Chemometrics and Intelligent Laboratory Systems 133 (2014) 33–41

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Evaluation of trends in residuals of multivariate calibration models by permutation test

Paulo R. Filgueiras [a], Júlio Cesar L. Alves [a], Cristina M.S. Sad [b], Eustáquio V.R. Castro [b], Júlio C.M. Dias [c], Ronei J. Poppi [a,*]

[a] Institute of Chemistry, University of Campinas — UNICAMP, P.O. Box 6154, 13083-970 Campinas, SP, Brazil
[b] Department of Chemistry, Federal University of Espírito Santo, Laboratory of Research and Development of Methodologies for Analysis of Oils, Av. Fernando Ferrari, 514, Goiabeiras, Vitória 29075-910, Espírito Santo, Brazil
[c] CENPES/PETROBRAS, Av. Jequitiba 950, Rio de Janeiro 21941-598, Brazil

ARTICLE INFO

Article history:
Received 2 December 2013
Received in revised form 6 February 2014
Accepted 11 February 2014
Available online 19 February 2014

Keywords:
Permutation test
Principal component regression
Partial least squares
Support vector regression (SVR)
Residuals in multivariate calibration

ABSTRACT

This paper proposes the use of a nonparametric permutation test to assess the presence of trends in the residuals of multivariate calibration models. The permutation test was applied to the residuals of models generated by principal component regression (PCR), partial least squares (PLS) regression and support vector regression (SVR). Three datasets of real cases were studied: the first dataset consisted of near-infrared spectra for animal fat biodiesel determination in binary blends, the second one consisted of attenuated total reflectance infrared spectra (ATR-FTIR) for the determination of kinematic viscosity in petroleum and the third one consisted of near infrared spectra for the determination of the flash point in diesel oil from an in-line blending optimizer system of a petroleum refinery. In all datasets, the residuals of the linear models presented trends that have been satisfactorily diagnosed by a permutation test. Additionally, it was verified that 500,000 permutations were enough to produce reliable test results.

1. Introduction

Residuals in multivariate calibration are estimates of the experimental error obtained by subtracting the reference from the predicted responses. The predicted response is calculated from the model after all regression parameters have been estimated from the experimental data. The residual is defined by Eq. (1):

$$e_i = y_i - \hat{y}_i \tag{1}$$

where $\hat{y}_i$ is the predicted value, y_i is the reference value and e_i is the residual for the sample i.

In correctly adjusted calibration models, it is expected that the residuals remain roughly uniform in size as the measured value increases and is normally distributed about zero. A poor fit of the model reflects trends in their residuals, where relevant information has not been incorporated into the model.

The presence of trends in regression residuals can be assessed using graphics of the residuals with the reference values. Fig. 1a provides an example of the residuals in a set of one hundred points in which no trends can be detected. It is possible to note that no systematic error is present when analyzing the distribution shown in Fig. 1b. The same dataset is presented in Fig. 1c, where a simple visual observation is enough to diagnose trends in the residuals. The difference between the datasets depicted in Fig. 1a and Fig. 1c is the order in which the residuals appear. A systematic residual test is not suitable for this comparison because the data appear to be symmetrically distributed around zero and the residuals follow a quadratic trend.

Due to the importance of residual analysis in multivariate regression models, this paper proposes a nonparametric permutation test [1,2] to assess the presence of trends in residuals. An advantage is that permutation tests are distribution-free, which means that no assumptions about normality or homoscedasticity are required. In recent years, permutation test methods have been increasingly applied to multivariate problems in analytical chemistry. These methods have been applied to identify significant effects in experimental designs [3], compare the predictive accuracy of different models [4] and conduct variable selection in multivariate calibration [5,6].

2. Methods and data

2.1. Permutation test

The permutation method is a repetitive reordering of N entries in the reference variable **y**. The elements in the original **y** variable are reordered, thereby creating new response variables just by switching

* Corresponding author. Tel.: +55 19 35213126; fax: +55 19 35213023.
E-mail address: ronei@iqm.unicamp.br (R.J. Poppi).

http://dx.doi.org/10.1016/j.chemolab.2014.02.002
0169-7439/© 2014 Elsevier B.V. All rights reserved.

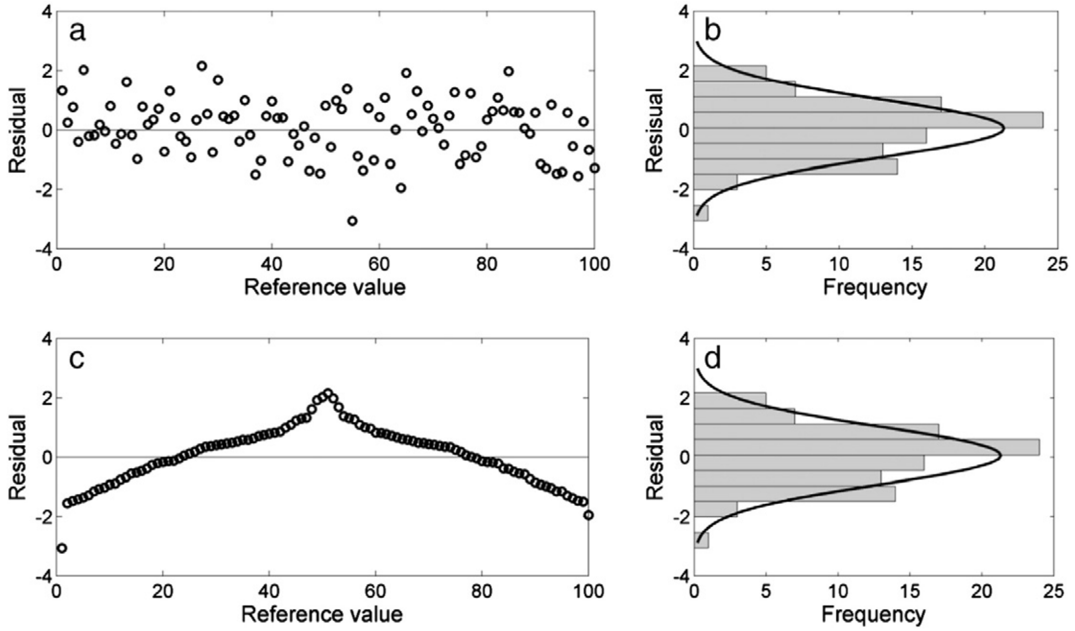

Fig. 1. Simulated residuals with a normal distribution, average of zero and unit variance: (a) no trends in residuals; (b) histogram of residuals without trend; (c) residuals presenting a quadratic trend; (d) histogram of residuals with trend.

their internal positions. The new permuted **y** variable should have no or very limited association with the predicted residual of the calibration [6]. Consider the following example: using only the numbers 1, 2 and 3, it is possible to form six sets by reordering the elements, and these sets will differ only in the order of their contents. The first is the orderly set $A = \{1, \ 2, \ 3\}$; the other five sets are $A_1^p = \{1, \ 3, \ 2\}$, $A_2^p = \{2, \ 1, \ 3\}$, $A_3^p = \{2, \ 3, \ 1\}$, $A_4^p = \{3, \ 1, \ 2\}$ and $A_5^p = \{3, \ 2, \ 1\}$. The possible number of permutations increases with the number of elements. The number of combinations is given by N! (where N is the number of elements). In a set with 4 elements, it is possible to obtain 24 permutations; with 10 samples, 32,628,800 permutations are possible.

Non-parametric tests have some advantages compared to parametric tests, such as the exemption of the assumption of data normality. Non-parametric tests are also simpler in execution, and the p-value is exact if the permutation is tested a reasonable number of times.

The non-parametric permutation test applied to evaluate trends in residuals is based on the randomization values of the vector **y** (variable with reference value) while keeping the order of **X** data constant (instrumental variables).

The first step is to define the hypotheses used to check for biased errors:

a) Null hypothesis or H_0: the prediction residuals e_i are independent of y_i;
b) Alternative hypothesis or H_1: the prediction residuals e_i are related with y_i according to the following equation:

$$e_i = g(y_i) + \varepsilon_i \tag{2}$$

where ε_i is an independent random error and $g(y_i)$ is some polynomial function that can model the relationship between the residuals and the reference values. It presents the following form:

$$e_i = b_n y_i^n + b_{n-1} y_i^{n-1} + \ldots + b_1 y_i + b_0. \tag{3}$$

In the alternative hypothesis, the dependence of the residuals with reference values is proposed, and it is assumed that the entire effect of randomness present in y_i is due only to the random variableε_i.

There will be evidence of trends in the residuals if the highest polynomial coefficient b_n in Eq. (3) is greater than zero if b_n is positive or less than zero if b_n has negative value at the significance level of probability assumed (we arbitrarily use the value of 0.05). For any equation $g(y_i)$, only the highest polynomial coefficient is tested. The algorithm is summarized as follows:

i. calculate the b_n polynomial coefficient from the adjustment of the original residuals in the function of the reference values. Here, it is called b_n^*.
ii. randomly permute only the vector **y**;
iii. calculate the b_n^k coefficient for the kth permuted **y**;
iv. compare b_n^* with b_n^k;
v. repeat steps (ii) to (iv) K times.

If the distribution of residuals is random around zero, the b_n^* coefficient belongs to the random distribution of b_n^i calculated from the permuted **y**. The p-value of the test is determined by the proportion of the number of times where $b_n^* > b_n^i$. When the p-value for the test is smaller than the level of significance adopted ($\alpha = 0.05$), there is no evidence to accept H_0, and the residuals are not random. Otherwise, H_0 is accepted, and there are no trends in the residuals. The Matlab code for the proposed permutation test algorithm is presented in Appendix A.

2.2. Multivariate calibration methods

The permutation test was applied in the residuals of three real-world datasets from three multivariate calibration methods: principal component regression (PCR) and partial least squares (PLS) regression, and support vector regression (SVR).

2.2.1. Principal component regression

The general form for the linear models (PCR and PLS) can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{4}$$

where **X** is the matrix of instrumental data (spectra), **b** is the vector of the regression coefficients and **y** is the vector of the reference values.