Chemometrics and Intelligent Laboratory Systems 180 (2018) 1–14

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Variable selection optimization for multivariate models with Polar Qualification System

Shikhar Mohan [a], Bruce R. Buchanan [b], Glen D. Wollenberg [b], Benoît Igne [c], James K. Drennen III [a,c], Carl A. Anderson [a,c,*]

[a] Graduate School of Pharmaceutical Sciences, Duquesne University, Pittsburgh, PA, USA
[b] S2I, LLC, 216 Virginia Avenue, Shenandoa, VA, USA
[c] Duquesne Center for Pharmaceutical Technology, Duquesne University, Pittsburgh, PA, USA

ARTICLE INFO

Keywords:
Polar qualification system
Variable selection
Genetic algorithm
Transmission Raman spectroscopy

ABSTRACT

Multivariate models are used in many fields to predict a response from a set of variables having an undetermined covariate structure. Variable selection often improves multivariate model performance by removing information not related to the response of interest. Many variable selection methods exist for this purpose. This study investigates Polar Qualification System (PQS) as a tool for variables selection. A Raman transmission dataset of tablets containing Niacinamide (active pharmaceutical ingredient) and Niacin (degradant) was modeled for degradant weight concentration using Partial Least Squares (PLS) regression. Three variable selection techniques were compared for the development of a stability indicating method: specific peak selection (manual selection), genetic algorithms (GA-PLS), and a newly developed PQS-Hadamard method. The model performance of these techniques was compared to a model developed with the whole spectrum. All models built with selected variables showed reduced prediction error compared to model created with the full variable range. However, the PQS-Hadamard method was demonstrated to be more computationally efficient compared to GA-PLS. Further, it is a potentially automatable process, unlike the specific peak selection, which requires expert selection of variables.

1. Introduction

Multivariate models are often complex due to the unknown relationships among the variables and the response. The elimination of covariate and non-informative variables enhances model interpretation. Also, it is commonly accepted that the predictive ability of the model is improved if the non-informative variables are removed [1,2]. Examples of variable selection have been demonstrated for spectroscopic data such as Near Infrared (NIR) and Raman [2–8]. In spectroscopic data, variable selection is referred to as wavelength selection. Raman and NIR spectra have responses at wavelength regions associated with specific functional groups. If a quantitative NIR/Raman model is built for a specific component, selecting only wavelength regions or Raman shifts associated with the functional groups in that component can improve model performance [7].

However, there are disadvantages with variable selection. Including all variables can enhance model robustness in spectroscopic based models. In addition, removing wavelength bands not corresponding to parameter/analyte of interest doesn't always improve model accuracy and robustness [9]. A reason being that the wavelength region corresponding to the analyte of interest can still include overlapping interference that are difficult to isolate. Including all variables is a simplistic approach for model robustness, however, extra cost is required to generate all the relevant samples featuring expected interference. Therefore, variable selection methods that appropriately identify significant variables associated with the analyte of interest and remove interfering variables are desirable.

There are various types of variable selection techniques. They can be categorized as 'Manual', 'Univariate', 'Sequential', and 'Multivariate' [10]. The 'Manual' technique involves the selection of variables based on prior knowledge; typically requiring expert knowledge of the data. Selection of wavelength regions specific to the analyte of interest is an example of manual variable selection for spectroscopy. In many cases however, extensive prior knowledge of the data is not available.

* Corresponding author. Duquesne Center for Pharmaceutical Technology, Duquesne University, Pittsburgh, PA 15282, USA.
E-mail address: andersonca@duq.edu (C.A. Anderson).

https://doi.org/10.1016/j.chemolab.2018.06.002
Received 14 December 2017; Received in revised form 29 May 2018; Accepted 1 June 2018
Available online 4 June 2018

'Univariate' methods involve calculating a measure of correlation between each variable and the response, then selecting variables based on highest correlation [10]. This method will provide information on the most significant individual variables but will not provide information on the multivariate structure of variables. Combining collinear variables that are informative with respect to the response typically reduces noise associated with individual variables thus improving the model performance [11,12]. 'Sequential' and 'Multivariate' methods seek to take advantage of combination of variables. 'Sequential' methods include algorithms such as backwards elimination and forward selection [13]. 'Multivariate' methods use multivariate model statistics to locate the collection of variables which give the best model performance. Examples include iterative Partial Least Squares (iPLS) [12], Genetic Algorithms using partial least-squares (GA-PLS) [14–16], uninformative variable elimination (UVE) [11], and significance tests of model parameters [10]. In this work, a new automated variable selection technique (Polar Qualification System-Hadamard (PQS-Hadamard) method) is presented.

Genetic algorithms combined with partial least-squares (PLS) regression is widely used as a wavelength selection technique for spectroscopic data sets [17]. The GA-PLS method is preferred over other 'multivariate' methods due to its heuristic search algorithm. It effectively generates combinations of variables that are iteratively optimized using PLS model statistics to ensure no loss of predictive ability [12,15,16,18–20]. 'Multivariate' methods such as GA-PLS are advantageous as they require no prior knowledge of the data and are readily automated. However, these methods are computationally demanding and require users to select a large number of parameters (starting/ending conditions), limiting its use to expert practitioners. These user-defined inputs need to be carefully selected as they influence the final selected variables. As a variable selection technique, GA-PLS is susceptible to overfitting [17]. However, it is accepted as a variable selection tool and is used as a standard to compare new variable selection techniques [12]. The method proposed in this article (PQS-Hadamard) has the advantages of an automated method while reducing user defined inputs, improving calculation efficiency, and minimizing the potential for overfitting.

A transmission Raman dataset of multi-component tablets was analyzed in this work. These tablets contained Niacinamide, the active pharmaceutical ingredient (API), and Niacin, the degradant, along with various excipients. Niacinamide is a form of vitamin B3 (Niacin), which is used to prevent pellagra, and is usually preferred over Niacin due to less severe side effects. Quantitative prediction models were generated to predict the degradant (Niacin) percent weight.

The goal of this study was to compare quantitative Niacin prediction performance with variables selected by PQS-Hadamard method to variables selected by genetic algorithms and specific peak selection on a transmission Raman dataset.

2. Theory

2.1. Genetic algorithms

Genetic algorithms are primarily used for variable selection optimization [21]. The methodology of genetic algorithm is based on Darwin's theory of evolution where new generation/population of variables are created by combining variables with 'good fitness'. Genetic algorithms seek to enhance 'fitness' between each generation until an end-point criterion is reached. For GA-PLS, 'fitness' enhancement is typically based on minimization of PLS model statistics such as cross validation error or prediction error. This paper specifically used cross validation error for 'fitness' assessment.

Genetic Algorithms involve four main steps [10] explained below:

- Step 1: Assign a random binary value to a variable (or window). This vector is randomly selected. The next vector is again a randomization of the binary code. This process is repeated 'm' times resulting in 'm' vectors. The 'm' term is a user defined input and is referred to as the initial population size.
- Step 2: Generate PLS models from variables selected from each vector. The root mean square error of cross validation (RMSECV) is then used to assess the 'fitness' of each vector. The vectors with 'good fitness' (low RMSECV) in this step are referred to as the 'parent' vectors and are carried on to the next step.
- Step 3: Undergo crossover. The 'parent' vectors from step 2 are combined with one another to create a new population (or generation). The combination occurs by taking two 'parent' vectors and splitting them at the same one (single crossover) or two (double crossover) randomly chosen point(s). The resulting segments are then crossed (only middle section crossed for double crossover) to create two 'offspring' vectors. The idea is that these 'offspring' vectors have 'better fitness' than their 'parent' vectors. Double crossover is usually used because the 'offspring' vectors are more similar to the 'parent' vectors
- Step 4: Perform mutation. In each new generation vector, a small probability (mutation rate) of a change to each variables' binary code is added. This is important because if a variable is not selected in the initial population then it may never be selected. Mutation allows for these non-selected variables to have a chance to be considered. Mutation rates are typically low in order to advance the algorithms to an end-point.

This algorithm is repeated until convergence criteria are met. This criterion is user defined and is often a function of the percent of identical variables selected in each vector and the minimization of cross-validation error.

2.2. Polar Qualification System (PQS)

Spectral data processing is computationally extensive; chemometrics tools such as Principal Component Analysis (PCA) and PLS continue to require appreciable processing power. To simplify spectroscopic analysis, and reduce computational demands, Polar Qualification System (PQS) was introduced by Kaffka and Gyarmati in the 3rd International Conference on Near Infrared Spectroscopy in Brussels [22–24].

This technique involves representing spectra in polar coordinates (polar spectra) to obtain one value (center of mass of the polar spectra) representing an entire spectrum. The reason to use polar coordinates is to introduce geometrical considerations, as the center of mass between samples will move towards wavelengths/variables with high variance. This method has already been applied as a wavelength selection technique and used for quality control purposes [24,25] however this work applies additional procedures to efficiently automate the PQS method.

Polar Qualification System involves three steps listed below. Similar to chemometric techniques such as PLS, appropriate spectral preprocessing is applied to enhance signal to noise ratio before performing these three steps.

- Step 1: Spectral transformation from Cartesian space into polar space.
- Step 2: Calculation of centers of mass.
- Step 3: Classification of centers of mass.

The first step is to represent Raman spectra in polar space. In a polar space, a point is defined by an angle and a radius. The angle represents the variable so the first variable will be the first angle, the second variable will be the second angle and so on. The radius then represents the value