Microchemical Journal 109 (2013) 112–116

Contents lists available at SciVerse ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

A comparison of three procedures for robust PCA of experimental results of the homogeneity test of a new sodium diclofenac candidate certified reference material

Werickson Fortunato de Carvalho Rocha *, Raquel Nogueira, Gisele Estevão Baptista da Silva, Suzane Maio Queiroz, Gabriel Fonseca Sarmanho

National Institute of Metrology, Quality and Technology (Inmetro), Directorate of Industrial and Scientific Metrology, Chemical Metrology Division, 25250-020, Xerém, Duque de Caxias, RJ, Brazil

ARTICLE INFO

Article history:
Received 29 November 2011
Received in revised form 5 March 2012
Accepted 29 March 2012
Available online 5 April 2012

Keywords:
Certified reference materials (CRM)
Robustness
Principal components analysis (PCA)
Active pharmaceutical ingredients (API)
Sodium diclofenac

ABSTRACT

This paper compares classic and robust multivariate methods for evaluation of experimental results from the homogeneity test of a new sodium diclofenac candidate certified reference material (CRM). The results showed that the robust principal component analysis (PCA) based on projection pursuit was the most effective method for identification of outliers compared to the classic method and to the two other robust approaches: the ROBCA algorithm and the spherical PCA, when the concentrations of all active pharmaceutical ingredient (API) impurities were considered simultaneously. The PCA based on projection pursuit was able to identify six outliers, while the other methods identified only five. Through the use of these statistical tools, it was possible to reduce the value of the standard uncertainty due to between-bottle (in)homogeneity (u_{bb}) and to guarantee an accurate result of the combined standard uncertainty of the certified reference material (u_{CRM}).

1. Introduction

Exploratory data analysis and classification methods have always been important tools of multivariate data chemical analysis [1,2]. The application of these methods has expanded in recent years due to, among other things, an increased emphasis on high throughput chemical analysis, where researchers are often interested in differentiating among different chemical states of materials.

Exploratory methods, such as principal components analysis (PCA) or hierarchical cluster analysis (HCA), are unsupervised, so any class separation that is observed is likely to be real [3]. PCA has dominated as a method to visualize high dimensional data in lower dimensional spaces. In the classic PCA approach [4], the first principal component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the data points projected on it. If this procedure is further used, all the principal components can be obtained, which correspond to the eigenvectors of the empirical covariance matrix. Unfortunately, both the classic variance (which is being maximized) and the classic covariance matrix (which is being decomposed) are very sensitive to anomalous observations. Consequently, the first components are often attracted toward outlying points and thus may not capture the variation of the regular observations. Therefore, data reduction based on classic PCA becomes unreliable if outliers are present in the data.

This problem can be circumvented through robust PCA methods which are based on robust statistical analysis [4]. The goal of robust PCA methods is to obtain principal components that are not much influenced by outliers. A first group of methods is obtained by replacing the classic covariance matrix by a robust covariance estimator, such as the reweighed minimum covariance determinant (MCD) estimator [5]. The second approach to robust PCA uses projection pursuit (PP) and calculates directly the robust estimates of the eigenvalues and eigenvectors without passing by robust covariance estimation [6–8].

The third approach to robust PCA has been proposed by Hubert et al. [9] and is called ROBPCA. This method combines ideas of both projection pursuit and robust covariance estimation. The projection pursuit part is used for the initial dimensional reduction. Some ideas based on the MCD estimator are then applied to this lower dimensional data space.

Another approach to robust PCA is the spherical principal components procedure, which was first proposed by Locantore et al. [10] as a method for functional data analysis. The idea is to perform classic PCA on the data, projected onto a unit sphere. For details about spherical principal components see references [11,12].

Exploratory methods are being used in chemical metrology to evaluate different types of results. Lima et al. [13], for instance, used exploratory methods for homogeneity evaluation of a wheat flour reference material. Gonçalves et al. [14] evaluated different pH glass electrodes by an interlaboratory comparison using PCA and HCA, while Rocha and

* Corresponding author at: Instituto Nacional de Metrologia, Qualidade e Tecnologia (Inmetro), Diretoria de Metrologia Científica, Divisão de Metrologia Química, Av. Nossa Senhora das Graças 50, Xerém, Duque de Caxias, RJ, 25250-020, Brazil. Tel./fax: +55 2126799069.
E-mail address: wfrocha@inmetro.gov.br (W.F.C. Rocha).

0026-265X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.microc.2012.03.028

Nogueira [15] used PCA and HCA to evaluate outliers in homogeneity test results for certification of two pharmaceutical reference materials.

Certified reference materials (CRM) are widely used in analytical chemistry, especially for equipment calibration, assessment of measurement methods, and establishment of metrological traceability of measurement results [16]. The certification of a candidate CRM is carried out according to the ISO Guide 30 series [16–21]. As stated in the ISO Guide 35:2006 [16], all uncertainty sources related to the certified property value have to be estimated, including characterization uncertainty, uncertainty due to between-bottle (in)homogeneity (u_{bb}), and uncertainties due to short- and long-term stability studies (u_{sts} and u_{lts}), which are then combined according to the law of propagation of uncertainties to obtain the combined standard uncertainty of the CRM property value (u_{CRM}). The homogeneity evaluation is carried out not only between different units of the candidate CRM batch, but also within the units [16].

In this paper, we compare three procedures for robust PCA with the classic PCA used in chemical metrology for an easy and quick evaluation of homogeneity test results of a new sodium diclofenac candidate CRM, followed by selection of the replicates to be considered for estimation of the uncertainty due to between-bottle (in)homogeneity (u_{bb}).

2. Material, methods and experimental design

2.1. Instrumentation, methods, and reagents

A high performance liquid chromatography (HPLC) system (Shimadzu, Kyoto, Japan), consisting of a LC-20AT quaternary pump, a DGU-20A$_3$/DGU-20A$_5$ on-line degasser, a SIL-20A/20AC auto-sampler, a SPD-20A photodiode array detector, and a CBM-20A/20A interface, was used for determination of organic impurities (related substances).

The following experimental conditions were used for the sodium diclofenac HPLC analysis, based on the Brazilian Pharmacopeia IV [22] monograph: column C8, 15 cm × 4.6 mm i.d., 5 μm, Phenomenex Luna (Phenomenex, Torrance, CA, USA), pre-column Phenomenex Security Guard C8, 4 × 3 mm, 5 μm, mobile phase methanol–0.005 mol L^{-1} potassium dihydrogen phosphate pH 2.5 (70:30, v/v), flow-rate 1.0 mL min^{-1}, UV detection at 254 nm, injection volume 50 μL.

The organic impurities mass fraction (w_{org}) was determined by HPLC peak area ratio, after HPLC analysis of solutions containing 750 μg g^{-1} and 7.5 μg g^{-1} of sodium diclofenac, prepared gravimetrically using an analytical balance (Shimadzu), model AUW 220D, with resolution of 0.01 mg. Each solution was prepared in triplicate and injected three times into the HPLC system.

The volatiles mass fraction (w_{vol}) was estimated by loss on drying at 105 °C/3 h, while the inorganic impurities mass fraction (w_{inorg}) was determined using an Elan DRC ICP-MS (Perkin Elmer, Waltham, MA, USA), after microwave digestion.

The studies were carried out using the Inmetro first batch of sodium diclofenac candidate CRM. The reagents were HPLC grade methanol (Tedia, Rio de Janeiro, Brazil), *o*-phosphoric acid 85% w/w p.a. (Merck, São Paulo, Brazil), and potassium dihydrogen phosphate p.a. (Sigma Aldrich, St. Louis, USA).

2.2. Certification of candidate reference materials

The certification of the sodium diclofenac candidate CRM was carried out according to the ISO Guides 34:2009 [22] and 35:2006 [16]. Briefly, the certification involved the material characterization, homogeneity testing, short- and long-term stability studies, and uncertainties estimation. The mass fraction of the active pharmaceutical ingredient (w_{API}) of the CRM, expressed in g 100 g^{-1}, was determined by mass balance according to Eq. (1).

$$w_{API} = 100 - \sum w_{org,i} - \sum w_{inorg,i} - \sum w_{vol,i} \quad (1)$$

where $\sum w_{org,i}$, $\sum w_{inorg,i}$ and $\sum w_{vol,i}$ are the respective sums of mass fractions of organic, inorganic, and volatiles, expressed in g 100 g^{-1}.

2.3. Homogeneity evaluation according to the ISO Guide 35:2006

The homogeneity test was carried out using 17 flasks of the candidate CRM selected at random and analyzed as described above. Then, the single-factor analysis of variance (ANOVA) of measurement results was carried out to evaluate the homogeneity. According to the ISO Guide 35:2006 [16], the uncertainties due to between-bottle (in)homogeneity (u_{bb}) can be calculated by Eq. (2) or (3):

$$u_{bb} = \sqrt{(MS_{between} - MS_{within})/n_0} \quad (2)$$

$$u_{bb} = \sqrt{MS_{within}/n_0} \cdot \sqrt[4]{2/(\nu MS_{within})} \quad (3)$$

where $MS_{between}$ is the mean square between CRM batch units (ANOVA), MS_{within} is the mean square within CRM batch units (ANOVA), and n_0 is the number of replicates. If $MS_{between}$ is larger than MS_{within}, either Eq. (2) is used, or both equations are applied and the largest u_{bb} value is chosen. On the other hand, if $MS_{between}$ is smaller than MS_{within}, which indicates poor repeatability of the measurement method, only Eq. (3) can be used.

2.4. Homogeneity evaluation according to the exploratory analysis

From the homogeneity results, one matrix (153 × 6) was constructed for the new sodium diclofenac candidate CRM. The total number of rows corresponded to the number of results: 153 (17 flasks, three replicates of each flask, three HPLC injections of each replicate). The columns corresponded to the mass fractions of each of the six organic impurities observed, expressed in g 100 g^{-1}.

Initially, the results were pretreated using autoscaling and mean centering. Both pretreated and not pretreated results were then submitted to the multivariate statistical analysis methods of principal component analysis (PCA) using the statistical software environment R, freely available at http://cran.r-project.org/ using the package chemometrics of Filzmoser and Varmuza [23] and rrcov of Todorov [24], which can also be downloaded from the same web page.

3. Results and discussion

The single-factor analysis of variance (ANOVA) of homogeneity test results (sodium diclofenac mass fraction w_{API} obtained by mass balance) was carried out, and the following results were obtained: $MS_{between} = 1.779673 \times 10^{-4}$ (g 100 g^{-1})^2 and $MS_{within} = 2.185600 \times 10^{-5}$ (g 100 g^{-1})^2. Considering the number of replicates (n) equal to 9, the u_{bb} value calculated by Eq. (2) was 0.0041666 g 100 g^{-1}.

The contribution of the (in)homogeneity (C) to the studied material was calculated by Eq. (4):

$$C(\%) = u_{bb} x 100/\bar{x} \quad (4)$$

Where $\bar{x}$ is the average of all homogeneity test results.

The C value indicates the impact of the homogeneity uncertainty in the value of the studied property value. For the sodium diclofenac candidate CRM, the C value was estimated as 0.0041666%, indicating a good agreement between the flask results. The homogeneity of a reference material batch is one of the certification requirements according to the ISO Guide 35 [16]. Therefore, the degree of (in)homogeneity of the material has to be well known. Exploratory methods were further used to better evaluate and understand the homogeneity results.

Experimental results usually suffer any type of pretreatment prior to using exploratory methods. The best pretreatment methods are those that ultimately produce a robust model with the most accurate predictive ability [25]. For the construction of exploratory methods,