Chemometrics and Intelligent Laboratory Systems 120 (2013) 142–153

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A consensus modeling approach to update a spectroscopic calibration

Parviz Shahbazikhah, John H. Kalivas *

Department of Chemistry, Idaho State University, Pocatello, ID 83209, United States

ARTICLE INFO

Article history:
Received 22 March 2012
Received in revised form 8 June 2012
Accepted 19 June 2012
Available online 25 June 2012

Keywords:
Tikhonov regularization
Partial least squares
Consensus modeling
Multivariate calibration
Calibration maintenance
Calibration transfer
Model updating

ABSTRACT

A spectroscopic calibration model is only valid to predict samples within the current calibration sample space span. This space is characterized by the calibration spectral variances set by the sample matrix properties and instrument environment (the primary conditions). Prediction samples commonly have new spectral variances (the secondary conditions) and dynamic model maintenance is needed. Previous work has shown that variants of Tikhonov regularization (TR) are capable of accomplishing this task by updating the primary model with only a few secondary condition samples (a current standardization set). An aspect of the TR variants is a weight (tuning parameter) for the small standardization set augmented to the full primary calibration sample set. In past work, this tuning parameter in combination with a second regularization parameter is graphically assessed to select a single updated model. Developed in this paper is a novel graphical consensus approach that selects a family of models across a range of tuning parameter values. Thresholds on model merit values are used to identify appropriate updated models. Model merits can be R^2, intercept, and slope for the primary calibration and standardization sets, root mean square error (RMSE) terms, and/or the model vector magnitude. Two previously used TR variants are studied. One TR modification requires analyte reference values and the other variation uses no reference values. The TR consensus approach with reference samples is applied to updating a laboratory based near infrared (NIR) primary calibration model to predict active pharmaceutical ingredient (API) tablet concentrations from NIR spectra measured on tablets produced in the full production secondary conditions. The TR approach without reference samples updates a NIR pure component analyte model at one temperature to new temperature dependent sample matrix conditions. In both studies, consensus models predict equivalently to individual models selected by previously developed graphical approaches. The consensus approach is also applied to model updating by partial least squares (PLS). While PLS and TR predict similarly, PLS can be limited by the discrete factorization process. The described consensus approach is also applicable to just primary calibration.

1. Introduction

A significant problem in multivariate calibration using spectroscopic data is the changing measurement conditions. Specifically, a calibration set of samples is measured on an instrument and a calibration model is formed. The calibration sample matrix and the instrument operating circumstances define the "primary conditions" for the calibration model. A new prediction sample commonly has a different sample matrix and/or is measured in a new instrumental state and hence, the measured spectrum lives in the "secondary conditions". If the primary calibration model is not adapted to the new chemical, physical, instrumental, or environmental changes, it will not accurately determine the calibrated chemical or physical property for samples measured in the new conditions. Thus, methods are needed to solve this calibration maintenance problem.

A goal of calibration maintenance, also referred to as calibration transfer or standardization, is to maintain the primary model for accurate determinations of samples measured in secondary conditions. Various approaches have been presented in order to accomplish calibration maintenance and are reviewed [1–4]. One approach is model updating where new variances not present in the primary conditions can be accounted for, including new spectrally responding species. This aspect is not possible with many other calibration maintenance processes.

For model updating, spectra from samples measured in the secondary condition are commonly augmented to the primary set of calibration samples. Many samples are typically needed to span the new conditions in order to offset the many samples already existing in the primary set. A simpler approach is to augment with only a few new samples to desensitize the model to the new condition or instrument [5–23]. Critical to these approaches is appropriate weighting of the new samples from the secondary conditions. With proper weighting of a few new samples, the large number of primary calibration samples does not have to be used and if used, does not bias the updated model.

A variety of new approaches based on the fundamentals of Tikhonov regularization (TR) have been developed to model-update with weighting [17–21]. These approaches mostly focus on updating

* Corresponding author. Tel.: +1 2082822726.
E-mail address: kalijohn@isu.edu (J.H. Kalivas).

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2012.06.006

a primary calibration by augmenting the primary calibration reference set with a few new reference samples measured in the secondary conditions (the standardization set) and then appropriately weighted. Some TR variants include model updating based on minimizing 2-norms (L_2 or Euclidean norm), 2- and 1-norm (L_1) for simultaneous model updating with wavelength selection (updated sparse models), and 1-norm for robustness to the standardizations set. Recent work has advanced these 2- and 1-norm variants to use no reference samples from the primary or secondary conditions [24]. Specifically, a pure component analyte spectrum is updated to the current secondary matrix and instrument state by augmenting with a few weighted non-analyte spectra measured in the new conditions. These and other TR variants are further overviewed in reference [21].

For the TR variants with or without reference samples, two tuning parameters need to be determined. One is for model stability (to minimize over- or under-fitting in the model) and the other is the weight for the standardization set. Past work has used empirical systematic graphical methods to select "a" model from the large number of models formed with the TR variants. It was suggested in reference [17] that a consensus approach could be used to identify a "collection" of models rather than selecting one model. Presented in this paper is a study of a novel consensus approach developed to select a collection of models over a range of tuning parameter values. The process is more mechanical than choosing one model. Consensus models are selected based on those satisfying model merits with natural target thresholds. Example merits include R^2, intercept, and slope values from plotting predicted calibrated analyte property values against reference values for the primary calibration and standardization sets. Other possible model merits with more empirical target thresholds are respective root mean square errors (RMSE) of prediction for the primary and standardization sets in conjunction with the model vector magnitudes (2-norm or 1-norm as the case may be).

Two previously studied data sets are evaluated. One data set involves reference samples to update a primary laboratory near-infrared (NIR) calibration model for predicting tablet active pharmaceutical ingredient (API) concentrations for tablets produced in secondary full production conditions [17,25,26]. With this data set, the TR variants using reference samples based on 2- and 1-norms are studied for consensus modeling. The consensus results are compared to selecting one model in previous work [17]. The second data set requires no reference samples and encompasses updating an ethanol pure component NIR analyte spectrum at 30 °C to predict in new secondary sample matrix conditions at 70 °C [24,26,27]. With this data set, the TR approach in 2-norm without reference samples is studied for consensus modeling and results are compared to selecting one model in previous work [24].

Many of the TR model updating representations can also be solved by partial least squares (PLS) or other processes [18,21]. Model updating using consensus PLS approaches represented in TR 2-norm formats are also evaluated for the two data sets.

2. Processes

2.1. Primary calibration with reference samples

A mathematical relationship for multivariate calibration is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ denotes an $m\times 1$ vector of quantitative reference information of the analyte for m calibration samples, $\mathbf{X}$ symbolizes the $m\times n$ matrix of respective spectra measured over n wavelengths, $\mathbf{b}$ represents an $n\times 1$ model vector, and $\mathbf{e}$ signifies the $m\times 1$ vector of normally distributed errors with mean zero and covariance matrix $\sigma^2\mathbf{I}$ with $\mathbf{I}$ being the $m\times m$ identity matrix. Multivariate calibration seeks to estimate $\mathbf{b}$, by $\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y}$ where $\mathbf{X}^{+}$ is a generalized inverse of $\mathbf{X}$. Several approaches can be used to form the generalized inverse including ridge regression (RR), PLS, etc. [28,29]. Once $\hat{\mathbf{b}}$ is determined, it is then used to predict new samples by $\hat{y} = \mathbf{x}^t\hat{\mathbf{b}}$ where the superscript t signifies the matrix algebra transpose operation. If $\mathbf{X}$ does not span new measurement and sample conditions, then predictions with the estimated model vector will be in error.

2.2. Calibration maintenance by model updating with reference samples

To form an updated model, the primary sample matrix $\mathbf{X}$ is augmented with samples measured in the new secondary conditions. Mathematically, Eq. (1) (ignoring the $\mathbf{e}$ term) is modified to

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{y_M}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \mathbf{M}\end{pmatrix}\mathbf{b} \tag{2}$$

where $\mathbf{M}$ represents an $s\times n$ matrix of s spectra measured in the new secondary conditions and $\mathbf{y_M}$ denotes the corresponding $s\times 1$ vector analyte reference values. If the number of secondary samples augmenting the primary calibration is large, then essentially a full recalibration is performed with no gain in efficiency. However, if the number of samples is small, then the new model vector will be biased towards the primary condition due to the majority of the samples spanning the primary condition. To resolve this bias issue, the small standardization set is weighted by a tuning or meta-parameter λ ($0\le\lambda$) to form

$$\begin{pmatrix}\mathbf{y}\\ \lambda\mathbf{y_M}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \lambda\mathbf{M}\end{pmatrix}\mathbf{b}. \tag{3}$$

Eq. (3) can be solved by PLS or another process [18,21]. With respect to PLS, the number of PLS latent variables (latent vectors, basis vectors, factors) is limited to the rank of the augmented matrix, typically the sum of the rank of $\mathbf{X}$ and $\mathbf{M}$. The size of the primary calibration set $\mathbf{X}$ is generally large thereby allowing PLS to characterize the sample matrix and instrument space.

2.3. Calibration maintenance by model updating with reference samples using TR variants TR2 and TR2-1

The TR2 and TR2-1 variants of TR have been well explained [17–21] and only a brief discussion is provided in this paper. Eq. (3) can be modified to a TR format by

$$\begin{pmatrix}\mathbf{y}\\ \eta\mathbf{I}\\ \lambda\mathbf{y_M}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \mathbf{0}\\ \lambda\mathbf{M}\end{pmatrix}\mathbf{b} \tag{4}$$

where η ($0\le\eta$) represents the TR regularization parameter (second tuning or meta-parameter), $\mathbf{I}$ signifies the identity matrix of size $n\times n$, and $\mathbf{0}$ denotes the $n\times 1$ zero vector. Essentially, Eq. (4) is in an RR format but now augmented with a standardization set. Specifically, RR is the solution to the minimization expression

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|_2^2 + \eta^2\left\|\mathbf{b}\right\|_2^2\right) \tag{5}$$

while TR for calibration maintenance using Eq. (4) is represented by

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|_2^2 + \eta^2\left\|\mathbf{b}\right\|_2^2 + \lambda^2\left\|\mathbf{Mb}-\mathbf{y_M}\right\|_2^2\right) \tag{6}$$

where $\|\cdot\|_2$ represents the Euclidean vector norm (L_2 or 2-norm). This TR variant is referred to as TR2. The regularization parameter η is not needed if PLS is used with Eq. (3) as latent variables replace the need for η, albeit an η could be used to form a ridge type PLS updated model.

Sparse models (models with wavelengths selected either individually or in bands) are often desired. Several TR variants can form sparse models where wavelength selection is part of the model forming