Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 66 (2003) 89–91
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemometrics

Software description

New Matlab software for wavelength selection

I.M. Baskir*, A.V. Drozd

Department of Chemical Metrology, Kharkov National University, 61077, Svobody sq. 4, Kharkov, Ukraine

Received 7 August 2002; received in revised form 3 December 2002; accepted 20 December 2002

Abstract

A toolbox of Matlab utilities intended for performing multivariate calibration and wavelength selection in spectroscopic methods of analysis is presented. The software provides graphical user interface (GUI), selection (possibly automated) of continuous spectral regions with a statistical test allowing to avoid overfitting the data.

Keywords: Wavelength selection; Calibration; Statistics

1. Introduction

Freely available Matlab software (toolbox with the main utility called OPTIMIZM) is proposed for multivariate calibration and wavelength selection in spectroscopic methods of analysis.

The software is destined primarily for selection of continuous regions in spectra, which are relevant for the determination of the component of interest. To avoid overfitted solutions, a statistical examination of reliability of elimination of data points is performed at each step of optimization. The calculations are managed through graphical user interface (GUI) shells. The software does not require a highly experienced user, since only limited number of parameters should be tuned or tracked.

The software, along with examples and a manual, can be sent by Internet. Please send your request to the E-mail addresses: baskir@univer.kharkov.ua and drozd@univer.kharkov.ua.

Distribution includes 48 source files (25 of them can be used from Matlab command line, these support help/cross-reference), MAT-files with illustrative datasets, manual (a text file with step-by-step instructions of using examples).

2. Requirements

Matlab 5 is required. The software uses Matlab and Stats toolboxes by MathWorks, and no third parties' utilities.

Wavelength selection is performed for the given calibration set. Performance depends on the number of samples rather than the initial number of variables, which is allowed to be both lower than 10 and higher than 1000. The software was tested on a 100 MHz Pentium computer under Windows (95, 98, NT 4.0). Optimization of wavelength set for typical problems (dozens of samples, hundreds of wavelengths) takes from minutes to dozens of minutes.

* Corresponding author. Tel.: +380-572-457248.
E-mail addresses: baskir@univer.kharkov.ua (I.M. Baskir), drozd@univer.kharkov.ua (A.V. Drozd).

0169-7439/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0169-7439(03)00002-9

3. Algorithms

Calibration by classical least squares, principal component regression or partial least squares (PLS) can be performed. For the latter two methods, the number of latent variables (NLV) is calculated automatically as in Ref. [1] or selected by user. In each run of optimization, software tries to decrease NLV.

The optimization algorithm is a variant of backward elimination [2] applied, as proposed in Refs. [3,4], to continuous regions in spectra rather than to discrete data points. This approach, although somewhat resembling the so-called "interval PLS" [5], differs from it in the following points: each run means elimination of one region, the width of regions is selected automatically, and statistics is used to examine heteroscedasticity rather than improvement of variance.

In each iteration of the optimization, one should find the way of dividing the spectrum (n data points) into several (k) regions, each containing the same number of points (e.g., for $n=300$, the possible variants are: two regions, each of 150 wavelengths, three regions, each of 100 wavelengths and so on until $k=n$). The lowest k is tried first. The value of a criterion to be minimized (see below) is calculated for each data subset being combination of all regions except one (so, for $n=300$, $k=3$, these subsets contain the points: 1–200 (No. 1), 1–100 and 201–300 (No. 2), 101–300 (No. 3)). Assuming that the values of the criterion, which is regarded as a variance with f degrees of freedom, constitute an array, Φ, of k elements, the probability, P, that the lowest variance does not belong to the population, is calculated as:

$$P = \mathrm{fcdf}((\mathrm{sum}(\Phi)/\mathrm{min}(\Phi) - 1)/(k-1), (k-1)\times f, f)^k, \quad (1)$$

where sum, min, fcdf are the corresponding Matlab functions. $P = 1 - \int_0^g \rho(x)dx$, where $g = \mathrm{min}(\Phi)/\mathrm{sum}(\Phi)$, a modification of Cochran's statistics [6], $\rho(x)$ is its probability density function. If P is greater than the stated confidence and decreases for higher k (i.e., for more narrow separate regions), the given division is accepted, and the region, elimination of which provides min(Φ), is excluded. The process is automated and repeated cyclically, as long as proper k can be found, until the criterion is minimized.

As the criterion, one of the following three quantities, listed in order of increase of f, can be selected by user: the variance of calibration; the cross-validated variance of calibration (SECCV^2, the recommended choice); the squared noise-to-signal ratio, $(N/S)^2$, defined as

$$(N/S)^2 = \|\delta \mathbf{A}_{\mathrm{net}}\|^2 / \|\mathbf{A}_{\mathrm{net}}\|^2, \quad (2)$$

where $\|\ \|$ denotes Euclidean norm, δ denotes error, $\mathbf{A}_{\mathrm{net}}$ is the net analyte signal matrix found for the calibration set. The $\mathbf{A}_{\mathrm{net}}$ components corresponding to each sample constitute a vector, calculated [7] as:

$$\mathbf{a}_{\mathrm{net}} = C\mathbf{b}^+, \quad (3)$$

where C is the predicted quantity (concentration) estimate; $\mathbf{b}^+$ is the pseudo-inverse of the calibration vector. The $\delta\mathbf{A}_{\mathrm{net}}$ matrix (of the same size as that of $\mathbf{A}_{\mathrm{net}}$) contains differences between the corresponding $\mathbf{A}_{\mathrm{net}}$ components (found using all calibration samples) and the quantities found by formula (3) applied by turns to the samples left out from calibration during jackknifing as in the known approaches to determination of uncertainties of calibration vector components [8,9]. Although it is difficult to assign f for $(N/S)^2$, we found it reasonable, for mean-centered data, $f=(m-1)\times n/k$, where m is the number of samples.

The statistical test on heteroscedasticity can be performed for any confidence level, the recommended values are 0.95 and 0.05. The latter choice means [3] that namely heteroscedasticity is the null hypothesis in the statistical trial. The optimization should be stopped, if the homoscedasticity is proven with the probability ≥ 0.95. We recommend such variant for optimization in infrared spectra.

4. Functionality features

The recommended way of using the software is that through launching the OPTIMIZM utility (GUI shell). Its window contains the plot, spectra vs. wavelengths, where the points selected currently are marked, along with the GUI controls allowing to choose the calibration method, the criterion of optimality and confidence. If the points cannot be divided into regions in a number of ways (that requires divisibility of n), the software prompts to the user to cut "extra" data points manually (either typing the indices of pre-selected

points in Matlab notation (e.g., 1:100) in a text box or dragging "rubber box(es)" in the plot by mouse). At the same time, any user's intervention is solely optional, since default solutions are specified. Once the optimization is performed, the final model (NLV and selected points) is shown in the corresponding GUI controls. It can be optimized additionally, e.g., for a different criterion or confidence.

If needed (either before or after the optimization), user can launch a utility, which predicts concentrations in unknown samples using the current model.

Some options (whether to scale the data; how to split the set of samples in cross-validation/jackknifing; whether to examine NLV cyclically, and some others) are controlled by global variables, which can be changed by experienced user.

The most extensive data set in our distribution, the protein determination in wheat [10] ($m = 70$, $n = 701$, wavelengths in the range 1100–2500 nm), exemplifies the facilities of the software. Using PLS, sequential optimization of $SECCV^2$ and $(N/S)^2$ at 0.05 confidence selects 48 wavelengths (four regions: 1164–1198, 1242–1252, 1266–1300, 1314–1324 nm), NLV = 8, root-mean-squared error of prediction (calculated for an independent set of 20 samples), RMSEP = 0.29 vs. full spectrum: NLV = 9, RMSEP = 0.67.

5. Validation

In order to review the software, it has been independently tested by Dr. O.Ye. Rodionova, Semenov Institute of Chemical Physics, RAS, Moscow (Russia). The report in part is quoted:

> I inform that I have installed and tested the Optimizm software (Matlab chemometrics toolbox by I.M.Baskir and A.V.Drozd). The test included the data input, calibration model selection and validation, wavelength selection. The software runs according to the features described in the documentation correctly. I would not like to conclude about advantages or disadvantages of Optimizm with respect to other chemometrics software. However, no doubt that it is very good in removing "bad" variables.

Acknowledgements

The authors acknowledge Dr. Riccardo Leardi (Di. C.T.F.A.—Genova, Italy) for valuable discussion and recommendations.

References

[1] E.V. Thomas, D.A. Haaland, Anal. Chem. 60 (1988) 1193–1202.
[2] A. Junker, G. Bergmann, Z. Anal. Chem. 272 (1976) 267–275.
[3] A.V. Drozd, I.M. Baskir, Kharkov Univ. Bull. Chem. 3 (1999) 77–80.
[4] I.M. Baskir, A.V. Drozd, Kharkov Univ. Bull. Chem. 5 (2000) 67–70.
[5] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[6] W.G. Cochran, Ann. Eugen. 11 (1941) 47–52.
[7] K. Faber, Anal. Chem. 70 (1998) 5108–5110.
[8] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, London, 1994.
[9] V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851–3858.
[10] J.H. Kalivas, Chemom. Intell. Lab. Syst. 37 (1997) 255–259.