Front. Optoelectron. 2013, 6(2): 216–223
DOI 10.1007/s12200-013-0320-3

RESEARCH ARTICLE

Particle size regression correction for NIR spectrum based on the relationship between absorbance and particle size

Jinrui MI[1,2], Luda ZHANG[2], Longlian ZHAO[1], Junhui LI (✉)[1]

1 College of Information and Electrical Engineering, China Agriculture University, Beijing 100083, China
2 College of Science, China Agriculture University, Beijing 100083, China

Abstract Based on the effect of sample size on the near-infrared (NIR) spectrum, the absorbance (log(R)) in any wavelength is divided into two parts, and one of them is defined as non-particle-size-related spectrometry (nPRS) because it is not influenced by particle size. To study the relationship between the absorbance and particle size, the experiment material including nine samples with different particle size was used. According to the regression analysis, the relationship was studied as the reciprocal regression model, $y = a + bx + c/x$. Meanwhile, the model divides absorbance into two parts, one of them forms nPRS. According to the nPRS, a new correction method, particle size regression correction (PRC) was introduced. In discriminate analysis, the spectra from three different samples (rice, glutinous rice and sago), pretreated by PRC, could be directly and accurately distinguished by principal component analysis (PCA), while by the traditional correction method, such as multiplicative signal correction (MSC) and standard normal variate (SNV), could not do that.

Keywords near-infrared diffuse reflectance spectrometry (NIRDRS), regression analysis, non-particle-size-related spectrum (nPRS), particle-size regress correction (PRC)

1 Introduction

Near-infrared diffuse reflectance spectrometry (NIRDRS) is an important technique for the measuring and analyzing of sample. NIRDRS has an advantage that the sample need not complex pretreatment, but the measurement of NIRDS is affected by physical properties of the sample, such as size and shape, packing, surface, and color [1]. To overcome the negative effect, methods for scatter corrections have been developed, for example, multiplicative signal correction(MSC) [2], piecewise multiplicative signal correction (PMSC) [3], and standard normal variate (SNV) [4]. These methods have been widely used in varied fields, such as agriculture, medical science, and food science [5–7]. But all these methods can only be employed for scatter corrections of spectra without any information about the sample. Due to the significant influence of physical property of sample on the NIR spectra, the analysis result could not be satisfied by using these methods.

In 2003, extended multiplicative signal correction (EMSC) for diffused spectra was introduced by Martens et al. [8]. And it has been applied to the research of agricultural and food science [9,10]. Different from these traditional methods described above, a near-ideal chemical spectrum was used in the correction process of the EMSC. And then, Liu et al. compared the EMSC with these traditional methods, such as MSC and SNV. It was found that EMSC-pretreated data not only well accessed the chemical information, but also consistently led to the overall best prediction of the chemical composition [11]. Another analysis method for diffused spectra is based on the rule of photon migration in biologic tissue based on Monte Carlo simulation [12]. This method has been widely used in medicine [13], as well as in agriculture science by Wang et al. [14–16]. Both methods could restrain the influence of physical property effectively, and the result is satisfactory. But the disadvantage is the limited range of application. The difficulty of obtaining chemical spectrum makes EMSC only suitable for the sample with a simple chemical structure, while the rule of photon migration is difficult to be applied in the conventional NIR analysis. In this study, a correction method, simple in principle and easy to realize, was introduced.

Received February 20, 2013; accepted March 18, 2013

E-mail: caunir@cau.edu.cn

2 Non-particle-size related spectrum (nPRS)

The research of this thesis is based on a premise hypothesis that the absorbance is the sum of information related to physical factor and that not related to physical factor in any wavelength, as shown in Eq. (1).

$$A_\lambda = I_{\lambda\_\text{phy}} + I_{\lambda\_\text{nonphy}} = f(x) + g(.), \quad (1)$$

where I_{λ_phy} and $I_{\lambda\_\text{nonphy}}$ represent two types of information at λ (cm^{-1}). The I_{λ_phy} is associated with the physical properties of the sample, expressing as $f(x)$, x of which is the physical properties, and in this research x means the particle size, while the $I_{\lambda\_\text{nonphy}}$ is not related to the physical properties, expressing as $g(.)$.

There are three kinds of agriculture produce (rice, glutinous rice and sago), as experiment materials in the research, which are dried, milled. And then the materials' flour is sifted in turn by 10 test sieves, the mesh sizes of which are different. The sieve mesh and the mesh size are one-to-one relationship (Table 1). The materials' flour is sifted in turn by 10 test sieves, and there are 11 samples for each material. Two of them with particle size of more than #20 (sieve mesh) and less than #200 (sieve mesh) are not used in the research, because it is difficult to measure size. In this way, each of materials is separated into nine samples with different particle sizes, in all 27 samples.

Table 1 Relationship between sieve mesh and mesh size

sieve mesh	mesh size/mm	sieve mesh	mesh size/mm
20	0.71	120	0.125
40	0.45	140	0.105
60	0.28	160	0.098
80	0.18	180	0.09
100	0.154	200	0.076

The spectra of 27 samples are obtained in k/s mode from 4000 to 12000 cm^{-1} on a Bruker, MPA FT-NIR instrument, equipped with an integral sphere, and a PbS detector. The nominal resolution is 8 cm^{-1} and 64 scans were co-added. The data interval is 4 cm^{-1}. The samples are measured in a sample cup, and the powder was slightly compressed with a spatula before the measurement. The result spectrum saves as log(R). For reducing the extraneous factor influence, each of samples collects five spectra and calculates one mean spectrometer, in all 27.

The trend of the glutinous rice's k/s distribution is showed in Fig. 1 at two wavelengths (7726 and 4127 cm^{-1}), randomly chosen. The distribution trend would be described by some mathematical model at these wavelengths. Once the model is selected, it will be applied for the full spectral region to study whether it could be suitable. The trend has a "√" shape in Fig. 1(a), while a log or liner trends are showed in Fig. 1(b). Therefore, four types of regression models are selected for analysis. They are polynomial model (Eq. (2)), logistic model (Eq. (3)), reciprocal model (Eq. (4)), and exponential model (Eq. (5)).

$$y = a_1 + a_2 x + a_3 x^2, \quad (2)$$

$$y = a + bx + c\ln x, \quad (3)$$

$$y = a + bx + c/x, \quad (4)$$

$$y = a + bx + ce^x. \quad (5)$$

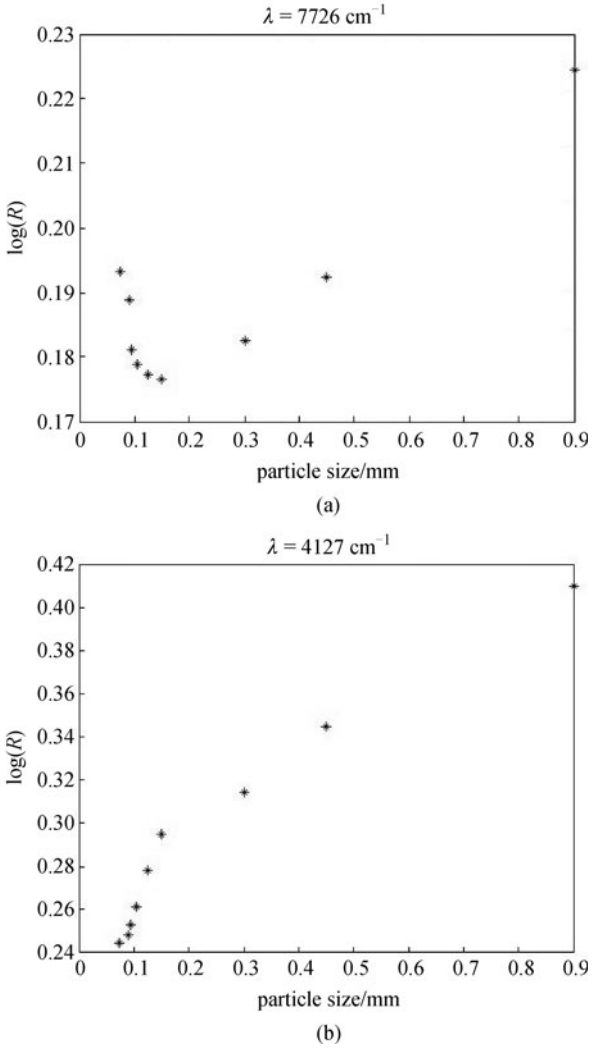

Fig. 1 log(R) distribution at 7726 (a) and 4127 (b) cm^{-1}

Figure 2 is the results (R^2 and RMSE (root-mean-square error)) of four regression models in the infrared wavelength range for glutinous rice. The R^2 of reciprocal

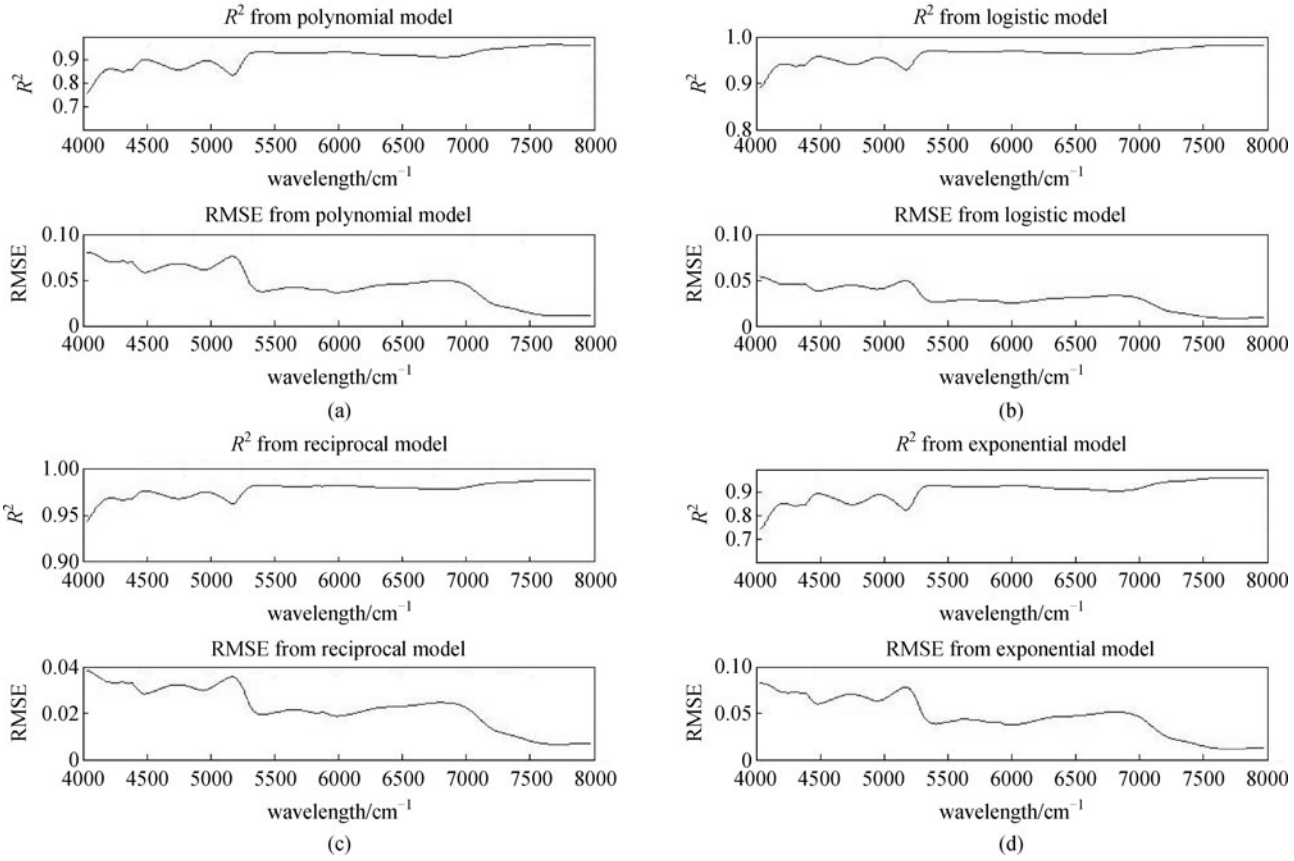

Fig. 2 R^2 and RMSE of four regression models by using quadric polynomial model (a); logistic model (b); reciprocal model (c) and exponential model (d)

regression model reaches more than 0.94 in the 4000–8000 cm^{-1}, higher than the other three models, while the RMSE of it is lower than the others. The same regression analysis is done to the other two materials, and the results are similar. The information is shown in Table 2. It follows that the regression results of the reciprocal regression are the best in all.

Based on the hypothesis model before, in the model, $y = a + bx + c/x$, the y, absorbency, is divided into two parts. One of them, coefficient a in the model, is not associated with x, particle size, which is not in line with the physics information, mentioned in the hypothesis model before. Hence this part is called non-particle-size-related information. And coefficient a in all wave band constitutes non-particle-size-related spectrum (nPRS). Likewise, the other part which is due to x, is named as particle-size-related information, and constitute particle-size-related spectrum (PRS).

To calculate the nPRS, n spectra (column vector $\boldsymbol{Y}_i$, i = 1,2,…,n) of simples with different particle size and the particle size, x_i, should be collected ($n \geqslant 3$). Each of them is respectively stored in the spectra matrix $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n]$ and particle size vector, $\boldsymbol{x} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_n]$. According to the format of regression model, $y = a + bx + c/x$, the regress matrix, $\boldsymbol{R} = [1; \boldsymbol{x}; 1/\boldsymbol{x}]$, is created. Therefore, the model can be written as

$$\boldsymbol{Y} = \boldsymbol{AR}, \tag{6}$$

where matrix $\boldsymbol{A} = [\boldsymbol{a}\ \boldsymbol{b}\ \boldsymbol{c}]$, is m-by-3 (m = the length of spectrum).

A versatile solution for the model is the least squares estimator (Eq. (7)).

$$\boldsymbol{A} = \boldsymbol{YR}^{\mathrm{T}}(\boldsymbol{RR}^{\mathrm{T}})^{-1}. \tag{7}$$

The first column elements of matrix $\boldsymbol{A}$ is the nPRS of samples, while the column vector, $\boldsymbol{b}$ and $\boldsymbol{c}$, is used to calculate PRS, by $bx_i + c/x_i$. The glutinous rice's nPRS, original spectrum and PRS are shown in Fig. 3.

By the result of the research above, the NIR spectrum could be divided into two parts, the non-particle-size-related spectrum and the particle-size related spectrum. The relationship between the particle size (x) and particle-size related spectrum could be expressed as a function model, $y = bx + c/x$. Therefore, the NIR k/s spectrum (row vector $\boldsymbol{z}$) could be express as (Eq. (8))

$$\boldsymbol{z}_i = \boldsymbol{a} + \boldsymbol{b}_i x + \frac{\boldsymbol{c}_i}{x}, \tag{8}$$

where the row vector $\boldsymbol{a}$ expresses the non-particle-size-related spectrum, the row vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ express the coefficients of particle-size-related spectrum, and i denotes the ith spectrum.

Table 2 R^2 and RMSE from different regression models

rice	R^2			RMSE		
	max	min	mean	max	min	mean
Eq. (2)	0.9667	0.7578	0.9141	0.0799	0.0114	0.0449
Eq. (3)	0.9825	0.8903	0.9617	0.0537	0.0082	0.0300
Eq. (4)	0.9886	0.9440	0.9788	0.0384	0.0066	0.0223
Eq. (5)	0.9650	0.7437	0.9089	0.0822	0.0116	0.0463
glutinous rice	R^2			RMSE		
	max	min	mean	max	min	mean
Eq. (2)	0.9473	0.7409	0.8846	0.0844	0.0099	0.0440
Eq. (3)	0.9766	0.8825	0.9513	0.0569	0.0066	0.0285
Eq. (4)	0.9848	0.9499	0.9775	0.0371	0.0053	0.0192
Eq. (5)	0.9439	0.7267	0.8774	0.0867	0.0102	0.0454
sago	R^2			RMSE		
	max	min	mean	max	min	mean
Eq. (2)	0.9627	0.8563	0.9444	0.0215	0.0057	0.0121
Eq. (3)	0.9917	0.9527	0.9824	0.0123	0.0045	0.0063
Eq. (4)	0.9976	0.9648	0.9914	0.0052	0.0030	0.0038
Eq. (5)	0.9592	0.8457	0.9398	0.0222	0.0058	0.0126

3 Particle size regression correction (PRC)

Based on the particle size regression model, a method is proposed for the scattering correction, named as particle size regression correction (PRC).

If the coefficients in Eq. (8) had been known theoretically, or estimated perfectly, then the PRC correction

$$\boldsymbol{z}_{i\_\text{corrected}} = \boldsymbol{z}_i - \boldsymbol{b}_i \cdot x - \boldsymbol{c}_i / x, \tag{9}$$

would remove the particle-size related information, yielding corrected spectrum with only non-particle-size related information left: $z_{i\_\text{corrected}} \approx$ non-particle-size related spectrum. Ideally, it would then be advantageous to replace the measured spectrum z_i with $z_{i\_\text{corrected}}$ in subsequent multivariate calibration, since the latter reduce the particle-size influence to the spectrum.

It is assumed that an ideal spectrum is divided into particle-size related spectrum and non-particle-size related spectrum, and it has a good linear relation with any other sample spectrum. It means that its vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ have good linear relation with those from the other sample spectra, shown as (Eqs. (10)–(12))

$$\boldsymbol{a}_j = \alpha_{1j} \cdot \boldsymbol{a} + \alpha_{2j}, \tag{10}$$

$$\boldsymbol{b}_j = \beta_{1j} \cdot \boldsymbol{b} + \beta_{2j}, \tag{11}$$

$$\boldsymbol{c}_j = \chi_{1j} \cdot \boldsymbol{c} + \chi_{2j}, \tag{12}$$

where j denotes the jth spectrum.

Taking Eqs. (10)–(12) into the particle-size-related spectrum model, Eq. (8) can be rewritten as (Eq. (13))

$$\boldsymbol{z}_i = (\alpha_{1i} \cdot \boldsymbol{a} + \alpha_{2i}) + (\beta_{1i} \cdot \boldsymbol{b} + \beta_{2i}) \cdot x + (\chi_{1i} \cdot \boldsymbol{c} + \chi_{2i})/x, \tag{13}$$

which can be simplified as PRC model and is shown as (Eq. (14))

$$\boldsymbol{z}_i = \alpha \cdot \boldsymbol{a} + \beta \cdot \boldsymbol{b} + \chi \cdot \boldsymbol{c} + \delta, \tag{14}$$

where

$$\begin{cases} \alpha = \alpha_{1i}, \\ \beta = \beta_{1i} x, \\ \chi = \chi_{1i}/x, \\ \delta = \alpha_{2i} + \beta_{2i} x + \chi_{2i}/x. \end{cases}$$

The vectors $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ could be estimated from ideal spectrum, mentioned before. Once these vectors are estimated, the PRC model can be rewritten as the matrix form and its least square estimator is shown as (Eq. (15))

$$\boldsymbol{z}_i = \boldsymbol{M} \cdot \begin{bmatrix} \alpha \\ \beta \\ \chi \\ \delta \end{bmatrix} \xrightarrow{\text{least squares estimate}}$$

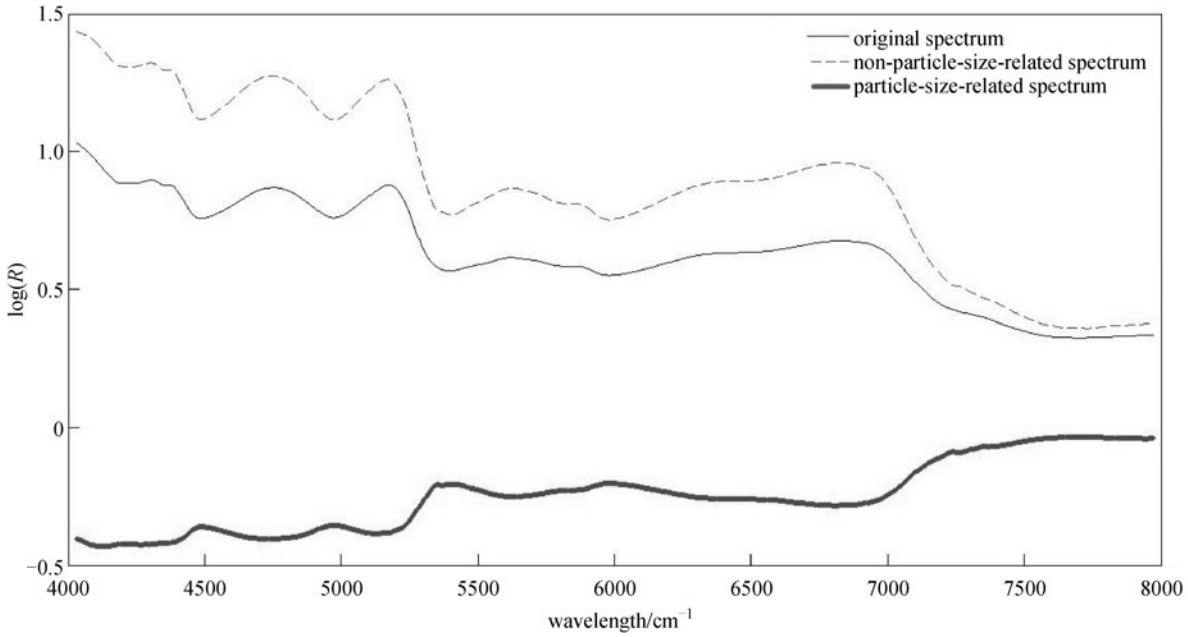

Fig. 3 Results of nPRS, mean k/s spectrum and PRS. nPRS: non-particle-size-related spectrum; PRS: particle-size-related spectrum

$$\begin{bmatrix} \alpha \\ \beta \\ \chi \\ \delta \end{bmatrix} = (\boldsymbol{M}' \cdot \boldsymbol{M})^{-1} \cdot \boldsymbol{M}' \cdot \boldsymbol{z}_i, \tag{15}$$

where

$$\boldsymbol{M} = [\boldsymbol{a}\ \boldsymbol{b}\ \boldsymbol{c}\ 1].$$

And then the PRC correction result, Eq. (16), can be obtained

$$\boldsymbol{z}_{i\_\text{corrected}} = (\boldsymbol{z}_i - \beta \cdot \boldsymbol{b}_i - \chi \cdot \boldsymbol{c}_i - \delta)/\alpha, \tag{16}$$

which removes the difference, due to the particle size.

The way, particle size regression correction (PRC), is used in the analysis to the spectra from experimental materials (rice, glutinous rice and sago). The spectra are pretreated by different correction methods, MSC, SNV and PRC. Thereinto, the coefficient vectors ***a***, ***b*** and ***c*** are the mean vectors from the three kinds of experiment materials, shown as Table 3. The pretreated spectra and first-derivative spectra are shown in Fig. 4. It is shown that the PRC could smooth out the difference between spectra, and the effect of correction from PRC is better than that from MSC and SNV by direct observations.

After visual analysis, the principal component analysis (PCA) is introduced in this research. The first-derivative spectra are projected into low dimensional PCA space. And the spatial distribution maps of the first principal component are shown in Fig. 5.

It is shown that the first-derivative original spectra, analyzed by PCA, cannot be discriminated by category from the first principal component, neither the spectra pretreated by MSC or SNV can. But Fig. 5(d) shows that the spectra, treated by PRC, could be directly discriminated by the first principal component.

Table 3 Vectors ***a***, ***b*** and ***c***

vector	meaning
a	$\boldsymbol{a} = (\boldsymbol{a}_{\text{rice}} + \boldsymbol{a}_{\text{glutinous\_rice}} + \boldsymbol{a}_{\text{sago}})/3$
b	$\boldsymbol{b} = (\boldsymbol{b}_{\text{rice}} + \boldsymbol{b}_{\text{glutinous\_rice}} + \boldsymbol{b}_{\text{sago}})/3$
c	$\boldsymbol{c} = (\boldsymbol{c}_{\text{rice}} + \boldsymbol{c}_{\text{glutinous\_rice}} + \boldsymbol{c}_{\text{sago}})/3$

By comparing with the effects from other correction method, it shows that PRC could not only make the spectra's discrepancies from the particle size decrease, but also effectively improve the identification result to the experimental materials (rice, glutinous rice and sago). In addition, the PCA to all spectra, pretreated by PRC, without first-derivative is done, as Fig. 6. The result shows it is discriminated completely into three categories by first two principal components information at least. At the same time, the discrimination results from spectra, not pretreated by PRC, without first-derivative are worse. Additional other estimation solutions of coefficient vectors look worth of the further research.

4 Conclusions

NIR diffuse reflection is an important method for collecting the spectra of solid samples. Different physical factors, such as particle size, shape, and color, contribute to

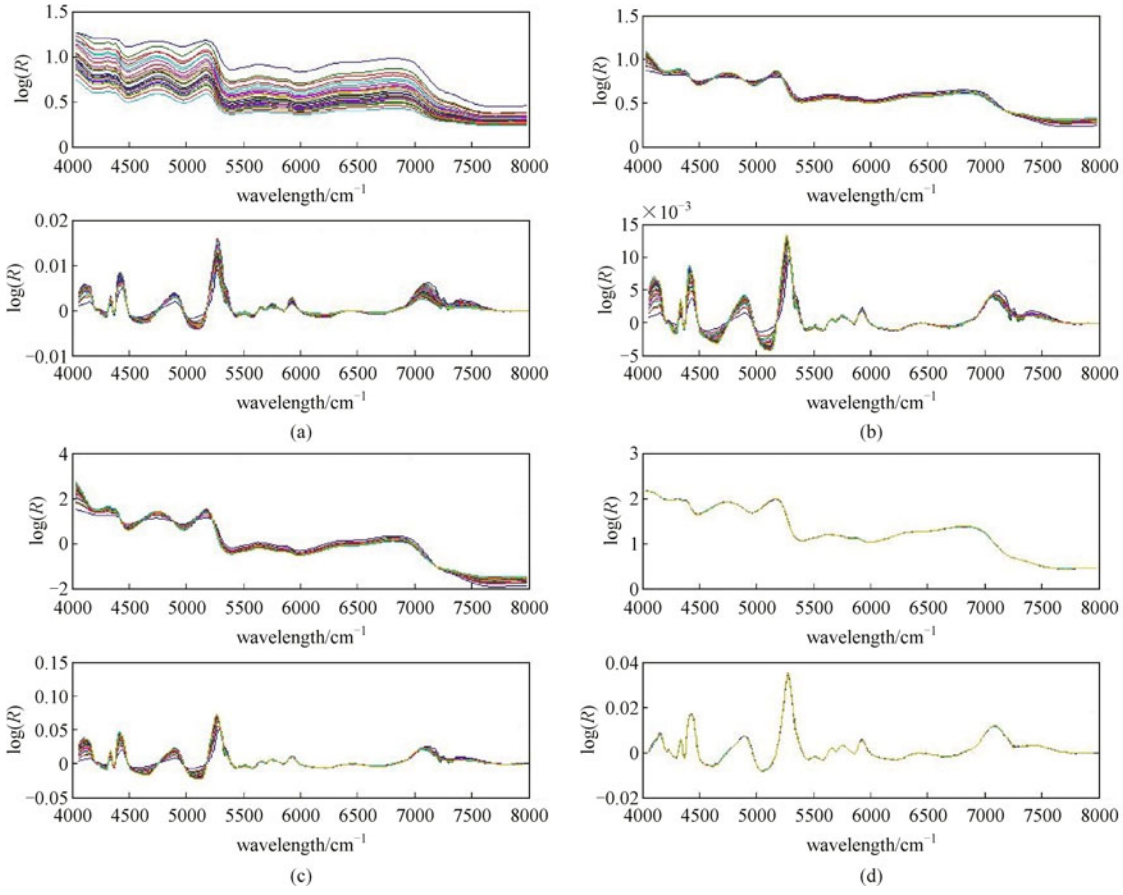

Fig. 4 NIR spectra (up) and first-derivative spectra (down) from original spectra (a); MSC (b); SNV (c) and PRC (d)

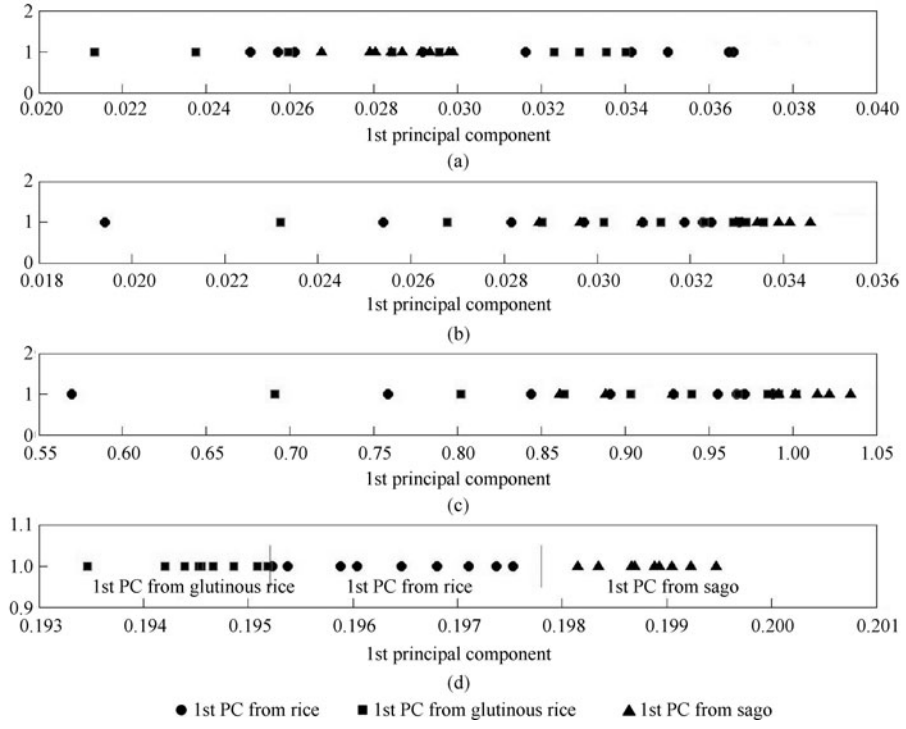

Fig. 5 Spatial distribution maps of the first principal component from first-derivative original spectra (a); MSC (b); SNV (c) and PRC (d). PC: principal component

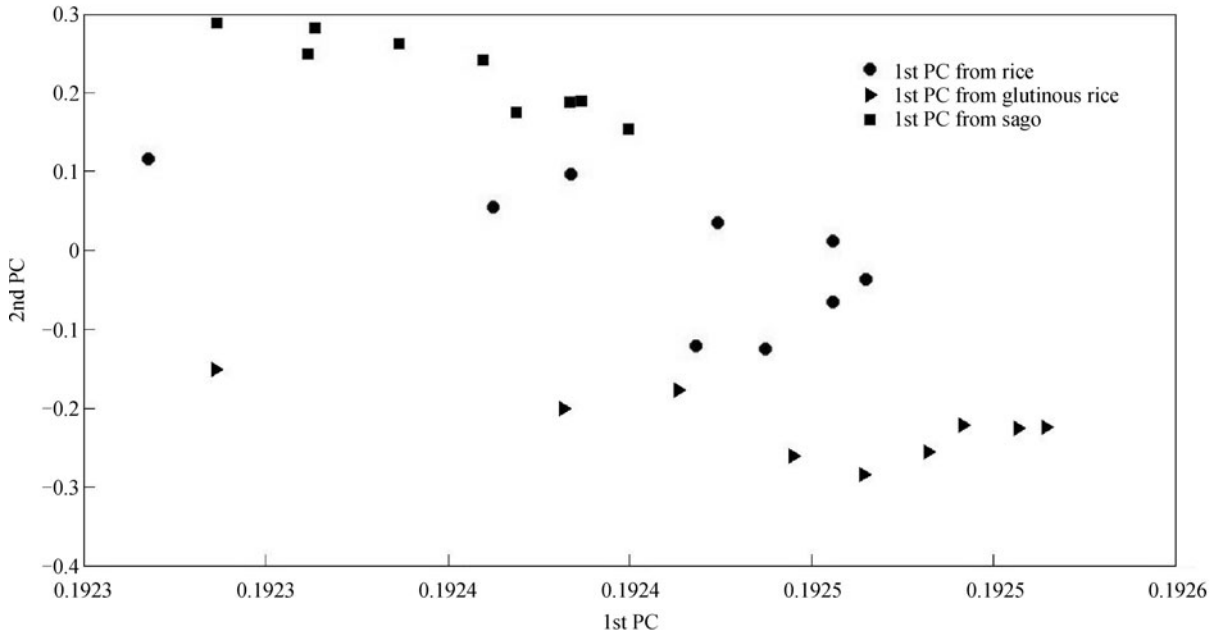

Fig. 6 Spatial distribution of the first two principal components from PRC_EACH without first-derivative. PC: principal component

different spectra and also affect the result from spectra analysis. Therefore, many methods for spectra correcting have been introduced. However, traditional methods, such as MSC and SNV, are mathematical methods for spectra only, lack of physical meaning, and satisfactory effects cannot be obtained by traditional methods sometimes. Thus, it is necessary that other information is used during the correction process to achieve better spectra correcting.

These results from the particle size regression correction method in this paper showed that each type of spectra, pretreated by PRC, could be directly identified by PCA presumably due to the ability of absorbance modeling to distinguish information unrelated to particle size from these effects related to particle size. Some correction coefficients should be estimated for the PRC. PRC is designed for the powder samples and particle samples, simple in principle and easy to realize. PRC can be widely used in the NIR analysis.

Acknowledgements The work was made possible with support from two research projects by the National Natural Science Foundation of China (Grant Nos. 61144012 and 31101289).

Rerferences

1. Burns D A, Ciurczak E W. Handbook of Near-Infrared Analysis. 3rd eds. Boca Raton: CSC Press LLC, 2006, 23–26
2. Martens H, Jensen S A, Geladi P. Multivariate linearity transformation for near-infrared reflectance spectrometry. In: Proceedings of the Nordic symposium on applied statistics. 1983, 205–234
3. Tomas I, Bruce K. Piese-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products. Applied Spectroscopy, 1993, 47(6): 702–709
4. Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for nir-infrared reflectance spectra of meat. Applied Spectroscopy, 1985, 39(3): 491–500
5. Tomas I, Naes T. Effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. Applied Spectroscopy, 1988, 42(7): 1273–1284
6. Lu Q Y, Chen Y M, Mikami T, Kawano M, Li Z G. Adaptability of four-samples sensory tests and prediction of visual and near-infrared reflectance spectroscopy for Chinese indica rice. Journal of Food Engineering, 2007, 79(4): 1445–1451
7. Xu K X, Qiu Q J, Jiang J Y, Yang X Y. Non-invasive glucose sensing with near-infrared spectroscopy enhanced by optical measurement conditions reproduction technique. Optics and Lasers in Engineering, 2005, 43(10): 1096–1106
8. Martens H, Nielsen J P, Engelsen S B. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Analytical Chemistry, 2003, 75(3): 394–404
9. Bruun S W, Søndergaard I, Jacobsen S. Analysis of protein structures and interactions in complex food by near-infrared spectroscopy. 1. Gluten Power. Journal of Agricultural and Food Chemistry, 2007, 55(18): 7234–7243
10. Bruun S W, Søndergaard I, Jacobsen S. Analysis of protein structures and interactions in complex food by near-infrared spectroscopy. 2. Hydrated Gluten. Journal of Agricultural and Food Chemistry, 2007, 55(18): 7244–7251
11. Lui L, Ye X P, Arnold M. Saxton, Womac A I. Pretreatment of near infrared spectral data in fast biomass analysis. Journal of Near

Infrared Spectroscopy, 2010, 18(5): 317–331
12. Prahl S A, Keijzer M, Jacques S L, Welch A J. A Monte Carlo model of light propagation in tissue. In: SPIE Proceeding of Dosimetry of Laser Radiation in Medicine and Biology. 1989, 102–111
13. Prince S, Malarvizhi S. Monte Carlo simulation of NIR diffuse reflectance in the normal and diseased human breast tissues. BioFactors, 2007, 30(4): 255–263
14. Hou R F, Huang L, Wang Z Y, Xu Z L. Preliminary study of the light migration in farm product tissue. Transactions of the Chinese Society of Agricultural Engineering, 2005, 21(9): 12–15 (in Chinese)
15. Xu Z L, Wang Z Y, Huang L, Liu Z C, Hou R F, Wang C. Double-integrating-sphere system for measuring optical properties of farm products and its application. Transactions of the Chinese Society of Agricultural Engineering, 2006, 22(11): 244–249 (in Chinese)
16. Wang Z Y, Hou R F, Huang L, Xu Z L, Wang C, Qiao X J. Light transport in multi-layered farm products by using Monte Carlo simulation and experimental investigation. Transactions of the Chinese Society of Agricultural Engineering, 2007, 23(5): 1–7 (in Chinese)