Chemometrics and Intelligent Laboratory Systems 174 (2018) 45–55

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Alternate deflation and inflation of search space in reweighted sampling: An effective variable selection approach for PLS model

Biswanath Mahanty

Department of Biotechnology, Karunya Institute of Technology and Sciences, Karunya Nagar, Coimbatore 641114, India

ARTICLE INFO

Keywords:
ADISS
Binary matrix sampling
Partial least square regression
Space shrinkage
Variable selection

ABSTRACT

Based on assessment of randomized sub-model populations generated through reweighted binary matrix sampling (BMS), an innovative variable selection strategy for PLS regression model, called alternate deflation and inflation of search space (ADISS) is proposed. Normalized regression coefficients of best PLS sub-models population is used to formulate the weight vector for re-weighted BMS. Unlike the most existing algorithm, ADISS alternatively shifts between forward selection (inflation) and backward elimination (deflation) of variable space, minimizing the risk of accidental loss of informative variables. Compared with methods such as competitive adaptive reweighted sampling (CARS), variable iterative space shrinkage approach (VISSA), or Monte Carlo uninformative variable elimination (MC-UVE), proposed method showed lower cross-validation or prediction error for two different benchmark NIR data sets. ADISS frequently selects nearly the same sets of variables across multiple independent runs, that signifies stability of the output. The unsupervised execution, termination and projection of final variable set from the algorithm is important advantage while considering for large scale data.

1. Introduction

Multivariate calibration models are increasingly being adopted to extract quantitative or qualitative information from highly collinear spectroscopic data of complex biological and environmental samples [1–3]. Partial least squares (PLS) regressions models provides excellent selectivity and prediction accuracy even when contribution of interfering substances remains unpredictable or cannot be accounted in priori [4,5]. However, model developed with large-number of variables from relatively fewer samples, commonly referred as "large *p* small *n* problem", can deteriorate the PLS regression [6–8]. Removal of uninformative variables while keeping some "useful redundancy" is the key to have a parsimonious model with improved prediction efficiency and interpretability [9].

Variable selection strategies in PLS either identifies variable subset directly from regression output (filter method), or pipe-back the selected subset into refitting algorithm (wrapper-methods), or integrate variable selection strategy into core of PLSR algorithm (embedded method) [10,11]. Though faster and easier to implement, variable subset selection in filter methods is somewhat arbitrary without assessing their final performance. On other hand, the wrappers methods select variables in an iterative way while evaluating performance (e.g. cross validation error) of the PLS models based a randomized population [11]. A large number of randomized wrappers based variable selection strategies has been proposed in this decade such as, uninformative variable elimination (UVE) [12], Monte Carlo based-UVE MC-UVE [13], recursive weighted PLS (*r*PLS) [14], competitive adaptive reweighted sampling (CARS) [15], variable iterative space shrinkage approach (VISSA) [16], variable combination population analysis (VCPA) [17], sub-window permutation analysis (SPA) [18], iteratively retaining informative variables (IRIV) [19]. Strategic comparison among few of those algorithms have also been reviewed, though albeit non-exhaustively [10,20].

The fundamental framework across those randomized wrappers methods remains the same, *i.e.* generation of a population of sub-models, followed by formulation of weight vector (from the best performing subsets) for next round of weighted resampling. Performance of these algorithm differs in the way the weight vector is created, used in resampling, or the variable space is altered. For example, use of normalized PLS regression vector to formulate the weight in CARS should be more effective in capturing the relative importance of the variables than the binary weight scheme in VISSA, where equal importance is given to each of the retained variables in a sub-model [14,19]. However, weighted Monte Carlo re-sampling in CARS selects some variables more frequently than others, resulting in distribution of variables in population different from the original weight vector [19]. UVE, MC-UVE, SPA or CARS are computationally efficient, but they select the variables

E-mail address: bmahanty@karunya.edu.

https://doi.org/10.1016/j.chemolab.2018.01.005
Received 12 November 2017; Received in revised form 6 January 2018; Accepted 17 January 2018

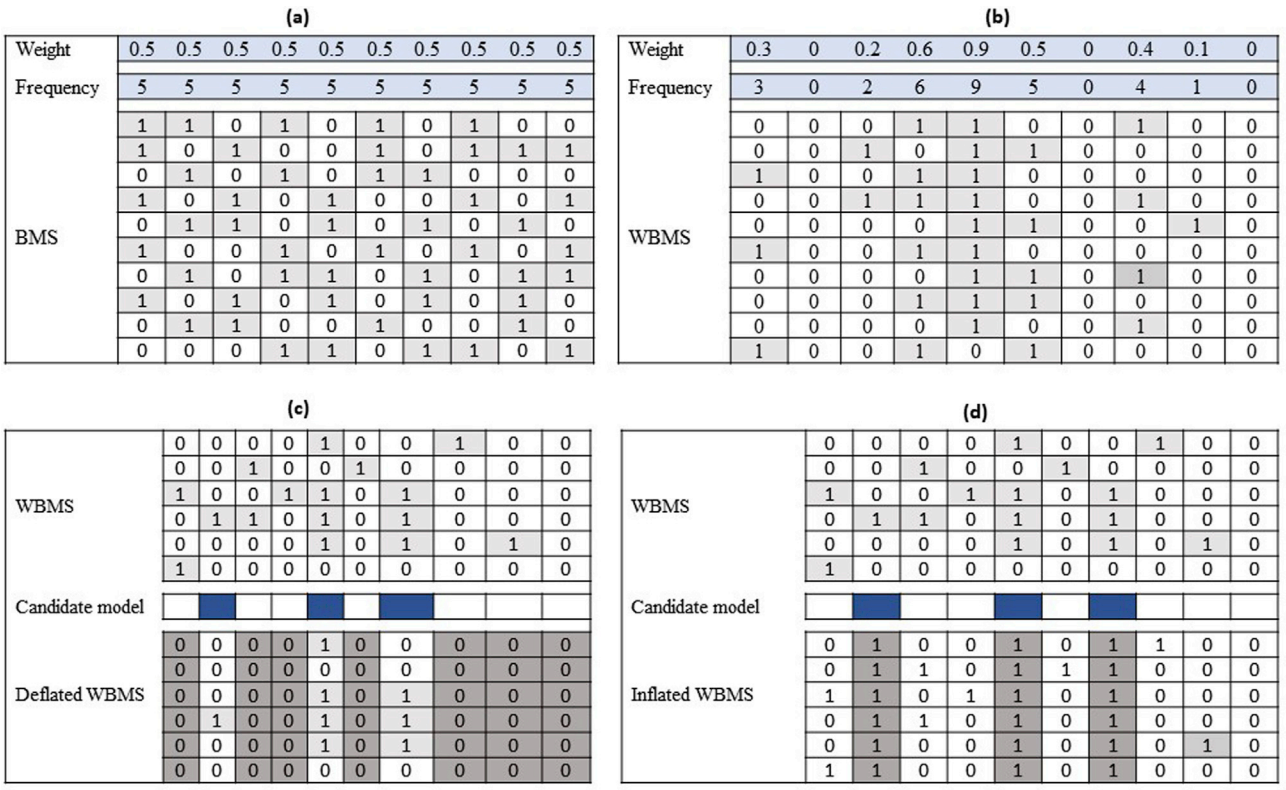

Fig. 1. The processes of (a) binary matrix sampling (BMS) and (b) weighted binary matrix sampling (WBMS). The number "1" in each row represents the variable that is selected for that sub-model while the number "0" represents not selected. Frequency of a variable across the population is calculated by multiplying the population size (number of rows) by corresponding variable weight (fraction). Each column of the binary matrix is then allotted "ones" in a number equals to its frequency and column elements are permuted. The variable combination with lowest RMSECV from a round BMS or WBMS is referred as "candidate model". For subsequent run a regular WBMS is (c) inflated to retain all variables of the "candidate model" or (d) deflated to exclude all variables not in "candidate model".

individually ignoring their combination effect.

The low weight variables are quickly (and forcibly) discarded based on an exponentially decreasing function in VCPA or CARS during initial iterations [17]. In VISSA, non-improvement of performance for the best sub-models across iterations triggers a softer space shrinkage, where variables except those in the best sub-model are eliminated from next round [16]. However, there is no way to get back any variable once irreversibly eliminated at initial phase of the algorithm, even if it could be a part of globally optimal model.

Based on the core idea of binary matrix sampling (BMS), a novel variable selection strategy, called Alternate Deflation and Inflation of Search Space (ADISS) is proposed. In line of VISSA, when performance of the best model stops improving in iterations, the variable space is alternatively deflated or inflated in ADISS. During deflation, all columns of BMS except those best sub-model variables from last iteration are forced zero. In inflation cycle, variables in best sub-model are always retained. Unlike unidirectional space shrinkage in VISSA, this approach protects against inadvertent loss of informative variables.

In this work, application of ADISS algorithm for selection spectral variables in PLS regression model was investigated, and results were compared with few commonly encountered strategies. Though applied for spectral data set, the approach can be incorporated into any multivariate problems.

2. Theory and algorithm

2.1. Brief introduction of the existing methods

The algorithms inherent to nearly all variable selection strategy necessarily incorporate a method of sampling the variable space (or sample space), generating a subset of models, assessing the fitness of the candidate models and implementing a criteria for selection or convergence to optimal solution [21]. The statistical assessment of output distribution from a large population of sub-models would have comprehensive information content on the underlying data and is central idea of model population analysis (MPA) [20]. The different variable selection algorithms based on MPA framework diverges based on choice of random sampling techniques (e.g. MC sampling, bootstrapping or BMS), the space being sampled (sample or variable space) or the output being considered.

2.1.1. Monte Carlo uninformative variable elimination (MC-UVE)

Absolute regression coefficient reflects relative importance of variables, is the fundamental framework of MC-UVE [22]. A fraction of samples (e.g., 80%) is randomly selected from calibration dataset as a training subset on which a PLS model is developed [13]. This procedure is repeated N (e.g., 500) times resulting in a matrix ($N \times p$) of calculated regression coefficients (β). From distribution of β, a reliability index (RI), defined as the ratio of the mean to the standard deviation of this distribution, is generated that is used to rank the variables in order of importance. These sorted variables are sequentially added to establish PLS models until RMSECV gets minimum in cross-validation. The RI corresponding to the final variable selected is set as the threshold. All variables that are related with a RI lower than this threshold value can be removed.

2.1.2. CARS

Like UVE, CARS also adopts absolute regression coefficients to scale importance of the variables. For each sampling run, a constant ratio (e.g., 80%) of samples is first randomly selected to build a calibration model [15]. Variables are sorted on absolute regression coefficients and only the best set of variables are retained depending on limit imposed by