JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Technology Report

Developing and Implementing an R Shiny Application to Introduce Multivariate Calibration to Advanced Undergraduate Students

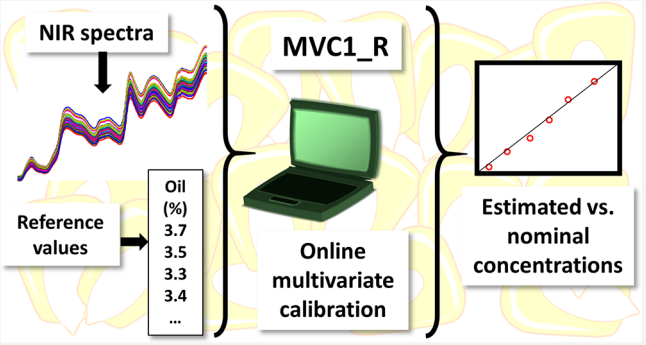
Tomás M. Antonelli and Alejandro C. Olivieri*

Cite This: *J. Chem. Educ.* 2020, 97, 1176–1180

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: During a short chemometrics course in the seventh semester of the chemistry undergraduate program, students receive a brief theoretical introduction to multivariate calibration, focused on partial least-squares regression as the most commonly employed data processing tool. The theory is complemented with the use of MVC1_R, an easy-to-use software developed in-house as an R Shiny application. The present report describes student activities with the latter software in the development of mathematical models to predict quality parameters of corn seeds from near-infrared spectra. Subsequently, an experimental project is carried out involving near-infrared spectral measurements, which are widely used in several industrial fields for quality control. To process the obtained data, students apply the knowledge acquired during the theoretical/software sessions.



KEYWORDS: *Upper-Division Undergraduate, Graduate Education/Research, Analytical Chemistry, Computer-Based Learning, Chemometrics, IR Spectroscopy, Calibration*

Multivariate calibration involves a series of mathematical models for processing partially selective instrumental signals, particularly near-infrared (NIR) spectra,[1] to determine analyte concentrations or sample properties in complex samples. These models are trained by establishing a relationship between a sample set with known property values (calibration phase), which is then applied to unknown specimens to estimate their properties (prediction phase). Since there are no NIR specific wavelengths for a given property or analyte, full spectra at many different wavelengths are processed, hence the name multivariate for this calibration format, in contrast to the classical univariate calibration based on a single wavelength (e.g., UV−vis spectrophotometry).

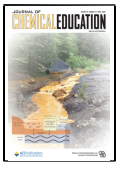
Despite its popularity in many industrial fields,[2] multivariate calibration is not usually taught in chemistry careers.[3] Previous papers in this *Journal* have described examples with data sets produced by students in experimental courses.[4−10] The most common model for NIR spectra is partial least-squares (PLS),[11] which is incorporated in all computers operating NIR spectrometers. The theory of PLS has been described in the relevant literature.[11] During a typical session of a chemometrics course, the basics of PLS are briefly reviewed with students previous to software operation.

Usually, computer programs for students are either commercial,[8−10] or developed in-house and installed in their computers.[5] Issues with commercial software include cost and time lag for being updated. In-house software is kept in a server, which has to be continuously checked for updates. An appropriate response to the above issues is the production of free software which can be operated online and will automatically reflect the updates. In this context, R is a free software environment for statistical computing and graphics,[12] which can be compiled and run on a wide variety of platforms. There are many R applications for multivariate calibration, most of them requiring computer skills.[13] It was thus decided to employ Shiny, an R package allowing researchers to build interactive web apps straight from R.[14] Shiny combines the computational power of R with the interactivity of the modern web, allowing standalone Shiny apps to be hosted on a webpage. Details on the presently developed software used by students are provided below. Students use it while simultaneously receiving theoretical information, before conducting a laboratory practice based on NIR spectroscopy, which requires the skills obtained by using the software.

OVERVIEW OF THE ACTIVITY

MVC1_R, a Shiny R application for multivariate calibration, can be easily and intuitively employed online through a simple

Received: September 10, 2019
Revised: February 8, 2020
Published: February 28, 2020



https://dx.doi.org/10.1021/acs.jchemed.9b00850
J. Chem. Educ. 2020, 97, 1176–1180

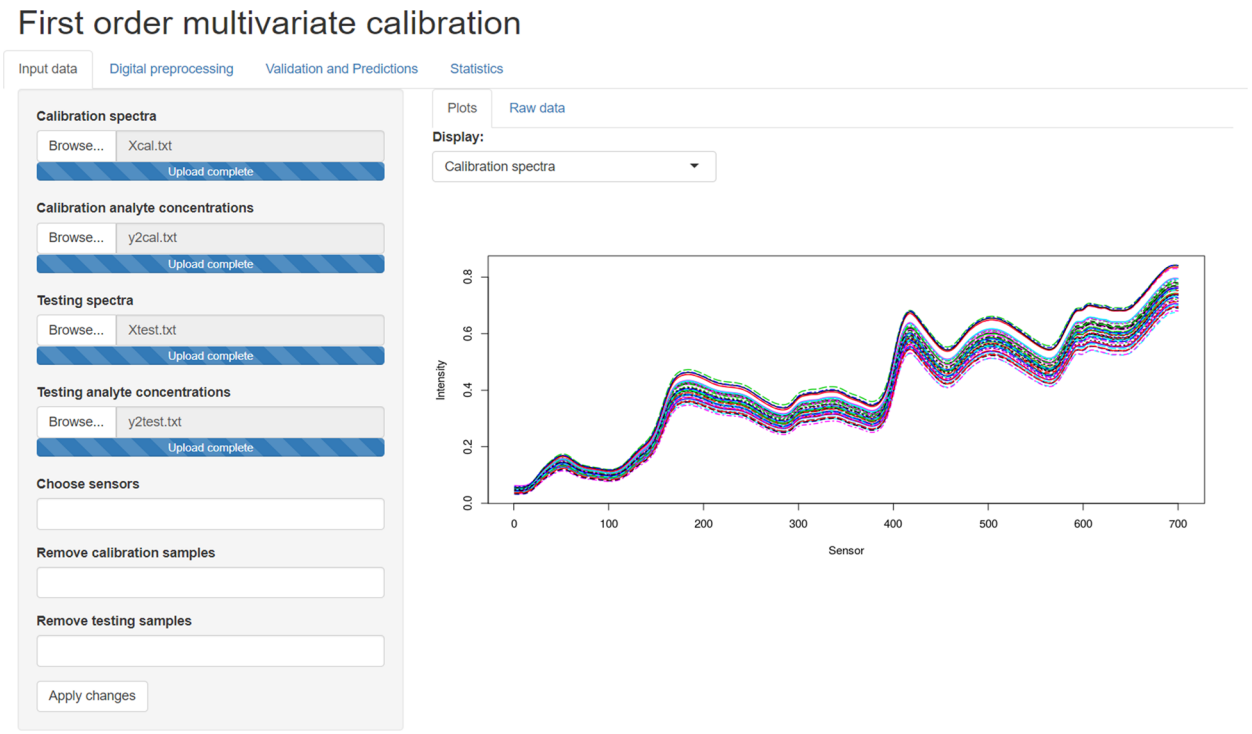


Figure 1. "Data input" tab of MVC1_R, showing the browsers for loading the data, and plot of the calibration spectra which appears when clicking "Apply changes".

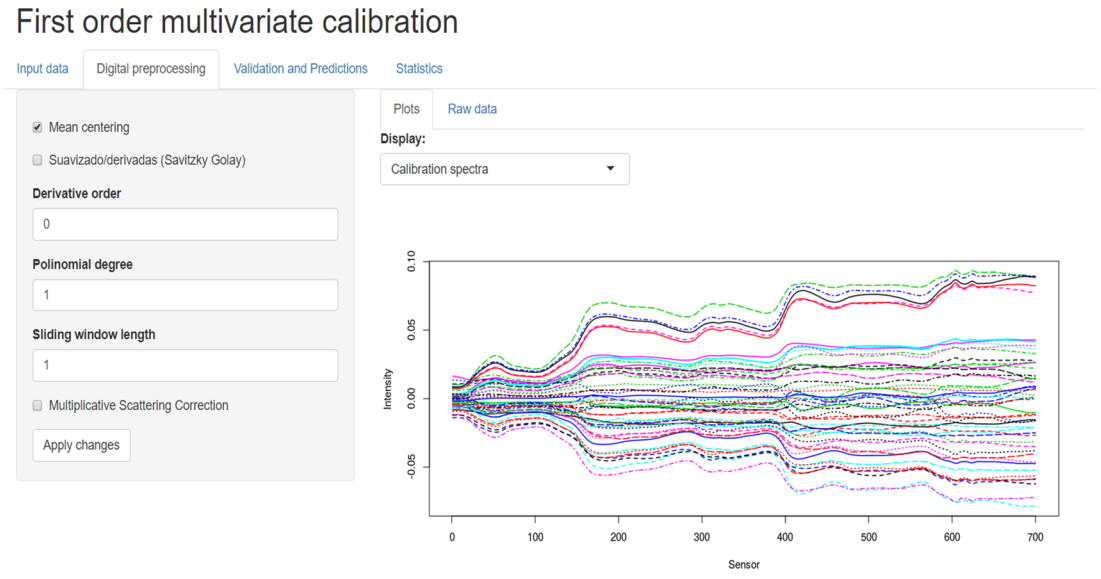


Figure 2. "Digital pre-processing" tab of MVC1_R and plot resulting from the application of mean-centering on the calibration spectra.

series of windows. It allows students to load the data, digitally preprocess them, build a PLS regression model from the training data, and apply the model to test samples. The operation of MVC1_R is illustrated with NIR data downloaded from the Internet, aimed at the determination of quality parameters of corn seed samples.

SOFTWARE DETAILS

MVC1_R was written using R-software, version 3.4.2 (R Core Team, 2012). Shiny is available through the RStudio webpage.[15] MVC1_R can be directly invoked by connecting to the Internet site where it is located.[16] Four tabs are available: "Data input", "Digital preprocessing", "Calibration and prediction", and "Statistics" (Figure 1). They should be sequentially accessed to complete a PLS model building and application to future samples, as described below in detail. Manual and examples are available in the Supporting Information.

STUDENTS RESULTS

Description of the Data Set

The CORN data can be freely downloaded from the Internet.[17] It includes NIR spectra of a set of 80 corn seeds and their corresponding quality parameters (moisture, oil,

https://dx.doi.org/10.1021/acs.jchemed.9b00850
J. Chem. Educ. 2020, 97, 1176−1180

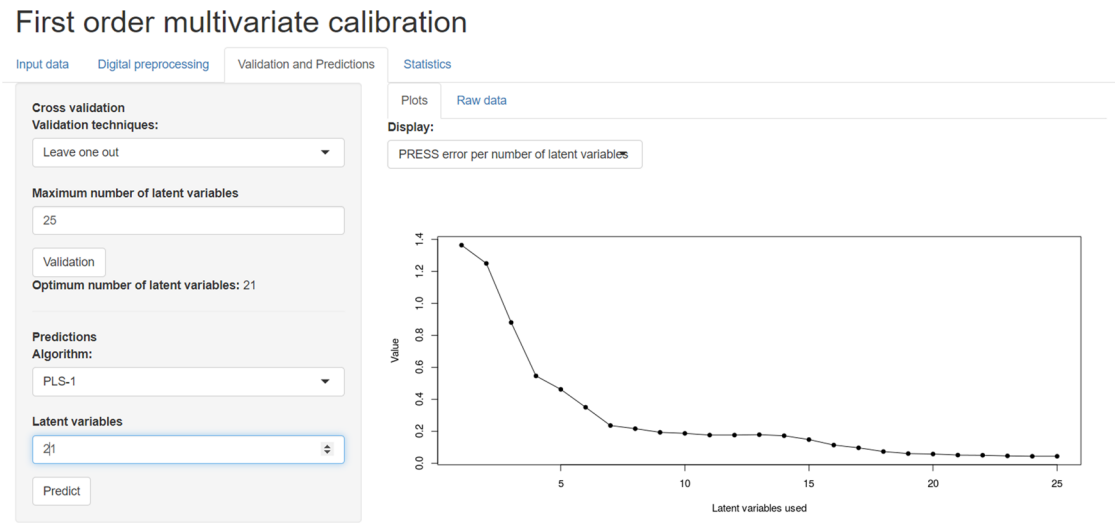


Figure 3. "Calibration and prediction" tab of MVC1_R and cross-validation results.

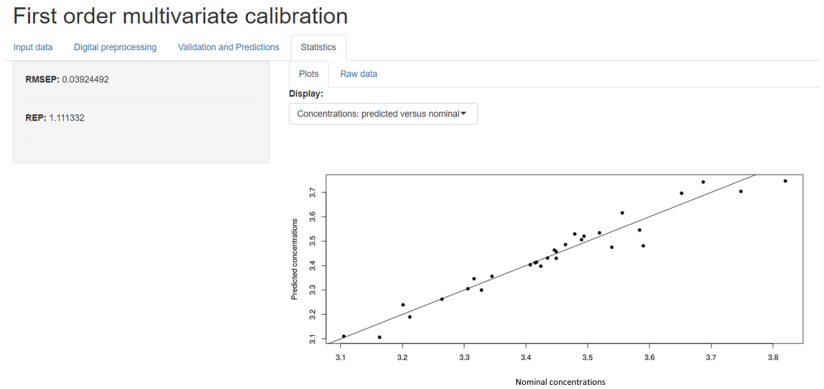


Figure 4. "Statistics" tab of MVC1_R, showing the prediction errors, and plot of predicted vs nominal values.

protein, and starch). The spectra were measured in the wavelength range 1100−2498 at 2 nm intervals (700 wavelengths). The set was divided at random into a calibration and a test set with 50 and 30 samples, respectively, producing two spectral data tables of size 700 × 50 and 700 × 30. The property values were distributed in four tables, each of size 50 × 1 for calibration, and four additional ones, each of size 30 × 1, for testing the models.

Data Input

To load the CORN data from the hard disk with MVC1_R, students use the tab "Data input". Convenient browsers allow them to load the four required files containing:

(1) the matrix of calibration spectra (Xcal.txt),
(2) each of the vectors of calibration values for moisture, oil, protein or starch respectively (y1 cal.txt, y2 cal.txt, y3 cal.txt or y4 cal.txt),
(3) the matrix of test spectra (Xtest.txt), and
(4) the corresponding test property values (y1test.txt, y2test.txt, y3test.txt, or y4test.txt).

Figure 1 shows the MVC1_R screen for processing the oil content. Notice that, for future unknown samples, the property values are not available, in which case the corresponding browser is not used. Clicking in "Apply changes" activates plots (Figure 1) and provides access to spectral and concentration values. Three additional windows for selecting spectral regions, and for removing calibration or test samples, are available. The purpose and use of these three activities is explained in the software manual (Supporting Information) and elsewhere.[2]

Digital Preprocessing

The next tab gives the possibility of digitally preprocessing the spectra (Figure 2). Mean-centering is almost always applied and involves subtracting from all spectra the mean calibration spectrum, and from all properties the mean calibration value. This preprocessing has the effect of removing constant background signals from all spectra and makes further analysis simpler. For liquid samples, no additional activities are needed, but solid samples usually require mathematical preprocessing in order to remove the contribution of the NIR dispersion to the spectra, which is not related to chemical composition.

In MVC1_R, preprocessing includes spectral derivatives and multiplicative scattering correction (MSC). Derivatives (first or second) remove background linear signals (first-derivative is a constant) or parabolic ones (second-derivative is a constant). MSC removes effects which are proportional to the mean spectrum, as is often the case with dispersion signals. Further information can be found in the literature.[2] In any case, after checking a digital preprocessing box, clicking in "Apply changes" refreshes the page and updates the spectra.

To compare the effect of different preprocessing methods, students were divided in four groups, each applying a specific method before model building and prediction.

https://dx.doi.org/10.1021/acs.jchemed.9b00850
J. Chem. Educ. 2020, 97, 1176−1180

Calibration and Prediction

Building a PLS calibration model requires setting the optimum number of latent variables.[2] They are abstract combinations of spectra (loadings), and relative component concentrations (scores), which are needed to represent the raw data by compression into a few variables.[2] This is usually done using a procedure called cross-validation, as described elsewhere,[2] and implemented in MVC1_R in the "Calibration and prediction" tab (Figure 3). For cross-validation, after setting a maximum number of latent variables to be probed (suggestion: half the number of training samples) and clicking "CV", the optimum value is displayed, as well as a plot of the cross-validation sum of square errors as a function of the number of latent variables and the corresponding statistics (see Supporting Information). The second step is to set the number of latent variables for prediction and click in "Predict", which will show the estimated property values in the test samples (Figure 3).

Statistics

Finally, the "Statistics" tab is useful when the reference values for the test properties are available (Figure 4). It shows a measure of the average prediction error, called root-mean-square error of prediction (RMSEP) in concentration or property units, and the relative error of prediction (REP), in percent with respect to the mean calibration property.

The specific results for oil, probably the most important oilseed quality parameter, are shown in Table 1. Each student group applied a different digital preprocessing, and thus, they differed in the optimum number of latent variables and average prediction error. The results triggered interesting discussions on which preprocessing activity should be preferred. Some students argued, based on its lowest prediction error, that first-derivative is the best preprocessing (Table 1). Others recalled what they learned in the theoretical sessions and argued that if errors are not too large, lesser latent variables are preferable because they lead to simpler models. This is the case with the second-derivative (Table 1). An even more interesting debate followed: Is the REP for first-derivative (1.0%) really smaller, in statistical terms, than the REP for second-derivative (1.3%)? Students are asked to search the literature for performing such a comparison, finding various tests, all of them suggesting that the difference is not significant, and thus that second-derivative preprocessing is to be preferred.

Table 1. Student Results on the Estimation of Oil in the CORN Data Set Using NIR-PLS Analysis

Student Group	Preprocessing	Optimum Number of Latent Variables	RMSEP (%)	REP (%)
1	Mean-centering only	21	0.040	1.1
2	First-derivative and mean-centering	14	0.036	1.0
3	Second-derivative and mean-centering	9	0.046	1.3
4	MSC[a] and mean-centering	19	0.069	2.0

[a]MSC = acronym for multiplicative scattering correction.

After processing the data for the remaining three seed parameters, using second-derivative followed by mean-centering as preprocessing, the students found the results collected in Table 2. A discussion then followed on various issues: (1) the number of latent variables for the four parameters are similar, although not identical, and (2) the relative errors (REP%) are all reasonably low. Students are asked to rationalize the estimated number of latent variables on the basis of their previous theoretical knowledge on the matter, i.e., that the PLS model is specific for each target property. Regarding the average errors, students are asked to search in the official literature for the accepted error levels to estimate the seed parameters using classical methodologies. This would provide a hint as to the potential application of NIR-PLS analysis for this type of samples.

Table 2. Students Results on the Estimation of Four Quality Parameters in the CORN Data Set Using NIR-PLS Analysis

Parameter	Optimum Number of Latent Variables[a]	RMSEP (%)	REP (%)
Moisture	6	0.12	1.2
Oil	9	0.046	1.3
Protein	8	0.16	1.8
Starch	10	0.26	0.40

[a]Preprocessing: second-derivative and mean-centering.

CONCLUSIONS

Practical aspects of multivariate calibration based on near-infrared spectroscopy can be conveniently learned using free online software for partial least-squares analysis. An R Shiny application was developed for this purpose and successfully adapted to a short chemometrics course for advanced undergraduate chemistry students.

ASSOCIATED CONTENT

Supporting Information

Two files are provided: (1) 'MVC1_R Manual.docx'. (DOCX), and (2) 'Data_set.zip'. , explaining the nature and purpose of these text files (ZIP). The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.9b00850.

Document explaining the use of the software (PDF, DOCX)

Compressed file containing the text files Xcal.txt, y1cal.txt, y2cal.txt, y3cal.txt, y4cal.txt, Xtest.txt, y1test.txt, y2test.txt, y3test.txt, and y4test.txt (see Data Input section), and Data_help.txt (ZIP)

AUTHOR INFORMATION

Corresponding Author

Alejandro C. Olivieri − *Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Rosario S2002LRK, Argentina;* orcid.org/0000-0003-4276-0369; Email: olivieri@iquir-conicet.gov.ar

Author

Tomás M. Antonelli − *Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Rosario S2002LRK, Argentina;* orcid.org/0000-0002-0765-5638

https://dx.doi.org/10.1021/acs.jchemed.9b00850
J. Chem. Educ. 2020, 97, 1176−1180

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jchemed.9b00850

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Universidad Nacional de Rosario (Project BIO-521), CONICET, and ANPCyT (Project PICT-2016-1122) are gratefully acknowledged for financial support. T.M.A. thanks Instituto Politécnico Superior Gral. San Martín for a stay at IQUIR-CONICET.

■ REFERENCES

(1) Pasquini, C. Near infrared spectroscopy: a mature analytical technique with new perspectives - a review. *Anal. Chim. Acta* **2018**, *1026*, 8−36.
(2) Olivieri, A. C. *Introduction to Multivariate Calibration. A Practical Approach*; Springer-Nature: Berlin, 2018. DOI: 10.1007/978-3-319-97097-4.
(3) Öberg, T. Introducing chemometrics to graduate students. *J. Chem. Educ.* **2006**, *83*, 1178−1181.
(4) Charles, M. J.; Martin, N. W.; Msimanga, H. Z. Simultaneous determination of aspirin, salicylamide, and caffeine in pain relievers by target factor analysis. *J. Chem. Educ.* **1997**, *74*, 1114−1117.
(5) Ribone, M. E.; Pagani, A. P.; Olivieri, A. C.; Goicoechea, H. C. Determination of the active principle in a syrup by spectrophotometry and principal component regression (PCR) analysis. An advanced undergraduate experiment involving chemometrics. *J. Chem. Educ.* **2000**, *77*, 1330−1333.
(6) Msimanga, H. Z.; Elkins, P.; Tata, S. K.; Smith, D. R. A chemometrics module for an undergraduate instrumental analysis chemistry course. *J. Chem. Educ.* **2005**, *82*, 415−424.
(7) Wanke, R.; Stauffer, J. An advanced undergraduate chemistry laboratory experiment exploring NIR spectroscopy and chemometrics. *J. Chem. Educ.* **2007**, *84*, 1171−1173.
(8) Wang, L.; Mizaikoff, B.; Kranz, C. Quantification of sugar mixtures with near-infrared Raman spectroscopy and multivariate data analysis. A quantitative analysis laboratory experiment. *J. Chem. Educ.* **2009**, *86*, 1322−1325.
(9) Pierce, K. M.; Schale, S. P.; Le, T. M.; Larson, J. C. An advanced analytical chemistry experiment using gas chromatography−mass spectrometry, MATLAB, and chemometrics to predict biodiesel blend percent composition. *J. Chem. Educ.* **2011**, *88*, 806−810.
(10) de Oliveira, R. R.; das Neves, L. S.; de Lima, K. M. G. Experimental design, near-infrared spectroscopy, and multivariate calibration: an advanced project in a chemometrics course. *J. Chem. Educ.* **2012**, *89*, 1566−1571.
(11) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(12) Crawley, M. J. *The R Book*; Wiley: Chichester, UK, 2013. DOI: 10.1002/9781118448908.
(13) Mevik, B. H.; Wehrens, R. The pls package: principal component and partial least squares regression in R. *J. Stat. Softw.* **2007**, *18*, 1−24.
(14) Beeley, C. *Web Application Development with R Using Shiny*; Packt Publishing Ltd.: Birmingham, UK, 2013.
(15) Shiny. http://shiny.rstudio.com/ (accessed Feb 8, 2020).
(16) First-order multivariate calibration https://atmunr.shinyapps.io/MVC1_R/ (accessed Feb 8, 2020).
(17) NIR of corn samples http://www.eigenvector.com/data/Corn (accessed Feb 8, 2020).