Analytica Chimica Acta 921 (2016) 28–37

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Fusion strategies for selecting multiple tuning parameters for multivariate calibration and other penalty based processes: A model updating application for pharmaceutical analysis

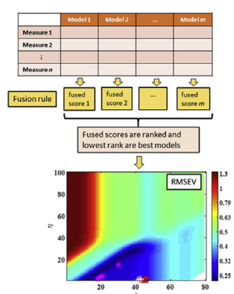
Alister J. Tencate [a], John H. Kalivas [a, *], Alexander J. White [b]

[a] Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA
[b] Department of Physics and Optical Engineering, Rose-Hulman Institute of Technology, Terre Huate, IN 47803, USA

HIGHLIGHTS

- Multiple measures of model quality are simultaneously assessed.
- Multiple tuning parameter values (penalty weights) can be automatically selected.
- Selected tuning parameter values balance bias and variance.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 November 2015
Received in revised form
11 March 2016
Accepted 29 March 2016
Available online 7 April 2016

Keywords:
Sum of ranking differences
Multivariate calibration
Partial least squares
Tikhonov regularization
Calibration maintenance
Model updating
Fusion
Multicriteria optimization
Penalty regression

ABSTRACT

New multivariate calibration methods and other processes are being developed that require selection of multiple tuning parameter (penalty) values to form the final model. With one or more tuning parameters, using only one measure of model quality to select final tuning parameter values is not sufficient. Optimization of several model quality measures is challenging. Thus, three fusion ranking methods are investigated for simultaneous assessment of multiple measures of model quality for selecting tuning parameter values. One is a supervised learning fusion rule named sum of ranking differences (SRD). The other two are non-supervised learning processes based on the sum and median operations. The effect of the number of models evaluated on the three fusion rules are also evaluated using three procedures. One procedure uses all models from all possible combinations of the tuning parameters. To reduce the number of models evaluated, an iterative process (only applicable to SRD) is applied and thresholding a model quality measure before applying the fusion rules is also used. A near infrared pharmaceutical data set requiring model updating is used to evaluate the three fusion rules. In this case, calibration of the primary conditions is for the active pharmaceutical ingredient (API) of tablets produced in a laboratory. The secondary conditions for calibration updating is for tablets produced in the full batch setting. Two model updating processes requiring selection of two unique tuning parameter values are studied. One is based on Tikhonov regularization (TR) and the other is a variation of partial least squares (PLS). The three fusion methods are shown to provide equivalent and acceptable results allowing automatic selection of the tuning parameter values. Best tuning parameter values are selected when model quality measures used with the fusion rules are for the small secondary sample set used to form the updated models. In this model updating situation, evaluation of all possible models, thresholding, and iterative SRD performed equivalently for the three fusion rules with TR and PLS performed worse. While the application is

* Corresponding author.
E-mail address: kalijohn@isu.edu (J.H. Kalivas).

http://dx.doi.org/10.1016/j.aca.2016.03.046

model updating, the fusion processes are applicable to other situations requiring selection of multiple tuning parameter values.

1. Introduction

A growing trend in developing new multivariate calibration methods and other data analysis modeling processes is including additional penalty terms in the model. Each additional penalty term typically involves a tuning parameter requiring optimization for an acceptable model. Since the goals of many calibration models are accurate and precise predications for new samples, it seems reasonable to use a model quality measure based on prediction error, such as from a cross-validation process. However, assessing the model quality with one measure based on prediction error when multiple tuning parameters are involved increases the risk of overfitting. To counter the chance of overfitting, additional model quality measures should be evaluated as tuning parameters are adjusted. With the increase of model quality measures comes the need for decision making tools to evaluate the multiple model quality measure allowing automatic selection of multiple tuning parameters. This situation is becoming increasingly relevant for a wide range of disciplines.

One approach to optimizing a multipenalty based model using multiple measures of model quality is multicriteria (multiresponse) optimization [1–7]. Broadly speaking, multicriteria optimization involves "Making a systematic and rational decision of the best alternative among several candidates when multiple (and often conflicting) criteria are present." [1]. The difficult part is how to combine (evaluate) and possibly weight the multiple criteria (measures of model quality in this paper) to form the final selection. Several approaches have been used to try and solve this issue. Popular are desirability functions [4]. However, there are several functions to choose from as well as numerous weighting schemes. An approach to avoid evaluation functions and weighting schemes is target optimization. However, like some desirability functions, target values for the model quality measures are needed and often this is not possible. Some graphical empirical approaches have been suggested, but these require expert knowledge and experience with the graphical interpretation [8–13].

Data fusion is becoming more common place across multiple disciplines in order to provide improved prediction accuracy for quantitative (e.g., concentrations) and qualitative (e.g., classification) purposes [14]. In the broadest meaning, "data fusion is a multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from multiple sources." [15]. Three levels of fusion are generally considered: low, mid, and high. A recent review paper characterizes these categories and provides a thorough review of fusion methods [16]. Briefly, low level fusion involves concatenation of data arrays to a single array. Mid-level indicates that features are extracted from each data array and then the data arrays are concatenated to a single array. The third level is termed high where multiple modeling processes are used for one or more data arrays and the final prediction result is a combination of individual results from using each model individually. The approach taken in this paper for tuning parameter selection probably fits high level fusion best. In this case, the final result (selected tuning parameter values and hence, model) is based on a fusion of the multiple inputs (model quality measures) where each input could be used separately to make the final decision. No weighting scheme is used for the fusion processes.

The method of sum of ranking differences (SRD) [17,18] has recently been used as a process allowing automatic tuning parameter selection for the calibration methods partial least squares (PLS) and the Tikhonov regularization (TR) method known as ridge regression (RR) [19]. The SRD method uses a collection of model quality measures to form a consensus model ranking by comparing each model quality measure value to an assigned target direction, such as minimum, maximum, etc., but a specific numerical value can also be used. The smaller the assigned rank given a model, the closer the corresponding model meets the respective targets of the different model quality measures. A single model or a collection of models can then be selected based on the SRD ranks.

While SRD prefers at least seven model quality measures for meaningful rankings, fusion rules such as mean, sum, stacking, and others [20–23] are not constrained in this way. However, these fusion rules lack the directional targeting used with SRD in the ranking process. The SRD process also includes a comparison to random ranking [18] in order to ascertain if a model ranking is any different than randomly ranking that model. For this study, the three fusion rules SRD, sum, and median are evaluated.

To test the three fusion rules and fusion processes developed in this paper for multiple tuning parameter selection, a model updating situation is studied. Common in model updating, a large calibration sample set characterizing the primary conditions is used to build a calibration model for the primary conditions. A small updating set of samples from new secondary conditions are then used to update the primary calibration model to focus on predicting new samples from the secondary conditions. This approach to model updating usually necessitates optimization of at least two tuning parameters [9–11,24].

Two model updating methods are studied and both require selection of two tuning parameter values. One is a Tikhonov regularization (TR) method [9–11]. The other method is based on partial least squares (PLS) [24]. A near infrared (NIR) pharmaceutical data set requiring model updating is used to evaluate the three fusion rules for selecting respective model tuning parameter values. In this data set, the primary conditions are for calibration of the active pharmaceutical ingredient (API) of tablets produced in a laboratory and the secondary conditions requiring model updating are tablets to be predicted for API content produced in a full batch setting.

2. Methods

2.1. Model updating

2.1.1. TR

The TR model updating process involves a variation on the standard form of TR, ridge regression (RR) [9–11]. For TR model updating, the RR model for the primary conditions is updated by augmenting the primary $\mathbf{X}$, $\mathbf{y}$ set with the secondary set $\mathbf{M}$, $\mathbf{y_M}$ composed of a few samples measured in the secondary conditions as shown in equation (1).

$$\begin{pmatrix} \mathbf{y} \\ 0 \\ \eta \mathbf{y_M} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \lambda \mathbf{I} \\ \eta \mathbf{M} \end{pmatrix} \mathbf{b} \tag{1}$$