Table of Content Abstract

Principal Component Regression that minimises the sum of the squares of the relative errors: Application in multivariate calibration models

O. Valencia[a], M.C. Ortiz[b*], L.A. Sarabia[a]

[a]Department of Mathematics and Computation, [b]Department of Chemistry,

Faculty of Sciences, Universidad de Burgos

Plaza Misael Bañuelos s/n, 09001 Burgos (Spain)

Short Abstract

Relative errors are used in Chemometrics to evaluate the performance of multivariate models. However, these models are not obtained through the criterion of minimising relative errors, as would be expected in models whose response is a concentration.

This work proposes the use of a Principal Component Regression that minimises the sum of the squares of the relative errors (wPCR). This model has been applied to several datasets, 9 of which are multivariate calibrations from complex mixtures based on instrumental signals coming from different analytical techniques. Compared to the usual PCR, the wPCR model results in a decrease in relative errors, particularly for the smallest concentrations.

*Corresponding author. Telephone number: +34-947-259571. *E-mail address*: mcortiz@ubu.es (M.C. Ortiz).

Principal Component Regression that minimises the sum of the squares of the relative errors: Application in multivariate calibration models

O. Valencia[a], M.C. Ortiz[b*], L.A. Sarabia[a]

[a]Department of Mathematics and Computation, [b]Department of Chemistry,

Faculty of Sciences, Universidad de Burgos

Plaza Misael Bañuelos s/n, 09001 Burgos (Spain)

Abstract

Relative errors are typically used in Chemometrics to evaluate the performance of a multivariate predictive model. However, these models are not obtained through the criterion of minimising relative errors, as would be expected in a model whose response is the concentration of an analyte.

There are no studies in Chemometrics on the use of a Principal Component Regression that minimises the sum of the squares of the relative errors. This work proposes a model, which serves this purpose. The suggested model, wPCR, has been applied to 7 datasets with 12 predicted responses, 10 of which are multivariate calibrations of analytes in complex mixtures based on instrumental signals coming from various analytical techniques.

As PCR and wPCR are methods seeking to optimize different criteria, each one achieves a better performance with respect to its own criterion. Therefore, the new model wPCR leads to better results insofar as the relative errors are considered, especially for the smallest responses. In this sense, the wPCR model also outperforms PCR with logarithmic transformation of the response (logPCR).

In addition, as for the performance of the method using Joint Confidence Regions for the intercept and the slope of the accuracy line, it is shown that the application of wPCR does not

introduce bias, neither constant nor proportional for the models built, nor a systematic alteration of the achievable accuracy.

Keywords: relative error; Principal Component Regression; accuracy line; Joint Confidence Region; multivariate calibration

1. Introduction

The Least Squares criterion does not take account of the internal distribution of errors, that is, the amount and location of individual errors. However, the way errors are distributed is relevant, since two errors similar in size can clearly differ in percentages, i.e., when compared to different observed values of the response variable. This issue is particularly applicable in Chemometrics as to the multivariate calibration methods in which the response variable is a concentration and the importance of a certain error strongly depends on whether it occurs at a small or a large concentration.

Multivariate calibration models based on the least squared criterion, such as Multivariate Linear Regression, Principal Component Regression or Partial Least Squares Regression, are meant to predict the concentration of one or several chemical analytes from a multivariate recorded signal (UV-visible spectroscopy, molecular excitation-emission fluorescence, infrared spectroscopy, polarography…). An overall adequate model can be achieved, but predicted concentrations far away in percentage points from the actual concentrations of the calibration standards can still occur particularly at low concentration samples.

In the statistical literature on multivariate regression by least squares, the subject of response transformations has been widely addressed (e.g., chapter 13 in Ref. [1]). In particular, the logarithmic transformation would naturally up-weight the low values of the response and would be able to reduce relative errors.

In Least Squares Relative Errors, LSRE, a model that minimizes the sum of squares of relative errors, $\sum((y-\hat{y})/y)^2$, is fitted. In the univariate regression case, Ferreira et al.[2] have obtained explicit formulae as quotients of determinants for coefficients and their variance while pointing to the connection between weighted least squares, wLS, and LSRE. The analysis extends to the multivariate case in the Ref. 3. The authors provide a unique solution and a matrix expression is obtained for the calculation of coefficients. Additionally, it is shown to be equivalent to a weighted multivariate regression whose weights equal the inverse of the squared values of the response.

Virtually all textbooks devoted to regression models include a chapter on weighted regression as it is the standard procedure for getting residuals with equal variances[1,4]. Generally, Least Squares Relative Errors, LSRE, inherits all the theoretical properties from the least squared regression, especially when relative errors follow a normal distribution.

Concerning univariate calibration, linear regression by least squares has been used with the response weighted by the variance of the experimental signal[5]. Recently[6,7,8], different weights have been suggested ($1/y^2,\ 1/y, 1/y^{0.5}, 1/x^2, 1/x\ , 1/x^{0.5}$ being x the concentration and y the signal) aiming at selecting the one that best fit the data. Nevertheless, this does not result in the univariate version of LSRE, since relative errors are not those of the concentration but those of the experimental signal.

Regarding the field of multivariate calibration models, even though relative errors are not included in the objective function to be minimised, results are usually interpreted in terms of relative or percentage errors. But as far as the authors know, relative errors of concentrations have never been used as a fit criterion. In this work, we suggest a procedure focused on relative errors for the estimation of multivariate calibration models. More specifically, we propose minimizing the sum of squared relative errors as a criterion to develop a multivariate calibration model through Principal Component Regression. The procedure designed, which happens to be based on a specific weighted regression, seems to be suitable for dealing with poor prediction

problems. Section 2 details the suggested methodology and Section 3 the software used. In section 4, we provide different case studies for the multivariate calibration model (subsection 4.1). As they are intended to determine the analyte concentrations in complex mixtures, a soft regression model is required. Finally, the work concludes with some discussion about the calibration procedure proposed in terms of-accuracy (subsection 4.2), predictive ability (subsection 4.3) and a comparison when a logarithmic transformation of response. (subsection 4.4).

2. Methodology

Multivariate calibration models are usually based on a Least Squares procedure. Let X_1, X_2, ..., X_p denote the *p* independent variables and **y** the dependent variable. The Least Squares Regression model with *n* observations and *p* predictors can be written as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where **y** is an *n* x 1 vector of concentrations, **X** is an *n* x (*p*+1) matrix whose first column is a vector of ones and the remaining columns comprise the values of the *p* independent variables and **β** is a (*p*+1) *x* 1 vector of unknown coefficients that will be estimated. Lastly, **ε** is an *n* x 1 vector of underlying random errors, supposedly uncorrelated, with zero mean and equal variance σ^2, i.e., $E(\boldsymbol{\varepsilon})=\mathbf{0}$, $V(\boldsymbol{\varepsilon})=\mathbf{I}\sigma^2$. (See Ref. 1 for a detailed explanation).

The Least Squares (LS) criterion to estimate **β** is minimizing the Sum of Squared estimate of errors (*SSe*)

$$SSe = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

which gives an estimator $\widehat{\boldsymbol{\beta}}_{LS}$ as follows

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{3}$$

As mentioned above, our goal here is not to focus on errors as such, but on relative errors (re_i) defined as the errors over the corresponding actual values of the dependent variable

$$re_i = \frac{e_i}{y_i} = \frac{y_i - \hat{y}_i}{y_i} \quad (4)$$

where $\mathbf{y}$ is assumed to be a positive response variable ($y_i > 0$, i=1, …,n)

Thus, the new optimization criterion consists of minimizing the Sum of Squared relative errors (SSre)

$$SSre = \sum_{i=1}^{n} re_i^2 = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i}\right)^2 = \sum_{i=1}^{n} \left(\frac{1}{{y_i}^2}\right) (y_i - \hat{y}_i)^2 \quad (5)$$

As Eq. (5) points out, this is a kind of optimization problem, which can be transformed into, or tackled by means of, a Weighted Least Squares (wLS) procedure, with specific weights $w_i = \frac{1}{{y_i}^2}$.

Let $\mathbf{W}$ be an n x n matrix with the w_i on the diagonal and zeros everywhere else. The problem posed In Eq. (5) addresses the estimation of new coefficients $\widehat{\boldsymbol{\beta}}_{wLS}$ which minimise the weighted sum of squared residuals by means of Eq. (6).

$$\widehat{\boldsymbol{\beta}}_{wLS} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} \quad (6)$$

Nevertheless, multivariate calibration models built through least squares regressions, weighted or not, must deal with two inherent difficulties:

i. Datasets where the number of observations n is lower than the number of predictor variables p, i.e., the unknown coefficients outnumber the equations, which implies an underdetermined system.

   This is rather frequent in multivariate calibration, where the number of available calibration samples is often less than the number of instrumental signals, for instance in molecular spectroscopic signals (UV-visible, emission fluorescence, near infrared, …).

ii. Collinearity between the predictor variables, which leads to almost singular $\boldsymbol{X}^T\boldsymbol{X}$ matrices, and therefore unstable matrix inversions and $\widehat{\boldsymbol{\beta}}$ coefficients. 'Near-multicollinearity' is often found in spectral measurements and other instrumental signals[9].

Principal Component Regression PCR is a suitable calibration technique in 'soft calibrations' considering that spectra have a latent structure related to a few chemical species. In this regard, PCR provides the advantages of 'data compression' methods[9]: *p* predictors are turned into *a* principal components orthogonal to each other, thus supplying non-redundant information. These components, computed by a Principal Component Analysis (PCA), are calculated in descending order according to the amount of variance captured so a suitable choice of *a* accounts for the main patterns in the data. Once *a* Principal Components (PCs) have been chosen, PCR regresses the response variable $\mathbf{y}$ onto their corresponding PCs scores.

Formally, it turns out that the PCA is based on a decomposition of the data matrix $\mathbf{X_0}$ into two orthogonal matrices, $\boldsymbol{X_0} = \boldsymbol{T_0}\boldsymbol{P^T}$, $\boldsymbol{P^T}$ being an *a* x *p* loadings matrix and $\boldsymbol{T_0}$ the *n* x *a* principal component scores matrix. Notice that, in this decomposition, subscript 0 refers to a matrix without the column of ones.

Adding the column of ones to matrix $\boldsymbol{T}_0$, to obtain matrix $\boldsymbol{T}$, the PCR uses $\boldsymbol{T}$ as regressors, which leads to LS estimators for PCR, $\widehat{\boldsymbol{\beta}}_{LS}^{PCR}$

$$\widehat{\boldsymbol{\beta}}_{LS}^{PCR} = \boldsymbol{P}^T\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{y} \tag{7}$$

Using PCR with the purpose of minimizing the relative errors, new regression coefficients $\widehat{\boldsymbol{\beta}}_{wLS}^{PCR}$ can be computed as weighted least squared estimators:

$$\widehat{\boldsymbol{\beta}}_{wLS}^{PCR} = (\boldsymbol{T}^T\boldsymbol{W}\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{W}\boldsymbol{y} \tag{8}$$

matrix $\mathbf{W}$ being the specific weighting matrix

$$\boldsymbol{W} = \mathrm{diag}\left(\frac{1}{{y_1}^2}, \frac{1}{{y_2}^2}, \ldots, \frac{1}{{y_n}^2}\right) \tag{9}$$

As in the first part of the Eq. 7, we can write

$$\widehat{\boldsymbol{\beta}}_{wLS}^{PCR} = \boldsymbol{P}^T \widehat{\boldsymbol{\beta}}_{wLS} \quad (10)$$

and replacing $\widehat{\boldsymbol{\beta}}_{wLS}$ with its equivalent in Eq. (8)

$$\widehat{\boldsymbol{\beta}}_{wLS}^{PCR} = \boldsymbol{P}^T (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} \quad (11)$$

the coefficients of the proposed model are expressed from $\mathbf{X}$, $\mathbf{y}$, the weighting matrix $\mathbf{W}$ and the loadings from the PCA of $\mathbf{X_0}$.

However, the prediction power of PCR is closely related to the number *a* of components finally kept. If *a* is chosen considering a fixed proportion of variance explained in $\mathbf{X}$-space, chances are that either 1) it retains just a few PCs with the highest eigenvalues and discards some PCs with lower eigenvalues strongly related to the response, or 2) it retains components up to particular minor PCs relevant to the response, thus including all the preceding PCs (with larger eigenvalues), some of which may not have explanatory power on $\mathbf{y}$. Furthermore, when a cross-validation procedure is used, unimportant PCs may be included in the final model if they have comparatively large variance[10]. Either way, it may lead to an unsuitable selection of *a*.

Beyond these conventional procedures, a series of proposals to achieve the best choice of components in PCR may be found in the literature[11,12], also including the use of Bootstrap methods as an alternative to cross-validation[13], the use of model reduction methods through similarity transformations[14] or the selection of PCs by a penalized least squares minimization[10]. To our knowkedge, all the methods cited focus on minimising errors as such not relative errors.

To minimize the relative errors in 'soft' calibration models, we suggest a specific PCR as to the number of PCs considered and their selection procedure in the regression model. From an initial PCA with the largest possible number of PCs *a*=min (*n*-1, *p*), we have discarded the latest PCs when capturing a negligible percentage of the variance of $\mathbf{X_0}$. The calibration model through PCR is built by Backward elimination (section 15.3 of Ref. **¡Error! Marcador no definido.**), with p-to

remove ≥ 0.05, p-to-enter ≤ 0.05. A forward stepwise has also been run to assess the stability of the model concerning the stepwise procedure. Whenever these two methods substantially differ in the PCs selected, an ordinary least squares method has been computed and PCs with significant coefficients retained.

As outlying observations strongly affect all of the least squares methods, including PCR, even if some weighted least squares procedure is used[15], the dealing of these samples, when necessary, has been done according to their studentized residuals and their characterization as influential points.

Although Ref. 15 puts forward an exhaustive procedure to adressing outlying observations, which leads to a robust PCR regarding outliers, the procedure here discards observations when showing an external studentized residual larger than 3 (in absolute terms) or even larger than 2.5 if they are highly influential. The latter feature has been measured through an absolute value of DFFITS[1], the difference between the predicted values when the model is fit with and without the *i*-th data point**,** higher than $2p/\sqrt{n}$.

When a sample is discarded in PCR, it is also ruled out in building the wPCR model, so as to avoid the ocurrence of additional principal components derived from the presence of this particular observation.

Therefore, the multivariate calibration models derived from the minimization of squared errors, hereafter PCR, are computed in three stages:

i. PCA of $\boldsymbol{X_0}$, $\boldsymbol{X_0 = T_0 P^T}$ computing scores and loadings of the largest possible number of components *a*=min (*n*-1, *p*).
ii. Fine-tuning the number of PCs kept by discarding some PCs with virtually no explanatory power on $\boldsymbol{X_0}$**.** This leads to a smaller number of initial PCs which allows a cumulative percentage of explained $\boldsymbol{X_0}$ above 99.9% while each component explains at least 0.2%. These PCs will be potential predictors in the calibration model.

iii. PCR, regressing the $\mathbf{y}$ variable onto the scores of the potential PCs obtained in stage ii through a Backward Stepwise method.

Conversely, the multivariate calibration models coming from the minimization of *squared relative errors* become, as described above, a specific weighted principal component regression, henceforth denoted as wPCR. To compute them, we have just to include the weighting matrix $\mathbf{W}$ of Eq. (9) in stage iii.

In the case of considering the logarithmic transformation of the response (Section 4.4), Principal Component Regression will be noted as logPCR.

Once PCR and wPCR coefficients are estimated and their respective models validated, relative errors, re_i, *i*=1,.., *n*, for both models have been calculated and then *PCR* relative errors have been compared to wPCR relative errors. The rationale of the comparison lies in including overall summaries of performance, such as average and variability of relative errors, adjusted R^2 of the relative error-based models, as well as some insight on individual performance, namely re_i and their distribution across the samples (observations).

Along these lines, overall indicators are the Mean Absolute relative error (MAre)

$$MAre = \frac{\sum_{i=1}^{n}|re_i|}{n} \tag{12}$$

and the Root Mean Squared relative errors (RMSre)

$$RMSre = \sqrt{\frac{\sum_{i=1}^{n} {re_i}^2}{n}} \tag{13}$$

The adjusted R^2 of the models based on relative errors is computed with the known expression

$$R^2_{adj} = 100\left[1 - \left(\frac{n-1}{n-p}\right)\frac{SS_E}{SS_T}\right] \tag{14}$$

but the sum of squares include the weights used. That is, the total sum of squares is given by $SS_T = \sum_{i=1}^{n} w_i(y_i - \bar{y})^2$, whereas the residual sum of squares in the suggested models, the loss function to be minimised, is calculated as

$$SS_E = \sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2 \quad (15)$$

Individual comparisons of relative errors and their distribution are displayed by a Multiple Barchart with PCR and wPCR relative errors for each observation (sample). Multiple Barcharts of re_i are displayed in increasing order of dependent variable, the observed concentration.

Furthermore, to check the accuracy of the suggested method, wPCR, the predicted values by means of wPCR, fitted concentration $\hat{y}_i$, have been regressed onto the true values of concentration y_i , getting a regression line that can be regarded as an 'accuracy line'[16]. When predicted values are close enough to observed values, this line is expected to have a zero intercept and a slope equal to 1. As there is a negative strong correlation between the LS estimates of the intercept and the slope, a Joint Confidence Region for β_0 and β_1 at level 100(1-α) is used[5,16]. If point (0, 1) belongs to this region, the calibration method is considered true (unbiased). So Joint Confidence Regions, JCR, for the regressions 'estimated concentration vs. true concentration' have been computed for the proposed calibration method wPCR. Additionally, the representation of regions for both PCR and wPRC allows to compare the precision of these two methods: the larger the area of the confidence region the less precise the calibration method.

Moreover, the position of the intercept and the slope of the accuracy line relative to the point (0,1) is an indication of similarity between the two models.

In summary, given a matrix with *p* predictor variables $\mathbf{X_0}$ and a response vector **y**, the proposed calibration procedure, wPCR, consists of:

a) Performing a PCA while keeping as many components, *a*, as possible, with the two criteria described in steps (i) and (ii) above mentioned, thus obtaining the scores matrix **T.**

b) Using Eq. (8) along with the weigths in Eq. (9), computing $\widehat{\boldsymbol{\beta}}_{\mathrm{wLS}}^{\mathrm{PCR}}$ by means of a weighted least squares regression of **y** onto **T**, through a backward stepwise method for variable selection, while identifying possible outliers.

c) Assessing the absence of bias in the calibration model obtained by verifying that the point (0,1) belongs to the Joint Confidence Region for the independent term and the slope of the accuracy line.

3. Software

The regression models are fitted and validated with the statistical program STATGRAPHICS Centurion XVIII[17]. To obtain the JCR, a home made program has been written in MATLAB[18].

4. Case studies on Multivariate Linear Calibration by PCR

To evaluate the performance of our proposal on relative errors, a series of datasets are presented here. Their descriptions and references are summarized in Table I.

Our main results for every dataset are summarized in Tables II and III**.** Table II shows a comparative list of the principal components kept by each model for every case study evaluated. The performance indicators are gathered in Table III.

4.1 Comparative analysis PCR versus wPCR

4.1.1 Dichromate-Permanganate dataset

The calibration models to predict the concentration of both potassium dichromate and potassium permanganate have been based on six potential PCs accounting for 99.99% of the X variance in each one. As two outliers have been ruled out, the models have been built from 16 samples (first and second row in Table II).

The PCR to predict the concentration of potassium dichromate keeps just three PCs, namely the first, second and fifth one. The distribution of relative errors is displayed in Figure 1a. The most outstanding feature is the larger relative errors of the first samples depicted (those with extremely low concentration of potassium dichromate), particularly samples number 2 and 3. This calibration model leads to an average relative error of 7.8% (Table III) but the use of wPCR shortens it while reducing the RMSre (39%), thus resulting in a more balanced distribution of relative errors.

Regarding the concentration of potassium permanganate, the regular *PCR* calibration model takes four PCs: the first two, fourth and fifth. In this case, predictions are rather close to the observed concentrations, MAre being below 2%, although samples number 1 and 3, with low concentration of potassium permanganate, show the highest relative errors. The calibration computed by wPCR selects just the first two PCs, leading to a substantial reduction in the average relative error together with a remarkable fall of the RMSre, about 37% (Tables II and III). The prediction of the concentration in the above-mentioned samples strikingly improves so that relative errors become more uniformly distributed (Figure 1b), none of them exceeding 3%.

Table I. Datasets studied

	Dataset	Objects	Predictors	Responses	Ref.
1	Dichromate-Permanganate	Binary mixtures	UV-vis absorbance, 380-520 nm each 20 nm	Analyte concentrations. Table S1	[19]
2	Polycyclic Aromatic Hydrocarbons	Mixture of 10 compounds	UV-vis absorbance, 220-350 nm each 5 nm	Pyrene, benzanthracene and phenanthrene concentrations. Table S2	[20]
3	Vintages Port	Portuguese wines aged from 4 to 27 years	41 oenological parameters	Age of vintages. Table S3	[21]
4	Zinc	Calibration samples	Polarography. Current registered, 0.095 and -1.117 V	Zinc concentration. Table S4	[22]
5	Food colorants	Binary mixtures	UV-vis absorbance, 340-570 nm each 5 nm	Tartrazine (E-102) and sunset yellow (E-110) concentrations. Table S5	Our lab
6	Fraudulent whisky	Ternary mixtures of water, low-quality and high-quality whisky	UV-vis absorbance, 220-400 nm each 4 nm	Percent of each whisky. Table S6	[23]
7	Chromium in toys	Derivatized samples of hexavalent chromium	UV-vis absorbance, 472-628 nm each 4 nm	Analyte concentration. Table S7	[24]

4.1.2 Polycyclic Aromatic Hydrocarbons (PAHs) dataset

Table II. Performance of suggested model wPCR vs regular model PCR in datasets. Number of observations by number of predictors (n x p), Number of initial Principal Components (PCs) considered, Percentage of X variance captured by the initial PCs, Regression method, Number of PCs selected, PCs selected and Number outliers removed. Abbreviations used: Back: Backward regression, OLS: Ordinary Least Squares Regression, Py: Pyrene, Benz: Benzanthracene, Phen: Phenanthrene

Datasets	n x p	No. initial PCs	% X	Regression method	No. PCs selected		PCs selected		No. outliers removed	
					PCR	wPCR	*PCR*	wPCR	PCR	wPCR
Potassium dichromate	18 x 9	6	99.99	Back	3	3	1, 2, 5	1, 2, 5	2	2
Potassium permanganate		6	99.99	Back	4	2	1, 2, 4, 5	1, 2	2	2
Py	25 x 27	16	99.99	Back	9	9	1-7, 10, 12	1-4, 6-7, 10-12	0	0
Benz		16	99.99	Back	11	9	1-7, 8-10, 12	1- 7, 11-12	0	0
Phen		16	99.99	OLS	8	8	1-3, 8, 10-11, 13-14	1-3, 8, 10-11, 13-14	1	1
Vintages	20 x 41	16	98.90	OLS	5	5	1, 2, 3, 4, 7	1, 2, 3, 4, 7	1	1
Zinc	27 x 34	12	99.20	Back	9	10	1- 7, 10, 12	1-7, 8, 10, 12	3	2
Tartrazine	12 x 47	5	99.98	Back	3	3	1, 2, 3	1, 2, 3	0	0
Sunset Yellow		5	99.97	Back	3	3	1, 2, 3	1, 2, 3	1	1
Low-quality whisky	24 x 91	6	99.96	Back	3	3	1, 2, 3	1, 2, 3	2	5
High-quality whisky		6	99.94	Back	3	3	1, 2, 3	1, 2, 3	0	0
Hexavalent chromium	17 x 79	6	99.99	OLS	3	3	1, 2, 5	1, 2, 5	5	5

Table III. Performance of suggested model wPCR vs regular model PCR in fitting and in prediction. Mean Absolute relative error (MAre), Root Mean Squared relative errors (RMSre), Improvement in Mare and RMSre (%). Abbreviations used: Py: Pyrene, Benz: Benzanthracene, Phen: Phenanthrene.

Datasets	In fitting						In prediction					
	MAre			RMSre			MAre			RMSre		
	PCR	wPCR	Impr (%)	PCR	wPCR	Impr (%)	PCR	wPCR	Impr (%)	PCR	wPCR	Impr (%)
Potassium dichromate	7.88	5.63	28.6	0.1215	0.0733	39.7	9.19	6.08	33.8	0.1145	0.0646	43.6
Potassium permanganate	1.53	1.32	13.7	0.0228	0.0144	36.8	1.85	1.23	33.5	0.0208	0.0146	29.8
Py	7.08	5.05	28.7	0.1045	0.0617	41.0	9.02	6.13	32.0	0.1059	0.0708	33.1
Benz	2.54	2.19	13.8	0.0357	0.0282	21.0	5.00	2.84	43.2	0.0630	0.0404	35.9
Phen	5.33	4.33	18.8	0.0668	0.0596	10.8	8.56	8.10	5.4	0.0995	0.0948	4.7
Vintages	13.00	10.90	16.2	0.1632	0.1410	13.6	15.54	11.37	26.8	0.1817	0.1562	14.0
Zinc	1.25	1.22	2.4	0.0200	0.0153	23.5	2.92	2.45	16.1	0.0418	0.0413	1.2
Tartrazine	0.79	0.78	1.3	0.0114	0.0093	18.4	1.37	0.95	30.7	0.0201	0.0097	51.7
Sunset Yellow	0.64	0.60	6.3	0.0082	0.0077	6.1	1.14	0.96	15.8	0.0141	0.0131	7.1
Low-quality whisky (old)	2.11	0.80	62.1	0.0410	0.0100	75.6	2.58	1.72	33.3	0.0297	0.0202	32.0
High-quality whisky (Chivas)	1.78	1.27	28.7	0.0228	0.0165	27.6	5.73	5.54	3.3	0.0779	0.0659	15.4
Hexavalent chromium	1.76	1.59	9.7	0.0225	0.0193	14.2	20.3	18.12	10.7	0.2864	0.1884	34.2

These models have been built from sixteen potential PCs accounting for 99.99% of the X variance in each of them. Except for the *Phen* case, where an outlier has been detected, the calibration models have been obtained using all available samples (third to fifth rows in Table II)

The relative errors of predictions have been computed and summarized in Table III (third to fifth rows). In the three PAHs studied, the wPCR calibration allows relative errors to decrease, both in terms of MAr*e* and in terms of variability, markedly in the prediction of *Py*, whose RMSre achieves a reduction of 41%.

As to the PCs kept (Table II), the suggested method wPCR retains a similar number of components to that of *PCR*, or even less (in the Benz case) still resulting in better outcomes.

Figure 2a, concerning Py predictions, shows *PCR* relative errors above 20 or even 30% in samples with the lowest concentrations (especially, samples 1 to 3), which decrease by the application of wPCR at the cost of a slight increase of relative errors in some higher concentration samples, all but one with records below 10% in absolute terms.

As to the pattern displayed for *Benz* (Figure 2b), predictions are drastically improved for samples number 1 to 6, again those with the smallest concentrations. The wPCR proposed model gets a more homogeneous distribution of relative errors, most of them ending up below 5%.

Regarding the prediction of *Phen*, the stepwise methods have shown some differences in the PCs selected so an ordinary least squares method have been used and 8 PCs with significant coefficients selected. As Figure 2c displays, the samples with the lowest concentrations highly advantage from the application of wPCR, as relative errors sharply fall in samples number 1 to 10, although the overall behaviour of the variability of relative errors is not as enhanced as in the previous cases.

4.1.3 Vintages Port dataset

The proposed method to optimize relative errors may be suitable in a variety of contexts other than calibration models, i.e. when running a Principal Component Regression as long as the computation of relative errors makes sense. For illustrative purposes, we present here a dataset, taken from Ref. 29, concerning a regression model to predict the age of a wine. The dataset is made up of samples from 20 vintages with ages ranging from 4 to 27 years and 41 oenological variables.

When builiding PCR, as described in section 2, an object is discarded as outlier and a final model is built on 19 vintages (Table II, sixth row). The PCR model draws five PCs as regressors, specifically the first four and the seventh one and gives rise to absolute errors below 3 years in most of the cases. However, this results in a MAre above 13% (Table III, sixth row), since relative errors for some younger vintages approach 30%, thus meaning poor predictions. Moreover, as Figure 3 displays, there is a noticeable variability among the wines, with better predictions in older ones, namely, those aged over 20 years.

The wPCR suggested method selects a subset of analogous PCs but achieves better predictions in percentage terms, especially for vintages of five or less years, where relative errors fall sharply. Roughly speaking, vintages up to ten years (wines number 1 to 10) get lower relative errors without substantially worsening predictions for older vintages (more than 20 years). Although not all the vintages benefit from the wPCR model, the new average relative error falls to about 10% and a small reduction of *RMSE* takes place, thus improving the predicted ages and reaching a more equalized distribution (Figure 3).

4.1.4 Zinc dataset

The analysis is carried out from 12 initial PCs capturing 99.2% of **X** variance. Regular *PCR* selects ten PCs (from first to seventh, plus tenth and twelfth) which leads to a small average relative error (seventh row in Tables II and III). Nevertheless, a few samples with low concentration of zinc,

particularly number 1 and 4, show predictions whose relative errors exceed 5%. When applying the wPCR model, which takes an aditional PC (the eighth one), predictions for these samples improve without affecting samples with high concentrations meaninfully. Since the RMSE decreases by 23%, the overall distribution of relative errors becomes more balanced (Figure S1 in supplementary material).

4.1.5 Food colorants dataset

The calibration models to predict the concentration of both *Tartrazine* and *Sunset Yellow* have been based on five potential PCs representing about 99.98% of the **X** variance in each additive (eighth and ninth rows in Table II). Predicted concentrations of the two additives have been performed from the first three PCs, both in PCR and wPCR, resulting in small relative errors, lower than 1% in average (Table III). Nonetheless, wPCR achieves even better predictions mainly in samples with lower concentrations (Figure 4a). Regarding Tartrazine, the suggested calibration model gets a decrease of 19% in RMSre. Concerning Sunset Yellow, a slight reduction in RMSre is seen (eighth and ninth rows in Table III) but still a visible improvement for samples number 1 and 3 (Figure 4b)

4.1.6 Fraudulent whisky dataset

The percentages of high-quality and low-quality whisky have been predicted from 6 initial PCs which capture about 99.95% of the **X** variance in both cases. The calibration models select the first three PCs, PCR and wPCR alike (tenth and eleventh rows in Tables II and III).

When predicting the percentage of high-quality whisky, the wPCR model reduces the average relative error as well as the RMSre (by almost 30%). In addition, the predictions for whiskies with less than 50% of high-quality whisky (samples 1 to 9) improve notably, without having an important effect over the predictions for samples with large percentages of high-quality whisky (see Figure 5a), any relative error surpassing 5% (samples 10 to 24).

As to the percentage of low-quality whisky, where five outlier samples have been found and discarded (Table II), MAre and RMSre fall sharply, as Table III shows. This results in a decrease of relative errors, especially in samples 1 and 3, and a more homogeneous distribution with all relative errors virtually under 2% (Figure 5b)

4.1.7 Chromium in toys dataset

The calibration model to predict the concentration of hexavalent chromium has been built from an initial set of 6 PCs which represent 99.99% of the **X** variance. Again, five outlier samples have been ruled out and since the stepwise procedures (backward and forward) do not retain a similar number of components, an Ordinary Least Squares (OLS) regression has been conducted. Thus, the predition model keeps 3 PCs, namely, the two first and the sixth one (twelfth row in Table II).

The outcome of wPCR slightly improves that of PCR, in terms of MAre and RMSre (twelfth row in Table III), but still gets a better prediction for low concentration samples, as Figure S2 in the Supplementary material displays.

Overall, the analysis of the results of Table II shows that in all the cases analysed there is a reduction of MAre ranging from 1 to 28.7% and even greater for RMSre ranging from 6.1 to 41.0 %, excluding the case of Low-quality whisky in which reductions in MAre and RMSre are 62.1% and 75.6% respectively. The criterion of minimising relative errors supplies smaller MAre values as expected. The decrease of RMSre caused by a more uniform distribution of relative errors in the predicted response is even more remarkable.

4.2 Accuracy lines

Finally, the performance of the method in terms of accuracy has been evaluated, the accuracy of the suggested method, wPCR and also that of PCR has been checked for all the datasets. The parameters of these regressions 'predicted concentration versus true concentration' are shown in Table IV. The JCR for the parameters of the accuracy line, all of them built at 95% confidence, are

showed in Figure S3 of Supplementary material, where each sub-figure includes the representation of both the PCR and the wPCR ellipses for comparative purposes. As shown, the point (0,1) falls inside the corresponding JCR in every sub-figure, thus meaning that the intercept and the slope are not significantly different from 0 and 1 jointly. Therefore, the wPCR method has neither constant nor proportional bias.

Table IV. Accuracy line. Estimates for Intercept b_0 and slope b_1, Residual standard deviation Sy/x, explained variance R^2(%). Abbreviations used: Py: Pyrene, Benz: Benzanthracene, Phen: Phenanthrene

	PCR				wPCR			
Datasets	b_0	b_1	Sy/x	R^2	b_0	b_1	Sy/x	R^2
Potassium dichromate	0.0319	0.9925	0.2152	99.25	0.1211	0.9594	0.2329	99.06
Potassium permanganate	0.0066	0.9995	0.1531	99.95	0.0499	0.9950	0.2227	99.89
Py	0.0082	0.9820	0.0298	98.20	0.0012	0.9922	0.0350	97.58
Benz	0.0033	0.9980	0.0360	99.80	0.0253	0.9811	0.0532	99.53
Phen	0.0091	0.9833	0.0347	98.33	0.0188	0.9562	0.0359	98.11
Vintages	0.6201	0.9550	1.6496	95.24	0.6147	0.9179	1.6891	94.62
Zinc	-0.0380	1.0102	0.1409	99.39	0.0359	0.9933	0.1096	99.62
Tartrazine	0.0024	0.9994	0.0439	99.94	0.0011	0.9996	0.0500	99.92
Sunset Yellow	0.0006	0.9998	0.0234	99.98	-0.0080	1.0021	0.0255	99.98
Low-quality whisky	0.0348	0.9994	0.7330	99.94	0.2954	0.9946	0.5846	99.95
High-quality whisky	0.1111	0.9982	1.1899	99.82	0.1792	0.9960	1.3127	99.78
Hexavalent chromium	0.0006	0.9989	0.0111	99.89	0.0009	0.9977	0.0124	99.86

From an analytical point of view, it is important to figure out the effect on calibration of using the adjustment criterion of minimizing the sum of squared relative errors instead of the usual sum of squared errors. In this sense, the orientation and size of the JCR obtained with wPCR and PCR for each case study show that, they are mostly similar, with no discernible pattern linked to the type of regression used.

For example, in Benzanthracene (Figure S3d), PCR calibration (magenta ellipse) is more accurate than wPCR (blue ellipse) but the opposite occurs in the Zinc case, Figure S3g.

A detailed study of the results allows to note that changes in the accuracy line are independent of the percentage reduction of RMSre. For instance, the JCR of both models (PCR and wPCR) in the case studies of Phenanthrene, Vintages and Hexavalent chromium (Figures S1e, S1f and S1l) differ

in various ways, whereas the percentage reductions of RMSre are quite similar (10.8, 13.6 and 14.2% respectively)

A similar remark may be done for MAre, which is virtually the same for analytes Benzanthracene and Potassium permanganate, 13.8 and 13.7 % respectively, although the type of calibration, wPCR or PCR, impacts on the JCR differently, as seen in Figures S3b) and S3d).

4.3 Comparison of PCR and wPCR in prediction

The comparative evaluation of two methods can be done through a test set independent of the training set or by crossvalidation, CV. Despite the popularity of CV, its use has been criticized[25].

Moreover, in most of our case studies, calibration samples are mixtures of analytes so that samples with identical response (concentration) correspond to different signals due to the presence of different amounts of another analyte. In other words, there is no redundant information, but each sample provides significantly different information. In this sense, discarding samples can lead to lower quality designs affecting the validity of the calibration model, regardless of whether the regression is weighted or not. In line with this concern, we have decided to limit the potential adverse effect derived of constantly changing samples on the calibration models by using an external dataset instead of a cross-validation.

Therefore, to evaluate the effect in prediction of wPCR versus PCR for each dataset, a test set has been defined by randomly selecting about 25% of the dataset samples (or observations). Subsequently, a PCA on the training set has been conducted following the criteria described in section 2 of methodology. Finally, both unweighted and weighted backward stepwise regressions (PCR and wPCR) were computed in the test set, thus allowing to calculate the relative errors in prediction.

The results in terms of Mare and RMSre of these relative errors are gathered in Table III (columns 8-9 and 11-12, respectively). As shown, relative errors in prediction are always higher than in

fitting. However, relative errors in prediction computed by wPCR are always lower than those computed by PCR. The improvement, in percentage, for MAre varies between 3.3 and 43.2 while for RMSre, it ranges from 1.2 to 51.7%.

The distributions of relative errors in the test set samples (observations), PCR vs wPCR, have been displayed in Figures S4 to S10 of the supplementary material. In general, the pattern of smaller relative errors in samples of lower concentration, arises again except for the Phenanthrene calibration (Figure S5c). Concerning this case study, it is worth noting that the spectra of the PAHs dataset correspond to mixtures of 10 analytes. The effect of having the same response (concentration to be predicted) in samples with different amounts of other analytes can be clearly seen in the same Figure S5 of the supplementary material.

4.4 Effect of the logarithmic transformation of the response

Logarithmic transforming the response is one of the most used methods to stabilize a least squares regression and is particularly suitable in case of multiplicative errors. Since the difference between the logs of two values is relatively smaller when they are large than when they are small, it can be thought that using logPCR might lead to smaller relative errors in low concentration samples.

To explore this possibility, once the PCA has been done, the logarithm of the response has been regressed onto the **T** variables, with a backward stepwise procedure for variable selection. The results in terms of relative errors are presented in Table S8 of the supplementary material. As for the 12 cases studied, Figure 6 shows that the value of MAre obtained with logPCR is worse than with PCR which, in turn, is worse than the one from wPCR. RMSre exhibits the same pattern, as shown in table S8 of the supplementary material.

CONCLUSIONS

In the 12 predicted variables, it has been shown that a regression model based on relative errors, avoids the tendency to increase relative errors in small values of the response relative to the large ones, which is typical of the usual regression models. This is relevant in multivariate calibrations where the response is an analyte concentration and the relative error in its determination is a common analytical criterion for assessing calibration performance. To address the problems of collinearity and correlation of multivariate analytical signals, the wPCR proposed is a Principal Component Regression along with the criterion of adjustment of relative errors.

The reduction in the mean of the absolute values of relative errors (MAre) achieved by wPCR over PCR ranges from 1 to 62.1% (3% to 43% in prediction) and 6.1 to 75.6% (1.2% to 52% in prediction) in its standard deviation (RMSe). The wPCR model also outperforms PCR with logarithmic transformation of the response (logPCR).

The regression model attained by wPCR has no bias and its effect on the parameters of the accuracy line is not significant, nor is it related to the level of reduction in MAre and/or RMSre. Therefore, the change of criteria for achieving the Principal Component Regression does not introduce systematic trends.

Acknowledgments

The authors thank the financial support provided by Spanish MINECO (AEI/FEDER, UE) through project CTQ2017-88894-R and Consejería de la Junta de Castilla y León (BU052P20), both co-financed with European Regional Development Funds. Junta de Castilla y León and Fondo Social Europeo.

REFERENCES

1. Draper NR, Smith H, *Applied Regression Analysis*. 3rd edition New York NY: John Wiley and Sons; 1998.

2. Ferreira JM, Caramelo L, Chhabra RP. The use of relative residues in fitting experimental data: an example from fluid mechanics. International Journal of Mathematical Education in Science and Technology. 2000; 31: 545-552 DOI: 10.1080/002073900412651.

3. Tofallis C, Least Squares Percentage Regression. *Journal of Modern Applied Statistical Methods* 2008. 7:526-534 DOI: 10.22237/jmasm/1225513020.

4. Montgomery DC, Peck EA, Vining GG, *Introduction to Linear Regression Analysis, 5th edition Jhon Wiley and Sons*; 2012.

5. Ortiz MC, Sánchez MS, Sarabia LA, Quality of Analytical Measurements: Univariate Regression. In Brown SD, Tauler R, Walczak B, ed. *Comprehensive Chemometrics. Chemical and Biochemical Data Analysis*, 2nd ed. Amsterdam, Elsevier: 2020.

6. Almeida AM, Castel-Branco MM, Falcao AC, Linear regression for calibration lines revisited: weighting schemes for bioanalytical methods; *Journal of Chromatography B*, 2002 774 215–222.

7. Alladio E, Amante E, Bozzolino C, et al. Effective validation of chromatographic analytical methods: The illustrative case of androgenic steroids, *Talanta* 2020; 215 120867; DOI /10.1016/j.talanta.2020.120867.

8. Sánchez J.M., Linear calibrations in chromatography: The incorrect use of ordinary least squares for determinations at low levels, and the need to redefine the limit of quantification with this regression model, *Journal of Separation Science*, 2020;1–10.

9. Naes T, Isaksson T, Fearn T, Davies T. *A User-friendly Guide to Multivariate Calibration and Classification.* NIR Publications, 2002.

10. Lee H, Park YM, Lee S. Principal Component Regression by Principal Component Selection. *Commun. Stat. Appl. Methods* 2015; 22(2):173-180. DOI: 10.5351/CSAM.2015.22.2.173.

11. Næs T, Martens H. Principal Component Regression in NIR Analysis: Viewpoints, Background Details and Selection of Components. *J Chemom.* 1988; 2.2: 155–167.

12. Sutter JM, Kalivas JH, Lang PM. Which Principal Components to Utilize for Principal Component Regression. *J Chemom.* 1992, 6.4: 217–225.

13. Wehrens R, Van Der Linden WE. Bootstrapping Principal Component Regression Models. *J Chemom.* 1997; 11.2: 157–171.

14. Ergon R. Finding Y-relevant part of X by use of PCR and PLSR model reduction methods. *J Chemom.* 2007, 21: 537–546.

15. Hubert M, Verboven S. A robust PCR method for high-dimensional regressors. *J Chemom.* 2003, 17:438-452.

16. Mandel J, Linning FJ. Study of Accuracy in Chemical Analysis Using Linear Calibration Curves. *Anal. Chem*. 1957, 29.5: 743-749.

17. STATGRAPHICS Centurion XVIII Version 18.1.11 (64 bit), Statgraphics Technologies, Inc., The Plains, VA, USA, 2018.

18. MATLAB, Version 9.7.0.1190202, The Mathworks, Inc., Natick, MA, USA, 2019.

19. Ortiz MC, Herrero A, Sanllorente S, Reguera C. La calidad de la medida analítica. *The quality of the information contained in chemical measures*, Servicio de Publicaciones de la Universidad de Burgos, Burgos, Spain, 2005. ISBN: 84-96394-26-1.

20. Brereton RG. *Chemometrics: Data Driven Extraction for Science,* 2nd ed. Chichester, UK: John Wiley & Sons Ltd; 2018.

21. Ortiz MC, Sarabia LA, Symingtonc C, Santamaría F, Íñiguez M. Analysis of Ageing and Typification of Vintage Ports by Partial Least Squares and Soft Independent Modelling Class Analogy. *Analyst* 1996; 121:1009-1013.

22. Herrero A, Ortiz MC, Multivariate calibration transfer applied to the routine polarographic determination of copper, lead, cadmium and zinc, *Analytica Chimica Acta* 1997; 348: 51-59.

23. Lucio Dallo FJ, Identificación y cuantificación de la adulteración de un whisky de calidad, End Degree Project, University of Burgos, 2013.

24. Real García, BD, Optimización del funcionamiento de procedimientos analíticos en cromatografía y espectroscopia mediante el uso de diseño de experimentos y quimiometría, PhD, University of Burgos, 2010.

25. Esbensen K H, Geladi P, Principles of Proper Validation: use and abuse of re-sampling for validation, J. Chemometrics 2010; 24:168–187, DOI: 10.1002/cem.1310.

FIGURE CAPTIONS

Fig. 1 *Dichromate-Permanganate* dataset. Relative error distributions compared: PCR (in blue) vs wPCR (in orange). 1a) Potassium Dichromate, 1b) *Potassium permanganate.* Sample codes and concentrations are shown in Table S1.

Fig. 2 *Polycyclic Aromatic Hydrocarbons (PAHs)* dataset. Relative error distributions compared: PCR (in blue) vs wPCR (in orange). 2a) Py, 2b) Benz, 2c) Phen. Abbreviations used: Py (Pyrene), Benz (Benzanthracene) and Phen (Phenanthrene). Sample codes and concentrations are shown in Table S2.

Fig. 3 *Vintages* dataset. Relative error distributions compared: PCR (in blue) vs wPCR (in orange). Vintage codes and ages are shown in Table S3.

Fig. 4 *Food colorants* dataset. Relative error distributions compared: PCR (in blue) vs wPCR (in orange). 4a) Tartrazine (E-102), 4b) Sunset Yellow (E-110). Sample codes and concentrations are shown in Table S5.

Fig. 5 *Fraudulent whisky* dataset. Relative error distributions compared: PCR (in blue) vs wPCR (in orange). 5a) High-quality, 5b) Low-quality. Sample codes and concentrations (%) are shown in Table S6.

Fig. 6 MA*re* computed by means of logPCR, PCR and wPCR for the 12 responses studied.