Contents lists available at SciVerse ScienceDirect

International Journal of Pharmaceutics

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

Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection

Sanghong Kim[a], Manabu Kano[a,*], Hiroshi Nakagawa[b], Shinji Hasebe[a]

[a] Dept. of Chemical Engineering, Kyoto University, Kyoto 6158510, Japan
[b] Formulation Technology Research Laboratories, Daiichi Sankyo Co., Ltd., Hiratsuka 2540014, Japan

ARTICLE INFO

Article history:
Received 20 June 2011
Received in revised form
20 September 2011
Accepted 2 October 2011
Available online 7 October 2011

Keywords:
PAT
NIR
API content estimation
Locally weighted partial least squares
Wavelength selection

ABSTRACT

Development of quality estimation models using near infrared spectroscopy (NIRS) and multivariate analysis has been accelerated as a process analytical technology (PAT) tool in the pharmaceutical industry. Although linear regression methods such as partial least squares (PLS) are widely used, they cannot always achieve high estimation accuracy because physical and chemical properties of a measuring object have a complex effect on NIR spectra. In this research, locally weighted PLS (LW-PLS) which utilizes a newly defined similarity between samples is proposed to estimate active pharmaceutical ingredient (API) content in granules for tableting. In addition, a statistical wavelength selection method which quantifies the effect of API content and other factors on NIR spectra is proposed. LW-PLS and the proposed wavelength selection method were applied to real process data provided by Daiichi Sankyo Co., Ltd., and the estimation accuracy was improved by 38.6% in root mean square error of prediction (RMSEP) compared to the conventional PLS using wavelengths selected on the basis of variable importance on the projection (VIP). The results clearly show that the proposed calibration modeling technique is useful for API content estimation and is superior to the conventional one.

1. Introduction

In the pharmaceutical industry, the documents on quality by design (QbD) and process analytical technology (PAT) (FDA, 2004; ICH, 2005, 2005, 2008) were published by Food and Drug Administration (FDA) and International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). Since then, online process monitoring and control technologies have attracted much attention. Near infrared spectroscopy (NIRS) is a powerful online monitoring method because of its noninvasiveness and short measuring time; the researches on estimation of many kinds of material attributes such as water content during granulation, blend uniformity, content uniformity and coating thickness by using NIR spectra have been actively conducted (Roggo et al., 2007; Reich, 2005). In this paper, the estimation objective is active pharmaceutical ingredient (API) content in granules for tableting, which is generally not measured. If API content in granules can be estimated by using a PAT tool, the operation condition of the following processes can be changed to make API content in the final products satisfy the specification.

Most of the past researches used linear regression methods such as partial least squares (PLS) to construct estimation models (Moes et al., 2008; Berthiaux et al., 2006; Cogdill et al., 2005; Wu et al., 2009; Li and Worosila, 2005; Berntsson et al., 2002; Sulub et al., 2009; Virtanen et al., 2007). However, linear models cannot always estimate material attributes accurately because physical and chemical properties of a measuring object have a complex effect on NIR spectra, which are inputs of estimation models. To solve this problem, non-linear regression methods such as artificial neural network (ANN) and locally weighted regression (LWR) are used in the agriculture and food industry (Pérez-Marin et al., 2007). Another key issue is how to cope with changes in process characteristics. In the chemical industry, model maintenance is recognized as the most important problem concerning soft-sensors (Kano and Ogawa, 2010). This problem is quite important not only in the chemical industry but also in other industries including the pharmaceutical industry. In this research, LWR is investigated to cope with changes in process characteristics as well as non-linearity. In LWR (Cleveland and Devlin, 1988), a local model is constructed by prioritizing samples in a database according to the similarity between a query sample and them. In general, the similarity is defined on the basis of the Euclidean distance or the Mahalanobis distance (Cleveland and Devlin, 1988; Centner and Massart, 1998). In addition, the similarity which takes account not only of the distance between samples but also of the estimates of output derived by a global model (Wang et al., 1994; Chang et al., 2001) and the similarity based not only on the distance but also on the correlation among variables (Fujiwara et al., 2009, 2010) have been proposed.

[*] Corresponding author. Tel.: +81 0 75 383 2687; fax: +81 0 75 383 2657.
E-mail address: manabu@cheme.kyoto-u.ac.jp (M. Kano).

0378-5173/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpharm.2011.10.007

In this research, in order to construct high performance estimation models, a new similarity measure is proposed and locally weighted partial least squares (LW-PLS) models are constructed. In the proposed method, LW-PLS models are first constructed by using the conventional similarity based on the Euclidean distance, then the LW-PLS models are reconstructed by using the new similarity based on the weighted Euclidean distance. The absolute values of the regression coefficients of the first LW-PLS models are used as the weights for input variables. Furthermore, a statistical wavelength selection method which quantifies the effect of API content and other factors on NIR spectra is proposed. In the present situation, wavelengths are selected by using engineering knowledge and by trial and error. Such conventional approaches are time-consuming and not theoretically well supported. Although advanced methods such as genetic algorithm (Jouen-Rimbauda and Massart, 1995; Arakawa et al., 2011), interval PLS (Nørgaard et al., 2000), and moving window PLS (Jiang et al., 2002) have been proposed, these methods are computationally intensive because they need iterative calculations. The proposed method can select important wavelengths without iterative calculations.

This paper is organized as follows. In Sections 2 and 3, LW-PLS and the proposed wavelength selection method are explained, respectively. In Section 4, the performance of the proposed methods is evaluated through applying them to a blending process. Finally, this research is concluded in Section 5.

2. Locally weighted PLS

The nth sample of input and output variables is denoted by

$$\boldsymbol{x}_n = [x_{n1}, x_{n2}, \ldots, x_{nM}]^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{y}_n = [y_{n1}, y_{n2}, \ldots, y_{nL}]^{\mathrm{T}} \tag{2}$$

where M is the number of input variables, L is the number of output variables and superscript T denotes the transpose of a vector or matrix. $\boldsymbol{X} \in \Re^{N\times M}$ and $\boldsymbol{Y} \in \Re^{N\times L}$ are the input and output variable matrices whose nth rows are $\boldsymbol{x}_n^{\mathrm{T}}$ and $\boldsymbol{y}_n^{\mathrm{T}}$, respectively. N is the number of samples.

In LW-PLS, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are stored in a database. When an output estimation is required for a query sample $\boldsymbol{x}_q$, the similarity ω_n between $\boldsymbol{x}_q$ and $\boldsymbol{x}_n$ is calculated and a local PLS model is constructed by weighting samples with a similarity matrix $\boldsymbol{\Omega} \in \Re^{N\times N}$ defined by

$$\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{\omega}) \tag{3}$$

$$\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_N]^{\mathrm{T}} \tag{4}$$

where diag($\boldsymbol{a}$) denotes a diagonal matrix whose diagonal elements are $\boldsymbol{a}$.

It is important to appropriately define the similarity to achieve the high estimation accuracy by using LW-PLS. In the past researches, many kinds of the similarities have been investigated (Cleveland and Devlin, 1988; Centner and Massart, 1998; Wang et al., 1994; Chang et al., 2001; Fujiwara et al., 2009, 2010). The proposed method utilizes a new similarity measure based on the weighted Euclidean distance

$$d_{2,n} = \sqrt{(\boldsymbol{x}_n - \boldsymbol{x}_q)^{\mathrm{T}}\boldsymbol{\Theta}(\boldsymbol{x}_n - \boldsymbol{x}_q)} \tag{5}$$

where $\boldsymbol{\Theta} \in \Re^{M\times M}$ is a weighting matrix.

$$\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta}) \tag{6}$$

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_M]^{\mathrm{T}} \tag{7}$$

θ_m is defined as the absolute value of the mth variable's regression coefficient of an LW-PLS model in which the normal Euclidean distance

$$d_{1,n} = \sqrt{(\boldsymbol{x}_n - \boldsymbol{x}_q)^{\mathrm{T}}(\boldsymbol{x}_n - \boldsymbol{x}_q)} \tag{8}$$

is used to construct the model. In this research, two types of similarities

$$\omega_{i,n} = \exp\left(-\frac{d_{i,n}}{\sigma_{d_i}\varphi}\right) \quad (i = 1, 2) \tag{9}$$

are investigated, where σ_{d_i} is standard deviation of $d_{i,n}$(n = 1, 2, ..., N) and φ is a localization parameter; the similarity decreases steeply when φ is small and gradually when φ is large. The accuracy of the proposed LW-PLS model is higher than or at least the same as that of linear PLS model if the localization parameter φ is tuned properly because LW-PLS becomes equivalent to linear PLS when $\varphi = \infty$. In other words, LW-PLS includes PLS as a special case. This relationship has some analogy with the relationship between PLS and multiple regression analysis (MRA); PLS includes MRA as a special case. The definition of similarity in Eq. (9) is inspired by the work of (Shigemori et al., 2011), in which θ_m is defined as the absolute value of the mth variable's regression coefficient of a global multiple linear regression model.

The output estimate $\hat{\boldsymbol{y}} \in \Re^{L}$ is calculated as follows.

1. Set i = 1 and determine the number of latent variables R.
2. Set r = 1.
3. Calculate the similarity matrix $\boldsymbol{\Omega}_i$ by using Eqs. (5)–(9).

$$\boldsymbol{\Omega}_i = \mathrm{diag}(\boldsymbol{\omega}_i) \tag{10}$$

$$\boldsymbol{\omega}_i = [\omega_{i,1}, \omega_{i,2}, \ldots, \omega_{i,N}]^{\mathrm{T}} \tag{11}$$

4. Calculate $\boldsymbol{X}_i$, $\boldsymbol{Y}_i$ and $\boldsymbol{x}_{qi}$

$$\boldsymbol{X}_i = \boldsymbol{X} - \boldsymbol{1}_N[\bar{x}_{i,1}, \bar{x}_{i,2}, \ldots, \bar{x}_{i,M}] \tag{12}$$

$$\boldsymbol{Y}_i = \boldsymbol{Y} - \boldsymbol{1}_N[\bar{y}_{i,1}, \bar{y}_{i,2}, \ldots, \bar{y}_{i,L}] \tag{13}$$

$$\boldsymbol{x}_{qi} = \boldsymbol{x}_q - [\bar{x}_{i,1}, \bar{x}_{i,2}, \ldots, \bar{x}_{i,M}]^{\mathrm{T}} \tag{14}$$

$$\bar{x}_{i,m} = \frac{\sum_{n=1}^{N}\omega_{i,n}x_{nm}}{\sum_{n=1}^{N}\omega_{i,n}} \tag{15}$$

$$\bar{y}_{i,l} = \frac{\sum_{n=1}^{N}\omega_{i,n}y_{nl}}{\sum_{n=1}^{N}\omega_{i,n}} \tag{16}$$

where $\boldsymbol{1}_N \in \Re^{N}$ is a vector of ones.

5. Set $\boldsymbol{X}_{i,r} = \boldsymbol{X}_i$, $\boldsymbol{Y}_{i,r} = \boldsymbol{Y}_i$, $\boldsymbol{x}_{qi,r} = \boldsymbol{x}_{qi}$ and $\hat{\boldsymbol{y}}_q = [\bar{y}_{i,1}, \bar{y}_{i,2}, \ldots, \bar{y}_{i,L}]^{\mathrm{T}}$.
6. Derive the rth latent variable of $\boldsymbol{X}_i$

$$\boldsymbol{t}_{i,r} = \boldsymbol{X}_i\boldsymbol{w}_{i,r} \tag{17}$$

where $\boldsymbol{w}_{i,r}$ is the eigenvector of $\boldsymbol{X}_{i,r}^{\mathrm{T}}\boldsymbol{\Omega}_i\boldsymbol{Y}_{i,r}\boldsymbol{Y}_{i,r}^{\mathrm{T}}\boldsymbol{\Omega}_i\boldsymbol{X}_{i,r}$ which corresponds to the maximum eigen value.

7. Derive the rth loading vector of $\boldsymbol{X}_i$

$$\boldsymbol{p}_{i,r} = \frac{\boldsymbol{X}_{i,r}^{\mathrm{T}}\boldsymbol{\Omega}_i\boldsymbol{t}_{i,r}}{\boldsymbol{t}_{i,r}^{\mathrm{T}}\boldsymbol{\Omega}_i\boldsymbol{t}_{i,r}} \tag{18}$$

and the regression coefficient vector

$$\boldsymbol{q}_{i,r} = \frac{\boldsymbol{Y}_{i,r}^{\mathrm{T}}\boldsymbol{\Omega}_i\boldsymbol{t}_{i,r}}{\boldsymbol{t}_{i,r}^{\mathrm{T}}\boldsymbol{\Omega}_i\boldsymbol{t}_{i,r}} \tag{19}$$

8. Derive the rth latent variable of $\boldsymbol{x}_{qi}$

$$t_{qi,r} = \boldsymbol{x}_{qi,r}^{\mathrm{T}}\boldsymbol{w}_{i,r} \tag{20}$$

9. If $i = 2$, change $\hat{\boldsymbol{y}}_q$ to $\hat{\boldsymbol{y}}_q + t_{qr}\boldsymbol{q}_{i,r}$.
10. If $i = 2$ and $r = R$, finish estimation. Otherwise, set

$$\boldsymbol{X}_{i,r+1} = \boldsymbol{X}_{i,r} - \boldsymbol{t}_{i,r}\boldsymbol{p}_{i,r}^{\mathrm{T}} \tag{21}$$

$$\boldsymbol{Y}_{i,r+1} = \boldsymbol{Y}_{i,r} - \boldsymbol{t}_{i,r}\boldsymbol{q}_{i,r}^{\mathrm{T}} \tag{22}$$

$$\boldsymbol{x}_{qi,r+1} = \boldsymbol{x}_{qi,r} - t_{qi,r}\boldsymbol{p}_{i,r} \tag{23}$$

11. If $r = R$, go to the next step. Otherwise, set $r = r + 1$ and return to step 6.
12. Set $i = i + 1$ and return to step 2.

At step 4, the weighted average of each variable is subtracted from each column of $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{x}_q^{\mathrm{T}}$ to make the query sample near to the origin of the multidimensional space. The estimation accuracy may be improved by updating $\boldsymbol{\Omega}$ more than once; however, it makes the computational load heavier. Therefore, $\boldsymbol{\Omega}$ is updated only once in this paper.

3. Statistical wavelength selection

The estimation accuracy strongly depends on the wavelength selection when spectra data are used as model inputs (Andersen and Bro, 2010). Therefore, it is crucial to select an appropriate subset of wavelengths to optimize the model performance. In this research, a statistical wavelength selection method is proposed under the assumption that spectra data are obtained from multiple lots with different API content. This assumption is generally satisfied in practice. The concept of the proposed method is that the selected wavelengths must have the following two features: small absorbance variance in the same lot and large absorbance variance between different lots. Thus, each wavelength is evaluated by the ratio of between-lots variance to within-lot variance.

The nth measurement of absorbance at the mth wavelength in the kth lot is denoted by $x_{nmk}(n = 1, 2, \ldots, N_k,\ m = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, K)$, where N_k, M and K denote the number of samples in the kth lot, the number of wavelengths and the number of lots, respectively.

The proposed statistical wavelength selection procedure is as follows.

1. Calculate mean and variance of x_{nmk} at the mth wavelength in the kth lot.

$$\bar{x}_{*mk} = \frac{1}{N_k}\sum_{n=1}^{N_k} x_{nmk} \tag{24}$$

$$\mathrm{V}_n(x_{nmk}) = \frac{1}{N_k - 1}\sum_{n=1}^{N_k}(x_{nmk} - \bar{x}_{*mk})^2 \tag{25}$$

2. Select the wavelengths at which the following condition is satisfied.

$$\eta = \frac{\mathrm{V}_k(\bar{x}_{*mk})}{\sum_{k=1}^{K}\mathrm{V}_n(x_{nmk})} > \lambda \tag{26}$$

$$\mathrm{V}_k(\bar{x}_{*mk}) = \frac{1}{K - 1}\sum_{k=1}^{K}(\bar{x}_{*mk} - \bar{x}_{*m*})^2 \tag{27}$$

$$\bar{x}_{*m*} = \frac{1}{K}\sum_{k=1}^{K}\bar{x}_{*mk} \tag{28}$$

where λ denotes a threshold for wavelength selection.

Table 1
Experimental data.

Lot number	Number of samples	Mean of API content [%]
1	90	68.1
2	86	83.0
3	100	88.7
4	20	97.4
5	10	98.6
6	90	107.7
7	90	113.8
8	90	128.3
9	10	73.9
10	10	94.0
11	10	96.8
12	10	98.3
13	10	98.8
14	10	99.5
15	10	100.1
16	10	122.9
17	10	96.0
18	10	100.0

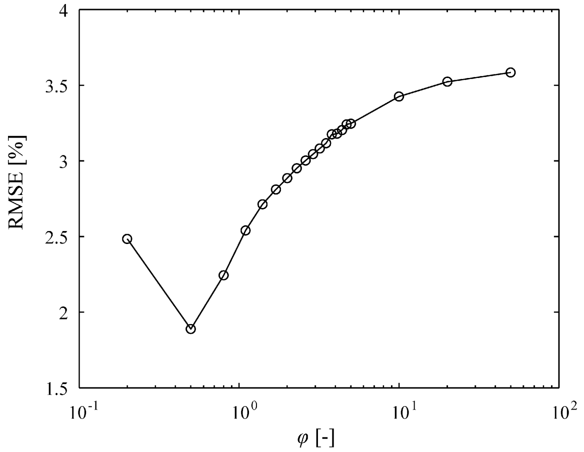

Fig. 1. The relationship between RMSE and the localization parameter φ when LW-PLS 2 is applied (λ = 10 and R = 9).

When the effect of the difference in API content in the same lot on the spectra is negligible, $\mathrm{V}_n(x_{nmk})$ indicates the effect of the factors other than API content on the spectra. Therefore, the wavelengths with small within-lot variance $\mathrm{V}_n(x_{nmk})$ should be selected. In addition, $\mathrm{V}_k(\bar{x}_{*mk})$ indicates the effect of API content on the spectra and the wavelengths with large between-lots variance $\mathrm{V}_k(\bar{x}_{*mk})$ should be selected. Thus, the suitable wavelengths for estimation can be selected on the basis of η.

Table 2
Search range of the parameters.

Parameter	Search range
φ	0.2, 0.5, 0.8, 1.1, 1.4, 1.7, 2, 2.3, 2.6, 2.9, 3.2, 3.5, 3.8, 4.1 4.4, 4.7, 5, 10, 20, 30, 50
λ	0, 1, 2, 3, 4, 5, 10, 15, 20
μ	0, 0.5, 0.8, 1, 2, 3, 4, 5
R	1, 2, 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15

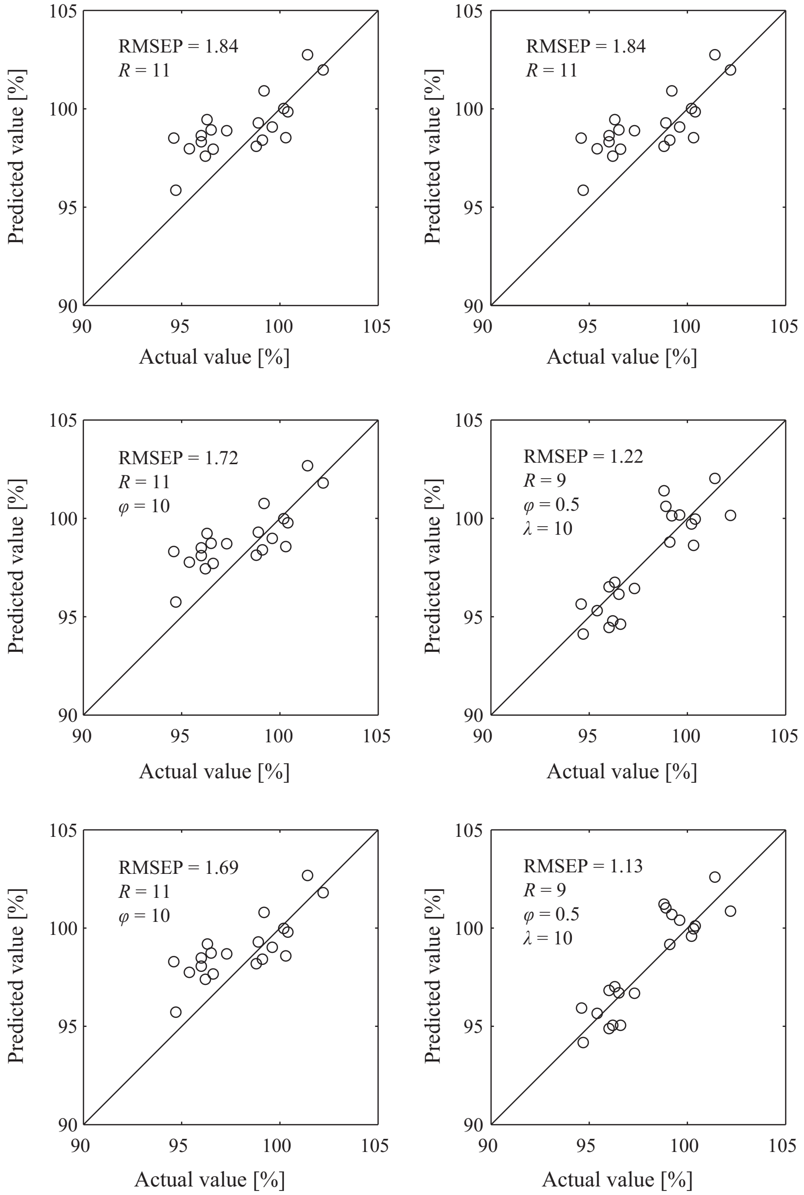

Fig. 2. Results of model validation. (Left-top) case 1: PLS with wavelengths selected on the basis of VIP, (right-top) case 2: PLS with wavelengths selected by the proposed method, (middle-top) case 3: LW-PLS 1 with wavelengths selected on the basis of VIP, (middle-top) case 4: LW-PLS 1 with wavelengths selected by the proposed method, (left-bottom) case 5: LW-PLS 2 with wavelengths selected on the basis of VIP, and (right-bottom) case 6: LW-PLS 2 with wavelengths selected by the proposed method.

4. Application to real process data

4.1. Experimental

The target drug products consist of six components. Eighteen blending experiments were conducted with different API content using a 3 L scale V-blender (Tsutsui Scientific Instruments Co., Ltd.). After each blending experiment, the granules for tableting were taken out and 200 mg of the granules were set in vials, NIR spectra (2203 points in 800–2500 nm) were measured with MPA (Bruker Optics K. K.), and API content was measured with Alliance Waters 2690 Separations Module (Waters Corporation). The overview of

Table 3
Comparison of the calibration modeling techniques.

Case	Model	Wavelength selection	♯	λ	μ	φ	R	RMSE	RMSEP
1	PLS	VIP	2087	–	0	–	11	2.15	1.84
2	PLS	Proposed	2087	0	–	–	11	2.15	1.84
3	LW-PLS 1	VIP	2087	–	0	10	11	2.14	1.71
4	LW-PLS 1	Proposed	259	10	–	0.5	9	1.96	1.22
5	LW-PLS 2	VIP	2087	–	0	10	11	2.13	1.69
6	LW-PLS 2	Proposed	259	10	–	0.5	9	1.89	1.13

♯: The number of selected wavelengths.

the experimental data is shown in Table 1. In this study, the data of lots from 1 to 8 are the calibration set, the data of lots from 9 to 16 are the test set, and the data of lots 17 and 18 are the prediction set.

4.2. Data analysis

Before the detailed data analysis, non-linearity between input and output variables was evaluated with the normal probability plot of the residuals, which can check whether the residuals are normally distributed or not. When linear PLS model was used, the normal probability plot showed that the residuals were not normally distributed. This result suggested that linear PLS model was not suitable to estimate API content and that LW-PLS might improve estimation accuracy. The detailed comparison procedure of modeling methods and variable selection methods is as follows.

1. Preprocessing
   Apply first order differential using Savitsky–Golay filter (Savitzky and Golay, 1964) and standard normal variate (SNV) to NIR spectra data. By using Savitsky–Golay filter, the effect of the noise on NIR spectra can be reduced. SNV can correct the variance in light path length caused by changes in the particle size and density (Barnes et al., 1989). In this application, the window size and the polynomial order in Savitsky–Golay filter were 117 and 5, respectively.
2. Wavelength selection
   Use the wavelengths selected by the proposed method or on the basis of variable importance on the projection (VIP) (Eriksson et al., 2001), a widely used variable selection measure, as model inputs. In each method, wavelengths which has larger η or VIP than a threshold are selected.
3. Model construction
   Construct estimation models by using conventional PLS, LW-PLS without updating Ω (LW-PLS 1) or LW-PLS with updating Ω (LW-PLS 2).

Six estimation models were constructed with respect to the selections in steps 2 and 3. Model parameters in each model, i.e. the localization parameter φ, the threshold for the proposed variable selection index λ, the threshold for VIP μ, and the number of latent variables R, were determined by using the calibration set (data of lots from 1 to 8) and the test set (data of lots from 9 to 16). Estimation models were constructed with different parameter sets, and API content of the test set was estimated by the models, then the parameters which derived the minimum estimation error were selected. The search range of the parameters is shown in Table 2.

4.3. Results and discussion

Table 3 shows the selected parameters, root mean square error of parameter tuning (RMSE) and root mean square error of prediction (RMSEP). Fig. 1 shows the relationship between RMSE and φ when LW-PLS 2 is applied (λ = 10 and R = 9); RMSE was large when φ was too small or too large. Overfitting occurred when φ was too small, and models were unable to cope with non-linearity between input and output variables when φ was too large. In addition, model validation results are shown in Fig. 2. When conventional PLS was used (cases 1 and 2), RMSEPs were the same because the proposed wavelength selection method and the VIP based method selected all wavelengths. The proposed wavelength selection method selected 259 wavelengths, which had index η larger than 10, when LW-PLS 1 (case 4) and LW-PLS 2 (case 6) were used. On the other hand, the VIP based method selected all wavelengths, when LW-PLS 1 (case 3) and LW-PLS 2 (case 5) were used. VIP (R = 11) and preprocessed API spectrum are shown in Fig. 3 (top). VIP mostly has a correlation with absorbance values of API spectra and it can be expected that RMSE becomes small when the threshold for VIP μ is large. However, the best RMSE was obtained by using all wavelengths when VIP was applied. In addition, the index η and preprocessed API spectrum are shown in Fig. 3 (middle), and $V_k(\bar{x}_{*mk})$ and $\sum_{k=1}^{K} V_n(x_{nmk})$ in Eq. (26) are shown in Fig. 3 (bottom). η dose not have a correlation with absorbance values of API spectra because η takes account not only of the effect of API content on NIR spectra but also of the effect of other factors on NIR spectra.The wavelengths around 1910 and 1970 nm were not selected although peak absorbance values of API and $V_k(\bar{x}_{*mk})$, which is the effect of API content on NIR spectra, were large. This is because $\sum_{k=1}^{K} V_n(x_{nmk})$, which is the effect of other factors on NIR spectra, was also large at these wavelengths. On the contrary, the wavelengths around 1120 and 1190 nm were selected although peak absorbance values of API and $V_k(\bar{x}_{*mk})$ were small. In addition, loading values for each input variable showed similar tendency as η. Thus, η should be suitable for variable selection. By using the proposed wavelength selection method, RMSEP was improved by 28.7% when LW-PLS 1 was used (cases 3 and 4) and by 33.1% when LW-PLS 2 was used (cases 5 and 6). Moreover, LW-PLS 2 (cases 5 and 6) was superior to PLS and LW-PLS 1 (cases 1–4). With the proposed wavelength selection method (cases 2, 4 and 6), LW-PLS 2 derived 38.6% and 7.4% less RMSEP than PLS (case 2) and LW-PLS 1 (case 4), respectively. The results of the case study demonstrate the usefulness of the proposed wavelength selection method and LW-PLS 2.

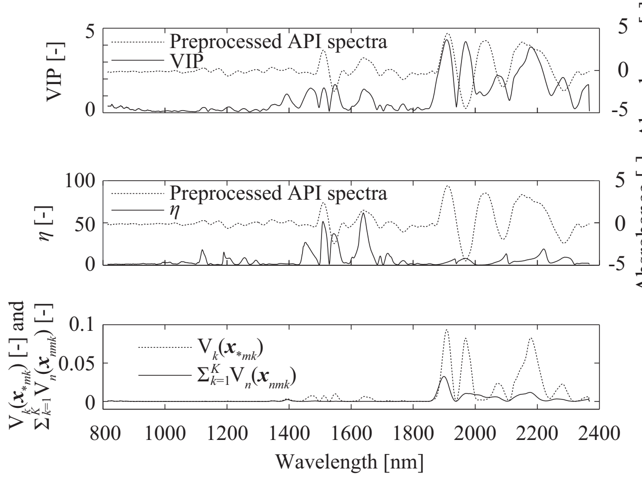

Fig. 3. (Top) VIP and preprocessed API spectra, (middle) wavelength selection index η and preprocessed API spectra, and (bottom) $V_k(\bar{x}_{*mk})$ and $\sum_{k=1}^{K} V_n(x_{nmk})$.

5. Conclusions

Locally weighted partial least squares, which utilizes the similarity based on the weighted Euclidean distance, was proposed to estimate API content in a blending process. The regression coefficients of the LW-PLS model using the normal Euclidean distance were used as weights for input variables. In addition, the statistical wavelength selection method which quantifies the effect of API content and other factors on NIR spectra was proposed. By using the proposed methods, the estimation accuracy was improved by 38.6% in RMSEP compared to the conventional PLS using wavelengths selected on the basis of variable importance on the projection. The results clearly show that the proposed calibration modeling technique is useful for API content estimation and is superior to the conventional one.

Acknowledgements

This work was partially supported by Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (C) 21560793.

References

Andersen, C., Bro, R., 2010. Variable selection in regression—a tutorial. J. Chemom. 24, 728–737.

Arakawa, M., Yamashita, Y., Funatsu, K., 2011. Genetic algorithm-based wavelength selection method for spectral calibration. J. Chemom. 25, 10–19.

Barnes, R., Dhanoa, M., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777.

Berntsson, O., Danielsson, L.G., Lagerholm, B., Folestad, S., 2002. Quantitative in-line monitoring of powder blending by near infrared reflection spectroscopy. Powder Technol. 123, 185–193.

Berthiaux, H., Mosorov, V., Tomczak, L., Gatumel, C., Demeyre, J., 2006. Principal component analysis for characterising homogeneity in powder mixing using image processing techniques. Chem. Eng. Process. 45, 397–403.

Centner, V., Massart, D., 1998. Optimization in locally weighted regression. Anal. Chem. 70, 4206–4211.

Chang, S., Baughman, E., McIntosh, B., 2001. Implementation of locally weighted regression to maintain calibrations on FT-NIR analyzers for industrial processes. Appl. Spectrosc. 55, 1199–1206.

Cleveland, W., Devlin, S., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83, 596–610.

Cogdill, R., Anderson, C., Delgado-Lopez, M., Molseed, D., Chisholm, R., Bolton, R., Herkert, T., Afnan, A., Drennen III, J., 2005. Process analytical technology case study part I: feasibility studies for quantitative near-infrared method development. AAPS PharmSciTech 6, 262–272.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S., 2001. Multi- and Megavariate Data Analysis. Principles and Applications. Umetrics Academy.

FDA 2004. Pharmaceutical cGMPs for the 21st century—a risk-based approach final report.

Fujiwara, K., Kano, M., Hasebe, S., 2010. Development of correlation-based clustering method and its application to software sensing. Chemom. Intell. Lab. Syst. 101, 130–138.

Fujiwara, K., Kano, M., Hasebe, S., Takinami, A., 2009. Soft-sensor development using correlation-based just-in-time modeling. AIChE J. 55, 1754–1765.

ICH 2005. ICH harmonised tripartite guideline—pharmaceutical development Q8 (R2).

ICH 2005. ICH harmonised tripartite guideline—quality risk management Q9.

ICH 2008. ICH harmonised tripartite guideline—pharmaceutical quality system Q10.

Jiang, J.H., James, R., Siesler, B., Ozaki, Y., 2002. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. Anal. Chem. 74, 3555–3565.

Jouen-Rimbauda, D., Massart, D.L., 1995. Genetic algorithms as a tool for wavelength selection in multivariate calibration. Anal. Chem. 67, 4295–4301.

Kano, M., Ogawa, M., 2010. The state of the art in chemical process control in Japan: good practice and questionnaire survey. J. Process Control 20, 969–982.

Li, W., Worosila, G., 2005. Quantitation of active pharmaceutical ingredients and excipients in powder blends using designed multivariate calibration models by near-infrared spectroscopy. Int. J. Pharm. 295, 213–219.

Moes, J.J., Ruijken, M.M., Gout, E., Frijlink, H.W., Ugwoke, M.I., 2008. Application of process analytical technology in tablet process development using NIR spectroscopy: blend uniformity, content uniformity and coating thickness measurements. Int. J. Pharm. 357, 108–118.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J., Munck, L., Engelsen, S., 2000. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. 54, 413–419.

Pérez-Marin, D., Garrido-Varo, A., Guerrero, J., 2007. Non-linear regression methods in NIRS quantitative analysis. Talanta 72, 28–42.

Reich, G., 2005. Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications. Adv. Drug Deliv. Rev. 57, 1109–1143.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., Jent, N., 2007. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. J. Pharm. Biomed. Anal. 44, 683–700.

Savitzky, A., Golay, M., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.

Shigemori, H., Kano, M., Hasebe, S., 2011. Optimum quality design system for steel products through locally weighted regression model. J. Process Control 21, 293–301.

Sulub, Y., Wabuyele, B., Gargiulo, P., Pazdan, J., Cheney, J., Berry, J., Gupta, A., Shah, R., Wu, H., Khan, M., 2009. Real-time on-line blend uniformity monitoring using near-infrared reflectance spectrometry: a noninvasive off-line calibration approach. J. Pharm. Biomed. Anal. 49, 48–54.

Virtanen, S., Antikainen, O., Yliruusi, J., 2007. Uniformity of poorly miscible powders determined by near infrared spectroscopy. Int. J. Pharm. 345, 108–115.

Wang, Z., Isaksson, T., Kowalski, B., 1994. New approach for distance measurement in locally weighted regression. Anal. Chem. 66, 249–260.

Wu, H., Tawakkul, M., White, M., Khan, M., 2009. Quality-by-design (QbD): an integrated multivariate approach for the component quantification in powder blends. Int. J. Pharm. 372, 39–48.