Bangladesh J. Bot. **46**(3): 1153-1160, 2017 (September) Supplementary

MELON SEED VARIETY IDENTIFICATION BASED ON HYPERSPECTRAL TECHNOLOGY COMBINED WITH DISCRIMINANT ANALYSIS

CUILING LI[1,2], PENGFEI FAN[1,2], KAI JIANG[1,2], XIU WANG[1,2*], QINGCHUN FENG[1,2] AND CHUNFENG ZHANG[1,2]

Beijing Research Center of Intelligent Equipment for Agriculture and National Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China

Keywords: Melon seed, Variety identification, Hyperspectral technology, Discriminant analysis

Abstract

Impurity of melon seed variety causes harm to melon production. Hyperspectral technology combined with discriminant analysis was used to identify melon seed varieties at a non-destructive faster rate. A simple hyperspectral imaging system was developed to collect hyperspectral information of melon seeds. It included a light source unit, a hyperspectral image acquisition unit and a data processing unit. Hyperspectral images of melon seeds were calibrated and reflective spectral information was extracted from images. Multiple scattering correction (MSC), standard normal variable (SNV) transform, first derivative (FD) and Savitzky-Golay (SG) convolution smoothing were carried out. Effects of pre-treatment method were obvious, they mainly eliminated the impacts of surface scattering and optical path change, removed baseline drift and fine white noise. Carrying out SG, MSC, SNV and FD pre-treatments at the same time contributed greatly to discriminant effect. Principal component analysis (PCA) reduced the dimensions of spectral data and extract principal components. Fisher discriminant analysis and distance discriminant analysis generated good and similar discriminant effects, their discriminant accuracies were both higher than 90.0% for validation set. Research results showed that using hyperspectral technology combined with discriminant analysis could realize fast and non-destructive identification of melon seed variety

Introduction

In the identification of seed variety, morphological, physiological and biochemical identification and molecular marker methods are well known (Jia and Mao 2013). The, morphological identification method requires skilled analysts and is susceptible to the influence of subjective factors, as well as a time-consuming process. There are many physiological and biochemical identification methods, but most of them need pre-processing of seed samples. In it seed destruction is inevitable as well as also time-consuming, complicated and tedious. Likewise, molecular marker method has slow speed, high cost and sample destruction inevitable (Liu *et al.* 2003). So far, traditional methods of seed variety identification are unable to meet the needs of the rapid development of seed industry market. So, seed variety identification method gradually develops in the direction of rapid, non-destructive, efficient identification (Sun *et al.* 2012, Chen 2006).

Spectral detection technology has the characteristics of rapid, effective, non-polluting and non-destructive detection. It draws more and more attention from researchers, food enterpreneurs

*Author for correspondence: <wangx@nercita.org.cn>. [1]Beijing Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China. [2]National Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China.

and farmers. Besides, the technology is also efficient in the identification of seed variety, purity identification, moisture content and vitality determination and pest detection (Min and Kang, 2003, Shetty *et al.* 2011, Esteve and Jr 2014). Spectral analysis in variety identification of rice (Kong *et al.* 2013, Li *et al.* 2008), pepper (Yu and Yong, 2003), bean (Zhu *et al.* 2010) and corn seeds (Su *et al.* 2009, Chen *et al.* 2008) showed that the technology can ensure a high recognition rate and achieve rapid identification. Hyperspectral technology is a kind of important spectral detection technologies having high detection accuracy and great advantage in basic research.

Discriminant analysis (Chen *et al.* 2016) is a kind of important statistical analysis method, it sums up classification regularity, establishes discriminant function and identifies new sample's category according to the information provided by category-known samples. Linear discriminant analysis is a convenient and effective method in discriminant analysis, and it has been widely used in the field of pattern recognition. According to discriminant functions' establishment criterions, linear discriminant analysis is divided into distance discriminant analysis (Xiang *et al.* 2016), Fisher discriminant analysis (Tan *et al.* 2016), Bayes discriminant analysis (Wen *et al.* 2016) and so on. This study chose distance, Fisher criterion and Bayes criterion as the criteria for discriminant function establishments, respectively in discriminant analysis.

This research used visible/near infrared hyperspectral technology combined with discriminant analysis to detect melon seed variety. The objective of this research was to explore the feasibility of using hyperspectral technology to identify melon seed variety.

Materials and Methods

Melon seed samples from varieties: Yi Te Bai, Yi Te Jin, Jing Mi NO.7, Jing Mi NO.11, Kai Men Hong NO.8 , Huo Ju, Tian Jiao and Yi Li Sha Bai were bought from a seed shop. They were all hybrids and originated from China with a purity > 96.0%, cleanliness > 99.0% and moisture content <13.0%. This study randomly selected 100 seeds from each variety of melon seed as test samples. A total of 800 seed samples were kept in a dry container under room temperature for test.

The used hyperspectral imaging system (Li *et al.* 2011) whose structure principle diagram is shown in Fig. 1 mainly consisted of a light source unit, a hyperspectral image acquisition unit and a data processing unit. The light source unit contained two halogen lights equipped with two regulated power supplies. The hyperspectral image acquisition unit contained a hyperspectral camera (SOC710 Enhanced, USA). Its sensitive wavelength ranged from 400 nm to -1000 nm with spectral resolution of 2.34 nm. Because of the presence of a scanner built in the hyperspectral camera and no other rotating optical component, it was not needed to equip another cradle head to realize vertical scanning, scanning speed and integral time matched automatically. The data processing unit whose function was to process hyperspectral data and develop discriminant models of melon seed variety, mainly included a computer. In addition, this system also included an optical bracket and a dark chamber, the optical bracket was used to fix the hyperspectral camera and two light sources and place samples, the dark chamber isolated the effect of outside light.

Spectral data from seed samples captured by the hyperspectral camera contained samples' information of their own, but it also included some other irrelevant information, such as electrical noises, background, stray light and so on. Therefore, it was necessary to preprocess the spectral data when using chemometrics methods to develop discriminant models (Seasholtz and Kowalski, 1993).

Spectral preprocessing overcame the impact due to differences in sample thicknesses and optical scattering, and made the spectral information of samples allowing good correlation with reference values. Spectral preprocessing methods usually include MSC, SNV, first derivative, etc.

(Geladi *et al.* 1985, Isaksson and Næs, 1988, Kaihara *et al.* 2002, Pizarro *et al.* 2004, Windig *et al.* 2008).

This research adopted the MSC, SNV, FD and SG smoothing pre-processing methods to deal with the spectral data. MSC firstly put forward by Martens, etc. (Geladi *et al.* 1985) was mainly used to eliminate the impact on NIR diffuse reflection spectra because of uneven distribution of solid particles and particle size, surface scattering as well as optical path change (Chu *et al.* 2004). SNV is usually used after MSC pre-treatment, its functions were similar to MSC (Geladi *et al.* 1985, Chu *et al.* 2004). FD eliminated baseline drift and other background interference, distinguished overlapping peaks, and improved sensitivity and resolution (Ma *et al.* 2007). SG smoothing mainly eliminated baseline drift and tilt noise, in this method, derivative order number was set to zero, polynomial number was equal to '5', and smooth points was '25'.

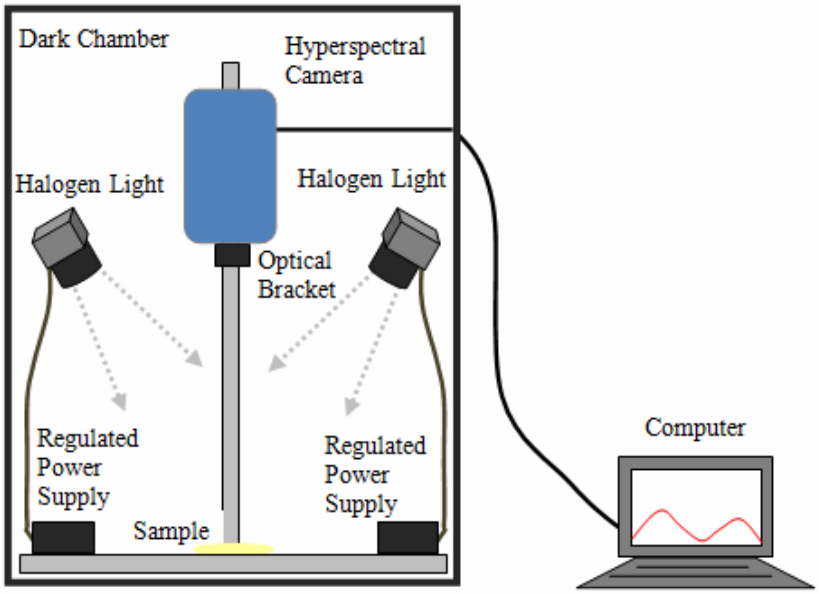

Fig.1. Structure principle diagram hyperspectral imaging system.

Original spectral data of melon seeds were preprocessed with MSC, SNV, FD and SG smoothing methods, therefore, the influences of baseline drift, sample particle size, background and stray light were eliminated, but there still contained 260 variables in the wavelength range from 387 nm - 1043 nm, there still existed some correlation between variables, if used all of these variables to develop models directly, not only the amount of data was large and the computing speed was slow , but also model stability and accuracy would be reduced (Rodríguez-Pulido *et al.* 2013). PCA can recombine numerous variables to a new set of mutually independent variables. It converts multiple variables into several comprehensive variables (principal components) on the premise of losing a little information, each principal component is a linear combination of original variables, and all principal components are independent of each other, this makes principal components have more superior performance than original variables.

This study utilized PCA to reduce the data dimension and eliminated the correlation between variables. In this method, contribution rates of principal components were sorted from big to small,

if the cumulative contribution rate of ahead principal components reached 85%, these components would be selected as the principal components. At last, the number and scores of principal components were determined so as to realize reduction of the data dimension and elimination of the correlation between variables, principal component numbers are shown in Table 1 below.

This research utilized Kennard-Stone algorithm (Woody *et al.* 2004) to divide seed samples into training set and validation set at the ratio of 3:1, the training set contained 600 samples, while the validation set contained 200 samples. The purpose of using Kennard-Stone algorithm is to ensure that the training set samples distributed evenly according to their space distances.

In distance discriminant analysis, since mahalanobis distance has statistical significance, and it is used more than Euclidean distance, this research selected mahalanobis distance as the discriminant function. First of all, according to category-known samples' property values, geometric gravity center of each category was calculated, and then, mahalanobis distances between the object to be classified and the geometric gravity centers of all categories were calculated severally. Finally, according to the criterion of minimum distance, the category which had the minimum mahalanobis distance with the object was taken as the classified result. In this study, two melon seed samples' spectral data vectors were *x* and *y*, $x = (x_1, x_2, \ldots, x_{260})^T$ and $y = (y_1, y_2, \ldots, y_{260})^T$, their mahalanobis distance calculation formula is shown as equation bellow:

$$d\ (x, y)\ = \sqrt{(x-y)\Sigma^{-1}(x-y)^{\mathrm{T}}}$$

where, $d\ (x, y)$ is the mahalanobis distance between sample *x* and *y*;$\sum$ is the covariance matrix of melon seed samples' spectral data matrix; $\boldsymbol{\Sigma^{-1}}$ is the inverse matrix of $\sum$.

The basic idea of Fisher discriminant analysis is that it projects samples of high-dimensional pattern to the best identification vector space so as to extract the classification information and compress feature space dimension. It can guarantee that pattern samples in the new subspace have the biggest distances with other classes and have the smallest distance in their own classes after projection. In this research, a melon seed sample's spectral data vector was x, $x = (x_1, x_2, \ldots, x_{260})^T$, its projection formula is shown as equation bellow:

$$f\ (x)\ = \sum_{i=1}^{N} c_i\, x_i$$

where, $f\ (x)$ is projection function; *i* is equal to 1,2 ,3 ,…, 260; x_i is one-dimension variable; c_i is the projection coefficient of x_i.

Bayes discriminant analysis is a multivariate statistical analysis method based on Bayes criterion. The Bayes statistics thought is that it assumes to cognize something about the study objects, commonly, prior probability is utilized to describe the cognition, and then, one sample is selected and its information is used to revise the existed cognition (prior probability), getting the posteriori probability, statistical inferences are carried through posterior probability. In this study, a melon seed sample's spectral data vector was x, $x = (x_1, x_2, \ldots, x_{260})^T$, Bayes discriminant method calculated conditional probabilities (posteriori probabilities) of sample *x* belonging to every category firstly, then compared these conditional probabilities, and classified sample *x* to the category which appeared the biggest conditional probability.

Results and Discussion

This study used the developed hyperspectral imaging system to collect hyperspectral images of melon seeds, and the data processing software SRAnal710 supplied by American SOC company was used to carry out calibration for hyperspectral images shown as Fig. 2. Software SRAnal710 converted DN values of hyperspectral image to reflectivity values, getting original

reflectance spectrums. Regions of interest (ROIs) of each hyperspectral image were selected through software ENVI 4.3, and the average reflective spectrums of ROIs were taken as the reflective spectrums of the corresponding seeds. 800 melon seed samples' reflectance spectrums are shown in Fig. 3. Spectral curves pre-treated by MSC, SNV, FD and SG smoothing methods are shown in Fig. 4. It can be seen from Fig. 4 that pre-treatment methods' effect was obvious;, they mainly eliminated the impacts of surface scattering and optical path change, removed baseline drift and fine white noise.

Fig. 2. Images of melon seed samples from 8 varieties. 1. Yi Te Bai, 2. Yi Li Sha Bai, 3. Yi Te Jin, 4. Tian Jiao, 5. Jing Mi NO.11, 6. Kai Men Hong NO.8, 7. Jing Mi NO.7 and 8. Huo Ju.

Total of 800 melon seed samples were divided into 8 classes, they were marked '1', '2', '3', '4', '5', '6', '7', and '8', class '1' included melon seed samples of variety "Huo Ju"; class '2' included melon seed samples of variety "Jing Mi NO.7"; class '3' included melon seed samples of variety "Jing Mi NO.11"; class '4' included melon seed samples of variety "Kai Men Hong NO.8"; class '5' included melon seed samples of variety "Tian Jiao"; class '6' included melon seed samples of variety " Yi Li Sha Bai "; class '7' included melon seed samples of variety "Yi Te Bai"; class '8' included melon seed samples of variety "Yi Te Jin". This research carried out discriminant analysis through MATLAB software, and the discriminant effects of distance discriminant analysis, Fisher discriminant analysis and Bayes discriminant analysis were compared, discriminant results of melon seed samples in validation set are shown in Table 1, it can be found from Table 1 that spectral data pretreatment methods obviously improved the discriminant effects of identification models. Carrying out SG, MSC, SNV and FD pretreatments at the same time generated the best discriminant effect. Fisher discriminant analysis method generated the best discriminant result with the accuracy of 96.0%, while Bayes discriminant analysis method generated the worst discriminant result with the accuracy of 88.0% for validation set.

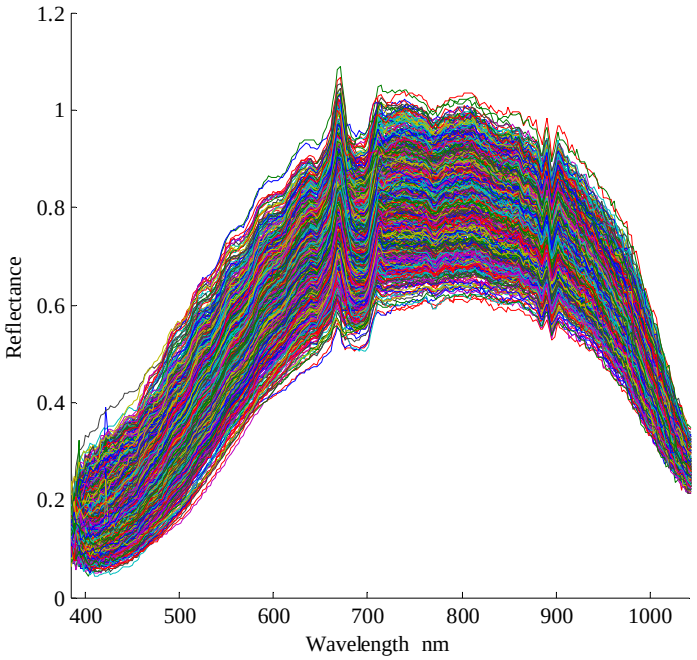

Fig.3. Reflectance spectrums of melon seed samples.

.

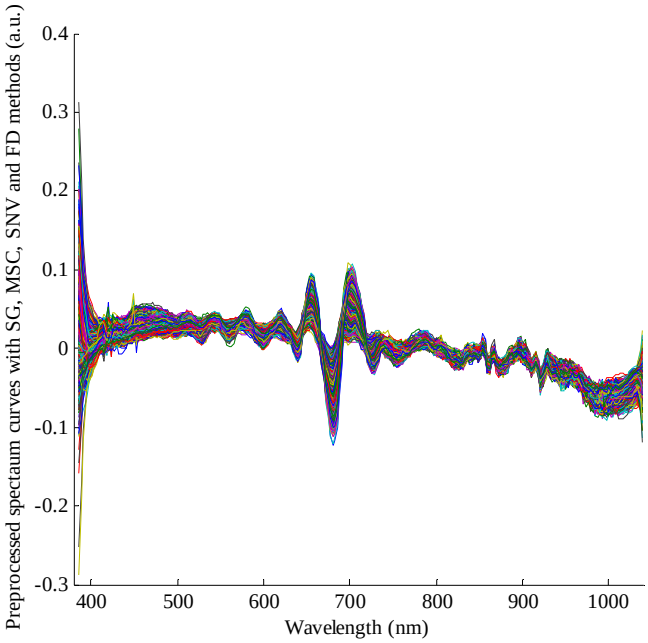

Fig. 4. Preprocessed spectrum curves with SG, MSC, SNV and FD methods.

Table 1. Numbers of principal components and discriminant results for validation set.

Preprocessing method	n_P	R_{A1} (%)	R_{A2} (%)	R_{A3} (%)
OD	2	39.0	34.0	34.5
MSC	3	75.0	77.0	76.0
SNV	3	75.5	77.0	76.5
SG + MSC	3	79.0	79.5	78.5
SG + SNV	3	80.5	80.5	78.5
SG + FD	13	92.5	93.5	85.0
SG + MSC + SNV	3	80.5	80.5	80.0
SG+ MSC + FD	17	94.0	94.0	87.0
SG + MSC + SNV + FD	17	92.5	96.0	88.0

"OD" means the original spectral data; n_P stands for the number of principal components; R_{A1} denotes discriminant results of sample variety for validation set using distance discriminant analysis; R_{A2} denotes discriminant results of sample variety for validation set using Fisher discriminant analysis; R_{A3} denotes discriminant results of sample variety for validation set using Bayes discriminant analysis.

This study adopted hyperspectral technology in combination with discriminant analysis to identify melon seed varieties fast and non-destructively. Discriminant analysis method produced the best discriminant result with the accuracy of 96% for validation set. Research results showed that using hyperspectral technology combined with discriminant analysis could realize fast and non-destructive identification of melon seed variety. In the following study, the spectral information will be integrated with image information to improve the recognition accuracy of melon seed varieties.

Acknowledgements

The authors are grateful for supports from the project of 2016 Special Projects of Innovation Ability Construction of Beijing Academy of Agriculture and Forestry Science (KJCX20151410), and the project of the Young Core Personal Project & Beijing Outstanding Talent Training Project (2015000020060G134) for this study.

References

Chen J 2006. Non-destructive test for the quality of agricultural product. J. Changchun Univ. Technol. (Nat. Sci. Ed.) **27**(3): 262-266.

Chen J, Chen X, Li W, Wang JH and Han DH 2008. Study on discrimination of corn seed based on near-infrared spectra and artificial neural network model. Spectrosc. Spectr. Anal. **28**(8): 1806-1809.

Chen H, Zhao XM, Tan QL, Zou X, Qian SY and Jiang JH 2016. Detection methods of tissue lesion based on linear discriminant analysis and ultrasonic image features. Chin. J. Med. Imaging Technol. **32**(11): 1757-1760.

Chu XL, Yuan HF and Lu WZ 2004. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique. Progress in Chem. **16**(4): 528-542.

Esteve AL and Jr HC 2014. Limitations and current applications of near infrared spectroscopy for single seed analysis. Talanta **121**: 288-299.

Geladi P, MacDougall D and Martens H 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Appl. Spectrosc. **39**(3): 491-500.

Isaksson T and Næs T 1988. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. Appl. Spectrosc. **42**(7): 1273-1284.

Jia W and Mao PS 2013. Review on the near infrared spectroscopy in seed quality testing research. Seed **32**(11): 46-51.

Kaihara M, Takahashi T, Akazawa T, Sato T and Takahashi S 2002. Application of near infrared spectroscopy to rapid analysis of coals. Spectrosc. Lett. **35**(3): 369-376.

Kong WW, Zhang C, Liu F, Nie PC and He Y 2013. Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. Sensors **13**(7): 8916-8927.

Li JB, Rao XQ and Ying YB 2011. Detection of common defects on oranges using hyperspectral reflectance imaging. Comput. Electron. Agr. **78**(1): 38-48.

Li XL, Tang YM, He Y and Ying XF 2008. Discrimination of varieties of paddy based on Vis/NIR spectroscopy combined with chemometrics. Spectrosc. Spectr. Anal. **28**(3): 578-581.

Liu YD, Ying YB, Cheng F 2003. Research of machine vision in purity inspection of seed. Trans. Chin. Soc. Agric. Mach. **34**(5): 161-163.

Ma G, Fu XP, Zhou Y, Ying YB, Xu HR, Xie LJ and Lin T 2007. Nondestructive sugar content determination of peaches by using near infrared spectroscopy technique. Spectrosc. Spectr. Anal. **27**(5): 907-910.

Min TG and Kang WS 2003. Nondestructive separation of viable and non-viable gourd (Lagenaria siceraria) seeds using single seed near infrared reflectance spectroscopy. J. Korean Soc. Hortic. Sci. **44**(5): 545-548.

Pizarro C, Esteban-Díez I, Nistal AJ and González-Sáiz JM 2004. Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. Analytica Chimica Acta **509**(2): 217-227.

Rodríguez-Pulido FJ, Barbin DF, Sun DW, Gordillo B, González-Miret ML and Heredia FJ 2013. Grape seed characterization by NIR hyperspectral imaging. Postharvest Biol. Tec.**76**: 74-82.

Seasholtz MB and Kowalski B 1993. The parsimony principle applied to multivariate calibration. Anal. Chim. Acta. **277**(2): 165-177.

Shetty N, Min TG, Gislum R, Olesen M and Boelt B 2011. Optimal sample size for predicting viability of cabbage and radish seeds based on near infrared spectra of single seeds. J. Near Infrared Spectrosc. **19**(6):451- 461.

Su Q, Wu WJ, Wang HW, Wang K and An D 2009. Fast discrimination of varieties of corn based on near infrared spectra and biomimetic pattern recognition. Spectrosc. Spectr. Anal. **29**(9):2413-2416.

Sun Q, Wang Q, Xue WQ, Ma HX, Sun BQ and Xie ZM 2012. Advance in non-destructive detection of seed quality. J. China Agric. Univ. **17**(3): 1-6.

Tan C, Dai B, Liu HR, Gong JS, Dai Z and Yang CJ 2016. Application of Fisher's discriminant analysis to discriminate different varieties of black tea and Dianhong tea cream produced by different methods. Food Science **37**(7): 62-65.

Wen CP, Bai YY, Zeng JJ and Su W 2016. Bayes discriminant analysis method of natural grassland classification. Chin. J. Grassland **38**(03):50-55.

Windig W, Shaver J and Bro R 2008. Loopy MSC: a simple way to improve multiplicative scatter correction. Appl. Spectrosc. **62**(10):1153-1159.

Woody NA, Feudale RN, Myles AJ and Brown SD 2004. Transfer of multivariate calibrations between four near-infrared spectrometers using orthogonal signal correction. Anal. Chem. **76**(9): 2595-2600.

Xiang SY, Xing HM and Xu DJ 2016. Spatial entity determination of spatial points based on distance discriminant analysis, Science of Surveying and Mapping **41**(6): 40-43.

Yu W and Yong KL 2003. Study on the characteristic and significance of capsaicinoid. Food Science **24**(11): 105-108.

Zhu DZ, Wang K, Zhou GH, Hou RF and Wang C 2010. The NIR spectra based variety discrimination for single soybean seed. Spectrosc. Spectr. Anal. **30**(12): 3217-3221.

(Manuscript received on 3 May, 2017; revised on 5 September, 2017)