Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 85 (2007) 269–277
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Software description

# TOMCAT: A MATLAB toolbox for multivariate calibration techniques

Michał Daszykowski [a], Sven Serneels [b], Krzysztof Kaczmarek [a], Piet Van Espen [b], Christophe Croux [c], Beata Walczak [a,*]

[a] Department of Chemometrics, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[b] Micro and Trace Analysis Centre, Universiteit Antwerpen, Universiteitsplein 1, B-2610 Wilrijk, Belgium
[c] Faculty of Economics and Applied Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

Received 8 February 2006; received in revised form 27 February 2006; accepted 18 March 2006
Available online 9 May 2006

**Abstract**

We have developed a new user-friendly graphical interface for robust calibration with a collection of m-files, called TOMCAT (*TO*olbox for *M*ultivariate *CA*libration *T*echniques). The graphical interface and its routines are freely available and programmed in MATLAB 6.5, probably one of the most popular programming environments in the chemometrics community. The graphical interface allows a user to apply the implemented methods in an easy way and it gives a straightforward possibility to visualize the obtained results. Several useful features such as interactive numbering of the displayed objects on a plot, viewing the content of the data, easy transfer of the data between the toolbox and the MATLAB workspace and vice versa, are also implemented. Among the implemented methods there are Principal Component Analysis and its robust variant, Partial Least Squares, Continuum Power Regression, Partial Robust M-Regression, Robust Continuum Regression and Radial Basis Functions Partial Least Squares.


*Keywords:* Partial Robust M-Regression; Robust Continuum Regression; Multivariate calibration; Nonlinear modeling; Radial Basis Functions Partial Least Squares

## 1. Introduction

Chemical data sets are usually multidimensional, complex and often contain more measured parameters than observations. Spectroscopic data are typical examples of such data. This is why in chemometrics latent variables methods are used to explore the information contained in the data. The most popular among them are Principal Component Analysis, PCA, allowing data compression and latent variables modeling techniques such as Principal Component Regression, PCR, and Partial Least Squares, PLS [1–3]. One difficulty that may arise while exploring and modeling the chemical data is a presence of outlying observations. In general, outlying observations are objects that have unique characteristics compared to the data majority. The outlying objects strongly affect all of the least squares methods, including PCA, PCR and PLS [4,5]. Therefore, it is important to detect them, and if necessary, to remove from the data. Another possibility to handle outliers in the data is to use so-called robust approaches that can provide reliable estimates even if outliers are present in the data [4]. Over several years, many robust versions of the classical chemometrical approaches have been proposed such as robust PCA [6–9], robust PCR [10,11], robust PLS [12–14], etc. Although in statistics robust methods have been widely accepted, their use in chemistry is rather limited. Therefore, the goal of our work is to popularize recently proposed robust methods, in particular Partial Robust M-Regression (PRM) [14], by offering to the public a collection of several classical and robust approaches. Another library of robust routines, called LIBRA, has been provided by Verboven et al. [15,16]. To facilitate the use of the implemented methods, we have developed a graphical interface that gives a user possibility to apply the methods in a straightforward way and to visualize the results.

In this article, we focus our attention on presenting the graphical interface and its features. Among others, users will find different methods including classical and robust PCA, linear calibration models such as PLS and its robust variant PRM,

* Corresponding author.
*E-mail address:* beata@us.edu.pl (B. Walczak).

0169-7439/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2006.03.006

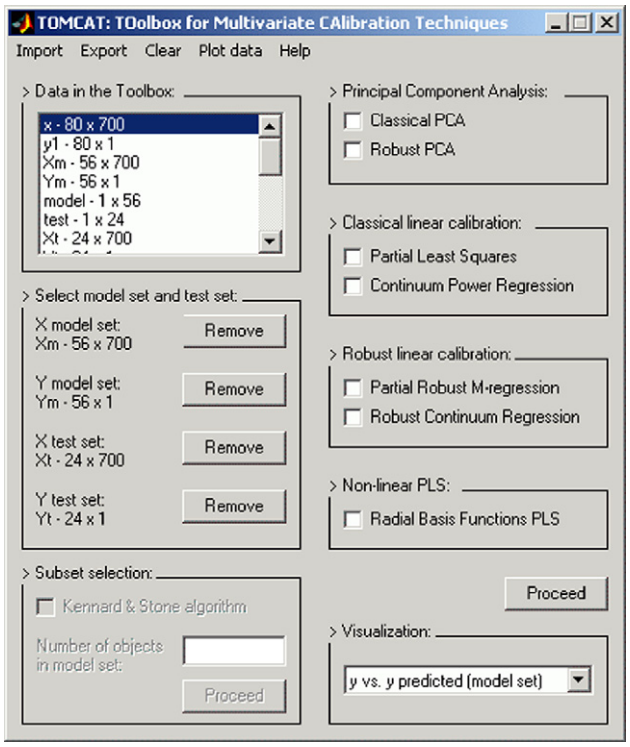


Fig. 1. The graphical user interface for robust calibration.

Continuum Power Regression (CPR), Robust Continuum Regression (RCR) and nonlinear calibration approach called Radial Basis Functions Partial Least Squares (RBF-PLS) [17,18]. The description of the implemented methods and the obtained results are reported in the related publications.

## 2. Software specifications and requirements

The routines for robust calibration and the graphical interface have been developed under MATLAB 6.5 (release 13) [19]. Dependent on the user's knowledge about MATLAB, either the m-files or the graphical interface can be used.

The interface routine, called 'TOMCAT', requires for its proper functioning a specific structure of the catalogues where the routines are located. Apart from the main interface, additional interfaces are designed for defining inputs and custom options for the applied methods. It ought to be mentioned that they are a part of the interface only. The following catalogues with m-files are automatically made when the zip file is extracted: 'Calibration', 'Data', 'Interfaces', 'PCA', 'Preprocessing' and 'Subset_Selection'.

To initialize the graphical interface, the current directory of MATLAB should be changed to the directory where the graphical interface is installed, by using the command 'cd'. Typing the command 'TOMCAT' in the MATLAB command window executes the graphical interface. When the graphical interface starts, the information about the location of the required m-files is added automatically into the MATLAB paths. Executing the graphical interface file results in displaying the interface window on the screen of the computer. After selecting the data and applying any of the implemented methods in the toolbox, the graphical interface looks similar to the one presented in Fig. 1.

The toolbox for robust calibration, is available from the two internet sources [20,21] as a compressed zip-file. Additional information about the implemented methods in the toolbox can be found at [22,23].

## 3. Collection of the implemented methods

The collection of m-files (implemented in the graphical interface methods) covers aspects of the data analysis such as classical and robust data preprocessing, subset selection, classical, robust as well as nonlinear calibration.

Among the classical data preprocessing methods there are mean column centering, column autoscaling, and Standard Normal Variate [24]. For robust preprocessing purposes, median column centering, L1-median column centering [25] and a standardization based on Sn and Qn scale estimators [26] are included.

Data compression and visualization can be performed with Principal Component Analysis (working always on the smaller data dimension, which speeds up the computations) [27] and its robust version based on the Projection Pursuit algorithm [28].

For classical calibration, a user may choose between PLS (for tall and wide data matrices, WIM-PLS and SIM-PLS) [29] and CPR [30,31]. The WIM-PLS, SIM-PLS [32] and CPR [33] routines are included into the toolbox upon an agreement of their authors. When outliers are present in the data, robust calibration methods such as Partial Robust M-Regression [14] and Robust Continuum Regression [34] can be applied.

In order to evaluate the complexity of the constructed calibration model, Cross-Validation routine is supplied with two options: a standard leave-n-out Cross-Validation and Monte Carlo Cross-Validation [35,36].

To model a nonlinear relationship between **X** and **y**, Radial Basis Functions Partial Least Squares approach is proposed.

## 4. Working with the graphical interface

The graphical interface is composed of eight panels and an upper window menu. There are the following panels: 'Data in the Toolbox', 'Select model set and test set', 'Subset selection', 'Principal Component Analysis', 'Classical linear calibration', 'Robust linear calibration', 'Nonlinear PLS' and 'Visualization'. The upper menu of the window contains at most five sub-menus, dependent on the data content in the graphical interface. These folders are 'Import', 'Export', 'Clear', 'Plot data' and 'Help'.

The graphical interface has been designed in such a way that certain options or methods can only be used if appropriate data are available and selected by a user. For instance, the PCA can be performed on the set of independent variables, **X**. Only if both **X** and a dependent variable **y** are present in the toolbox, and if both are selected, then the calibration methods can be used. There are also several checking procedures preserving unwanted action of a user, for instance, selecting twice the same data, setting the inputs of the methods at unacceptable levels, etc.