Chemometrics and Intelligent Laboratory Systems 148 (2015) 85–94

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Efficient wavenumber selection based on spectral fluctuation dividing and correlation-based clustering for calibration modeling

Takuya Miyano [a,*], Koichi Fujiwara [b], Manabu Kano [b], Hideaki Tanabe [a], Hiroshi Nakagawa [a], Tomoyuki Watanabe [a], Hidemi Minami [a]

[a] Formulation Technology Research Laboratories, Pharmaceutical Technology Division, Daiichi Sankyo Co., Ltd., 1-12-1, Shinomiya, Hiratsuka, Kanagawa 254-0014, Japan
[b] Department of Systems Science, Kyoto University, Yoshida-honmati, Sakyo-ku, Kyoto 606-8501, Japan

ARTICLE INFO

Article history:
Received 12 May 2015
Received in revised form 10 September 2015
Accepted 13 September 2015
Available online 18 September 2015

Keywords:
Near-infrared (NIR) spectroscopy
Wavenumber selection
Process analytical technology (PAT)
Nearest correlation spectral clustering
Calibration model

ABSTRACT

This study proposes an efficient wavenumber selection method to develop calibration models based on near-infrared (NIR) spectroscopy. First, spectral fluctuation dividing (SFD) divides a whole NIR spectrum into multiple spectral intervals based on a spectral fluctuation profile, which consists of the standard deviation of spectral intensities at each wavenumber. Then, nearest correlation spectral clustering (NCSC) clusters those spectral intervals into spectral interval groups based on the correlation of the spectral intensities among the spectral intervals. Finally, the proposed method builds a partial least squares (PLS) model using the spectral intensities in each spectral interval group, and selects several spectral interval groups based on the estimation accuracy of each PLS model. This method was named SFD–NCSC–PLS. In developing calibration models to estimate water and drug contents in granules, SFD–NCSC–PLS achieved higher estimation accuracy than the commonly-used interval PLS, searching combination moving window PLS, and the methods using either SFD or NCSC. The results show that SFD–NCSC–PLS can properly select wavenumbers that reflect the target response. In addition, SFD–NCSC–PLS took only less than half the computation time compared with the wavenumber selection methods using either SFD or NCSC. Thus, the proposed SFD–NCSC–PLS is a promising wavenumber selection method.

1. Introduction

The Food and Drug Administration (FDA) [1], the European Medicines Agency (EMA) [2–4], and the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) [5–7] have published a series of guidance about Quality by Design (QbD) and process analytical technology (PAT). QbD and PAT have been recognized as a paradigm to assure a higher quality of drug products. QbD is aimed at enhancing the understanding of a manufacturing process to consistently deliver an intended product quality. In realizing QbD, PAT is useful to design, analyze, and control the manufacturing process through timely measurements of quality attributes. A number of pharmaceutical companies have already introduced QbD and PAT to improve the product quality and the production efficiency [8,9].

Near-infrared (NIR) spectroscopy has been used as a PAT tool to monitor a wide range of quality attributes in a rapid and non-destructive manner [10–12]. Water content in granules during a granulation process [13,14] and drug content in granules during a blending process [15,16] have been investigated as the typical quality attributes to which NIR-based monitoring is applicable. Such NIR-based monitoring requires, in general, a calibration model that describes the relationship between a quality attribute and NIR spectra. To develop a calibration model, partial least squares (PLS) is widely used as a linear regression method that can cope with multicollinearity. The most important performance indices of the calibration model are its estimation accuracy and robustness, which depend on the choice of a regression method, a preprocessing method, calibration samples, and wavenumbers. Various approaches to improve the accuracy and robustness have been extensively studied by focusing on the regression methods [17–19], the calibration sample selection [20–22], the preprocessing methods [23–25], and the wavenumber selection [26].

Wavenumber selection is one of the critical steps in development of an accurate and robust calibration model [27] because using wavenumbers irrelevant to the target quality attributes deteriorates the accuracy and the robustness. It is difficult, however, to accurately identify the irrelevant wavenumbers using spectroscopic knowledge alone because of the complicated nature of an NIR spectrum such as absorption band overlap based on functional groups and/or physical attributes. Although wavenumbers relevant to the target quality attributes can be selected by trial and error, comprehensive evaluation of numerous

* Corresponding author. Tel.: +81 463 31 6954; fax: +81 463 31 6475.
E-mail addresses: miyano.takuya.iu@daiichisankyo.co.jp (T. Miyano), fujiwara.koichi@i.kyoto-u.ac.jp (K. Fujiwara), manabu@human.sys.i.kyoto-u.ac.jp (M. Kano), tanabe.hideaki.xd@daiichisankyo.co.jp (H. Tanabe), nakagawa.hiroshi.w5@daiichisankyo.co.jp (H. Nakagawa), watanabe.tomoyuki.ax@daiichisankyo.co.jp (T. Watanabe), minami.hidemi.-em@daiichisankyo.co.jp (H. Minami).

http://dx.doi.org/10.1016/j.chemolab.2015.09.009
0169-7439/

wavenumbers is impractical. Some statistical wavenumber selection methods were proposed to solve this issue [17,26]. One strategy is to rank individual wavenumbers according to the relevance to a variation in the target quality attribute, and then to select wavenumbers one-by-one. This strategy is represented by PLS-beta and variable influence on projection (VIP). Another strategy is to make groups of wavenumbers and then to select wavenumbers group-by-group. The wavenumber group can be derived as the wavenumbers in a spectral interval or the wavenumbers clustered on the basis of their similarity. Among interval-based methods, the simplest method is interval PLS (iPLS), which divides a whole spectrum into spectral intervals in equal width [28]. Although iPLS was originally developed to visualize useful spectral intervals in terms of estimation accuracy, iPLS has been also used as a wavenumber selection method. Useful spectral intervals can be selected by other interval-based methods such as moving window PLS (MWPLS) [29], changeable size MWPLS (CSMWPLS) [30], searching combination MWPLS (SCMWPLS) [30], and genetic algorism–PLS [31,32]. As a clustering-based method, nearest correlation spectral clustering (NCSC)-based variable selection was proposed [33]. NCSC was originally developed to cluster calibration samples in designing a soft-sensor [34,35]. NCSC clusters wavenumbers into wavenumber groups based on the correlation of spectral intensities among wavenumbers. This correlation-based clustering seems a rational approach because variations in content of one chemical substance induce spectral fluctuations not only at one wavenumber but also at multiple wavenumbers based on the absorption bands of the substance.

A computational load is one of the practical issues in utilizing a wavenumber selection method. Both interval-based and clustering-based wavenumber selection methods can improve the estimation accuracy, but these methods often require a heavy computational load through numerous PLS modeling to determine the tuning parameters. In addition, the estimation accuracy might be deteriorated due to unexpected variations in process characteristics in routine manufacturing [36]. Therefore, ICH, EMA, European Pharmacopoeia, and United States Pharmacopeia require to maintain the calibration model, to continuously verify its estimation accuracy, and to update the calibration model if needed through the life cycle of a drug product [37–40]. When the calibration model needs to be updated, minimizing the workload and the period is practically important so as not to dislocate the production planning. Since the model update often involves wavenumber selection, the wavenumber selection method with a low computational load is crucial to complete the model update in a short period.

To reduce the computational load while improving the estimation accuracy, we previously proposed an interval-based method: spectral fluctuation dividing (SFD)–PLS [41]. SFD derives a spectrum fluctuation profile as the standard deviation of spectral intensities at each wavenumber in a calibration set, and then divides a whole spectrum into multiple spectral intervals at local minimal points of the fluctuation profile. SFD–PLS builds a PLS model using the spectral intensities in each spectral interval, and selects useful spectral intervals in terms of the estimation accuracy. Although SFD–PLS has achieved a lower computational load and higher estimation accuracy than commonly-used iPLS, further computational load reduction has been expected by combining a clustering method with SFD because SFD tends to generate the excessive number of spectral intervals.

The present research proposes to cluster many SFD-based spectral intervals into several spectral interval groups using NCSC. The proposed wavenumber selection method, named SFD–NCSC–PLS, builds a PLS model using the spectral intensities in each spectral interval group, and selects useful groups in terms of the estimation accuracy. To demonstrate the usefulness of the proposed SFD–NCSC–PLS in terms of reducing the computational load and improving the estimation accuracy in comparison with iPLS, SCMWPLS, SFD–PLS, and the NCSC-based method, PLS models were developed using all of these wavenumber selection methods to estimate water and drug contents in granules.

2. Materials and methods

The present study used exactly the same samples, measurement methods, raw data sets, regression method, and evaluation indices as the previous study [41].

2.1. Materials

Granules containing a drug substance (Daiichi-Sankyo, Japan) were used as an analyte in granulation and blending processes. In the granulation process, the drug substance and several excipients were granulated in a fluid bed granulator: NFLO-5 (Freund, Japan), Aeromatic Fielder (GEA Pharma Systems, Belgium), WSG-120 (Powrex, Japan), or GPCG-120 (Glatt, Germany) at 4-kg to 100-kg scales. In the blending process, the granules were blended with a lubricant using a blender: S-3-S (Tsutsui Scientific Instruments, Japan), TCV-10 (Tokuju, Japan), PM-1000 (Bohle, Germany), TB-1200 (Tanico, Japan), or PM-2000 (Kotobuki, Japan) at 0.4-kg to 500-kg scales. Granules during the granulation process and those after the blending process were sampled.

2.2. Near-infrared (NIR) and reference measurement

In the granulation process, the NIR spectrum of the granules was obtained every 1 min during the process using a Fourier-transform NIR spectrometer MPA (Bruker GmbH, Germany) or equivalent Matrix-F (Bruker GmbH, Germany) through a fiber-optic probe mounted in the fluid bed granulator. For the reference measurement, the granules were sampled from the fluid bed granulator every 10 min during granulation and at the end of both spraying and drying. Water content of the sampled granules was measured by the loss on drying (LOD) method using HR73 (Mettler-Toledo, Japan) or equivalent HR83 (Mettler-Toledo, Japan). The water content was associated with the corresponding NIR spectrum. The precision of the LOD method is 0.2% of standard deviation. In the blending process, ca. 0.2 g of the sampled granules was weighed into a dedicated vial, and the vial was measured to obtain the NIR spectrum using MPA. As the reference measurement, the drug content in the corresponding vial was measured by the high performance liquid chromatography (HPLC) method using an Alliance Waters 2695 Separations Module (Waters Corporation, US). The precision of the HPLC method is 0.4% of standard deviation.

The NIR measurement conditions are shown in Table 1. All the NIR spectra were recorded using OPUS 6.5 software (Bruker GmbH, Germany). There is a little difference in the number of the wavenumbers in the two data sets for the water and drug content estimations because periodical calibration of the NIR spectrometer caused a little

Table 1
Experimental conditions to prepare the calibration and validation sets used for constructing and validating PLS models.

	Water content estimation	Drug content estimation
NIR measurement with the diffuse reflectance method		
Wavenumber range [cm^{-1}]	12,500–4000	12,500–4000
Resolution [cm^{-1}]	8	8
No. of wavenumber points	2201	2202
Integration time	8 times	64 times
Spectral preprocessing	First derivative	First derivative + SNV
No. of NIR spectra		
Calibration set	96 (13 batches)	64 (64 batches)
Validation set	58 (7 batches)	40 (40 batches)
Reference measurement		
Measurement method	LOD	HPLC
Range of value [%]		
Calibration set	1.1–17.0	67.7–130.7
Validation set	1.7–15.6	73.1–124.2