Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 526 (2004) 139–146
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

# Building robust calibration models for the analysis of estrogens by gas chromatography with mass spectrometry detection

Inmaculada García[a], Luis Sarabia[b], M. Cruz Ortiz[a,*], J. Manuel Aldama[c]

[a] *Department of Chemistry, Faculty of Sciences, University of Burgos, Pza. Misael Bañuelos s/n. 09001 Burgos, Spain*
[b] *Department of Mathematics and Computation, Faculty of Sciences, University of Burgos, Pza. Misael Bañuelos s/n. 09001 Burgos, Spain*
[c] *Instituto de Ciencias de la Salud (Análisis Físico-Químicos), Ctra. Extremadura, Km 114, 45600 Talavera de la Reina, Toledo, Spain*

Received 30 April 2004; received in revised form 16 September 2004; accepted 16 September 2004
Available online 28 October 2004

**Abstract**

Five hormonal growth promotants (diethylstilbestrol, hexestrol, dienestrol, 17-β-estradiol and 17-α-ethynylestradiol) have been analysed by gas chromatography with mass spectrometry detection (GC/MS, SIM mode) for four non-consecutive days. The aim is to build models with stable predictions. The strategies applied are internal standardization and global models carried out by gathering signals recorded on several days. Two models were examined: univariate models (with standardized peak area) and PARAFAC2 (the analyte scores were standardized by the scores of the internal standard). Internal standardization has been proved to be efficient for both models of dienestrol and ethynylestradiol. The mean relative error in absolute value when samples recorded on a different day to the calibration set are quantified by PARAFAC2 is 7.00% and 7.11% for dienestrol and ethynylestradiol, respectively. For diethylstilbestrol and estradiol, internal standardization was combined with global calibration models built with signals recorded under several sources of variability (different days). Thus predictions become steadier over time and in the estradiol example, errors decrease from 33.10% to 9.76%. The mean relative error in absolute value with PARAFAC2 updated models oscillates between 6.34% for ethynylestradiol and 10.74% for diethylstilbestrol. For univariate updated models errors range from 6.42% to 14.19% for ethynylestradiol and estradiol respectively. The combination of both strategies has been proved to be efficient independently of the analyte and of the signal order.


*Keywords:* Calibration maintenance; Internal standardization; PARAFAC2; Robust calibration models

## 1. Introduction

Due to their anabolic properties, hormonal growth promotants (HGPs) have increasingly been used during the last decades to accelerate growth in livestock. That is why in February of 1998 an assessment [1] was initiated to evaluate the side effects of six substances with estrogenic, androgenic or gestagenic action. In April 1999, the potential adverse effects to human health from hormone residues and their metabolites were published. It was proved that the six substances considered may cause developmental, immunological, neurobiological, immunotoxic, genotoxic and carcinogenic effects and that of the various susceptible risk groups, children constitute the group of greatest concern. Furthermore, the recurring use of these hormonal substances to promote growth in livestock increases their levels in the environment which might have dangerous consequences not only for humans but also for wildlife. Directives 1996/22/EC and 2003/74/EC [1] have accordingly banned substances with estrogenic, androgenic or gestagenic effects for administering to farm or aquaculture animals. Their administration is exclusively authorised for the purpose of therapeutical or zootechnical treatment.

Hormones have been determined by several analytical methods [2] although mass spectrometry [3] is the most widely used technique mainly employed as detector, in gas chromatography, GC/MS [4–7], and liquid chromatography,

* Corresponding author. Fax: +34 947 258 831.
*E-mail address:* mcortiz@ubu.es (M. Cruz Ortiz).

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.09.044

LC/MS [5,8], and LC/MS/MS [5,9]. Regarding the sample nature, urine [4] is the most frequent matrix analysed to detect illegal administration in living animals as being a readily available sample. Moreover, hormonal residue levels are rising in river water [6,8], sewage effluents [9], etc., because of the increasing supply of both endogenous and synthetic estrogens to animals and humans. Methods for analysing residues in aquatic environmental samples are examined in Ref. [3].

In this paper five substances with estrogenic action have been determined by GC/MS, diethylstilbestrol (DES), hexestrol (HEX) and dienestrol (DIEN), belonging to the stilbene group and the steroids 17-β-estradiol ($E_2$) and 17-α-ethynylestradiol ($EE_2$). During the method validation [10], the robustness of the analytical method itself to operating and environmental conditions is often tested [11,12] but not the stability or robustness of the mathematical models for quantifying new samples over time. The calibration step in residue analysis is time consuming, which is why once the model has been validated it is of interest to maintain accurate (true and precise) predictions over a long period of time.

The sources of variability in GC/MS causing changes in the signal and therefore instability in the models can be related to variations in the chromatograph (column and flow meter ageing, deterioration of certain components, replacement of worn out parts, etc.) or in the mass detector (cleaning the ion source, tune the detector, etc.) or in the experimental conditions (temperature, humidity, analyst, preparation of the standards, material, chemicals, etc.) among others.

Internal standardization of the peak area has already been proved to be an efficient technique to improve the reproducibility of the GC/MS results. Several alternatives have been used to build robust and stable two-way calibration models and to correct the multivariate signal instability in mass spectrometry without a complete recalibration stage. Ref. [13] shows that internal standardization of multivariate signals (mass spectra) with the intensity of a sole ion is not effective enough to quantify standards arranged 1 month and 1 year after the model was performed. The predicted concentrations were biased towards greater concentrations because the signal stability is different for each mass fragment. Besides, mass spectrometry is a technique which provides many ions representing only noise and depending on the fragmentation pattern, even significant masses are affected by noise to a different extent.

Because of its uniqueness property (quantify as well as identify the chromatograms and the spectra of pure analytes in non-specific signals), PARAFAC2 [14] is often used for the analysis of biological matrices [15]. As PARAFAC2 models second-order data (mass spectra recorded at different elution times), in this paper we propose the standardization of the scores estimated by PARAFAC2 (rather than the raw second-order signal) by the scores computed for the internal standard. As the whole mass spectra is taken into account to compute the scores of both the analyte and the internal standard, the correction might be more effective.

For the sensitivity to be more stable, internal standardization will be combined with another technique based on updating calibration models with signals recorded under several sources of variability. Thus the extra-variability in the chromatograms will be accounted for and the calibration models (here called global or updated) become more robust to experimental changes. Because updated models incorporate information related to new states, the quantification does not depend so much on them but on the analyte itself and its fragmentation pattern. Thus the updating of models makes possible the quantification of samples measured on different days to the calibration set and leads to more accurate predictions.

Although the methodology based on building updated models has predominantly been applied to first-order signals from near-infrared (NIR) spectroscopy [16], the maintenance of signals from electrochemistry [17] and Raman spectroscopy [18] among others has also been explored. Other alternatives for maintaining the validity and applicability of two-way calibration models can be found in Ref. [19].

This paper is aimed at combining two strategies to build robust calibration models whatever the signal order. On one hand, second-order signals have been newly standardized through the scores of the PARAFAC2 model. Furthermore, including signals recorded on different days is another alternative to keep the predictions over time.

## 2. Theory

### 2.1. PARAFAC2 models

Chromatographic data will be decomposed according to the PARAFAC2 [14,20] model which can be expressed as follows:

$$X_k = A D_k (B_k)^{\mathrm{T}} + E_k \quad (1)$$

where the matrix $X_k$ is the $k$th slab with dimensions $I \times J$, $I$ mass fragments recorded at $J$ times during analyte elution, $A$ the loading matrix of the first mode, $D_k$ a diagonal matrix holding the $k$th row of the third mode, $B_k$ the $k$th profile of the second mode, and $E_k$ is the matrix of the residuals.

PARAFAC2, unlike PARAFAC [21], does not assume that the elution profiles of each component are invariant in each sample but the cross-product matrix $B^{\mathrm{T}}B$. This allows one to solve the problem of factor shifts disturbing trilinearity such us changes in the retention time and the shape of the chromatograms which makes the model more suitable for developing the strategy of updating calibration models. The PARAFAC2 model expressed in Eq. (1) constrained with constant $B^{\mathrm{T}}B$ [22] over $k$ is unique (estimates the true underlying profiles) under certain conditions described in Ref. [14].

## 3. Experimental

### 3.1. Chemicals and reagents

Analytical standards of diethylstilbestrol, hexestrol, dienestrol, 17-β-estradiol and 17-α-ethynylestradiol were purchased from Riedel de Haën (Deisenhofen, Germany). Methanol and *tert*-butyl methyl ether were obtained from Merck (Darmstadt, Germany).

### 3.2. Standard solutions and data sets (C, V and T)

Individual standard solutions of diethylstilbestrol, hexestrol, dienestrol, 17-β-estradiol and 17-α-ethynylestradiol were prepared at 0.3 g $l^{-1}$ in methanol. 3 mg $l^{-1}$ diluted solutions were arranged by diluting the standard solutions in methanol. The nine standards used to build the calibration curve were prepared as follows: 70, 110, 150, 190, 230, 270, 310, 350 and 390 μl of the diluted solution (3 mg $l^{-1}$) were evaporated to dryness under nitrogen stream at 40 °C. The dried residues were then dissolved with 200 μl of *tert*-butyl methyl ether obtaining working standard solutions at concentrations 1.1, 1.7, 2.3, 2.9, 3.5, 4.1, 4.7, 5.4 and 6.0 mg $l^{-1}$ of each analyte.

Each standard was injected twice (Section 3.3) so the data set recorded each day consists of 18 samples. Nine samples (from 1.1 to 6.0 mg $l^{-1}$) were used as the calibration data set (C) and the other nine replicates (from 1.1 to 6.0 mg $l^{-1}$) as the validation data set (V). Working standards were arranged and measured on three non-consecutive days (named as days 1–3). On a different day (day 4) nine standards (T) were prepared in the concentration range mentioned above and recorded once. There will consequently be three calibration sets (labelled C1, C2 and C3), three validation sets (V1, V2 and V3) and a nine-standard test set (T).

In order to include in the models the variability due to the registration of the signals on different days, signals recorded on 2 days were grouped into a global calibration set (there are three possible combinations for assembling 2-day signals which will be named C12, C13 and C23) and in a global validation set (V12, V13 and V23). Signals recorded on all 3 days were joined in a global calibration set (C123) and in a validation set (V123).

### 3.3. Instrumental analysis

Analyses were performed with the HP 5973N gas chromatograph from Hewlett Packard. The separation was achieved with the J&W DB-17 ms GC Column from Agilent (bonded phase, 50% phenyl–50% dimethyl arylene siloxane, 0.25 μm film thickness) with dimensions 30 m × 0.25 mm i.d. Injections were made in the splitless mode with 10 min of solvent delay and using helium as carrier gas at 2.6 ml $min^{-1}$. The injector was kept at 250 °C and the detector at 290 °C. The oven temperature was programmed as follows: the initial temperature was set at 140 °C for 2 min, increased from 140 to 260 °C at 30 °C $min^{-1}$, kept at 260 °C for 4 min, subsequently raised to 300 °C at 30 °C $min^{-1}$ and held for 6 min. The oven equilibration time was set at 0.50 min. Sample injection volume was 2 μl. Analyses were carried out in the electron impact (EI) ionization mode at 70 eV operating in the selected ion monitoring mode (SIM). Two groups of ions were registered. Group 1 (DES, HEX and DIEN) start time was 10 min and the fragments (14 ions) recorded were 107, 120, 121, 135, 145, 173, 210, 237, 239, 251, 253, 266, 268 and 270. Group 2 ($E_2$ and $EE_2$) starts at 14.85 min and the 15 ions registered were 124, 133, 145, 146, 160, 172, 202, 213, 228, 229, 245, 272, 284, 296 and 302. In both groups the dwell time per ion was 100 ms.

### 3.4. GC/MS data, internal standardization and calibration models

Internal standardization was performed with one of the hormones analysed. As the concentration of the internal standard is not the same in all the standards, its signal must be previously normalized by the concentration in the standard (Section 3.2). Faults in the preparation of the internal standard were proved to be not significant by performing the total least squares regression (TLS) [23] which handles errors in both the independent and the dependent variables. Similar results to those of least squares were obtained and the errors in the preparation of the internal standard can be consequently considered non-significant.

Univariate models were built with the area recorded at the following ions 268, 135, 266, 272 and 213 for DES, HEX, DIEN, $E_2$ and $EE_2$, respectively. Internal standardization was carried out by dividing the area of each peak between that of $EE_2$ (normalized by its concentration in the standard) recorded at 213. The area of $EE_2$ was standardized with the area of DIEN at 266.

The original chromatogram was divided into the five peaks corresponding to each analyte so that each hormone can be independently analysed. The dimensions of the GC/MS matrices are (8 × 14), (10 × 14), (10 × 14), (8 × 15) and (16 × 15) for DES, HEX, DIEN, $E_2$, and $EE_2$, respectively. The first dimension refers to the chromatographic profile (number of scans or of elution times recorded in each peak). The second dimension refers to the mass spectra (number of *m*/*z* fragments). For simplicity all the results reported in the paper are expressed in terms of scan number rather than elution time. Scans were numbered from 10 min (scan 1). Data corresponding to each standard are grouped into the third dimension of the tensor (scan × mass spectra × sample) which will have 27 objects (samples) for 1-day models, 45 standards for 2-day models and 63 objects for 3-day models.

Calibration with PARAFAC2 models consists of the following steps: (i) a set of representative calibration samples are collected which span all expected sources of variation; (ii) data decomposition according to Eq. (1); (iii) selection of the proper number of factors; (iv) standardization of the sample scores of the factor related to the analyte with the

scores of the internal standard; (v) build a univariate regression between the standardized scores and the concentration of the standards in the training set; (vi) apply this regression to quantify new samples (validation and test set). The first three stages refer to PARAFAC2 which is a decomposition model whereas the other ones refer to the calibration step. The advantage of using a decomposition model (PARAFAC2) over a regression model (for example, PLS) is that in the first one not only the calibration set (C) but also the validation (V) and the test (T) sets intervene in the phase of data decomposition (steps (i)–(iii)), so that the additional variability from T will be taken into account.

Several initialization methods and convergence criteria were checked to perform PARAFAC2. Similar results were obtained in all cases which prove the robustness and validity of the PARAFAC2 models. In other words changes in the modeling procedure will not affect the conclusions derived from the fitted model. The results presented here were obtained by applying the ALS algorithm without restrictions in any mode in combination with a line search every fifth iteration. All PARAFAC2 models were carried out with one factor.

Internal standardization of the second-order signal was performed by dividing the scores of the analyte by the normalized scores of $EE_2$ (scores from $EE_2$ divided by the concentration in the standard). $EE_2$ scores were standardized with DIEN scores.

### 3.5. Software

PARAFAC2 models (Section 3.4) were built with the PLS_Toolbox 3.0 for use with MATLAB® (Version 6.1). The univariate models (standardized area versus concentration or standardized scores versus concentration) were done with PROGRESS [24] which performs the least median squares (LMS) regression. Those data with a LMS standardised residual in absolute value greater than 2.5 deviate from the linearity and are removed from the calibration set.

The capability of detection, CCβ ($X_0 = 0$), of the three-way models was determined [25] with NWAYDET, available from the authors.

## 4. Results and discussion

Chromatograms of all five analytes are displayed in Fig. 1 at the following mass fragments ($m/z$): 268, 135, 266, 272 and 213 for DES, HEX, DIEN, $E_2$ and $EE_2$, respectively. Chromatograms of dienestrol (DIEN, third analyte eluting), are amplified to show the inter-day variability from the 4 days. There are variations in intensity, the retention times, the shape of the peaks, etc. The signal instability observed will be consequently transmitted to the calibration models and to the estimations derived from them.

In this paper the robustness of the models is guaranteed through the mean of the relative errors in absolute value for

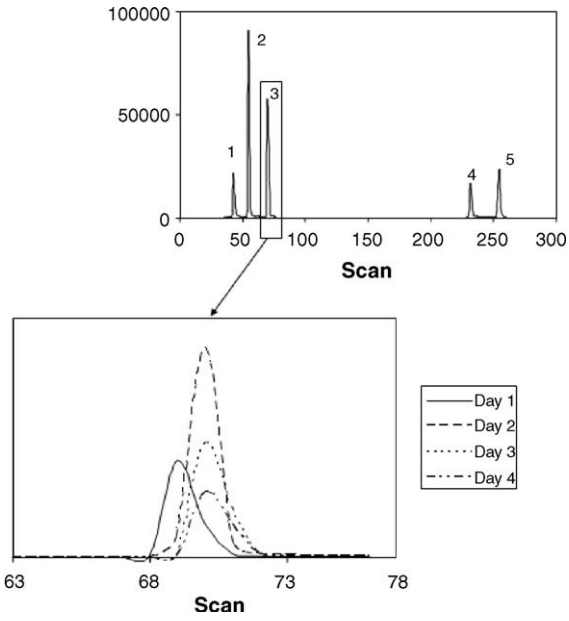


Fig. 1. Chromatogram of a standard containing diethylstilbestrol (1), hexestrol (2), dienestrol (3), 17-β-estradiol (4) and 17-α-ethynylestradiol (5). The chromatograms of dienestrol (3) registered on four different days have been enlarged to display the between day variability.

quantifying samples recorded on day 4 (T set did not intervene in the calibration step). The objective is to build models with small T errors, precise predictions (low standard deviations from the T set) as well as similarity between T and V errors (prediction of samples recorded on the same day as the calibration set).

Although the results obtained with both models are detailed in tables, most of the conclusions discussed here refer to PARAFAC2 due to the novelty of its standardization. Nevertheless differences between them are clearly stated.

### 4.1. Performance of the models without internal standardization

The performance of both univariate and PARAFAC2 models was firstly examined with non-standardized signals, that is calibration models were built with the peak area or the scores obtained from the PARAFAC2 decomposition. The results of the PARAFAC2 analysis are detailed for HEX and DIEN in Table 1 and are comparable to those obtained with univariate models (summed up in Table 2) because signals are specific. The same trend was observed for the rest of the compounds and their mean together with the standard deviation from the three 1-day models, the three 2-day models and the 3-day model are listed in Table 2. Among all five compounds, HEX is the most favourable case because T errors are not only the smallest but also the most precise. For the rest of the compounds errors are high and irreproducible so 1-day models without standardization cannot be used for quantifying standards recorded on a different day to the calibration set.

Table 1
Mean of the relative errors in absolute value from the validation (V) and the test (T) sets. PARAFAC2 models built with non-standardized signals were used for quantifying the concentration of hexestrol (HEX) and dienestrol (DIEN)

| Calibration day | HEX | | DIEN | |
|---|---|---|---|---|
| | V error (%) | T error (%) | V error (%) | T error (%) |
| 1 | 11.27 | 7.39 | 12.78 | 9.65 |
| 2 | 15.63 | 12.85 | 8.69 | 25.07 |
| 3 | 12.14 | 18.42 | 7.41 | 17.79 |
| 1 + 2 | 10.84 | 11.17 | 21.43 | 15.82 |
| 1 + 3 | 11.39 | 11.18 | 13.52 | 15.08 |
| 2 + 3 | 13.43 | 16.15 | 13.80 | 13.03 |
| 1 + 2 + 3 | 9.57 | 12.51 | 15.72 | 14.16 |

To solve this problem both univariate and PARAFAC2 global models were performed with signals recorded on different days. The validation of the univariate global calibration models conclude that residuals do not comply with the condition of normality [26] which prevents their application from quantifying future samples. Despite having normal residues and including several sources of variability, PARAFAC2 models do not succeed in decreasing prediction errors. It can be observed in Table 2 that T mean errors from the 2-day models are similar to those found for 1-day models. However, updating models has an advantage; predictions of the T set become more stable and the standard deviation decreases. The improvement of the precision is considerable for most of the analytes but above all for $E_2$, for which the standard deviation diminishes from 11.46% to 3.36%.

Differences in the precision of the analytes might be due to the different stability of the mass spectra. HEX is the compound with the most stable results not only in prediction but also in the detection limit. The standard deviation of the detection limit estimated by PARAFAC2 from C12, C13 and C23

Table 2
Mean and standard deviation (S.D.) of the relative errors in absolute value from the validation (V) and the test (T) sets from the three 1-day models, the three 2-day models and the 3-day model. PARAFAC2 models were performed with non-standardized signals

| Calibration | Analyte | V error (%) | | T error (%) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 1 day | DES | 6.18 | 1.44 | 20.21 | 9.99 |
| | HEX | 13.03 | 2.31 | 12.89 | 5.52 |
| | DIEN | 9.63 | 2.80 | 17.50 | 7.71 |
| | $E_2$ | 8.56 | 2.96 | 21.22 | 11.46 |
| | $EE_2$ | 7.07 | 1.87 | 18.52 | 13.44 |
| 2 days | DES | 15.85 | 7.67 | 13.82 | 9.15 |
| | HEX | 11.89 | 1.36 | 12.83 | 2.87 |
| | DIEN | 16.25 | 4.49 | 17.97 | 4.39 |
| | $E_2$ | 15.73 | 5.88 | 17.00 | 3.36 |
| | $EE_2$ | 15.47 | 6.41 | 18.75 | 4.93 |
| 3 days | DES | 11.94 | – | 3.80 | – |
| | HEX | 9.57 | – | 12.51 | – |
| | DIEN | 15.72 | – | 14.16 | – |
| | $E_2$ | 13.96 | – | 19.31 | – |
| | $EE_2$ | 16.87 | – | 16.69 | – |

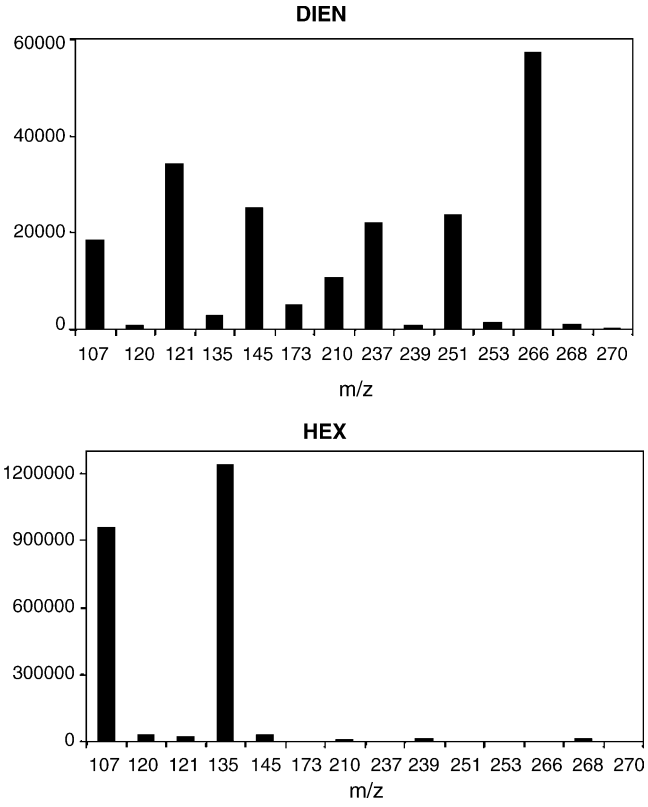


Fig. 2. Mass spectra of dienestrol (DIEN) and hexestrol (HEX) registered in the SIM mode and the same ionization conditions.

is 0.08 mg $l^{-1}$ of HEX whereas that of DIEN is 0.51 mg $l^{-1}$. Differences found between both compounds might be related to their fragmentation pattern shown in Fig. 2. Mass spectra of HEX have two major fragments (107 and 135) and several minor fragments. However, DIEN fragments to a greater extent which might be less reproducible over time. This property could be used for improving the stability of the models over time by selecting the proper mass fragments or even the ionization conditions (ion sources, electron energy, kind of derivatives, etc.) which makes the fragmentation pattern more robust over time.

### 4.2. Calibration models built with standardized signals

#### 4.2.1. One-day calibration models

Most of the changes in the sensitivity of the peak area of GC/MS can be corrected by using an internal standard. The main difficulty for standardizing multivariate signals is that the sensitivity of each mass fragment is affected in a different way and one ion cannot be used to correct the whole spectra. In this work we propose the standardization of the scores with the scores of the internal standard.

Despite the fact that several internal standards were tested for standardizing HEX signals, prediction errors, not presented in this paper, were worse than those displayed in Table 1. As stated in Section 4.1 the fragmentation pattern of HEX is stable and dividing its signal by that of the internal standard leads to poorer results.

Table 3
Mean of the relative errors in absolute value for predicting the concentration of the validation (V) and the test (T) sets of dienestrol (DIEN) and 17-α-ethynylestradiol ($EE_2$)

| Calibration day | DIEN | | | | $EE_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Univariate | | PARAFAC2 | | Univariate | | PARAFAC2 | |
| | V error (%) | T error (%) | V error (%) | T error (%) | V error (%) | T error (%) | V error (%) | T error (%) |
| 1 | 5.70 | 13.81 | 3.23 | 6.46 | 4.08 | 7.69 | 3.48 | 6.19 |
| 2 | 6.46 | 9.54 | 7.85 | 8.35 | 4.13 | 6.62 | 7.00 | 8.65 |
| 3 | 5.83 | 9.33 | 5.41 | 6.19 | 5.48 | 8.76 | 4.67 | 6.48 |
| 1 + 2 | 7.24 | 10.62 | 6.56 | 8.93 | 4.53 | 6.86 | 5.29 | 7.76 |
| 1 + 3 | 6.50 | 9.11 | 5.67 | 5.93 | 6.39 | 6.40 | 5.24 | 5.49 |
| 2 + 3 | 6.82 | 8.90 | 7.18 | 7.01 | 5.57 | 6.67 | 6.70 | 6.89 |
| 1 + 2 + 3 | 6.52 | 9.11 | 6.88 | 7.53 | 5.13 | 6.42 | 6.46 | 6.34 |

Univariate and PARAFAC2 models were performed with standardized signals.

The mean correlation coefficient from the three univariate 1-day models (C1, C2 and C3) is 0.94, 0.98, 0.99 and 0.99, whereas for PARAFAC2 is 0.98, 0.99, 0.99 and 0.99 for DES, DIEN, $E_2$ and $EE_2$, respectively. As models have correlation coefficients greater than 0.9, they explain most of the variability included in the standardized data.

Regarding the prediction ability of 1-day models, errors are listed in the first three rows of Table 3 for DIEN and $EE_2$ and of Table 4 for DES and $E_2$. The greatest differences between standardized and non-standardized signals have been found for DIEN and $EE_2$. Differences between V (predictions on the same day) and T (predictions on another day) errors are smaller with standardization than without it. This means that most of the variability due to the change of day has been removed by standardization and that the models and their estimations become more accurate. Specifically for DIEN, the averages of V and T errors from PARAFAC2 performed with non-standardized signals (Table 1, calibration day 2) are 8.69% and 25.07%, respectively; with standardized scores (Table 3) the mean from V errors, 7.85%, remains comparable but that from T errors decreases up to 8.35%.

On the other hand, not only has the mean from T errors decreased but also the standard deviation (more similar results between different days). Basically it involves models being more accurate. The improvement can be graphically observed in Fig. 3 which displays T errors for all seven PARAFAC2 models performed with standardized and non-standardized signals. The conclusion reached is important because by using an adequate technique for data pre-processing, the trueness and the precision of the analytical procedure significantly increases.

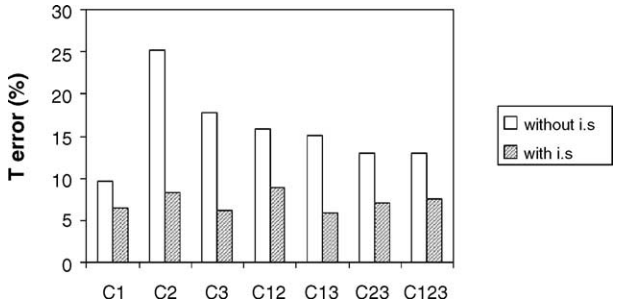


Fig. 3. Mean of the relative errors in absolute value for quantifying the test set (T) with 1-day (C1, C2 and C3), 2-day (C12, C13 and C23) and 3-day (C123) PARAFAC2 models. Comparison between non-standardized and standardized signals of dienestrol (DIEN).

Another aspect which has been revised is the sign of the relative errors to determine if the PARAFAC2 estimations are biased towards smaller or greater values. For the 1-day PARAFAC2 models as many positive errors have been

Table 4
Mean of the relative errors in absolute value for predicting the concentration of the validation (V) and the test (T) sets of diethylstilbestrol (DES) and 17-β-estradiol ($E_2$)

| Calibration day | DES | | | | $E_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Univariate | | PARAFAC2 | | Univariate | | PARAFAC2 | |
| | V error (%) | T error (%) | V error (%) | T error (%) | V error (%) | T error (%) | V error (%) | T error (%) |
| 1 | 9.27 | 8.32 | 3.64 | 4.68 | 8.69 | 34.25 | 3.54 | 8.90 |
| 2 | 10.67 | 7.94 | 8.23 | 13.04 | 4.01 | 19.74 | 4.29 | 33.10 |
| 3 | 10.29 | 10.21 | 5.92 | 4.80 | 5.27 | 10.35 | 3.78 | 2.67 |
| 1 + 2 | 9.14 | 8.73 | 10.56 | 14.25 | 5.07 | 29.05 | 12.98 | 18.42 |
| 1 + 3 | 10.69 | 8.22 | 15.04 | 5.29 | 10.96 | 12.34 | 5.93 | 5.73 |
| 2 + 3 | 10.81 | 9.90 | 10.31 | 15.08 | 11.24 | 13.62 | 14.04 | 14.53 |
| 1 + 2 + 3 | 8.54 | 8.65 | 12.96 | 10.74 | 10.95 | 14.19 | 8.84 | 9.76 |

Univariate and PARAFAC2 models were performed with standardized signals.

obtained as negative errors. For example, the $EE_2$ mean relative error in absolute value from C3 is 6.48% (Table 3) whereas the mean relative error is −0.93% and its median 1.03%. Both the mean and the median of the relative errors are statistically zero at a significance level of 0.05 and there is not a systematic error. In other words, the internal standardization has successfully corrected changes in the sensitivity of multivariate signals. This statement can be proved by comparing all three calibration curves (standardized scores versus concentration) which are $-0.12 + 1.07X$, $-0.00 + 0.99X$ and $-0.15 + 1.06X$ for C1, C2 and C3, respectively. Fixing the significance level at 0.05 all regression lines can be considered statistically equal which allows one to conclude that the stability of the standardized signals are extended to the calibration curves and consequently to the quantification of the T set over time. Because of taking into account the whole spectra of the internal standard, standardization of multivariate signals by means of the scores successfully corrects variations in the signals.

Results obtained for DES and $E_2$ (Table 4) were not so satisfactory in some cases. With regard to the univariate models built for $E_2$, T errors (34.25%, C1) are large compared with those of V errors (8.69%). This means that the model is valid for quantifying samples recorded on the same day as the calibration set (high correlation coefficient with small V errors) but the standardization of the area is not efficient enough to maintain accurate predictions over time. The calibration curves for the three 1-day models are $0.39 + 0.40X$, $0.18 + 0.52X$ and $0.62 + 0.46X$ for C1, C2 and C3, respectively. It can be observed that the model parameters change and standardization cannot handle alterations in the sensitivity of the $E_2$ signal.

Conclusions derived from univariate and PARAFAC2 models differ. The worst errors are obtained on the first day for univariate models and the second day for PARAFAC2. This implies that the same sources of variability do not equally affect univariate and three-way signals. The correction of univariate signals only takes into account one ion and the rest of the mass spectrum does not directly affect the results. If the $E_2$ ion is affected to the same extent by the changes in the experimental conditions as the fragment of the internal standard then the standardization will succeed in the correction (as happened to DIEN and $EE_2$). On the other hand the correction of the three-way signals was performed with the scores computed by PARAFAC2 which allows one to identify the factor related to the analyte and not to another sources of variability.

This conclusion was proved by standardising $E_2$ scores (computed from the three-way signal) with the peak area of $EE_2$ (univariate signal). The mean T error in absolute value from C2 is 45% which is worse than the error obtained with the signal standardised by the scores (Table 4, 33.10%). The reason for the worsening is that the signal of only one ion has been used to correct the full mass spectra.

And viceversa, univariate signals of $E_2$ were standardized with the scores of $EE_2$. The T mean error is 17% which is of the same magnitude as the error when the area of $E_2$ is standardized with the $EE_2$ area (Table 4, 19.74%). The standardization of the scores with the scores of the internal standard has been consequently proved to be more efficient than the standardization of the three-way signals with the univariate ones.

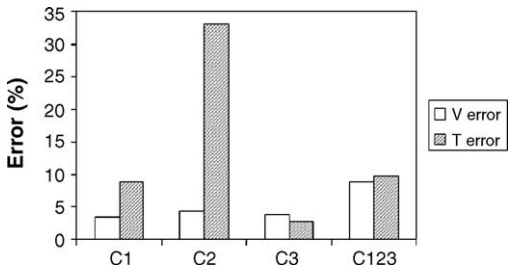


Fig. 4. Mean of the relative errors in absolute value for quantifying the validation (V) and the test (T) sets with 1-day (C1, C2 and C3) and 3-day (C123) PARAFAC2 models. 17-β-Estradiol ($E_2$) scores were standardized by those of 17-α-ethynulestradiol ($EE_2$).

The solution proposed here for minimising T errors is the use of global models.

### 4.2.2. Global calibration models

The results of updating models with the signals recorded on 2 and/or 3 days are listed in rows 4–7 of Tables 3 and 4. As internal standardization successfully corrects the changes for DIEN and $EE_2$ and global models do not significantly improve the outcome, in this section we will focus on their efficiency on DES and $E_2$ (Table 4). Specifically, the results achieved with the PARAFAC2 models of $E_2$ will be detailed.

It can be observed that global calibration models succeed in reducing T errors with respect to those of the second day (33.10%). The worst T error is 18.42% (C12) so the improvement is notable. It can also be observed that whenever signals recorded on the second day intervene in the model, T errors increase: 18.42 and 14.53% for C12 and C23, respectively. However, this effect is reduced by adding signals recorded on all 3 days (T error from C123, 9.76%). Fig. 4 shows both V and T errors estimated from 1- and 3-day models of $E_2$. T errors from C1 and C2 are significantly higher than V errors (prediction of samples recorded on the same day as the training set), that is, both models fail to quantify signals recorded on day 4. However, the updated model makes both errors not only equal but also more stable and therefore becomes more appropriate for keeping steady predictions. Additional variability not corrected by internal standardization can be modelled by global calibration models, thus maintaining steady and accurate predictions over time.

As is shown in Tables 3 and 4, similar results are obtained with univariate models. Internal standardization is effective for DIEN, $EE_2$ and DES whereas internal standardization combined with global models succeed in the $E_2$ case. This allows one to state that both strategies can be used to perform robust calibration models independently of the signal order.

## 5. Conclusions

It has been proved that data pre-processing is an important step in the performance of calibration models. An adequate technique, in this case internal standardization, might ameliorate and stabilize the estimations. The problem of standardizing three-way signals has been successfully solved by dividing the PARAFAC2 scores of the analyte between the scores calculated for the internal standard. In those cases in which internal standardization cannot deal completely with the changes in the sensitivity, updating calibration models by including signals recorded under several sources of variability has fixed the difficulty. The methodology here proposed worked for one-way (univariate) and three-way (PARAFAC2) models.

## Acknowledgments

This work has been partially supported by the Spanish Ministerio de Ciencia y Tecnología, DGI (BQU2003-07073) and Junta de Castilla y León (BU06/04). The authors acknowledge G. Tomasi and F. van den Berg for fruitful discussion. I. García thanks the Ministerio de Educación, Cultura y Deporte for the FPU Grant (Ref. AP2000-1314).

## References

[1] Amending Council Directive 96/22/EC concerning the prohibition on the use in stockfarming of certain substances having a hormonal or thyrostatic action and of β-agonist, Directive 2003/74/EC, Brussels, 2003.
[2] R.W. Giese, J. Chromatogr. A 1000 (2003) 401.
[3] M. Petrovic, E. Eljarrat, M.J. López de Alda, D. Barceló, J. Chromatogr. A 974 (2002) 23.
[4] L.C. Dickson, J.D. MacNeil, J. Reid, A.C.E. Fesser, J. AOAC Int. 86 (2003) 631.
[5] M.S. Díaz-Cruz, M.J. López de Alda, R. López, D. Barceló, J. Mass Spectrom. 38 (2003) 917.
[6] S. Nakamura, T.H. Sian, S. Daishima, J. Chromatogr. A 919 (2001) 275.
[7] M.H. Choi, K.R. Kim, B.C. Chung, Analyst 125 (2000) 711.
[8] T. Benijts, R. Dams, W. Günther, W. Lambert, A. de Leenheer, Rap. Commun. Mass Spectrom. 16 (2002) 1358.
[9] A. Laganà, A. Bacaloni, G. Fago, A. Marino, Rap. Commun. Mass Spectrom. 14 (2000) 401.
[10] Implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, 2002/657/EC Comission Decision, Brussels, 2002.
[11] I. García, M.C. Ortiz, L.A. Sarabia, C. Vilches, E. Gredilla, J. Chromatogr. A 992 (2003) 11.
[12] M.B. Sanz, L.A. Sarabia, A. Herrero, M.C. Ortiz, Talanta 56 (2002) 1039.
[13] J.L. Pérez, M. del Nogal, C. García, M.E. Fernández, B. Moreno, Anal. Chem. 75 (2003) 6361.
[14] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, J. Chemom. 13 (1999) 275.
[15] I. García, L. Sarabia, M.C. Ortiz, J.M. Aldama, Anal. Chim. Acta 515 (2004) 55.
[16] S. Macho, M.S. Larrechi, TRAC-Trend Anal. Chem. 21 (2002) 799.
[17] M.B. Sanz, L.A. Sarabia, A. Herrero, M.C. Ortiz, Electroanalysis 16 (2004) 748.
[18] H. Swierenga, A.P. de Weijer, R.J. van Wick, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 49 (1999) 1.
[19] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, Chemom. Intell. Lab. Syst. 64 (2002) 181.
[20] R. Bro, C.A. Andersson, H.A.L. Kiers, J. Chemom. 13 (1999) 295.
[21] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149.
[22] J.B. Harshman, UCLA Working Papers Phonet 22, 1972, p. 33.
[23] S. Van Huffel, J. Vandewalle, The Total Least Squares Problem: Computational Aspects and Analysis, SIAM, Philadelphia, PA, 1991.
[24] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[25] M.C. Ortiz, L.A. Sarabia, A. Herrero, M.S. Sánchez, B. Sanz, M.E. Rueda, D. Giménez, M.E. Meléndez, Chemom. Intell. Lab. Syst. 69 (2003) 21.
[26] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.