Quim. Nova, Vol. 44, No. 5, 655-658, 2021 http://dx.doi.org/10.21577/0100-4042.20170686

Educação

SIMULTANEOUS SPECTROPHOTOMETRIC DETERMINATION AND CLASSICAL LEAST SQUARES METHOD: A SIMPLE EXPERIMENT TO INTRODUCE THE CONCEPT OF MULTIVARIATE CALIBRATION

Jhonatas R. Carvalho[a], Larissa R. Lopes[a], Luciano N. Vidal[a] and Poliana M. Santos[a,*]
[a]Departamento Acadêmico de Química e Biologia, Universidade Tecnológica Federal do Paraná, 81280-340 Curitiba – PR, Brasil

Recebido em 13/08/2020; aceito em 05/11/2020; publicado na web em 09/12/2020

We herein present an experiment where the concentrations of tartrazine, sunset yellow and amaranth in samples containing these three food dyes are determined by system of equations (SE) and classical least squares (CLS) multivariate calibration methods using light absorption data. Firstly, concentrations are obtained by means of the well-known SE method, that is, by solving a set of three linear equations in which the Beer-Lambert's proportionality coefficients are obtained from analytical curves. Then, it is shown that the CLS method is a natural extension to SE, with an arbitrarily large number of equations. Nevertheless, within the CLS method, the unknown coefficients are found using mixtures with known concentrations of each dye. In order to introduce the students to the basics of algorithms and numerical computations, data treatment is performed in a command-line fashion using a freely available software. Advantages of multivariate calibration models over univariate ones are made clear, and the performance of the CLS and SE methods is compared based on the root-mean-square error.

Keywords: teaching Chemometrics; multivariate calibration; classical least squares.

INTRODUCTION

In the last few years, Chemometrics has become common in the majority of undergraduate and graduate chemistry curricula. One of the reasons for this is the ever-increasing trend in both industry and academic research applications. According to the International Chemometrics Society (ICS), Chemometrics is a chemical discipline that uses mathematical and statistical methods to design or select optimal measurement procedures and experiments, and to provide maximum chemical information by analyzing chemical data.[1] Principal component analysis (PCA) and partial least squares regression (PLSR) are probably the most widespread chemometric methods worldwide, with several laboratory experiments already reported.[2-9] Examples of their use include classification of elements within the periodic table,[2] classification of vegetable oils by infrared spectroscopy (FTIR)[3] and nuclear magnetic resonance (^{1}H NMR),[4] quantification of paracetamol in commercial tablets using near-infrared (NIR) spectroscopy,[5] quantification of biodiesel content in blends of biodiesel and conventional diesel by gas chromatography mass spectrometry (GC-MS),[6] and others.[7-9]

Another important chemometric method that has analytical applications,[10-12] although little explored in undergraduate and graduate chemistry laboratories, is the classical least squares (CLS) regression.[13] The CLS approach is grounded on the assumption that in a measurement, a system's *response* (e.g. absorbance) is given by summing the individual response of each *active constituent* (e.g. absorbing specie). It is further assumed that the individual response of every (active) constituent is proportional to its amount in the sample. Therefore, there is a linear relation between measured response and the concentration of each sample's constituent. In the experiment proposed herein, the concentrations of three food dyes (tartrazine, sunset yellow and amaranth) in aqueous solutions are spectrophotometrically determined using the Beer–Lambert law. Under the assumption that the linear relationship between absorbance and concentration holds over a wide range of light wavelengths (e.g. visible region), the CLS model is built by minimizing the sum of squared differences between predicted and measured absorbances for that set of wavelengths. Such minimization results in a set of proportionality coefficients (i.e. molar absorptivities) giving the least value for those squared differences. Those coefficients are obtained using a group of reference samples (*calibration set*), where the concentration of each constituent is known. Once the CLS coefficients are determined, the model can be employed to predict the amount of each constituent in a sample from its measured absorbances. Because CLS is a full spectrum method, it can provide improvement in precision over methods that are restricted to a small number of wavelengths, such as the well known system of equations (SE) method, considered below.

For multi-constituent systems displaying the aforementioned relations of additivity and linearity between the observed response and concentrations, the amount of each specie in a sample can be determined by solving a minimum set of linear equations, with the number of variables being equal to the number of equations. That approach is usually termed the SE method. The SE matrix equation is identical to that of CLS (see Supporting Information for more details). Nevertheless, they differ in the way equations are solved. Within SE method, the proportionality coefficients are obtained from analytical curves of the pure constituents.

Since the 1960s, several laboratory practices using the SE method have been proposed.[14-17] Most of them involve 2-constituent systems whose constituents are readily determined by that method, although laboratory experiments for 3-constituent systems were reported as well. Instances of 3- constituent analysis are the experiment proposed by Harker III and colleagues,[14] for simultaneous spectrophotometric determination of the isomeric cresol (*o*-, *m*-, and *p*-cresol), and that by Sigmann and Wheeler,[15] for quantification of FD&C Yellow 5, FD&C Red 40 and FD&C Blue 1 color additives present in powdered drink mixes. Conversely, a limited number of laboratory experiments using CLS can be found in the literature and most of which are focused on comparing its performance to other multivariate regression methods.[18,19] Possibly, this is due to the advantages of the inverse models over the simplest direct CLS multivariate model. In the direct models, the signal is considered to be directly proportional to the concentration, as dictated, for example, by Lambert–Beer's law in

*e-mail: polianasantos@utfpr.edu.br

classical UV-visible spectroscopy. On the other hand, in the inverse models, the concentration is considered to be directly proportional to the signal.[13] Thus, classical models assume that the major source of errors are in the response measurements, whereas a more appropriate assumption is that errors are primarily related to the measurement of concentration. In general, the greatest source of errors is associated to sample preparation, such as dilution, weighing and extraction procedures, rather than the instrumental reproducibility.[20] However, CLS has a high pedagogical value due its simplicity, thus giving the students grounds to understand more advanced multivariate calibration methods. For example, the multivariate curve resolution-alternating least squares (MCR-ALS) method, which has been largely applied in the last years,[21] has a common theoretical basis with CLS.[10]

Thus, we present a laboratory experiment for Analytical Chemistry courses to introduce Chemometrics to undergraduate students using their previous knowledge of the Beer–Lambert law. After recording the absorption spectra of aqueous solutions containing tartrazine, sunset yellow and amaranth, students initially find the concentration of each food dye through the SE method. Next, by expanding the number of variables from three to an arbitrarily large number of wavelengths (comprised within the visible region), the CLS method is employed to find the unknown concentrations. All data analysis is performed step-by-step by the students using a freely available software for numerical computations. An important limitation of CLS – namely, the existence of unknown background constituents – is also illustrated here.

EXPERIMENTAL

The experiment is primarily aimed at upper-level undergraduate students of chemistry courses; however, it is also amenable to other courses where students have some knowledge of linear algebra and analytical chemistry. We suggest two lab sessions of 3-4 hours to perform the experiment. Students may be split into small groups of 2-3 members. During the first session, each group prepares solutions of the standard curves and records their spectra. Solutions corresponding to the calibration and validation sets are also prepared in that lab session. Additionally, the instructor may provide unknown samples to be analyzed by the groups. The first lab session ends after the spectra of all solutions have been recorded. The second lab session is focused on data analysis. Students may act in groups or individually, depending on the availability of working terminals at the computer lab. Class starts with a brief introduction to the language syntax of the software chosen for the algebraic computations. Subsequently, students determine the concentration of tartrazine, sunset yellow and amaranth in the validation samples (and optionally, for the unknown ones) through the SE and CLS methods. The numerical computations are carried out using the GNU Octave software,[22] following the instructions provided in the Supporting Information.

Samples and data collection

From stock solutions of 500 mg L^{-1}, in 0.1 mol L^{-1} HCl, containing a single dye (tartrazine, sunset yellow or amaranth), groups are required to prepare univariate analytical curves with final concentrations ranging from 5 to 25 mg L^{-1}. The univariate analytical curves should contain at least 5 different concentrations.

Eighteen solutions containing tartrazine, sunset yellow and amaranth were also prepared by dilution of the single stock solutions (see Table 1). From that table one notice that the concentration of each constituent within a sample can have one of the following values (mg L^{-1}): 5.00, 7.00, 9.00, 11.00 or 13.00. The corresponding concentration was randomly selected using the "randperm" Octave's built-in function. That function creates "random" permutations based on the Mersenne-Twister pseudo-random number generator. In order to minimize concentration correlation within the set of samples, the concentrations may also be chosen using an experimental design, keeping a common sense that the calibration mixtures should be representative, as much as possible, of the combination of concentrations to be found in the future unknown samples.[13]

Table 1. Concentration of each dye in the samples used for CLS analysis

Sample	Tartrazine (mg L^{-1})	Sunset yellow (mg L^{-1})	Amaranth (mg L^{-1})
1	9.00	9.00	7.00
2	11.00	11.00	13.00
3	7.00	7.00	11.00
4	5.00	5.00	5.00
5	7.00	5.00	13.00
6	11.00	11.00	7.00
7	9.00	9.00	5.00
8	5.00	11.00	11.00
9	11.00	5.00	13.00
10	11.00	5.00	5.00
11	7.00	7.00	9.00
12	9.00	11.00	11.00
13	5.00	9.00	7.00
14	13.00	13.00	13.00
15	5.00	7.00	11.00
16	7.00	9.00	5.00
17	13.00	13.00	9.00
18	9.00	7.00	7.00

The absorption spectra were recorded over the 400-620 nm range using a 0.1 mol L^{-1} HCl solution as blank. All measurements were performed in a UV-Vis spectrophotometer (Cary 50, Varian) equipped with a 1.0 cm quartz cell.

Data analysis

For CLS analysis, the recorded absorbances of the eighteen samples are arranged in a matrix having 18 rows (number of samples) and 221 columns (number of variables). Those samples were further separated into two sets. The first 14 samples were selected to be the calibration set. These samples were employed to set up the CLS model. The remaining 4 were used for external validation (validation set). The basic CLS theory is presented in the Supporting Information as well as the steps for the development of the model using GNU Octave (or MATLAB).

The procedure to perform the SE calculations using GNU Octave (or MATLAB) is also described in detail in the Supporting Information. First, the students need to determinate the wavelength of maximum absorbance for each food dye (λ_{max}) using the pure spectra of tartrazine, sunset yellow and amaranth. Then, the molar absorptivity can be determined using Beer–Lambert plots. Finally, the students can solve the equations to calculate the concentration of each food dye within the mixtures.

The predictive ability of each method and for each constituent was assessed in terms of the root-mean-square error (RMSE) of the validation set:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_v} (\hat{c}_i - c_i)^2}{N_v}} \quad (1)$$

In the above equation, $\hat{c}_i$ and c_i represent the predicted and reference values for the concentration of a dye at sample "*i*", respectively, in a set formed by N_v validation samples.

RESULTS AND DISCUSSION

The individual absorption spectra of tartrazine, sunset yellow and amaranth, as well as the spectrum of a solution containing these three dyes, are shown in Figure 1. Students will find that the maximum absorbance for tartrazine, sunset yellow and amaranth occurs at 429, 482 and 523 nm, respectively. It can also be seen that the mixture spectrum is highly overlapped, particularly at the wavelengths of 429 and 482 nm, where the molar absorptivity of all dyes is about of the same order of magnitude (see Table 2). At 523 nm, absorptions of sunset yellow and amaranth are both significant. Therefore, quantitative analysis by the single-equation approach of the Beer–Lambert law cannot be satisfactorily applied to this system. For that reason, the SE and CLS methods were proposed.

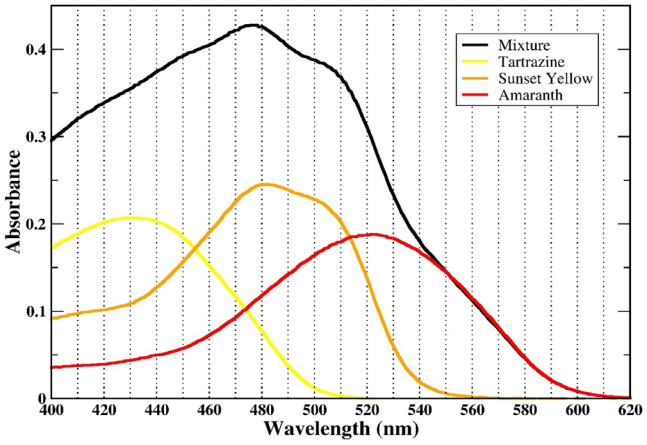

Figure 1. *Measured visible absorption spectra of tartrazine (5.00 mg L^{-1}), sunset yellow (5.00 mg L^{-1}) and amaranth (5.00 mg L^{-1}) in 0.1 mol L^{-1} of HCl. The mixture spectrum contains the same concentration of each dye*

Table 2. Molar absorptivity (10^{-2} L mol^{-1} cm^{-1}), obtained from the univariate analytical curves, of the pure dyes in aqueous solutions containing 0.1 mol L^{-1} of HCl

λ/nm	Tartrazine	Sunset Yellow	Amaranth
429	4.20	2.15	0.87
482	1.43	4.82	2.43
523	0.04	2.21	3.68

Following the instructions available as Supporting Information, students were able to develop the SE and CLS models. Table 3 shows the predicted concentrations for tartrazine, sunset yellow and amaranth obtained with SE and CLS models. In order to compare the accuracy of the methods, the samples listed (15-18) are those from the validation set of the CLS model. All models showed satisfactory results, with predicted concentrations close to the reference values. CLS gives the best predictive ability with RMSE for tartrazine, sunset yellow and amaranth equal to 0.122, 0.112 and 0,165 mol L^{-1}, respectively. In order to compare the performance of the models, an F-test was evaluated. The F-values were obtained by the ratio of the variances (RMSE squared), where $RMSE_1 > RMSE_2$. If $F_{calculated}$ is greater than the $F_{critical}$, the two RMSEs are significantly different. Results indicated that the differences in the RMSE values of the SE and CLS models were not significant for tartrazine ($F^{critical}_{v1=3,v2=3,\alpha=0.05} = 9{,}28 > F_{calculated} = 1{,}08$), sunset yellow ($F^{critical}_{v1=3,v2=3,\alpha=0.05} = 9{,}28 > F_{calculated} = 1{,}18$) and amaranth ($F^{critical}_{v1=3,v2=3,\alpha=0.05} = 9{,}28 > F_{calculated} = 2{,}73$), for a confidence level of 95%.

Students may also access the accuracy of the models through a "predicted *versus* measured" plot. Figure 2 shows that plot for tartrazine. From that figure, one can see that both data sets (calibration and validation) show good agreement with the reference values for the concentrations. Furthermore, such a comparison can be put into a more quantitative basis by using the correlation coefficient of the linear regression modeling. The SE and CLS models both yielded high correlation coefficients (> 0.99), thus indicating that they were able to predict accurate concentrations of the three food dyes. It is also worth of mention that the slopes of the corresponding linear regressions were near 1 whereas the intercepts were near 0.

The accuracy of the CLS model can be additionally checked by comparing the CLS-resolved spectra with the spectra of the standard

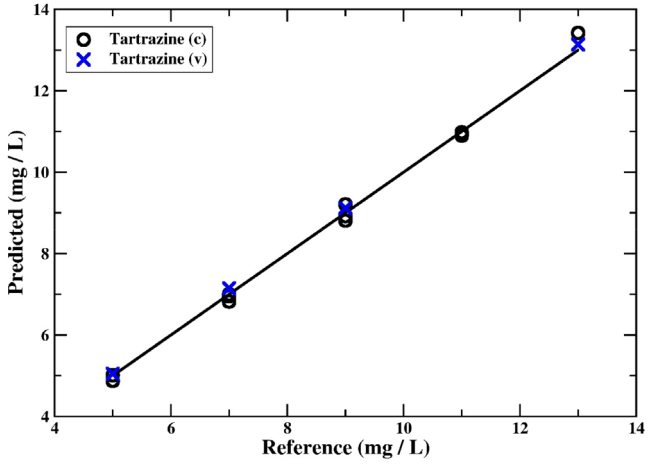

Figure 2. *Comparison between the concentrations of tartrazine predicted by CLS for the calibration (c) and validation (v) samples with the corresponding reference values. The exact agreement is represented by the solid line*

Table 3. Reference and predicted concentrations (mg L^{-1}) determined by the SE and CLS methods

Sample	Tartrazine			Sunset Yellow			Amaranth		
	Reference	SE	CLS	Reference	SE	CLS	Reference	SE	CLS
15	5.00	5.03	5.05	7.00	7.08	6.97	11.00	10.96	11.07
16	7.00	6.96	7.15	9.00	8.96	8.97	5.00	5.00	5.30
17	13.00	12.78	13.14	13.00	12.78	12.85	9.00	8.58	9.11
18	9.00	8.89	9.12	7.00	7.04	7.16	7.00	6.65	6.93
RMSE		0.127	0.122		0.122	0.112		0.273	0.165

pure compounds. As explained in the Supporting Information, the CLS method decomposes the unknown spectral measurement into its pure constituent spectra if they are known. According to that section, the pure spectrum of tartrazine in a given sample is estimated from the product between the first row of matrix $\hat{\boldsymbol{B}}$ and the predicted concentration of that constituent in the sample. On the other hand, the pure spectrum of sunset yellow is obtained when the second row of $\hat{\boldsymbol{B}}$ is multiplied by the concentration of that constituent, and the spectrum of amaranth is estimated using the third row of $\hat{\boldsymbol{B}}$ in a similar way. Results showed that the CLS-resolved spectra profile of the three food dyes were very similar to the visible absorption spectra of pure samples shown in Figure 1.

Although CLS is a powerful tool for studying mixtures of analytes, it cannot be successfully applied to all types of samples. The method works only if all active constituents within the sample are known. In order to demonstrate this drawback, students may build a new model using the same dataset but omitting the concentration information of a constituent in the sample. For example, when the CLS model is set up without the concentration of amaranth, the RMSE for tartrazine and sunset yellow become equal to 0.467 and 2.331 mg L^{-1}, respectively, whereas in the previous model, where the amount of all active constituents are known, the corresponding RMSE are 0.122 and 0.112 mg L^{-1} (see Table 3). In order to circumvent such a limitation, inverse calibration methods were proposed.[1] Within those methods, calibration can be performed without knowledge of all possible sample constituents in the calibration phase. In Chemometrics, the most widely used multivariate inverse calibration methods are PLS and PCR. Experiments exploring the PLS method are available from this journal.[5]

CONCLUSIONS

In chemical analysis, determination of chemical information is traditionally done using a minimal amount of output data generated through an experiment, for example, obtaining the concentration of an analyte by measuring light absorption in a single wavelength. By way of a simple and inexpensive experiment, students can learn that all data produced in a measurement may contain useful information that can be employed to build a more robust and reliable calibration model, through multivariate analysis. Due to its conceptual simplicity, the CLS method can be developed without the need for any specialized software package. This makes it a valuable model to introduce the students to the overall principles of multivariate calibration analysis. Furthermore, the hands-on approach adopted here to process the absorption data and to set up the calibration models allowed the students to gain knowledge on the basics of computer algorithms and numerical computations using high-level computer languages.

Finally, some systems are particularly well suited to be analyzed through CLS, such as the simultaneous quantification of analgesic and anti-inflammatory agents in pharmaceutical preparations.[12,23] Students may easily adapt the current experiment in order to handle those systems.

SUPPLEMENTARY MATERIAL

In the Supporting Information, available at http://quimicanova.sbq.org.br, the reader will find a brief description of the SE and CLS methods as well as the steps necessary for the development of those models using the GNU Octave (or MATLAB) software. The corresponding data required to perform the SE and CLS analysis are publicly available at "https://gitlab.com/lnvidal1/cls-data" website.

ACKNOWLEDGMENT

JRC thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), for a master's scholarship. PMS acknowledge the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the financial support of this work (Grant 409111/2016-3).

REFERENCES

1. Otto, M.; *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, 3rd ed., Wiley: New York, 2017.
2. Lyra, W. S.; Silva, E. C.; Araújo, M. C. U.; Fragoso, W. D.; Veras, G.; *Quim. Nova* **2010**, *33*, 1594.
3. Souza, A. M.; Poppi, R. J.; *Quim. Nova* **2012**, *35*, 223.
4. Anderson, S. L.; Rovnyak, D.; Strein, T. G.; *J. Chem. Educ.* **2017**, *94*, 1377.
5. Souza, A. M.; Breitkreitz, M. C.; Filgueiras, P. R.; Rohwedder, J. J. R.; Poppi, R. J.; *Quim. Nova* **2013**, *36*, 1057.
6. Pierce, K. M.; Schale, S. P.; Le, T. M.; Larson, J. C.; *J. Chem. Educ.* **2011**, *88*, 806.
7. Oliveira, R. R.; Neves, L. S.; Lima, K. M. G.; *J. Chem. Educ.* **2012**, *89*, 1566.
8. Valderrama, L.; Paiva, V. B.; Março, P. H.; Valderrama, P.; *Quim. Nova* **2016**, *39*, 245.
9. Sidou, L. F.; Borges, E. M.; *J. Chem. Educ.* (2020), DOI: 10.1021/acs.jchemed.9b00924
10. Amigo, J. M.; Ravn, C.; *Eur. J. Pharm. Sci.* **2009**, *37*, 76.
11. Sabin, G. P.; Rocha, W. F. C.; Poppi, R.; J.; *Microchem. J.* **2011**, *99*, 542.
12. Singh, V. D.; Daharwal, S. J.; *Spectrochim. Acta, Part A* **2017**, *171*, 369.
13. Olivieri, A.; *Introduction to Multivariate Calibration A Practical Approach*, 1st ed., Springer International Publishing: Switzerland, 2018.
14. Harker, G. G.; Huntington, J. L.; Gruber, T. A.; Hargis, L. G.; *J. Chem. Educ.* 1970, *47*, 712.
15. Sigmann, S. B.; Wheeler, D. E.; *J. Chem. Educ.* **2004**, *81*, 1475.
16. MacQueen, J, T.; Knight, S. O.; Reilley, C. N.; *J. Chem. Educ.* **1960**, *37*, 139.
17. Mehra, M. C.; Rioux, J.; *J. Chem. Educ.* **1982**, *59*, 688.
18. Gilbert, M. K.; Luttrell, R. D.; Stout, D.; Vogt, F.; *J. Chem. Educ.* **2008**, *85*, 135.
19. Ribone, M. É.; Pagani, A. P.; Olivieri, A. C.; *J. Chem. Educ.* **2000**, *77*, 1330.
20. Brereton, R. G.; *Analyst* **2000**, *125*, 2125.
21. Juan, A.; Jaumot, J.; Tauler, R.; *Anal. Methods* **2014**, *6*, 4964.
22. https://www.gnu.org/software/octave/doc/v4.4.1/, accessed November 2020.
23. Goicoechea, H. C.; Olivieri, A. C.; *J. Pharm. Biomed. Anal.* **1999**, *20*, 681.