MAD} = 1.483 *MAD*. The median and the robust measures of the data spread can be advantageously used in further data evaluations or transformations (e. g., for the often applied autoscaling).

2.3. Classical measures for characterizing a linear relationship between two variables (measurements) x_j and x_k are the (Pearson) correlation coefficient, r_{jk} (range -1 to +1), and the covariance c_{jk}, with $r_{jk} = c_{jk} /(s_j\, s_k)$. Robust correlation measures are the Spearman rank correlation and the Kendall's tau correlation (both with ranges -1 to +1). For the covariance matrix (a basic object in multivariate data analysis) several approaches have been suggested for a robust estimation; however, most require more objects than variables, which is rarely fulfilled in chemistry-related data; for a summary see (Varmuza et al. 2009, page 43 f.).

2.4. In multivariate data analysis, the distance between objects in the variable space is considered as a measure of the similarity of the objects, and widely used methods are based on this concept (e. g., PCA, principal component analysis). The mostly used Euclidean distance (and also the Mahalanobis distance) are highly influenced by outliers. The latter is used for outlier recognition; in this case a robust estimation of the data center and the covariance matrix (necessary for the Mahalanobis distance) are essential.

2.5. Linear latent variables are the basic concept of the most used multivariate data analysis methods in chemometrics, such as PCA, PLS (partial least-squares regression), and LDA (linear discriminant analysis, including PLS discriminant analysis, PLS-DA). A linear latent variable (component, factor) is a linear combination of all (or selected) variables. The parameters (loadings, regression coefficients) of the linear combination are controlled by the aim of data analysis or modeling: (a) maximum variance of the scores (the values of the component) for PCA, (b) maximum covariance (or correlation coefficient) between the scores and a given y-property of the objects (samples) for PLS (or OLS, ordinary least-squares regression); (c) scores with maximum discrimination between two object classes (LDA, PLS-DA).

2.6. PCA calculates latent variables with maximum variance and is therefore sensitive to outliers. Robust versions of PCA use instead of the classical variance (s^2) a robust estimation (e. g., via MAD) for searching the principal components (projection pursuit). Another approach estimates a robust covariance matrix, followed by a classical PCA, e. g. by eigenvector computation of this matrix.

2.7. For multiple linear regression (MLR), a robust objective function can be used instead of the classical non robust sum of squared residuals, e. g. the M-estimate (Maronna et al. 2006). Robust principal component regression (PCR) combines a robust PCA with a robust MLR. Robust versions of the most used regression method in chemometrics, PLS, can be realized by using a robust measure for the covariance (for searching the PLS components), or by down weighting large absolute residuals (partial robust M-regression, PRM-PLS), (Serneels et al. 2005).

2.8. Robust methods in multivariate classification gain high attention in machine learning but have found only little notice in chemometrics up to now. Because classical PLS-DA is widely used by chemists, robust versions may be a first step to robustification, as shown below.

3. Example

The COSIMA instrument (Kissel et al. 2007) onboard of Rosetta collects dust particles (emitted from a comet) on metal targets of size 1 cm × 1 cm; typical diameter of these grains is 20 - 500 µm (Schulz et al. 2015). The primary ion spot of the TOF-SIMS instrument has a diameter of ca 70 µm, and a position accuracy (x-, and y-coordinates) of ca 80 µm. Consequently, some spectra are accidentally measured “On grain” and some “Off grain”. A primary task in data evaluation is an automatic recognition of potentially relevant spectra (not from background). For method development a meteorite grain has been investigated by a twin laboratory instrument; 63 spectra are from background (group 1), 155 from grain or near grain (group 2); no. of variables is 2612. Fig. 1 shows the classification result obtained by a robust PLS-DA, which considers the uncertain assignment of the group 2 spectra by an appropriate weighting (Hoffmann et al. 2015).

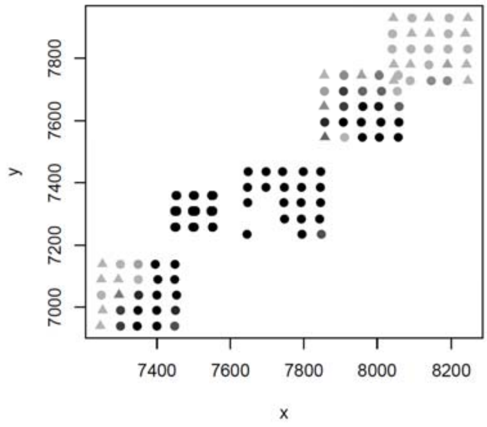

Fig. 1. Classification of 155 TOF-SIMS spectra measured on and near a meteorite grain. x, y, coordinates of measurements [µm]; circles indicate spectra assigned to the grain by a robust PLS-DA; triangles indicate assignments to background; the gray tone is the resulting weight of the spectra for being used as grain spectra (black for weight = 1). The area covered with dark circles indicates the (irregularly shaped) meteorite sample. This method is now successfully applied to spectra measured on cometary grains onboard of Rosetta. (Meteorite sample provided by Natural History Museum Vienna; spectra measured by M. Hilchenbach, Göttingen, Germany).

References

Brereton, R. G.: Chemometrics - Data analysis for the laboratory and chemical plant. Wiley, Chichester, United Kingdom, 2006.

Brereton, R. G.: Applied chemometrics for scientists. Wiley, Chichester, United Kingdom, 2007.

Brereton, R. G.: Chemometrics for pattern recognition. Wiley, Chichester, United Kingdom, 2009.

ESA, European Space Agency (2015) http://blogs.esa.int/rosetta/

Hastie, T., Tibshirani, R. J., Friedman, J.: The elements of statistical learning, 2nd ed.. Springer, New York, NY, USA, 2008.

Hoffmann, I., Filzmoser, P., Serneels, S., Varmuza, K.: Sparse and robust PLS for binary classification; TU Vienna, in preparation (2015).

Kissel, J. et al.: *Space Sci. Rev.*, 128 (2007) 823-867. COSIMA – High resolution time-of-flight secondary ion mass spectrometer for the analysis of cometary dust particles onboard ROSETTA.

Kowalski, B. R.: *J. Chem. Inf. Comput. Sci.* 15 (1975) 201-203. Chemometrics: views and propositions.

Mark, H., Workman, J.: Chemometrics in spectroscopy. Academic Press, New York, NY, USA, 2007.

Maronna, R., Martin, D., Yohai, V.: Robust statistics: Theory and methods. Wiley, Toronto, ON, Canada, 2006.

Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., Kaufmann, L.: Chemometrics: a textbook. Elsevier, Amsterdam, The Netherlands, 1988.

Massart, D. L., Vandeginste, B. G. M., Buydens, L. C. M., De Jong, S., Smeyers-Verbeke, J.: Handbook of chemometrics and qualimetrics: Part A. Elsevier, Amsterdam, The Netherlands, 1997.

Otto, M.: Chemometrics. Wiley-VCH, Weinheim, Germany, 2007.

Schulz, R. et al.: *Nature*, 518 (2015) 216-218. Comet 67P/Churyumov-Gerasimenko sheds dust coat accumulated over the past four years.

Serneels, S., Croux, C., Filzmoser, P., Van Espen, P. J.: *Chemom. Intell. Lab. Syst.* 79 (2005) 55-64. Partial robust M-regression.

Vandeginste, B. G. M., Massart, D. L., Buydens, L. C. M., De Jong, S., Smeyers-Verbeke, J.: Handbook of chemometrics and qualimetrics: Part B. Elsevier, Amsterdam, The Netherlands, 1998.

Varmuza, K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA, 2009.

Varmuza, K., et al.: (2015) *Conference on Solid State Analytics - FKA18*, Vienna, Austria. Recognition of relevant spectra in TOF-SIMS measurements on meteorite and comet grain samples by a chemometric approach. Poster presentation, http://www.lcm.tuwien.ac.at/vk/Manus/Poster-184-FKA-Vienna-2015.pdf

Wehrens, R.: *Chemometrics with R: Multivariate data analysis in the natural sciences and life sciences*, Springer, Berlin, Germany, 2011.

Wold, S.: *Kemisk Tidskrift* 3 (1972) 34-37. Spline functions, a new tool in data analysis.

Wold, S.: *Svensk Naturvetenskap* 201 (1974) 206. Chemometrics - chemistry and applied mathematics.

Acknowledgements. Thanks to Peter Filzmoser and Irene Hoffmann (TU Vienna) for support in statistics and R-programming. Supported by the Austrian Science Fund (FWF), project P 26871-N20.