Anal Bioanal Chem (2017) 409:3515–3525
DOI 10.1007/s00216-017-0288-8

RESEARCH PAPER

Analysis of multiple soybean phytonutrients by near-infrared reflectance spectroscopy

Gaoyang Zhang[1] · **Penghui Li**[1] · **Wenfei Zhang**[1] · **Jian Zhao**[1]

Received: 28 October 2016 / Revised: 23 January 2017 / Accepted: 28 February 2017 / Published online: 19 April 2017

Abstract Improvement of the nutritional quality of soybean is usually facilitated by a vast range of soybean germplasm with enough information about their multiple phytonutrients. In order to acquire this essential information from a huge number of soybean samples, a rapid analytic method is urgently required. Here, a nondestructive near-infrared reflectance spectroscopy (NIRS) method was developed for rapid and accurate measurement of 25 nutritional components in soybean simultaneously, including fatty acids palmitic acid, stearic acid, oleic acid, linoleic acid, and linolenic acid, vitamin E (VE), α-VE, γ-VE, δ-VE, saponins, isoflavonoids, and flavonoids. Modified partial least squares regression and first, second, third, and fourth derivative transformation was applied for the model development. The 1 minus variance ratio (1-VR) value of the optimal model can reach between the highest 0.95 and lowest 0.64. The predicted values of phytonutrients in soybean using NIRS technology are comparable to those obtained from using the traditional spectrum or chemical methods. A robust NIRS can be adopted as a reliable method to evaluate complex plant constituents for screening large-scale samples of soybean germplasm resources or genetic populations for improvement of nutritional qualities.

Keywords Soybean nutrients · Near infrared spectroscopy · Modeling · Quick germplasm screening

Electronic supplementary material The online version of this article (doi:10.1007/s00216-017-0288-8) contains supplementary material, which is available to authorized users.

✉ Jian Zhao
jzhao2@qq.com; 2948308723@qq.com

[1] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, No.1 Shizishan Street, Wuhan, Hubei 430070, China

Introduction

Soybean (*Glycine max* L.) is one of the major oilseed crops. Numerous constituents in soybean require thorough study before their utilization in industrial processing and daily diet, based on quick and reliable analysis [1, 2]. Due to many health beneficiary phytochemicals, such as flavonoids and isoflavones, saponins, oils, and vitamins, present in soybean seeds in various amounts, the quality and quantification of these nutrients are therefore needed for evaluation of soybean varieties and utilizations. The increasing demands for quality of soybean-derived foods with positive health-beneficial chemicals have created needs for developing fast and efficient analytical methods [1]. Except for traditional techniques for quantifying biologically active compounds in raw materials as well as in processed products, such as high performance liquid chromatography (HPLC) for vitamins and flavonoids and gas chromatography (GC) for fatty acids, some other technologies have been developed [1, 3, 4]. The spectroscopy methods provide useful qualitative and quantitative data about chemistry and biochemistry of these bioactive compounds. The near infrared reflectance (NIR), introduced in 1964, is based on the absorbance of light energy at a given frequency by molecules (or radicals) having a permanent dipole vibrating at the same frequency [3]. NIRS in contrast with traditional chemical analysis is considered fast, accurate, and nondestructive [1]. The NIRS is used to quantify many compounds such as polyphenols, carotenoids, and fatty acids (FAs) in vegetables, fruits, and many food products [4–6].

NIRS has been recognized as one of the most powerful analytical tools [7–9] and is widely used for the simple and rapid analyses of various agricultural and food products. The effect of optical length on Camellia oil adulteration [10], quantification of sugar to nitrogen ratio in wheat leaves [11], and maize carotenoid content [4] are evaluated using NIRS. Cross-

validation procedures show significant correlations between HPLC values and NIRS estimates, hence, indicating that NIRS is a well-established, widely applied technique to readily analyze bioactive compounds in plants. NIRS has also been used for the analysis of soybean constituents [12].

Vegetable oils are complex mixtures of organic compounds. Analytical techniques like chromatography, square wave voltammetry, Raman scattering, and fluorescence have been used for quality check of oils [13, 14]. Recently, da Costa and coworkers used NIRS for quality control of soybean oil. In this experiment, 30 expired and 20 non-expired samples based on expiration date mentioned on labels were used. The robustness of method motivated NIRS as a great analytical method for quality control of several complex compounds like food, fuel, drinks, and drugs [15]. Soybean seed quality was predicted by single seed NIRS. The partial least squares (PLS) regression gave accurate predictive models for oil, protein, volume, weight, density, and maximal cross-sectional area of the seed. PLS models for width and length were not predictive. Single seed NIRS facilitates broader adoption of this high-throughput, nondestructive, seed phenotyping technology [16]. Chinese yam samples were analyzed for sugar, polysaccharides, and flavonoids through NIR and mid-IR. The spectra were compared qualitatively using principal component analysis (PCA) and quantitatively using PLS and least squares-support vector machine (LS-SVM) models. The results indicated that NIR data performed somewhat better than the mid-IR data [17]. NIRS has been used extensively to analyze oil content and fatty acid profile in oilseeds such as sunflower [18], rapeseed [19], soybean [20], peanut [21], and castor [22], as well as in other commodities such as cereals [23], forages [24], corn [25], and animal products [26]. Conversely, NIRS has not yet been reported for evaluation of most soybean essential nutrients like tocopherols, saponins, FAs, and flavonoids. Here we report a combined NIRS measurement of the contents of multiple nutrients, including FAs, anthocyanins, proanthocyanidins, isoflavones, tocopherol, and saponin in diverse soybean germplasm resources collected from different provinces of China. Overlapping spectra were resolved using PLS method. The established methods with inexpensive and robust NIRS greatly facilitate our screening on various soybean germplasm resources and large numbers of populations for these quantitative trait loci (QTL) studies.

Materials and methods

Glycine max samples preparation

The soybean varieties grown in various provinces of China were collected and planted at the Chinese Academy of Agricultural Sciences. These varieties, collected from different regions, were planted in the same location and environmental backgrounds; they have some traits different from others and are consistent with properties within intraspecific, including wild species and cultivars. For FA analysis 237 samples, 224 samples for tocopherol, 491 samples for saponin, 249 samples for insoluble proanthocyanidins (PAs), 289 samples for anthocyanins, 268 samples for soluble Pas, and 269 samples for isoflavones were analyzed. The other 43 samples used in this study were collected from Shanxi Agricultural University, Oil Crop Research Institute and Chinese Academy of Agricultural Sciences. The collected samples were milled into fine powder at Huazhong Agricultural University, China, sealed in polyethylene bags (85 × 30 mm), and stored at −80 °C until analysis.

Chemicals and reagents

Methanol, hydrochloric acid (HCl), sulphuric acid (H_2SO_4), acetone, 1,2,3-triheptadecanoylglycerol, and rac-5,7-dimethyltocol as internal standard, methyl benzene, butylated hydroxytoluene (BHT), ethanol, hexane, acetic acid, n-heptane, n-butyl alcohol, p-dimethylaminocinnamaldehyde (DMACA), chloroform, (+)-catechin, vanillic aldehyde as well as perchloric acid were obtained from Sinopharm Chemical Reagent Co. Ltd., China. All reagents and solvents were of analytical grade or HPLC grade, and freshly doubly distilled water (ddH2O) was used throughout for aqueous solutions.

Analyses of oil contents and fatty acid components

The fatty acid composition was determined by the conventional GC method. For total fatty acid analysis, total oils were extracted and analyzed with GC as previously described [27]. Briefly, 30 mg soybean powder was added into the mix of 400 μL of methylbenzene and 1.5 mL of 2.5% H_2SO_4 in methanol (v/v, freshly prepared) containing 0.01% BHT. After incubation at 90 °C water bath for 1 h with nitrogen sealed cover, samples were cool down to room temperature. Then 1.8 mL of deionized water and 1 mL of hexane were added to each sample, followed by shaking vigorously and brief centrifugation, The supernatant was transferred to a 2.0 mL vial for analysis. Figure 1a shows the chromatogram obtained for a sample of powder subjected to the described procedure.

Determination of tocopherols

The 30–50 mg of finely ground soybean powder was extracted with 1000 μL of 9:1 (v/v) methanol:dichloromethane containing 0.01% BHT, and 5000 ng of 5,7-dimethyltocol (Matreya) as an internal standard. After sonication for 2 min and incubation for 30 min, the sample was centrifuged at 4 °C at 12,000 × *g* for 15 min, and 50 μL of the clear supernatant were collected for high-performance liquid chromatography (HPLC) as previously described [28, 29]. Figure 1b shows the

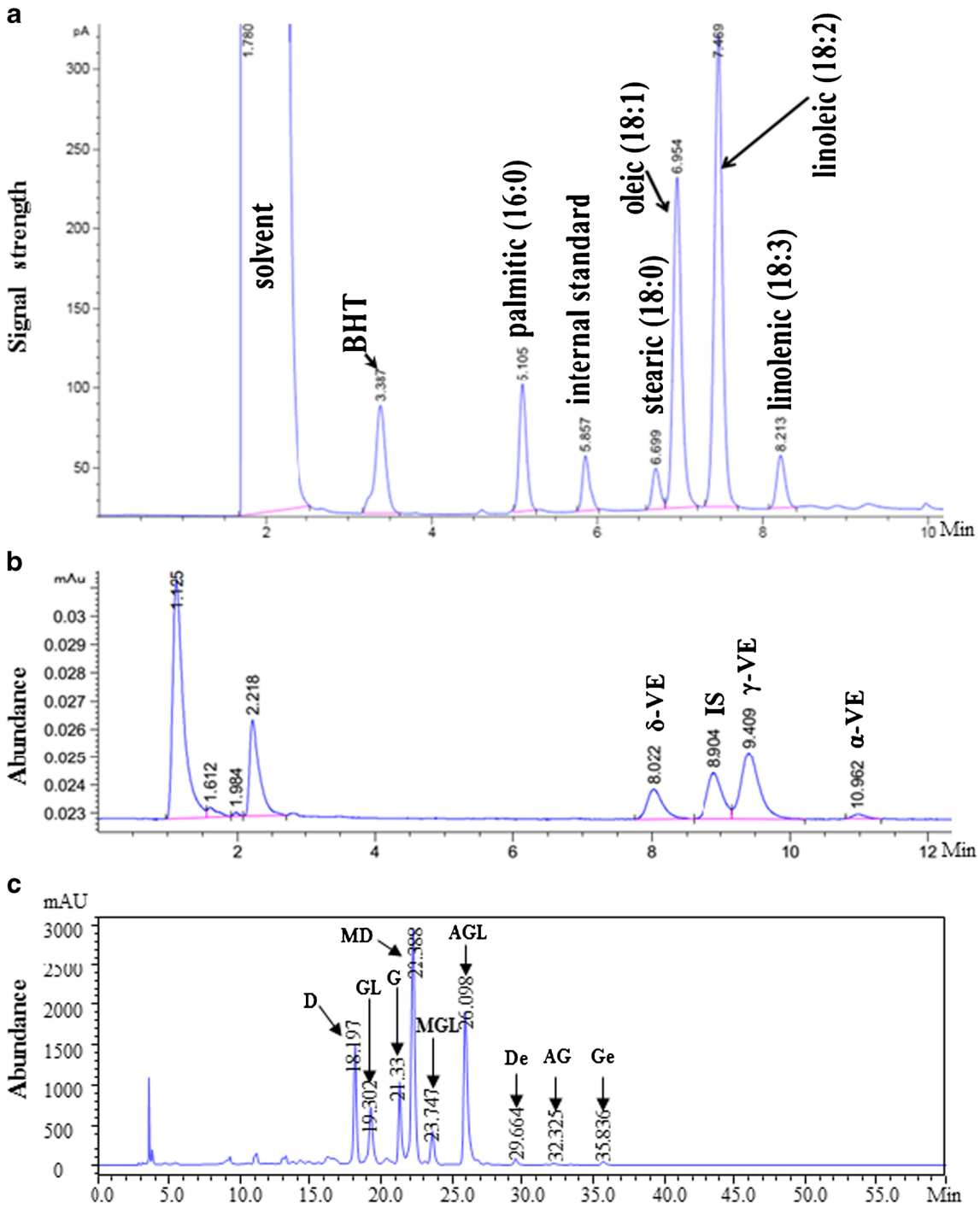

Fig. 1 Chromatograms obtained after application of a chemical analysis method to a standard solution. The unit of horizontal axis is minute. **(a)** GC-FID chromatographs for fatty acid analysis. **(b)** Tocopherol analysis using HPLC added fluorescence detector. **(c)** Isoflavones chromatographs obtained HPLC equipped with Diode array detector

chromatogram obtained for a sample of food subjected to the described procedure.

Determination of the isoflavones and flavonoid content

Anthocyanins, soluble and insoluble PAs, and isoflavones from soybean powder were extracted with the method described previously [30]. In brief, 3 mL 0.01% HCl/methanol was added to 0.03 g of ground samples, then sonicated for 30 min and rotated slowly overnight at 4 °C, followed by centrifugation. The supernatants were absorbed and recorded at 530 nm for anthocyanins analysis. Five mL of 70% acetone containing 0.5% acetic acid was used for soluble and insoluble PAs analysis, and chloroform and hexane were used in sequence for

supernatants extraction; after drying supernatants and residues separately, dried supernatants were suspended and determined at 640 nm after reaction with DMACA reagent. Contents of insoluble PAs were determined by reaction with 2 mL of butanol-HCl (95:5, v/v); the mixtures were sonicated and centrifuged, then measured at 550 nm. For determination of isoflavones, 2 mL of 80% methanol was mixed with 0.1 g of ground samples, then sonicated for 30 min, and extraction at 4 °C overnight. UV spectra were compared with standards and analyzed on the basis of chromatographic behavior. Figure 1c shows the chromatogram obtained for a sample of food subjected to the described procedure.

Quantification of total saponins

Saponins were extracted as previously described by Gu et al. [31]. Ground seed powder (0.3 g) was added with 5 mL ethanol and incubated in a water bath at 90 °C for 1 hour and allowed to cool down. After filtering and evaporating, the solution was dissolved in 2.5 mL water. The water was evaporated and samples were finally dissolved in 80% ethanol and absorbance was noted at 276 nm.

Near infrared spectroscopic analysis

The Foss NIRS DS 2500 analyzer with a standard 1.5 m 210/7210 bundle fiberoptic probe was used to measure the NIR reflectance spectra in the wavelength range of 400 to 2500 nm (from visible light to invisible light) at 0.5 nm intervals on each sample two times at ambient temperature (25 ± 2 °C). The probe uses a remote reflectance system and employs a ceramic plate as reference. The average spectra were calculated, and the analysis was carried out. The raw spectral data were collected with the use of the Mosaic program. The samples (10 g) were poured into a 5 × 5 cm rotating cup (Foss NIR Systems # 60053171) on a holder, which was gently compacted with quartz glass. The software used was Win ISI 1.50, installed on a Lenovo win7 computer.

Statistical analysis

The average spectra from 400–1108 nm was cut off by Win ISI 1.50 because of its instability. The presented procedure of statistical analysis is based on a procedure as described previously [32]. To obtain NIR equations,

Table 1 Reference values of NIR models for total oil (mg/g), palmitic (mg/g), stearic (mg/g), oleic (mg/g), linoleic (mg/g), linolenic acid (mg/g) and total tocopherol (μg/g),α-VE (μg/g), γ-VE (μg/g), δ-VE (μg/g), saponin (Abs/g), flavonoid (Abs/g), isoflavones (μg/g) in soybean

Components	Minimum	Maximum	Mean	Std. deviation	Total values	CV	Skewness	Kurtosis
Total FA	40.25	365.03	149.62	39.63	224	0.26	0.13	6.2
Stearic acid	1.41	13.84	5.26	1.73	224	0.33	0.58	4.83
Oleic acid	4.66	84.55	33.47	12.77	224	0.38	0.21	3.38
Linoleic acid	23.16	198.15	81.33	22.09	224	0.27	0.18	5.72
Linolenic acid	4.71	25.90	11.61	2.99	224	0.26	0.10	4.84
Palmitic acid	5.41	42.59	17.93	4.52	224	0.25	0.25	6.47
Total VE	39.57	860.81	224.10	79.67	185	0.36	2.71	23.41
αVE	1.07	18.52	3.83	2.34	233	1.19	1.36	4.98
δVE	4.85	99.65	46.07	17.02	376	0.49	4.03	35.63
γVE	14.41	177.9	86.21	29.91	377	0.35	1.49	11.02
Saponin	0.34	2.89	1.38	0.45	321	0.34	0.8	3.86
Anthocyanins	0.01	1.97	0.46	0.4	253	1.31	2.41	8.79
Insoluble PAs	0.1	25.14	5.21	5.88	242	1.22	2.23	8.31
Soluble PAs	0.01	21.46	0.78	1.66	263	2.15	8.01	94.23
Total PA	0.15	42.30	6.37	7.36	210	1.19	2.18	7.76
D	19.88	484.68	100.25	64.62	238	0.64	2.08	9.63
De	0.46	42.10	9.34	6.07	238	0.65	1.51	6.81
MD	15.9	1266	421.5	230.5	238	0.55	1.05	4.38
MGL	2.95	271.14	69.18	38.98	238	0.56	1.21	6.19
AGL	4.16	874.63	347.74	153.9	238	0.44	0.13	3
GL	8.47	234.20	95.73	36.2	238	0.38	0.92	4.59
AG	0.18	15.91	5.49	2.47	238	0.45	0.35	3.68
G	7.09	185.63	48	28.9	238	0.60	2.25	9.5
Ge	0.15	48.19	9.53	5.1	238	0.53	2.89	20.01
TIF	246.79	2511.65	1106.75	449.84	238	0.41	0.7	3.33

the modified partial least squares (MPLS) regression method was frequently used for all the parameters studied. Partial least squares (PLS) regression not only needs spectral information, which is similar to principal component regression (PCR), but also uses reference data (chemical, physical, etc.) to form the factors useful for fitting purposes [33]. MPLS, which possesses the advantage of stability and accuracy over the standard PLS algorithm, is often used. Before calculating the next factor, the NIR residuals at each wavelength from MPLS were obtained after each factor was calculated and standardized by dividing by the standard deviations of the residuals. Cross-validation is recommended in order to select the optimal number of factors and to avoid overfitting when developing MPLS equations. In NIR spectroscopy, global H (GH) quantifies how much a particular spectrum deviates from the average of all spectra in a collection in terms of standardized increments of variation about that average. The CENTER algorithm rank-orders all spectra in a collection based on GH and permits elimination of spectra with GH values exceeding a user-specified criterion, usually three standardized increments of variation. We excluded non-representative spectra, classified as those with a generalized Mahalanobis distance (GH) > 3, from calibration development using the CENTER algorithm [34].

We developed calibration models to predict chemistry using WinISI IIIv1.50 (Foss-Tecator, Infrasoft International LLC, State College, PA, USA) software. We chose a subset of spectra for each calibration to uniformly represent the full range of spectral qualities within the sample set of interest using the SELECT algorithm [34, 35] (math treatment, 1,4,4,1), where the first digit is the number of the derivative, the second is the gap over which the derivative is calculated, the third is the number of data points in a running average or smoothing, and the fourth is the second smoothing standard normal variate and detrend scatter correction), with a neighborhood Mahalanobis distance (GH) of 3. We paired each spectrum in the subset of spectra chosen for calibration and developed predictive models from the paired spectral and chemical data on a dry matter basis using mPLS regression or cross-validation. The calibration set is partitioned into several groups; each group is then validated using a calibration developed on the other samples. Finally, validation errors are combined into a standard error of cross-validation (SECV). It has been reported that the SECV is the best single estimate of the prediction capability of the equation and that this statistic is similar to the average standard error of prediction (SEP) from 10 randomly chosen prediction sets. The statistics used to select the best equations were RSQs (multiple correlation coefficients) and the standard error of cross-validation (SECV). We preliminarily fitted each model, eliminated outliers with residuals exceeding three times the standard error of calibration using the "Compare Predicted and Reference Values" function in WinISI, and created a new final version of the model.

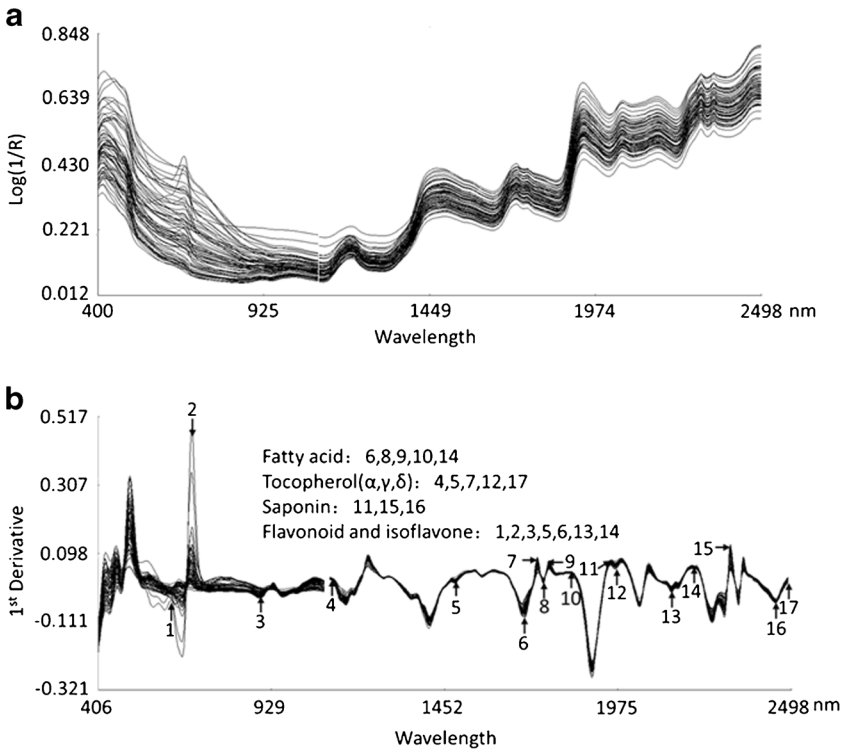

Fig. 2 Typical NIR spectra obtained from soybeans in powder samples. The datasets represent the near-infrared reflectance spectroscopy after scanning samples; these spectra are obtained from different samples. **(a)** Unprocessed spectra. **(b)** First derivative spectra. Absorption peaks for all components were indicated. Fatty acids (palmitic acid, stearic acid, oleic acid, linoleic acid, linolenic acid) had absorption peak in position 6, 8, 9, 10, 14, and so on for others

Result

Statistical description and correlation analysis

The range, mean and standard deviation of the concentration of different constituents studied in this experiment are summarized in Table 1. Electronicoc Supplementary Material (ESM) Fig. S1 shows their distribution. In the sample set, there was a wide variation in chemical composition, and the samples covered most of the variability. The soybean seed constituents showed significant correlation ($p < 0.01$, $p < 0.05$) except for relation between total FA and γ-VE; stearic acid and α-VE, insoluble PAs; α-VE and insoluble PAs; soluble PAs and anthocyanins; glycitin (GL) and daidzin (D); genistin (Ge) and daidzein (De); acetylgenistin (AG) and total PAs, insoluble PAs, stearic acid, α-VE; total PAs and insoluble PAs, α-VE. Positive and negative correlation analysis is listed in detail in ESM Table S1, e.g., total FA, palmitic acid, stearic acid, linoleic acid, and linolenic acid were negatively correlated with saponin and VE content ($p < 0.01$).

NIRS models

The soybean calibration sets typical and deviated spectra from scatter correction are shown in Fig. 2. The NIR spectral patterns of the samples along the X-axis were similar across the entire NIR wavelength region 400–2500 nm (Fig. 2), whereas the spectral intensity deviation of different samples along the Y-axis was clear. PLS regression models of the NIR data were built up on original calibration and cross-validation sets. Samples with large residuals were omitted; those that exhibit a GH value greater than three were considered as not belonging to the population from which the equations were

Table 2 Descriptive statistics of cross-validation for total oil (mg/g), palmitic (mg/g), stearic (mg/g), oleic (mg/g), linoleic (mg/g), linolenic acid (mg/g), and total tocopherol (μg/g), α-VE (μg/g), γ-VE (μg/g), δ-VE (μg/g), saponin (Abs/g), flavonoid (Abs/g), and isoflavone (μg/g) in soybean

Constituents	N	Mean	SD	Est. min	Est. max	SEC	RSQ	SECV	#	Scatter	Math treatment	1-VR	R^2
Total FA	40	149.29	25.28	73.45	225.13	6.13	0.94	8.76	136	NSVD	2,4,4,1	0.91	0.85
Palmitic acid	70	18.05	3.03	8.95	27.15	1.49	0.76	1.61	136	NSV	3,10,10,1	0.73	0.69
Stearic acid	47	4.97	1.09	1.69	8.25	0.46	0.82	0.54	136	None	1,4,4,1	0.76	0.68
Oleic acid	61	32.36	13.01	0	71.39	3.19	0.94	4.02	136	NSVD	2,4,4,1	0.91	0.90
Linoleic acid	65	79.97	17.27	28.17	131.78	5.54	0.9	6.72	136	DET	1,4,4,1	0.87	0.85
Linolenic acid	61	11.92	2	5.91	17.92	0.92	0.79	1.19	136	NSVD	1,4,4,1	0.69	0.63
Saponin	85	1.48	0.4	0.28	2.68	0.23	0.66	0.33	136	DET	3,4,4,1	0.34	0.35
Total VE	63	211.66	57.36	39.56	383.75	23.79	0.83	35.28	136	None	2,4,4,1	0.64	0.55
α VE	28	8.18	3.25	0	17.93	1.98	0.63	2.27	136	None	4,6,6,1	0.53	0.36
γ VE	78	79.47	28.83	0	165.95	11.22	0.85	16.92	136	None	2,10,10,1	0.67	0.42
δ VE	79	40.34	15.59	0	87.11	8.16	0.73	9.11	136	NSV	2,4,4,1	0.68	0.57
Anthocyanins	47	0.33	0.22	0	0.99	0.1	0.8	0.13	136	NSVD	2,4,4,1	0.66	0.37
Insoluble PAs	75	4.28	5.13	0	19.67	1.54	0.91	2.16	136	NSV	4,6,6,1	0.83	0.81
Soluble PAs	53	0.45	0.53	0	2.05	0.17	0.89	0.23	136	NSVD	2,10,10,1	0.82	0.78
Total PAs	53	3.4	2.54	0	11.03	0.74	0.91	1.27	136	NSVD	4,6,6,1	0.76	0.78
D	20	82.25	31.64	0	177.17	1.47	1	10.23	136	NSV	4,6,6,1	0.88	0.72
De	16	9.17	3.26	0	18.94	0.07	1	0.83	136	None	4,6,6,1	0.93	0.84
MD	20	378.55	116.12	30.21	726.9	10.82	0.99	42.53	136	DET	1,6,6,1	0.92	0.84
MGL	20	69.84	16.83	19.34	120.33	0.22	1	6.29	136	NSVD	4,3,3,1	0.87	0.74
AGL	20	362.54	99.25	64.78	660.29	2.41	1	30.8	136	NSVD	3,10,10,1	0.91	0.74
GL	30	93.06	26.7	12.95	173.16	1.57	1	8.65	136	None	3,4,4,1	0.89	0.76
AG	19	5.64	1.98	0	11.58	0.21	0.99	0.5	136	DET	1,10,10,1	0.94	0.86
G	20	40.27	9.19	12.72	67.83	2.8	0.91	4.53	136	DET	3,6,6,1	0.75	0.53
Ge	16	9.31	2.13	2.92	15.69	0.13	1	0.54	136	None	2,3,3,1	0.95	0.87
TIF	20	1006.05	435.76	0	2313.32	24.94	1	121.58	136	NSV	4,6,6,1	0.93	0.77

N: samples number used to develop the model; SD: original standard deviation; Est. min: estimated minimum; Est. max: estimated maximum; SEC: standard error of the calibration; RSQ: R squared; SECV: standard error of cross-validation; 1-VR: 1 minus the variance ratio; #: number of wavelengths; R^2: coefficient of determination; DET: detrend only; None: none scatter correction; SNV: standard normal variate; NSVD: a combination of SNV and DET; FA, fatty acid; VE: vitamin E; Pas: proanthocyanidins; D: daidzin; De: Daidzein; MD: malonyldaidzin; MGL: malonylglycitin; AGL: acetylglycitin; GL: glycitin; AG: acetylgenistin; G: genistin; Ge: genistein; TIF: total isoflavones

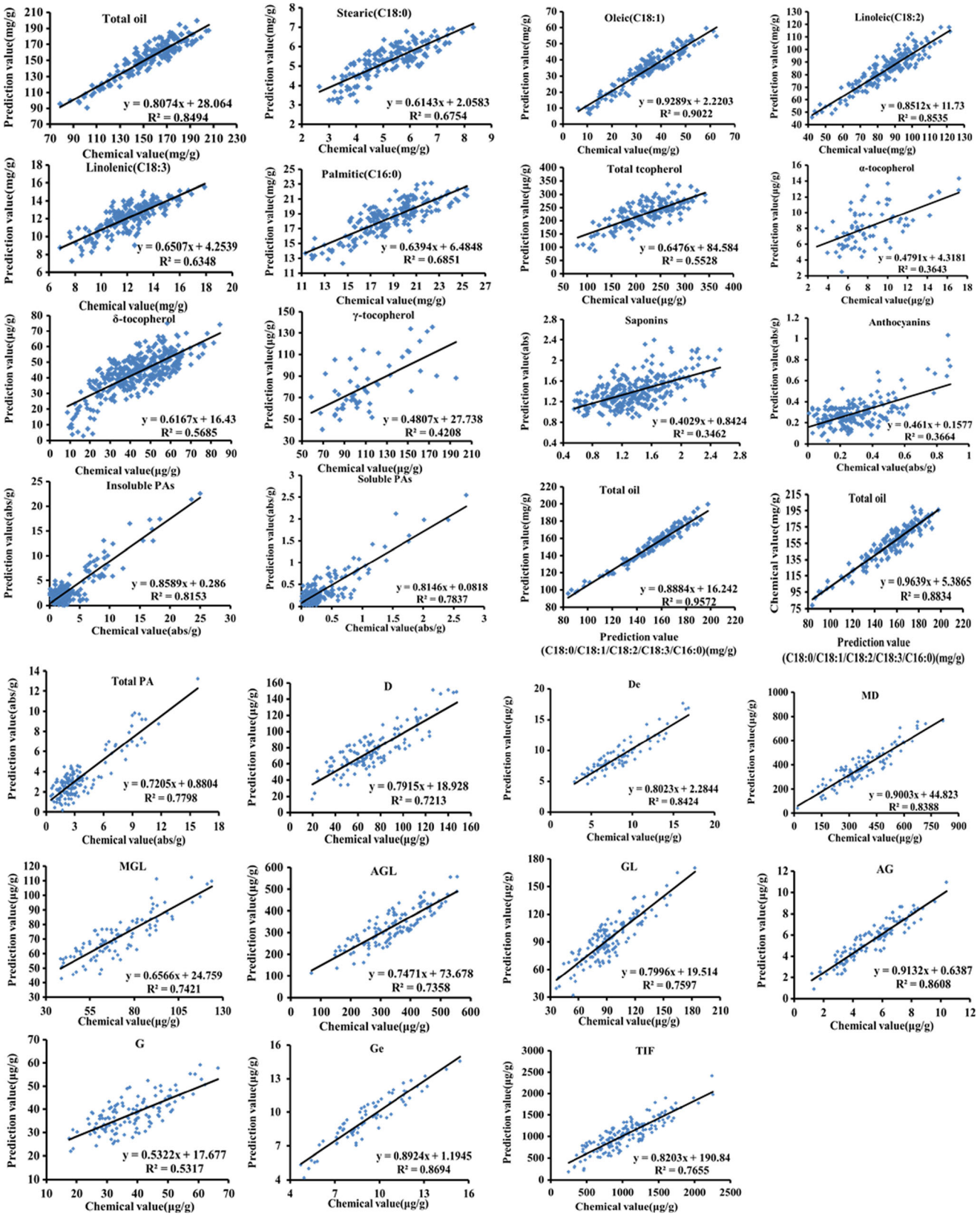

Fig. 3 Optimized regression line in cross-validation of MPLS models for the components of milling powder obtained by NIRs. Comparison of reference values with the values predicted by the calibration equations for fatty acid, tocopherol, saponin, isoflavones, and flavones. FA: fatty acid; VE: vitamin E; Pas: proanthocyanidins; D: daidzin; De: daidzein; MD: malonyldaidzin; MGL: malonylglycitin; AGL: acetylglycitin; GL: glycitin; AG: acetylgenistin; G: genistin; Ge: genistein; TIF: total isoflavones

Table 3 External validation values for total FA (mg/g), palmitic acid (mg/g), stearic acid (mg/g), oleic acid (mg/g), linoleic acid (mg/g), linolenic acid (mg/g) in soybean

Components	Minimum	Maximum	Mean	Std. deviation	Total values	CV	Skewness	Kurtosis
Total FA	6.35	98.23	72	19.78	43	0.27	−2.31	8.28
C18:0	0.42	9.08	5.44	1.97	43	0.36	−0.51	3.78
C18:1	1.74	49.89	31.09	10.55	43	0.34	−1.01	4.32
C18:2	1.10	15.11	11.66	3.29	43	0.28	−2.23	7.88
C18:3	7.26	106.99	78.57	21.95	43	0.28	−2.09	7.20
C16:0	1.81	23.72	17.25	4.64	43	0.27	−2.30	8.34

developed. The outliers were omitted from the calibration set. The statistics analysis for latent variables and outliers are given in ESM Table S2.

The PLS regression statistics of cross-validation are shown in Table 2. According to the PLS model, the Ge had the highest 1 minus the variance ratio (1-VR = 0.95) followed by De (1-VR = 0.93), other isoflavones D, 0.88; malonyldaidzin (MD), 0.92; malonylglycitin (MGL), 0.87; GL, 0.91; AG, 0.89; genistin (G), 0.75; total isoflavones (TIF), 0.93. The 1-VR value of fatty acid (FA) takes second place in our study (total FA, 0.91; oleic acid, 0.91; palmitic acid, 0.73; stearic acid, 0.76; linoleic acid, 0.87; and linolenic acid, 0.69). Tocopherol constituents had a lower 1-VR value (δ-VE, 0.68; γ-VE, 0.67; total VE, 0.64). Flavonoid constituents had proper 1-VR value (insoluble PAs, 0.83; soluble PAs, 0.82; total PAs (TPA), 0.76) except 0.66 from anthocyanins constituent. Conversely, cross-validation indicated poorer adjustment for concentrations of saponin (1-VR = 0.34) and α-VE (1-VR =0.53). The coefficient of determination is similar to 1-VR value (Table 2). The optimized regression lines of PLS models in the cross-validation of the constituents are represented in Fig. 3.

Ratio performance deviation (RPD) was often used to evaluate the prediction capacity of one model; it is the relationship between SD of chemical method and SEP in the NIRS model. The model will be predictive if the RPD value is greater than 2.5. We did a statistical analysis of RPD to evaluate prediction capacity of the model (ESM Table S2). Range comparison of lab values from chemical analysis and predicted values from NIR was executed in this study (ESM Table S2). These two values are almost identical. This data of ranges shows that the actual variation had overlapping with the estimated value, e.g., oleic acid, saponin, α-VE, δ-VE, and γ-VE, D, De.

External validation for oil

We checked the robustness of the method by applying NIRS technology to 43 new samples that did not belong to the calibration group and came from a new area (Shanxi Agricultural University) (Table 3). Their correlation analysis is listed in detail in ESM Table S3. Fatty acids had strong correlation with each other. The procedure was as follows: We recorded the spectra in triplicate and get the spectral mean. Then, we used the calibration equations obtained during the work to get

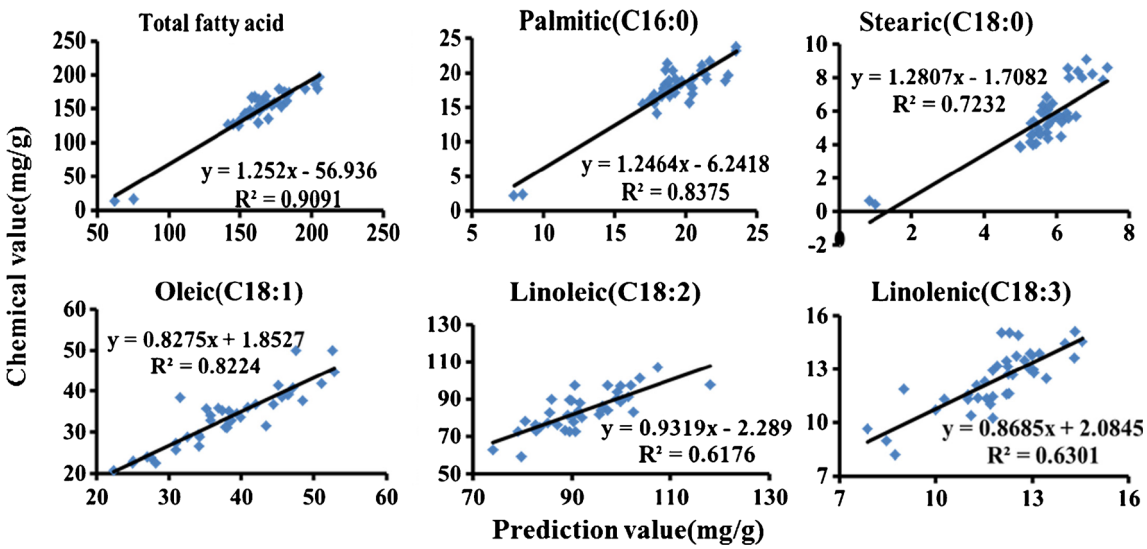

Fig. 4 External validation plots for seed oil content and concentrations of palmitic acid, stearic acid, oleic acid, and linoleic acid in a set of 43 flour samples of almond. RDP: ratio of SD of the validation set to the SEP

Table 4 Statistical comparison between standard GC and NIRS

Constituents	SEP	Means		Std. Dev		Bias	Bias limit	SEP(C)	SED(C) limit	RSQ
		NIRS	GC	NIRS	GC					
Total FA	92.06	72.00	162.90	19.78	25.98	−90.90	7.01	14.74	15.19	0.68
Palmitic acid	3.46	17.25	19.13	4.64	2.91	−1.88	0.97	2.93	2.10	0.63
Stearic acid	1.28	5.44	5.67	1.97	1.21	−0.23	0.33	1.27	0.71	0.61
Oleic acid	9.33	31.09	35.09	10.55	14.93	−4.01	2.41	8.52	5.23	0.69
Linoleic acid	17.04	78.57	89.45	21.95	16.19	−10.88	4.03	13.27	8.73	0.64
Linolenic acid	5.18	11.66	12.45	3.29	2.94	−0.79	0.71	5.18	1.55	0.14

SEP: standard error of prediction; SEP(C): prediction corrected standard errors

SED(C), standard error of deviation combined. Units and abbreviation are same as in Table 2

the predicted values, and the predicted values were compared with those obtained later using chemical analysis after GC-FID (flame ionization detector). The external validation plots were obtained (Fig. 4). Although the R^2 is affected by distribution of the values, they all had a higher R^2 value (total FA, 0.90; palmitic acid, 0.84; stearic acid, 0.72; oleic acid, 0.82) except 0.62 from linoleic acid and 0.63 from linolenic acid. Statistical comparison between standard GC and NIRS is shown in Table 4. It shows that mean fatty acids composition and standard deviations obtained from NIRS or GC is very similar.

Validation of NIRS with soybean germplasm survey

In validation experiments testing more than 500 soybean core germplasm resources for contents of different types of seed nutrients, we applied this technique to make a comprehensive analysis of all nutrients in soybean seed and tried to understand their relations. We found oleic acid (C18:1) and linoleic acid (C18:2) contents fluctuate much less in most soybean varieties, whereas total FA (TFA), palmitic acid (C16:0), stearic acid (C18:0), and linolenic acid (C18:3) contents showed higher diversity in these soybean varieties (ESM Figs. S1, S2) [36]. In contrast to contents of FAs, all isoflavonoids, including different types of isoflavones, varied greatly except AGL and AG in much lower levels [37]. As a protective antioxidant, vitamin Es (total VE, α-VE, γ-VE, or δ-VE) had large variations (ESM Figs. S1, S2), consistent with previous reports [38]. The contents of other ingredients, such as total saponins, anthocyanins, insoluble PAs, soluble PAs, and total PAs (TPA), also showed a larger and better diversity than the above components (ESM Figs. S1, S3). It seems that FA contents in these numerous varieties have scattered same trend with VE and isoflavonoid, but somewhat negative correlation with these of saponin, anthocyanins, insoluble PAs, soluble PAs, and TPAs (ESM Figs. S1, S2). These trends may be expected since it has been implicated in some observations [37, 39].

Discussion

A fast, accurate, noninvasive, and nondestructive detection technology and methodology is always overwhelmingly needed for research, particularly when dealing with a large number of samples. The fast and nondestructive technique NIRS has already gained wide acceptance for routine measurement of compounds in plants and agricultural products. Spectroscopic methods based on NIRS also offer potential savings for the industry and research in aspects related to the reduction of analysis time and cost. The environmentally friendly nature of the technology also positioned NIR spectroscopy as a very attractive technique. The NIRS models for oil, isoflavones, and tocopherols have been seen in some plants. In this study, NIRS provides reliable, high-throughput, and non-polluting measurement of multiple nutrients, such as FA, anthocyanin, proanthocyanidin, isoflavones, and saponins. The prediction capacity was excellent for total FA, stearic acid, linoleic acid, oleic acid, linolenic acid, insoluble PAs, total Pas, and isoflavones, with RPD values higher than 2.5, SECV values higher than 0.5, and 1-VR values higher than 0.7. The results presented here show that the contents of these important nutrients could be accurately and conveniently analyzed by the NIRS method. In practical application, we should be cautious on the prediction values of palmitic acid, tocopherol, anthocyanins, and soluble PAs because of their lower RPD, 1-VR, and SECV. For these components, it is best to extend our study with more samples from a wider geographical area in order to obtain a model with the greatest prediction. This is the first work that uses NIR spectroscopy for the determination of contents of fatty acids, isoflavones, PAs, anthocyanins, and tocopherols at the same time in soybean seeds. There exist a large number of germplasm resources in China for our screening for any specifically useful soybean variety for our study, our many different kinds of soybean populations for QTL and other genetic studies.

The rapid and accurate NIRS method established here is particularly suitable for large scale screening and is therefore greatly valued. Our results not only validate the rapidness and accuracy of our nondestructive measurement, but also show an establishment of a comparative assay and screening system for large soybean germplasm for maximal soybean resource utilization and genetic populations for possible QTLs and other purposes.

The present research shows that NIRS can be used for quick screening and estimating multiple essential nutrients in soybean, which facilitates breeders for better quality monitoring or researchers for improving the soybean seed nutritional quality.

Acknowledgements This work was supported by the Ministry of Science and Technology of China (grant 2016YFD0100504), and the Major State Basic Research Development Program of China (973 Program 2013CB127001).

Compliance with ethical standards

Conflict of interest Authors declare no conflict of interest

References

1. Cozzolino D. Infrared spectroscopy as a versatile analytical tool for the quantitative determination of antioxidants in agricultural products, foods and plants. Antioxidants (Basel). 2015;4(3):482–97.
2. Zhao J. Nutraceuticals—functional foods for improving health and preventing disease. In: Kayser O, Warzecha H editors. Pharmaceutical biotechnology – drug discovery and clinical applications (2nd Edition). Wiley-VCH Verlag GmbH & Co. KgaA; 2012. P. 599–628.
3. Pande R, Mishra HN. Fourier transform near-infrared spectroscopy for rapid and simple determination of phytic acid content in green gram seeds (*Vigna radiata*). Food Chem. 2015;172:880–4.
4. Brenna OV, Berardo N. Application of Near-Infrared Reflectance Spectroscopy (NIRS) to the Evaluation of Carotenoids Content in Maize. Food Chem. 2004;52(18):5577–82.
5. Lu XN, Rasco BA. Determination of antioxidant content and antioxidant activity in foods using infrared spectroscopy and chemometrics: a review. Crit Rev Food Sci Nutr. 2012;52(10):853–75.
6. Bittner LK, Schonbichler SA, Bonn GK, Huck CW. Near infrared spectroscopy (NIRS) as a tool to analyze phenolic compounds in plants. Curr Anal Chem. 2013;9(3):417–23.
7. Osborne BG, Fearn T, Hindle PT. Practical NIR spectroscopy with applications in food and beverage analysis. Harlow: Addison-Wesley Longman Ltd.; 1993.
8. Pandord JA, Williams PC, deMan JM. Analysis of oilseeds for protein, oil, fiber, and moisture by near-infrared reflectance spectroscopy. J Am Oil Chem Soc. 1988;65(10):1627–34.
9. Osborne BG. NIR Analysis of cereal products. In: Handbook of near-infrared analysis. 3rd ed., Practical Spectroscopy. CRC Press; 2007. p. 399–413.
10. Sun T, Wu YQ, Xu P, Wen ZC, Hu T, Liu MH. Effect of optical length on detection accuracy of camellia oil adulteration by near infrared spectroscopy. Spectrosc Spectr Anal. 2015;35(7):1894–8.
11. Yao X, Wang X, Huang Y, et al. Estimation of sugar to nitrogen ratio in wheat leaves with near infrared spectrometry. J Appl Ecol. 2015;26(8):2371–8.
12. Sato T, Equchi K, Hatano T, Nishiba Y. Use of near-infrared reflectance spectroscopy for the estimation of the isoflavone contents of soybean seeds. Plant Prod Sci. 2008;11(4):481–6.
13. Sikorska E, Górecki T, Khmelinskii IV, Sikorski M, Kozioł J. Classification of edible oils using synchronous scanning fluorescence spectroscopy. Food Chem. 2005;89(2):217–25.
14. Gambarra-Neto FF, Marino G, Araújo MCU, et al. Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis. Talanta. 2009;77(5):1660–6.
15. da Costa GB, Fernandes DDS, Gomes AA, de Almeida VE, Veras G. Using near infrared spectroscopy to classify soybean oil according to expiration date. Food Chem. 2016;196:539–43.
16. Hacisalihoglu G, Gustin JL, Louisma J, et al. Enhanced single seed trait predictions in soybean (*Glycine max*) and robust calibration model transfer with near-infrared reflectance spectroscopy. J Agric Food Chem. 2016;64(5):1079–86.
17. Zhuang H, Ni Y, Kokot S. A comparison of near- and mid-infrared spectroscopic methods for the analysis of several nutritionally important chemical substances in the Chinese yam (*Dioscorea opposita*): total sugar, polysaccharides, and flavonoids. Appl Spectrosc. 2015;69(4):488–95.
18. Pérez-Vich B, Velasco L, Fernández-Martínez JM. Determination of seed oil content and fatty acid composition in sunflower through the analysis of intact seeds, husked seeds, meal and oil by near-infrared reflectance spectroscopy. J Am Oil Chem Soc. 1998;75(5):547–55.
19. Hom NH, Becker HC, Möllers C. Nondestructive analysis of rapeseed quality by NIRS of small seed samples and single seeds. Euphytica. 2007;153(1):27–34.
20. Kovalenko IV, Rippke GR, Hurburgh CR. Measurement of soybean fatty acids by near-infrared spectroscopy: linear and nonlinear calibration methods. J Am Oil Chem Soc. 2006;83(5):421–7.
21. Sundaram J, Kandala CV, Butts CL, Chen CY, Sobolev V. Nondestructive NIR reflectance spectroscopic method for rapid fatty acid analysis of peanut seeds. Peanut Sci. 2011;38(2):85–92.
22. Fernández-Cuesta Á, Fernández-Martínez JM, Velasco L. Identification of high oleic castor seeds by near infrared reflectance spectroscopy. J Am Oil Chem Soc. 2012;89(3):431–5.
23. Roussel V, Branlard G, Vézine JC, Bertrand D, Balfourier F. NIRS analysis reveals temporal trends in the chemical composition of French bread wheat accessions cultivated between 1800 and 2000. J Cereal Sci. 2005;42(2):193–203.
24. Foster JG, Clapham WM, Fedders JM. Quantification of fatty acids in forages by near-infrared reflectance spectroscopy. J Agric Food Chem. 2006;54(9):3186–92.
25. Weinstock BA, Janni J, Hagen L, Wright S. Prediction of oil and oleic acid concentrations in individual corn (*Zea mays* L.) kernels using near-infrared reflectance hyperspectral imaging and multivariate analysis. Appl Spectrosc. 2006;60(1):9–16.
26. García-Olmo J, Garrido-Varo A, De Pedro E. Classification of real farm conditions Iberian pigs according to the feeding regime with multivariate models developed by using fatty acids composition or NIR spectral data. Grasas Aceites. 2009;60(3):233–7.
27. Chen B, Wang J, Zhang G, et al. Two types of soybean diacylglycerol acyltransferases are differentially involved in triacylglycerol biosynthesis and response to environmental stresses and hormones. Sci Rep. 2016;6:28541.
28. Yang W, Cahoon RE, Hunter SC, et al. Vitamin E biosynthesis: functional characterization of the monocot homogentisate geranylgeranyl transferase. Plant J. 2011;65(2):206–17.

29. Endrigkeit J, Wang X, Cai D, et al. Genetic mapping, cloning, and functional characterization of the BnaX.VTE4 gene encoding a gamma-tocopherol methyltransferase from oilseed rape. Theor Appl Genet. 2009;119(3):567–75.
30. Li P, Dong Q, Ge S, et al. Metabolic engineering of proanthocyanidin production by repressing the isoflavone pathways and redirecting anthocyanidin precursor flux in legume. Plant Biotechnol J. 2016;14(7):1604–18.
31. Gu L, Gu W. Spectrophotometric determination of soyasaponins. J Chin Cereals Oils Assoc. 2000;15(6):38–42.
32. González-Martín I, Hernández-Hierro JM, Bustamante-Rangel M, Barros-Ferreiro N. Near-infrared spectroscopy (NIRS) reflectance technology for the determination of tocopherols in alfalfa. Anal Bioanal Chem. 2006;386(5):1553–8.
33. Chu XL, Yuan HF, Lu WZ. Model transfer in multivariate calibration. Spectrosc Spectr Anal. 2001;21(6):881–5.
34. Zhang C, Shen Y, Chen J, Xiao P, Bao J. Nondestructive prediction of total phenolics, flavonoid contents, and antioxidant capacity of rice grain using near-infrared spectroscopy. J Agric Food Chem. 2008;56(18):8268–72.
35. Shenk JS, Westerhaus MO. Population structuring of near infrared spectra and modified partial least squares regression. Crop Sci. 1991;31(6):1548–55.
36. Clemente TE, Cahoon EB. Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol. 2009;151(3):1030–40.
37. Charron CS, Allen FL, Johnson RD, Pantalone VR, Sams CE. Correlations of oil and protein with isoflavone concentration in soybean [*Glycine max* (L.) Merrill]. J Agric Food Chem. 2005;53(18):7128–35.
38. Kim EH, Ro HM, Kim SL, Kim HS, Chung IM. Analysis of isoflavone, phenolic, soyasapogenol, and tocopherol compounds in soybean [*Glycine max* (L.) Merrill] germplasms of different seed weights and origins. J Agric Food Chem. 2012;60(23):6045–55.
39. Wang Z, Chen M, Chen T, et al. TRANSPARENT TESTA2 regulates embryonic fatty acid biosynthesis by targeting FUSCA3 during the early developmental stage of Arabidopsis seeds. Plant J. 2014;77(5):757–76.