Chemometrics and Intelligent Laboratory Systems 175 (2018) 74–81

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Stochastic cross validation

Lu Xu [a,b], Hai-Yan Fu [c,*], Mohammad Goodarzi [d], Chen-Bo Cai [e], Qiao-Bo Yin [c], Ya Wu [b], Bang-Cheng Tang [b], Yuan-Bin She [a,**]

[a] College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, Zhejiang, PR China
[b] College of Material and Chemical Engineering, Tongren University, Tongren, 554300, Guizhou, PR China
[c] The Modernization Engineering Technology Research Center of Ethnic Minority Medicine of Hubei Province, School of Pharmaceutical Sciences, South-Central University for Nationalities, Wuhan 430074, PR China
[d] Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, United States
[e] Department of Chemistry and Life Science, Chuxiong Normal University, Chuxiong 675000, PR China

ARTICLE INFO

Keywords:
Multivariate calibration
Model complexity
Partial least squares (PLS)
Cross validation (CV)
Stochastic cross validation (SCV)

ABSTRACT

Cross validation (CV) is by far one of the most commonly used methods to estimate model complexity for partial least squares (PLS). In this study, stochastic cross validation (SCV) was proposed as a novel CV strategy, where the percent of left-out objects (PLOO) was defined as a changeable random number. We proposed two SCV strategies, namely, SCV with uniformly distributed PLOO (SCV-U) and SCV with normally distributed PLOO (SCV-N). SCV-U is actually a hybrid of leave-one-out CV (LOOCV), *k*-fold CV and Monte Carlo CV (MCCV). The rationale behind SCV-N is that the probability of large perturbations of the original training set will be small. SCV is expected to provide more flexibility for data splitting to explore and learn from the data set and evaluate internally a built model.

SCV-U and SCV-N were used for PLS calibrations of three real data sets as well as a simulated data set and they were compared with LOOCV, *k*-fold CV and MCCV. Given a training and external validation set, different CV techniques were repeatedly used to evaluate the optimal model complexity and the prediction results were compared. The results indicate that SCV-U and SCV-N could provide useful alternatives to the traditional CV methods and SCV is less sensitive to the values of PLOO.

1. Introduction

As a cornerstone of chemometrics, partial least squares (PLS) is by far the most popular method for multivariate spectroscopic calibration due to its effectiveness and simplicity [1]. PLS can extract a few relevant components or latent variables (LVs) from the high-dimensional measured signal matrix, e.g., near-infrared (NIR) spectral data, to predict the response variable. When developing a PLS model, it is crucial to select a proper number of LVs. Keeping too few LVs will obtain a too simple model that underfits the data; while including too many LVs will increase the risk of overfitting and degrade the model generalization performance [2–5].

Numerous efforts have been devoted to estimation of PLS model complexity or the number of LVs [6–19]. Among the various methods proposed, cross validation (CV) might be the most commonly used [20–22]. Without requiring an external validation set, an ordinary CV proceeds in the following steps: (1) the training data set is repeatedly split into training and validation objects; (2) with different model complexity, a pool of PLS models are developed using the training objects to predict the validation objects; (3) certain indices, e.g., the root mean square error of cross validation (RMSECV) or Q^2 [23] are computed to estimate the prediction errors with different numbers of LVs; and (4) the proper model complexity can be determined by selecting the number of LVs to obtain the lowest RMSECV (or highest Q^2) or by some statistical tests like *F*-test [24] and permutation test [25]. There are different versions of CV, such as leave-one-out CV (LOOCV), Monte Carlo CV (MCCV), *k*-fold CV, etc., among which the major difference lies in the way the original training data set is split. The results obtained by CV usually vary with the specific methods and the CV parameters used, e.g., the number of folds in *k*-fold CV and the percent of left-out objects (PLOO) for prediction in MCCV [26,27]. Generally, for a complex data set, it is still not straightforward to select the optimal number of PLS LVs relying on a

* Corresponding author.
** Corresponding author.
E-mail addresses: fuhaiyan@mail.scuec.edu.cn (H.-Y. Fu), sheyb@zjut.edu.cn (Y.-B. She).

https://doi.org/10.1016/j.chemolab.2018.02.008
Received 9 October 2017; Received in revised form 11 February 2018; Accepted 14 February 2018
Available online 20 February 2018
0169-7439/

Abbreviations

CV	Cross validation
PLS	Partial least squares
SCV	Stochastic cross validation
PLOO	Percent of left-out objects
SCV-U	SCV with uniformly distributed PLOO
SCV-N	SCV with normally distributed PLOO
LOOCV	Leave-one-out cross validation
MCCV	Monte Carlo cross validation
LVs	Latent variables
NIR	Near-infrared
RMSECV	Root mean square error of cross validation
RMSEP	Root mean square error of prediction

single method and practical experience is required. Therefore, improved CV techniques will be useful to tackle this problem.

This work was motivated by a new idea of data splitting. Unlike in an ordinary CV, where PLOO is often fixed at a certain constant, a more flexible data splitting strategy was proposed in this work. The newly proposed stochastic cross validation (SCV) strategy avoids using the single mode of data splitting and is expected to provide more flexibility to explore and learn from the data set and evaluate internally a built model. Two different SCV methods were proposed and compared with their traditional counterparts such as LOOCV, *k*-fold CV and MCCV in estimating PLS model complexity using three real NIR spectral data sets.

2. Methods

2.1. k-fold CV and LOOCV

CV simulates the predictions of new objects by repeatedly splitting the original training data set into training and validation objects. A multivariate calibration model can be developed using the training objects and used to predict the validation objects. For the *k*-fold CV, the original training data set is divided into *k* subsets (as equally as possible). At each time, one of the *k* subsets is used as the validation set and the other *k*-1 subsets are put together to form a training set. The above procedure is repeated *k* times and each of the *k* subsets is predicted exactly once. Usually a 10-fold CV is used.

LOOCV can be seen as an extreme of the *k*-fold CV, where each subset contains only one object. Therefore, suppose the original training set has *n* objects, in LOOCV each object is predicted exactly once by the model trained using the other *n*-1 objects. For both LOOCV and *k*-fold CV, the average error across all the predictions, e.g., the root mean square error of cross validation (RMSECV), can be computed as:

$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)} \quad (1)$$

where $\hat{y}_i$ and y_i are the predicted value and reference value for the ith object, respectively.

To estimate the model complexity of PLS, RMSECV can be computed with different numbers of PLS LVs. Ideally, the optimal number of LVs can be selected to obtain the lowest RMSECV value; however, sometimes the RMSECV can be found to decrease by including more LVs in the PLS model. Therefore, practical experiences are usually needed to make a tradeoff between a low model complexity and a low RMSECV.

2.2. MCCV

MCCV [26] was proposed and used to avoid selecting too many PLS LVs and was also further improved for estimation of prediction errors [27]. In this work, MCCV was used as a method to select the proper number of PLS LVs. MCCV is based on random and repeated splitting of the original training data set into training and validation objects. It was suggested that as soon as the data structure could be retained, a higher percent of left-out objects (PLOO) should be adopted to avoid selecting too many PLS LVs than necessary. In other words, when the left-out objects for prediction could be sufficiently represented by the training data, as many as possible left-out objects should be adopted to avoid obtaining over-optimistic results. MCCV has been shown to be very effective to reduce the risk of overfitting. The root mean square error of MCCV (RMSEMCCV) can be calculated as:

$$RMSEMCCV = \sqrt{\frac{1}{B \times n_v}\sum_{i=1}^{B}\|\hat{y}_{vi} - y_{vi}\|} \quad (2)$$

where B and n_v are the times of data splitting and the number of validation objects in each splitting, respectively; $\hat{y}_{vi}$ and y_{vi} are the predicted and reference values of validation objects in the *i*th splitting, respectively. MCCV usually selects the PLS components to get the lowest RMSEMCCV value. PLOO in each splitting (n_v) can be set considering the size of the original data set. In this work, different PLOO values were studied for MCCV.

2.3. Two stochastic cross validation (SCV) methods

The main difference among different versions of CV lies in the way the original data set is divided into training and validation objects, e.g., how many objects are left out for validation at each data splitting. Generally, the proper data splitting should be data-driven and may depend on the internal data structure and the error level, which are usually not known *a priori*. Therefore, in this work, SCV was proposed with a changeable PLOO. Two different SCV methods were suggested, namely, SCV with uniformly distributed PLOO (SCV-U) and SCV with normally distributed PLOO (SCV-N). In SCV-U, PLOO was defined as a uniformly distributed random number:

$$PLOO_{SCV-U} \sim U\left(\frac{1}{n}, \alpha\right) \quad (3)$$

where n is the number of training objects and $\alpha(0<\alpha<1)$ is the maximum PLOO predefined.

In SCV-N, PLOO was proposed to have a normal distribution:

$$PLOO_{SCV-N} \sim N\left(0, \left(\frac{\alpha}{1.96}\right)^2\right) \quad (4)$$

In practical operations, the absolute value will be used considering possible negative values and when the randomly generated PLOO exceeds α, it will be set to be α. The factor 1.96 implies that about 5% of the randomly generated PLOO values would exceed α and will be set to be α.

SCV avoids using the single mode of data splitting and is expected to provide more flexibility to explore the data set. The root mean square error of SCV (RMSESCV) could be computed as:

$$RMSESCV = \sqrt{\frac{1}{N}\sum_{i=1}^{B}\|\hat{y}_{vi} - y_{vi}\|^2} \quad (5)$$

Where N is the total number of validation objects during data splitting. In this work, different maximum values of PLOO were studied for SCV-U and SCV-N.

2.4. Method comparison

In order to compare the performances of LOOCV, *k*-fold CV, MCCV,SCV-U and SCV-N in selecting the proper number of PLS LVs, for each data set, besides using the raw spectra, data preprocessed by taking second-order derivatives (D2) [28] and standard normal variate transformation (SNV) [29] were also used. With a given training and external