CHEMOMETRICAL PACKAGE FOR PC

Jiří Militký
Department of Textile Materials, Technical University
461 17 Liberec, Hálkova Street 6, Czech Republic
Milan Meloun
Department of Analytical Chemistry, Technical University
532 10 Pardubice, Čs. legií 565, Czech Republic
Karel Kupka
TriloByte, Ltd.,
530 02 Pardubice, U Sokolovny 21, Czech Republic

1. Introdution

Chemometrics is a relatively young discipline ranging over chemistry, mathematical statistics and informatics. Its important part comprises methods for the extraction of relevant information from chemical experiments. These methods make use of selected computer algorithms of mathematical statistics, which are discussed e. g., in Meloun *et al.* (1992) and Militký (1989). Some chemometrical computations can be performed by general statistical packages, namely SYSTAT, SAS, STATGRAPHICS. Extensive tests and studies of these and other packages have shown, however, that in many cases they do not support a researcher with suitable statistical methods nor are the numerical algorithms reliable enough (see Militký, 1990).

For these reasons, a new software package ADSTAT has been developed by the authors, who have a long-term experience in chemometrical data processing. ADSTAT consists of relatively independent modules each of which contains several programs or methods concerning one problem. Modules use a spreadsheet-like data editor with some special functions, and a graphical output. The program package is written in Turbo Pascal and makes use of the wide range of possibilities of IBM PC computers. ADSTAT version 2.0 has now started to be marketed by TriloByte, Ltd.

2. Structure of ADSTAT

ADSTAT requires 1 MB of RAM and 8.5 MB on hard disk. EGA and

VGA graphics are supported, as well as printing on matrix and laser printers. Output is provided in ASCII and TIFF formats. The presence of a mathematics coprocessor remarkably increases data processing speed. The program allows the user to work interactively or to analyze data files in batch operating mode.

The operation is simplified by a sophisticated data editor, disk file manager and dynamic graphic output together with general tabular output. Graphic output can be printed in presentation quality on various types of printers or stored into TIFF graphic format files for easy export into other program packages and text editors. ADSTAT uses extensively text windows, pull-down menus and interactive panels. The rules of their use are simple, direct and generally known from other programs. ADSTAT is controlled with keyboard and/or mouse.

Required settings are stored in file ADSTAT.CFG in the current directory. If this file is found, ADSTAT reads all settings and suggests them as implicit values. If this file is absent, ADSTAT creates a new one with start-up implicit settings. Configuration file *.CFG (implicitly ADSTAT.CFG) may be saved at any time. The following settings are saved in a configuration file:

1. Printer type and printing method, size, resolution, etc.;
2. Computing conditions separately for all methods, name and setup of text result file and graph selection setup.

2.1 *Data input*

The ADSTAT data editor is designed to make data input for statistical processing as easy as possible, to enable data file preprocessing and to allow data export and import in ASCII format (with .TXT implicit extension). The data editor uses three value types, which may be contained in a cell: numerical (-1E+99 through 1E+99), text (up to 8 characters) and empty mark.

The editor window contains work area, the local menu and the status line at the bottom with information about editor status, current block, cursor location, etc.

The work area consists of cells of a fixed size. Each cell has its own line and column coordinate.

Data can be entered in the editor in two ways:

(1) by reading data from disk in ADSTAT type file (*.ADD implicit extension) or in ASCII format;
(2) by keying-in the data directly from keyboard.

Data are saved in ADD format implicitly in compact binary code simultaneously with all information about the file, e.g. block definition, variable names. Editor capacity is 30.000 × 255 cells. Real effective capacity depends on memory and disk size.

Besides common character, word, row and block-oriented operations (such as delete, move, copy), it is possible to transpose a block, fill it with numbers with a constant increment and transform data according to a user function. Data can be stored to an ASCII-file on the disk.

2.2 *Program outputs*

All information on processing and results are written to a special result file during computations. This results file is stored in ASCII format, i.e. it is possible for the results to be displayed, printed out and processed with any text editor, as well. Results are arranged into logical paragraphs.

Nearly all methods are followed by graphic output. For some methods (exploratory analysis, residuals analysis) the graphic output is the central point of applied technique and has high confirmative value.

Graphs generated with the ADSTAT program are solely bivariate, because these are more commonly used and their interpretation is easier. Graphical information are saved into temporary file AD{GR}.BIN in internal binary form during the computation. After termination of the computation, the graphs are generated from this file with the Graphs item in menu. This file is also used for graphics in batch mode. It is possible to display up to 10 graphs at a time.

Graphs are displayed on the graphic screen. The graphics screen consists of a few components.

At the top, there is a bar with global icons. The largest area of the screen is occupied by windows with graphs. Each window consist of a frame, having a line with a graph name and graph local icons at the header, and graph body, with axes, curves and points, respectively.

At the lower status line, the most important hot keys for keyboard control are displayed.

The window with icons and graphs is the basic object on the screen. One of the windows is always active. The window can be reduced to the minimal size (iconize). It is then possible only to move it or restore it to its original size.

Graphs may be printed on a printer or written into file in commonly used TIFF black and white graphic format. The second feature is very useful for using the graphic output as report illustration, because most text editors and DTP software accept this format and can incorporate it into text.

2.3 *Organization of computation*

ADSTAT is a multiple windows system managed by a combination of pull-down menu and panels.

Windows may contain a horizontal local menu. Once activated, the windows may be reentered using their order numbers in the upper-left corner.

Interactive panels are controlled with the Tab key first of all, which moves the cursor among individual panel fields. Field items are selected with cursor arrows (the dot in parentheses is relocated in the selected item) or filling in values. Panels contain push-buttons allowing to save setup, recall setup, etc. Some panels contain a field with optional items or parameters, which may be tagged (a cross in a bracket [x]) to activate the item. Some items may have a vertical arrow pointing down one side. The arrow indicates it is possible to choose from a series of prechosen variants (e.g. selection of one of previously specified nonlinear regression models).

The program main menu is located on the first line of panel. The bottom line, or status line, is reserved for current information about the program run, brief help, error reports, etc. The whole area between the main menu and status line is reserved for windows, pull down menus and panels, where all the dialog with ADSTAT takes place. Context-sensitive help containing information about activated items can be recalled.

3. Modules of ADSTAT

In this part, brief outlines of particular modules are given. Attention is paid to the more interesting algorithms and techniques.

3.1 *Basic statistics*

This module includes programs suitable for all phases of one-dimensional sample analysis, comparison of two samples and determination of errors of an indirect measurment. Programs for both discrete and continuous exploratory analysis have special algorithms for detecting suspicious behaviour of data. Another program is used for comparing the sample distribution with ten theoretical distributions on the basis of the Q-Q and TDF graphs (Meloun *et al.*, 1992). Power transformations for an improvement of sample distribution can be provided by using special procedures (Meloun *et al.*, 1992).

In a routine data analysis only basic assumptions are usually required to be tested. Interval estimates of parameters or variances are computed by classical, robust and adaptive techniques.

One program will statistically compare two data samples. Besides classical tests, also more realistic adaptive and robust ones are available. For statistical characteristics of indirect measurements the Taylor series two-points technique and a simulation technique are used.

3.2 *Multivariate data*

In this module, programs are oriented towards processing multivariate chemometrical data i.e., estimation of parameters and their testing on the basis of one or two multidimensional samples. In a special program there is a choice of techniques for graphical representation of such data. Both symbolic and projection types of graph can be used. In the module there are also programs for basic statistical analysis of multivariate samples, correlation analysis and comparison of two multivariate samples.

3.3 *Calibration*

Calibration tasks have usually two phases. In the first one, a calibration model $f(x,a)$ itself is constructed using experimental data. In the second phase, the calibration model (or its inverse) is utilized to determine required response values with a confidence interval. In the case of a linear calibration model, both phases can be simply managed. For special cases of non-linear systems, with a complicated shape of the calibration curve, splines regression models can be successfully used. Another program in this module includes some ways to apply linear regression splines with fixed knots, which can be set up by the user according to three criteria. Quadratic and cubic splines are provided as well.

3.4 *Regression analysis*

This module serves to create linear and non-linear regression models, estimate their parameters and analyze them statistically. For linear regression, special algorithms have been implemented, and the least squares method is only a special case among a series of biased parameter estimations, controlled by a single variable. An analysis of linear and linearized models can be done. Before computations, the data can be transformed to polynomial form, by the Taylor expansion (up to quadratic terms) or generally (any variable is transformed by a user function). A powerfull SVD based algorithm is used (Militký, 1989). A variety of regression characteristics including partial regression graphs can be computed. For diagnostic purposes, a program with over 40 graphs for testing assumptions about the data model and least squares criterion has been included (Militký, 1989).

For non-linear models, an efficient algorithm MINOPT (Militký, 1989). has been developed for parameter estimation and statistical analysis. The regression model is written interactively in the usual algebraic form, with up to nine parameters. Some of the parameters can be set to constant, computation can be interrupted and restarted or aborted in any iteration.

3.5 *Smoothing*

Programs of this group are used for smoothing of experimental data with consecutive derivative or integral analysis. Two cubic spline smoothers are available, those of Reinsch and Späth (Meloun *et al.*, 1992). They differ in smoothing degree parameter and the effect of statistical weights given by the user. The algorithm of Savitzki-Golay is used for simple smoothing based on moving parabolas with adjustable length. Finally, data equidistantly distributed with respect to X-axis can be filtered by various types of digital filters.

3.6 *Analysis of variance*

Programs contained in this module provide one and two way ANOVA. All programs provide, besides a construction of a suitable ANOVA model, also verification of the assumptions on which the model has been constructed, and identification of influential points. One program performs one-way analyses for models with fixed effects and an unbalanced design. Two-way analyses can be performed using another two programs, for fixed and random effects.

A special program is designed for the two-way analysis of unrepeated measurements. Both the classical parameter estimation method and a robust one (median-polish) are used.

The programs also provide tests on the additivity of effects and a power transformation to improve the distribution of the data.

3.7 *Probability models*

In the program of this module, a 'maximum likelihood' method is used to estimate parameters of location, scale and shape for a chosen probability model. Parameter estimation and statistical analysis is realized for normal and two or three parameter Gamma, exponential, Weibull and lognormal distributions.

3.8 *Growth curves*

The growth curve module aims to assist the creation of the optimal type of growth curve and to estimate its parameters including statistical analysis. Nonlinear regression based on least squares is used for parameter estimation. The family of Schnutte's four-parameter growth curve models are offered, as well as logistic, Gompertz's, Mitzerlich's and Richards's models.

3.9 *Chemometric models*

This module contains basic non-linear chemometric models for calculation of dissociation and formation constants, processing of data from potentiometric titration and spectrofotometry. Programs make use of the optimization algorithm MINOPT (see regression analysis). Special programs are devoted to Debye-Hückel model and nonisothermal kinetics. Other chemometrical applications are constantly being added to this module.

4. Conclusion

The program package ADSTAT is built as an open system. Other modules will be added especially for specific chemometrical applications (principal components analysis, cluster analysis, PLS, etc.).

Bibliography

Meloun M., Militký J. and Forina M. (1992), *Chemometrics in Instrumental Laboratory*, Part 1, 2, Ellis Horwood Chichester.

Militký J. (1989), *Mathematical Model Building*, Mimeo, Ostrava.

Militký J., Kupka K. and Rudišar L. (1990), Proceedings Conf. COMPSTAT'90, Dubrovnik.