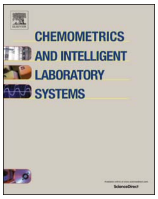
Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



A scoring metric for multivariate data for reproducibility analysis using chemometric methods

David A. Sheen[a,*], Werickson F.C. Rocha[b], Katrice A. Lippa[a], Daniel W. Bearden[c]

[a] Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[b] National Institute of Metrology, Quality and Technology -INMETRO, Division of Chemical Metrology, 25250-020 Duque de Caxias, RJ, Brazil
[c] Chemical Sciences Division, National Institute of Standards and Technology, Hollings Marine Laboratory, 331 Fort Johnson Road, Charleston, SC 29412, USA

ARTICLE INFO

Keywords:
Nuclear magnetic resonance
Metabolomics
Interlaboratory comparison

ABSTRACT

Process quality control and reproducibility in emerging measurement fields such as metabolomics is normally assured by interlaboratory comparison testing. As a part of this testing process, spectral features from a spectroscopic method such as nuclear magnetic resonance (NMR) spectroscopy are attributed to particular analytes within a mixture, and it is the metabolite concentrations that are returned for comparison between laboratories. However, data quality may also be assessed directly by using binned spectral data before the time-consuming identification and quantification. Use of the binned spectra has some advantages, including preserving information about trace constituents and enabling identification of process difficulties. In this paper, we demonstrate the use of binned NMR spectra to conduct a detailed interlaboratory comparison. Spectra of synthetic and biologically-obtained metabolite mixtures, taken from a previous interlaboratory study, are compared with cluster analysis using a variety of distance and entropy metrics. The individual measurements are then evaluated based on where they fall within their clusters, and a laboratory-level scoring metric is developed, which provides an assessment of each laboratory's individual performance.

1. Introduction

Chemometrics is a field that refers to the application of a wide range of statistical and mathematical methods, including multivariate methods, to problems of chemical origin [1–3]. With the advance of analytical instrumentation in chemical metrology, an increasing amount of data can be generated which requires multiple approaches for extracting reliable information. This has required and enabled the development of improved analytical procedures for data analysis based on sound chemometric principles in order to reliably assess the properties of interest in a system under study.

In recent years, the importance of metrology in the world has grown significantly since its main focus is to provide reliability, credibility, universality and quality measurements. Since measurements are essential, directly or indirectly, in virtually all decision-making processes, the scope of metrology is immense, involving important areas of society such as industry, trade, health, safety, defense and the environment. It is estimated that about 4% to 6% of the gross domestic product of industrialized countries is dedicated to measurement [4]. In this context, the use of chemometrics in combination with metrology is a potential approach to the interpretation of data in decision making, providing improved industrial and technological development. One of the important metrological activities that can be highlighted is the participation and organization of interlaboratory quality assurance programs. Quality assurance includes interlaboratory studies used as an external evaluation tool and in the demonstration of the reliability of laboratory analytical results. It also serves to identify gaps in the analytical process and enable comparability improvement. Moreover, it is one of the items required for accreditation tests by ISO/IEC 17025: 2005 [5].

According to ISO 13528: 2015 [6] and ISO/IEC 17043 [7], there are several statistical tools to be used to assess the results of analytical laboratories participating in proficiency testing. Among them there are the Z-scores, Z'-scores, Zeta scores and E_n scores. The problem with these metrics is related to the fact that they can only be used for cases of univariate measurement results and have not been systematically extended to multivariate analyses. However, some studies have demonstrated efforts to analyze the quality of multivariate data. An example of this is in the field of metabolomics [8] in which principal components analysis (PCA) was used to evaluate data from an interlaboratory comparison. In other work [9], a metric called *Qp*-score is proposed to evaluate the performance of each laboratory for multivariate data. Other than these studies, there have been few attempts to perform interlaboratory comparisons on multivariate data. But even these

[*] Corresponding author.

http://dx.doi.org/10.1016/j.chemolab.2016.12.010
Received 5 August 2016; Received in revised form 15 November 2016; Accepted 20 December 2016
Available online 23 December 2016
0169-7439/ Published by Elsevier B.V.

studies have used spectral data to measure some property or set of properties and then determined standard univariate scores for these measurements. For instance, in Gallo et al. [9], the participating laboratories determined calibration curves for metabolite concentration with respect to nuclear magnetic resonance spectroscopy (NMR) spectral feature intensity and then determined a score from those curves. In Viant et al. [8], several significant features within each spectrum were identified and then univariate scores determined based on the intensities of those features. In each case, however, the scoring process reduces a vector of thousands of components to one possessing relatively few components, and it is still difficult to extract a comprehensive metric of "goodness" from this information.

Considering the lack of a multivariate metric in the ISO standards that address the subject, the objective of this study is to propose the application of algorithms already known in the literature that can be used for the evaluation of multivariate data in proficiency tests. We re-examine the data collected by Viant et al. [8], and extend their analysis by proposing a scoring metric that assumes a single value for each NMR spectrum, which can be further extended to a laboratory-level metric that can be used for quality control and analysis.

1.1. Brief statement of the interlaboratory comparison problem

The problem of laboratory-outlier detection in an intercomparison study can be expressed in the following way:

$$R(x,r,f) = T(x) + I(x,r,f) + \varepsilon, \tag{1}$$

where R is the measured response function, T is the underlying true value, I is some instrument function and epsilon is noise. R is a function of the experimental independent variables x (in this case, NMR chemical shifts and the static NMR field strength) but also depends on the replicate number r and the facility identifier f. This expression for R allows an explicit statement for how the response might vary based on the use of different measuring devices in different locations, and even run-to-run variability in the same device. The underlying truth T depends only on the independent variables, while I explicitly contains the variability among the measurements.

In the normal regression problem, I is treated as being part of the noise epsilon. For an interlaboratory study, however, the instrument function could actually contain a great deal of information about the individual laboratories that make the measurements.

The purpose of an interlaboratory comparison study is to identify those laboratories whose instrument function is sufficiently systematically different to indicate that those laboratories may be sampling from a different population. For instance, in the Viant et al. study [8], NMR spectra were taken at various magnetic field strengths. The individual spectra consist of magnetic field dependent features (chemical shifts) and magnetic field independent features (spin-spin couplings). As a result, spectra taken under different field strengths are not directly comparable. If the instrument function contains such systematic lab-to-lab variations, then, the performance of one laboratory relative to the others will be consistently different when compared across a range of many different values of the independent variables x, which in this case means many different samples.

It should be noted here that measurements taken of the same object at different laboratories by different analysts on different instruments are considered to be independent of each other, in the sense that there is no cross-communication between the different laboratories. Likewise, the measurements of different objects by the same laboratory will also be independent.

1.2. Multivariate metrics used for interlaboratory comparison

In chemometrics and information theory, there are several common metrics used for pattern recognition [10–16]. It is important to mention the Euclidean distance [17–20] and the Mahalanobis distance [21–26] as the most used, however, there are other metrics based on a probabilistic approach [27], for example, the Hellinger distance [28], the Kullback-Leibler divergence [29] and the Jensen-Shannon distance [30,31]. All these metrics may be used in metrological activities such as, for example, interlaboratory comparison.

1.2.1. Similarity measures based on vector distance: Euclidian and Mahalanobis distance

The Euclidean distance is defined by

$$d_{\mathrm{E}}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} = \left[\sum\nolimits_k (x_k - y_k)^2\right]^{1/2}, \tag{2}$$

where the $\mathbf{x}$ and $\mathbf{y}$ column vectors represent two spectra and x_k and y_k are the features of those spectra, with the sum taken over the elements of the vectors. The Euclidean distance gives greater weight to large differences between prominent features than to differences between small, but possibly clinically significant, features. However, it does not correct for correlation structures in the data. To resolve this issue, the Mahalanobis distance is often used, which is defined by

$$d_{\mathrm{M}}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})} \tag{3}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix which may be estimated in numerous ways.

1.2.2. Similarity measures based on probabilistic distance: Hellinger distance, Kullback-Leibler divergence, and the Jensen-Shannon distance

Alternatives to the Euclidean and Mahalanobis distances include metrics from information theory to analyze probability density functions. Interpreting an NMR spectrum in this way requires that the spectrum be non-negative and also that it integrate to unity. Under this interpretation, the spectrum indicates what fraction of the total oscillatory power is contained at each frequency. The metrics we discuss here are the Hellinger distance [28], the Kullback-Leibler (KL) divergence [29], and the Jensen-Shannon distance [30,32–34]. The Hellinger distance between two spectra is defined by

$$d_{\mathrm{H}}(\mathbf{x},\mathbf{y}) = 1 - \sum\nolimits_k \sqrt{x_k y_k} \tag{4}$$

and varies between 0 and 1. If $d_{\mathrm{H}} = 0$, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are identically equal, indicating similar performance of the two data sets from which $\boldsymbol{x}$ and $\boldsymbol{y}$ are obtained. If $d_{\mathrm{H}} = 1$, $\boldsymbol{x}$ is zero everywhere that $\boldsymbol{y}$ is positive and vice versa, indicating a divergence of the two data sets. In terms of an interlaboratory comparison, $d_{\mathrm{H}} \ll 1$corresponds to similar performance between laboratories, while $d_{\mathrm{H}} \approx 1$ represents divergence in the results of the laboratories.

The KL divergence, sometimes termed the relative entropy, is defined by

$$d_{\mathrm{KL}}(\mathbf{x},\mathbf{y}) = 1/\ln 2 \sum\nolimits_k x_k \ln\left(\frac{x_k}{y_k}\right). \tag{5}$$

The KL divergence is not symmetric, and so what is used here is the symmetrized KL (SKL) divergence, sometimes termed the Jeffreys divergence,

$$d_{\mathrm{SKL}}(\mathbf{x},\mathbf{y}) = d_{\mathrm{KL}}(\mathbf{x},\mathbf{y}) + d_{\mathrm{KL}}(\mathbf{y},\mathbf{x}). \tag{6}$$

Unlike the Hellinger distance, the SKL divergence varies between 0 and positive infinity, with positive infinity corresponding to divergence between the laboratories. Furthermore, if an element of $\mathbf{x}$ or $\mathbf{y}$ is zero anywhere where the other is nonzero, the SKL divergence will diverge to infinity.

The SKL divergence is not a distance metric because it does not satisfy the triangle inequality, so as an alternative we will also use the Jensen-Shannon (JS) distance [35–37], defined by

$$d_{\mathrm{JS}}(\mathbf{x},\mathbf{y}) = \sqrt{d_{\mathrm{KL}}(\mathbf{x},\mathbf{m}) + d_{\mathrm{KL}}(\mathbf{y},\mathbf{m})} \tag{7}$$

where $\mathbf{m}$ is the arithmetic mean of $\mathbf{x}$ and $\mathbf{y}$. Like the Hellinger distance,