Analytica Chimica Acta 948 (2016) 19–29

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

# A novel algorithm for spectral interval combination optimization

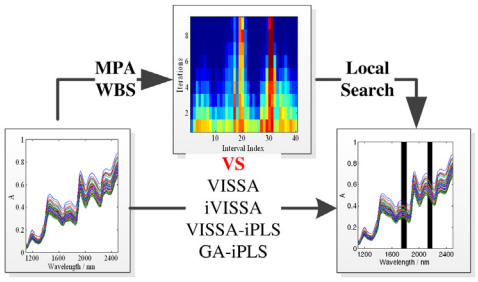
Xiangzhong Song [a], Yue Huang [a, b], Hong Yan [a], Yanmei Xiong [a, *], Shungeng Min [a, **]

[a] College of Science, China Agricultural University, Beijing, 100193, PR China
[b] Third Class Tobacco Supervision Station, Beijing, 101121, PR China

## HIGHLIGHTS

- A new wavelength interval combination optimization algorithm was proposed based on model popular analysis strategy.
- The combination of spectral intervals can be optimized in a soft shrinkage manner.
- Its computational intensity is economic benefit from fewer tune parameters and faster convergence speed.
- WBS was proved to be a more efficient sampling method than WBMS especially for implementing MPA strategy.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Article history:*
Received 25 August 2016
Received in revised form
26 October 2016
Accepted 28 October 2016
Available online 2 November 2016

*Keywords:*
Wavelength selection
Interval combination optimization (ICO)
Model population analysis (MPA)
Weighted bootstrap sampling (WBS)
Weighted binary matrix sampling (WBMS)

## ABSTRACT

In this study, a new wavelength interval selection algorithm named as interval combination optimization (ICO) was proposed under the framework of model population analysis (MPA). In this method, the full spectra are divided into a fixed number of equal-width intervals firstly. Then the optimal interval combination is searched iteratively under the guide of MPA in a soft shrinkage manner, among which weighted bootstrap sampling (WBS) is employed as random sampling method. Finally, local search is conducted to optimize the widths of selected intervals. Three NIR datasets were used to validate the performance of ICO algorithm. Results show that ICO can select fewer wavelengths with better prediction performance when compared with other four wavelength selection methods, including VISSA, VISSA-iPLS, iVISSA and GA-iPLS. In addition, the computational intensity of ICO is also economical, benefit from fewer tune parameters and faster convergence speed.



## 1. Introduction

Spectroscopic datasets collected by high throughput instruments are usually faced with the non-deterministic polynomial time (NP)-hard problem. This kind of datasets usually consists of large number of variables and relatively few samples due to the constraint of actual experimental conditions and costs. Multivariate calibration techniques such as principal component regression (PCR) and partial least squares regression (PLS) are usually employed to address this problem by extracting latent information from spectroscopic dataset. However, more and more researches have proved that variable selection is still beneficial for these multivariate calibration techniques from both experimental and theoretical aspects [1–5]. The benefits of variable selection can be

* Corresponding author.
** Corresponding author.
*E-mail addresses:* xiongym@cau.edu.cn (Y. Xiong), minsg@263.net (S. Min).

http://dx.doi.org/10.1016/j.aca.2016.10.041
0003-2670/

summarized in four main aspects: (1) the prediction ability of calibration model can usually be improved by eliminating uninformative or interfering variables; (2) new calibration model based on informative variables will be easier to interpret; (3) the computational speed of new model will be boosted; (4) low cost of dedicated online or inline analytical instrument with less spectral channels may be produced under the guide of variable selection [6].

In essence, variable selection is aimed to find an optimal combination of variables for the best prediction performance. However, as the number of variable combinations grows exponentially along with the increase of variables, the rough search is always impractical. Thus, a large number of variable selection methods have been proposed based on different strategies in the past decades, such as stepwise strategy, e.g. forward selection and backward elimination [7]; variable ranking strategy based on parameters of PLS model [8–10], e.g. loading weights [11,12], regression coefficients [13,14], variable in projection (VIP) [15], stability [16–19], and selective ratio [20]; optimization strategy based on artificial intelligent algorithms, e.g. genetic algorithm (GA) [21], simulated annealing (SA) [22,23], particle swarm optimization (PSO) [24] and ant colony optimization (ACO) [25]; projection strategy, e.g. successive projection algorithm (SPA) [26]. Besides, it is worth noting that model population analysis (MPA) strategy proposed by Liang's group can also be used for variable selection [27]. Based on this strategy, a series of variable selection methods has been proposed in recent years, such as iteratively retaining informative variables (IRIV) [28], variable combination population analysis (VCPA) [29], variable iterative space shrinkage approach (VISSA) [30,31], bootstrapping soft shrinkage (BOSS) [32].

As a general framework for designing new chemometrics or bioinformatics algorithms, MPA emphasizes that information should be extracted by analyzing a number of sub-models statistically, because the results or parameters of one single model are not always reliable. In detail, MPA usually contains three stages: (1) sub-datasets generation procedure, where random sampling method is applied to obtain a series of sub-datasets from variable or sample space, such as jackknife sampling [33], bootstrap sampling (BSS) [34], binary matrix sampling (BMS) [35]; (2) modeling procedure, where a series of sub-models are established based on sub-datasets generated in the previous step; (3) statistical analysis procedure, where interested outputs (e.g., RMSECV value) of all these sub-models are analyzed statistically.

Advantages of using MPA strategy to variable selection can be concluded in two aspects: (1) MPA extracts information from a large number of sub-models, which is beneficial for avoiding the uncertainty of one single model. (2) Synergistic or combination effects between different variables are more possible to be retained by MPA since random variable combinations are generated during the optimization process. Additionally, the strategy of soft shrinkage, which can avoid removing important variables by mistake, can also be regarded as an advantage of some new methods (e.g. VISSA and BOSS) developed from MPA. By this strategy, insignificant variables are not eliminated directly, but are assigned with a smaller sampling weight, ensuring that the process of optimization is implemented in the soft shrinkage way. Besides, weighted binary matrix sampling (WBMS) [30] and weighted bootstrap sampling (WBS) [36] are also two commonly used weighted random sampling methods. Up to now, there is no comparison of their performance yet.

Certainly, variable selection methods based on MPA have some drawbacks. First, their computational burden is much heavier than other methods, because they not only need to establish a large number of sub-models in each loop, but also require many loops to realize iteration convergence. Secondly, overfitting of these methods is at high risk due to the large number of variables combination [3]. Specially, WBMS generates sub-datasets too strictly depending on the sampling weights, even if the sampling weight of one variable becomes 1 by chance, it still has to be included in the future iterations.

Undoubtedly, for most kinds of spectral data, especially for near infrared spectroscopy, the selection of wavelength intervals seems more reasonable than single spectral points [3]. Because the informative variables within specific absorbing bands certainly contain similar information, which may lead some individual variable selections to chaos runs [37]. In contrast, interval selection methods can provide a more stable result. Chemical meaning can also be explained much easier. Furthermore, the selection of intervals can decrease the computational burden by reducing the number of possible combinations. It was more likely to avoid selecting single wavelengths in the noisy area which may have spurious correlations with the interested property [3]. Hence, there are a lot of spectral interval selection methods reported, such as interval partial least squares (iPLS) [38], moving windows PLS (MWPLS) [39] and many variants based on them [40–43]. Besides, some strategies commonly used for individual variable selection such as SPA [44], GA [45,46], ACO [47], etc. have also been modified for selecting informative intervals in recent years. However, MPA strategy and soft shrinkage strategy have rarely been applied to spectral interval selection.

New wavelength interval selection named as interval combination optimization (ICO) is proposed by coupling WBS with MPA, which can address drawbacks mentioned above together. In this study, three NIR datasets were applied to validate the performance of ICO. For comparison, four wavelength selection methods, including VISSA, interval VISSA (iVISSA), VISSA-iPLS and GA-iPLS, were also performed as references.

## 2. Theory and algorithm

### 2.1. Weighted binary matrix sampling (WBMS)

WBMS provides a random sampling strategy using a binary matrix [30]. In this K × P size binary matrix, K is the total sampling number and P is the number of objects. In each column of the binary matrix, “1” represents the object will be retained for modeling, while “0” represents the object will be excluded, and the ratio of “1” in each column will be updated according to the weight in each iteration. After the ranking order of each column is permutated, a new binary matrix is generated. In this new binary matrix, each row represents one random sampling procedure. Obviously, the greater the weight is, the greater the selected probability. And if the weight of one object is 1, it will be selected in every random sampling procedure, which means that it will have no possibility to be excluded. If the weight of one object is 0, it has no possibility to be retained by any random sampling procedure, which means that it will be eliminated.

### 2.2. Weighted bootstrap sampling (WBS)

WBS is a random sampling technique with replacement derived from BSS [36]. In WBS, one weight is allocated to one object firstly, which is between 0 and 1. Then WBS selects objects with a strategy like the roulette wheel. In this strategy, each object is corresponding to one slot on the roulette, and the size of which is proportional to the weight of the corresponding object. One object is selected in each run of this roulette. The theoretical selected probability of one object in each run can be calculated according to Equation (1). Therefore, even if the weight of one object reaches 1, it still has a chance to be excluded.