Chemometrics and Intelligent Laboratory Systems 107 (2011) 333–342

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Modified robust continuum regression by net analyte signal to improve prediction performance for data with outliers

Xiao-Yu Zhang, Qing-Bo Li, Guang-Jun Zhang *

Precision Opto-mechatronics Technology Key Laboratory of Education Ministry, School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 28 November 2010
Received in revised form 3 May 2011
Accepted 4 May 2011
Available online 10 May 2011

Keywords:
Modified robust continuum regression (mRCR)
Net analyte signal (NAS)
Outlying observation
Spectrometric quantization
Glucose concentration

ABSTRACT

Contaminated data exist in diverse situations, even in high quality surveys and experiments. If classical statistic models are blindly applied to data containing outliers, the results can be misleading at best. In this paper, a modified robust continuum regression (mRCR) method is proposed to improve prediction performance for data with outliers. The mRCR method constructs projection pursuit directions by using projection matrix for computing the net analyte signal (NAS) of the target analyte. This paper examines applications to the determination of glucose concentration by near-infrared (NIR) spectrometry, including aqueous solution with glucose experiment, plasma experiment in vitro, oral glucose tolerance test (OGTT) in vivo, to illustrate the advantages of mRCR for various kinds of outliers depending on the way of contamination. The results indicate that the mRCR method is entirely robust with respect to any type of outlying observations, and it yields smaller prediction errors for normal samples than other calibration methods.

1. Introduction

The probability that the data can be exactly originated from normal distribution is practically close to zero. The non-normal cases have been published in chemical literature: for example, in the immunogenicity assay, the assay response data for drug naive sera samples show a gamma distribution that is unimodal and skewed to the right [1]; due to the druglikeness features of launched drugs and clinical candidates, several distributions of physicochemical properties reported by the Oprea group obey the Pearson distribution that is asymmetry and fat-tailed [2]; the concentrations of components in light crude oil follow a log-normal distribution that is heavy-tailed [3]. Particularly, it is reasonable to expect that prevalent data would be contaminated with outliers inconsistent with the majority of observations and unlikely generated by the same model [4], because one often has less control over the execution of the experiment. Outliers incorporated into a classical multivariate calibration model can significantly degrade the performance of the model, since the classical multivariate linear regression is non-robust because of its extreme sensitivity to outliers in the data [5]. Many robust methods have been developed to offer protection against outliers. Most of the methods are based on robustness ideas and linear modeling. Although robustness is paid for by lower efficiency of the estimator and a higher computational effort, the resulting estimation will usually be more reliable for the data at hand [6]. Robust statistics has become a mainstay in any field of applied sciences. Chemometrics is no exception when regarding its vulnerability to possible outliers [7]. One might expect multivariate outlier detection technique prior to fitting non-robust model to solve this problem. However, not all the outliers can be detected. Classical tools are rarely able to detect all the multivariate outliers because they inherently cannot take into account the masking and swamping effects. Masking is the case where one or more outliers are incorrectly identified as normal samples because other outliers mask their presence; Swamping is the case where normal samples are made to appear to be outliers [8–10]. Outliers due to the change in the predictor factor cannot be detected because the difference in response level introduced by the predictor factor change is not detectable [9]. Outliers might be located in the range of the measurement noise, which makes them hard to be identified [11]. Moreover, the dimensionality increase makes the detection of outliers in complex data considerably more difficult [10,12]. Additionally, it is better to reserve the outlying observations when the sample size is very small. Therefore it is significant that the quantitative calibration model with high precision and good robustness can still be established for data with outliers.

In principal components regression (PCR), the latent variables are linear combinations of the predictor variables having maximal variance [13]. Partial least squares (PLS) constructs latent variables maximizing their covariance with the predictand [14]. A joint maximization criterion called continuum regression (CR) allows for a better fit and more predictive power by finding the optimal tuning parameter in a continuum range of models from OLS (ordinary least

* Corresponding author. Tel.: +86 10 82337787; fax: +86 10 82339671.
E-mail address: xiaoyuzhangzi@126.com (G.-J. Zhang).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.05.003

squares) over PLS to PCR [15]. In CR, the maximization criterion is approximated by a method called continuum power regression (CPR) [16]. Sundry practical applications in varying fields of science have shown that the application of CR introduced by Stone and Brooks [15] to non-contaminated data significantly improves prediction. Conversely, for contaminated data, CR may yield aberrant estimates due to its non-robustness to outliers. This drawback of CR has been heeded by Serneels et al. [17] who proposed robust continuum regression (RCR). RCR that robustifies the criterion which defines CR by using robust estimators of variance and covariance is a valuable alternative for the existing robust estimation methods. Whereas for RCR, only a limited number of projection pursuit (PP) [18] directions can be considered, so that the final solution obtained is only an approximation [17]. And RCR is time consuming. Additionally, RCR has randomicity. Accordingly, a modified robust continuum regression (mRCR) that constructs PP directions by using projection matrix for computing the net analyte signal (NAS) of the target analyte is proposed in this paper. It inherits the robustness properties of RCR and gives protection against various kinds of outliers, such as outliers in the response space (vertical outliers) and outliers in the predictor space (leverage points). The approach outperforms RCR in terms of precision and computational speed, and it is a determinate algorithm. The approach is capable of analyzing the data situation commonly found in certain biological applications where the number of variables is several orders of magnitude larger than the number of observations. Correspondingly, this paper focuses on several interesting and intensively studied near-infrared (NIR) spectra data sets for analysis of glucose concentration in order to illustrate the advantages of the proposed method.

The paper is organized as follows. Section 2 details the proposed method. Section 3 introduces NIR experiments, including NIR experiment of aqueous solution with glucose, NIR experiment of vitro plasma, and NIR experiment of human blood glucose noninvasive measurement in vivo. Section 4 presents NIR spectroscopy quantitative analysis to underline prediction performance and robustness properties of the proposed method for various types of outliers by comparing with other calibration methods, such as PLS, PCR, CPR and RCR, in terms of root mean squared error of prediction (RMSEP) and correlation coefficient. Finally, Section 5 briefly summarizes the results and gives concluding remarks.

2. Method

2.1. Robust continuum regression (RCR)

Firstly, let X be a spectra matrix with n rows, containing the observations, and p columns, containing the predictor variables. Let y be a column vector containing the n observations of the response variable. The vector of regression coefficients β is estimated in the linear model

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

with an error term ε. In near-infrared applications the number p of dimensions is typically several orders of magnitude larger than the number n of observations, so RCR performs a singular value decomposition on X^{T} such that [17]:

$$X^T = VDU^T \tag{2}$$

The matrices V and D take on the partitioned form:

$$V = \left(\tilde{V}\, 0_{p-n}\right) \tag{3}$$

and

$$D = \begin{pmatrix} \tilde{D} \\ 0_{p-n} \end{pmatrix} \tag{4}$$

The modified spectra matrix R_1 is presented as the formula

$$R_1 = U\tilde{D}^{\mathrm{T}} \tag{5}$$

Because R_1 is of size $n \times n$, the dimension is reduced to n instead of p.

Secondly, weighting vectors $w_i (i = 1, \ldots, h)$, where h is the number of latent variables, are calculated by

$$w_i = \arg\max_a \{Cov_\alpha(R_i a, y)^2 Var_\alpha(R_i a)^{\frac{\delta}{1-\delta}-1}\} \tag{6}$$

under the constraint

$$R_i = \left(I_n - \frac{t_{i-1}t_{i-1}^T}{t_{i-1}^T t_{i-1}}\right) R_{i-1} \quad (i > 1) \tag{7}$$

where a is a projection direction. The matrix R_i projected on a is given by $R_i a$. The tuning parameter δ can be chosen in the interval [0, 1]. PP can be applied to compute the weighting vectors. RCR constructs l projection pursuit directions ($l \geq n$) to be considered as l arbitrary linear combinations of the data points at hand (the first n directions being the directions given by the n observations available). By using α-trimmed variance and α-trimmed covariance, effects of leverage points on the estimation can be controlled. Similarly, effects of vertical outliers on the estimation can be controlled by using α-trimmed covariance. The trimming constant value in RCR is half of the percentage of outliers in observations. For subsequent projection directions the constraint Eq. (7) has to be fulfilled, which makes the estimated score vectors $t_i (i = 1, \ldots, h)$ uncorrelated. I_n is the unit matrix of size $n \times n$. t_{i-1} is the $i-1$th score vector [17].

Finally, a robust multivariate linear regression of y on score matrix T_{h} is performed by Huber M-regression [17,19]. Then the estimation of β is given by

$$\hat{\beta} = \tilde{V}\hat{\tilde{\beta}} \tag{8}$$

where $\hat{\tilde{\beta}}$ is the vector of regression coefficients relating R_1 and y.

The optimal values for δ and h are determined by a trimmed root mean squared error of cross-validation (RMSECV) [17]. This possesses a high resistance to outliers, because the optimal model parameters chosen by robust cross-validation aim at the majority of the observations, or even normal observations.

2.2. Modified robust continuum regression (mRCR)

NAS is defined as the part of a spectrum that is orthogonal to the subspace spanned by the spectra of all other components [20]. NAS only considers the part of the spectrum usable or available for quantization of the relevant component. Thus it can extract the useful information of the analyte of interest and eliminate the useless information of other components. As a matter of fact, in the mRCR method, projection matrix for computing NAS is the basis for further calculations such as weighting vector and regression coefficient.

We propose to construct PP directions with projection matrix for computing NAS of the analyte of interest. The projection matrix is employed to make the relationship between spectra and the analyte of interest closer by eliminating unwanted information, so more relevant and useful directions can be constructed for quantization of the analyte of interest and, furthermore, a better weighting vector can be obtained.