Journal of Innovative Optical Health Sciences
Vol. 11, No. 1 (2018) 1750009 (15 pages)

DOI: 10.1142/S1793545817500092

Improvement of NIR models for quality parameters of leech and earthworm medicines using outlier multiple diagnoses

Chunyan Wu, Jiashan Chen, Mengru Li, Yongjiang Wu
and Xuesong Liu*
College of Pharmaceutical Sciences
Zhejiang University, Hangzhou 310058, P. R. China
**liuxuesong@zju.edu.cn*

Received 6 October 2016
Accepted 12 February 2017
Published 27 March 2017

Leeches and earthworms are the main ingredients of Shuxuetong injection compositions, which are natural biomedicines. Near infrared (NIR) diffuse reflection spectroscopy has been used for quality assurance of Chinese medicines. In the present work, NIR spectroscopy was proposed as a rapid and nondestructive technique to assess the moisture content (MC), soluble solid content (SSC) and hypoxanthine content (HXC) of leeches and earthworms. This study goal was to improve NIR models for accurate quality control of leech and earthworm using outlier multiple diagnoses (OMD). OMD was composed of four outlier detection methods: spectrum outlier diagnostic (MD), leverage diagnostic (LD), principal component scores diagnostic (PCSD) and factor loading diagnostic (FLD). Conventional outlier diagnoses (MD, LD) and OMD were compared, and the best NIR models were those based on OMD. The correlation coefficients (R) for leech were 0.9779, 0.9616 and 0.9406 for MC, SSC and HXC, respectively. The values of relative standard error of prediction (RSEP) for leech were 2.3%, 5.1% and 9.0% for MC, SSC and HXC, respectively. The values of R for earthworm were 0.9478, 0.9991 and 0.9605 for MC, SSC and HXC, respectively. The values of RSEP for earthworm were 8.8%, 2.4% and 12% for MC, SSC and HXC, respectively. The performance of the NIR models was certainly improved by OMD.

Keywords: Leech; earthworm; near-infrared spectroscopy; outlier multiple diagnoses.

1. Introduction

Leeches and earthworms are hermaphroditic worms, belong to the phylum Annelida and possess luxuriant medicinal values. Leeches, such as the *hirudo medicinalis*, have been historically used in medicine to remove blood from patients.[1] The

*Corresponding author.

practice of leeching can be traced to ancient India and Greece, and continued well into the 18th and 19th centuries in both Europe and North America. In the 20th century, due to the anticoagulant hirudin in the leech's saliva, leech therapy has established itself in plastic and microsurgery as a protective tool against venous congestion and served to salvage the replanted digits and flaps. This novel therapeutic utilization of leeches resulted in more interest in isolation and characterization of the active constituents of leech saliva.[2] Subsequently, extensive researches on leech saliva unveiled the presence of a variety of bioactive peptides and proteins involving antithrombin (hirudin, bufrudin), antiplatelet (calin, saratin), factor Xa inhibitors (lefaxin), antibacterial (theromacin, theromyzin) and others. In 2004, the Food and Drug Organization (FDA) approved leeches for medicinal purposes.[3] Consequently, the use of leeches in modern medicine made a comeback as a new remedy for many chronic and life-threatening abnormalities, such as cardiovascular problems, cancer, metastasis, and infectious diseases.

Utilization of earthworms has began to extend gradually in the medicinal field, since Shizhen Li compiled the famous medical book Compendium of Materia Medica, in which the earthworm (Pheretima) was recorded as a drug prescribed for antipyretic and diuretic purposes in the form of a dried powder. Bioactive components with medicinal value from earthworms, known as green biomedicine, have already provoked increased attention in Asia and elsewhere in the world. As a result, earthworms have become an international medicine, as well as many medicinal components have been unveiled, including (1) earthworm proteases (lumbrokinase, collagenase, superoxide dismutase, cholinesterase, catalases, glycosidases); (2) metal-binding protein (metallothionein, calmodulin-binding protein); (3) other active proteins including those with proliferative improving activity like lysenin, eiseniapore, antitumor proteins, and glycoprotein; (4) active peptides (gut mobility regulation peptide, antibacterial peptide); (5) earthworm metabolites (carbamidine, lumbrinin, lumbrofobrin, terrestrolumbrolysin); (6) special organic acids (succinic acid, lauric acid, and unsaturated fatty acid) and (7) other components such as purin, vitamin B, tyrosine and Se.[4]

In conclusion, there is no doubt that leeches and earthworms, not only rich in bioactive peptides and proteins but also abundant in nucleosides and nucleobases, possess significantly medicinal values. Recently, nucleobases and nucleosides have been proven as important bioactive compounds involved in multiple biological activities such as anti-platelet aggregation,[5,6] anti-arrhythmic[7] and anti-seizure effects.[8] A research[9] on the identification and quantification of nucleosides and nucleobases in leeches and earthworms was carried out. Fourteen nucleosides and nucleobases were identified and quantified, namely cytosine, uracil, cytidine, guanine, hypoxanthine, xanthine, uridine, thymine, inosine, guanosine, thymidine, 2′-deoxyadenosine, 2′-deoxyinosine and 2′-deoxyuridine. Furthermore, hypoxanthine, uracil, xanthine and inosine were quantitatively determined as the main nucleosides in most earthworms (more than 70% of the total nucleosides and nucleobases) and leeches (more than 60% of the total nucleosides and nucleobases). More importantly, the hypoxanthine content (HXC) was the most among the main nucleosides.

Lots of analytical methods for determination of hypoxanthine were reported, such as hydrophilic-interaction chromatography (HILIC),[9] high-performance liquid chromatography coupled with diode array detection and evaporative light scattering detection (HPLC-DAD-ELSD),[10] high performance liquid chromatography electrospray ionization tandem mass spectrometry (HPLC-ESI-MS/MS),[11] ultra-performance liquid chromatography (UPLC),[12] capillary electrophoresis-mass spectrometry (CE-MS)[13] and capillary electrochromatography (CEC).[14] However, these sophisticated analytical techniques do not possess the rapid, nondestructive nature, pretreatment-simple and reagent-few features. Spectroscopic technologies have high measuring speed and requirement of less or even no sample preparations, such as near infrared (NIR), mid infrared, Raman, X-ray diffraction, etc.

NIR spectroscopy has been proven as an analytical technology with the development of chemometrics and successfully applied to many fields, such as routine chemical analysis and pharmaceutical industries.[15,16] To establish robust models and obtain more accurate predictions, chemometric methods, such as novel spectral pretreatments and original modeling algorithms, have attracted more and more attentions recently.[17] The elimination of outliers is useful to enhance the performance of the established models. Outlier was defined by Johnson and Wichern[18] as "an observation in a data set which appears to be inconsistent with the remainder of that set of data". Outliers existing in the data can

certainly affect the results. It is important to identify and, when appropriate, reject the abnormal data. Conventional outlier diagnoses include mahalanobis distance (MD) and leverage diagnostic (LD). In addition, principal component scores diagnostic (PCSD) and factor loading diagnostic (FLD) are also used for outliers diagnoses. Generally, conventional outlier diagnoses are the most common methods.[19,20] A proposal was made to integrate the four outlier detection methods for outlier diagnosis and this was named "outlier multiple diagnoses" (OMD). OMD can help with identifying and omitting outliers and can produce a more accurate diagnostic result. However, few studies have focused on improving the models via OMD.

This study aimed to improve the performance of NIR models using OMD. Testing involved analysis of HXC, soluble solid content (SSC) and moisture content (MC) in leeches and earthworms. SSC is the amount of active ingredients (soluble peptides, nucleosides, and nucleobases) dissolved in physiological saline solution. MC refers to the percentage of moisture loss at 105°C, excluding the surface MC.

2. Materials and Methods

2.1. *Materials*

The raw medicinal and unbroken leeches and earthworms were provided by Mudanjiang Youbo Pharmaceutical Co., Ltd. (Heilongjiang, China). They were collected from six and seven different plants, respectively. If one plant was regarded as one batch, every batch weighed about 50 g. Each batch was divided into three parts according to the head, the tail and the body. At the same time, the body parts of every batch were stochastically classified into five equal portions. Hence, every batch consisted of seven samples. A total of 42 leech samples and 49 earthworm samples were analyzed. Those samples were separately ground into powders and passed through an 80-mesh sieve. To minimize the effect of surface moisture, the powders were dried to constant weight at 60°C before analysis. For HXC and SSC analyses, all powders were first extracted with physiological saline solution at 10°C for 24 h. Then the extracts were centrifuged at 1500 rpm for 10 min, and the supernatant was stored at 4°C for subsequent analysis.

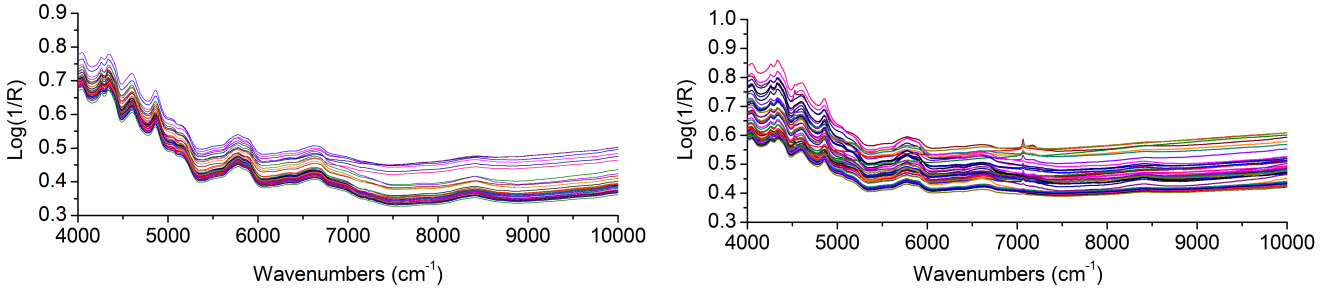
HPLC-grade acetonitrile and methanol were purchased from Merck (Darmstadt, Germany). The hypoxanthine standard (99.6% purity) was purchased from Chengdu Must Co., Ltd. (Chengdu, China). All other reagents were analytical grade.

2.2. *Collection of NIR spectra*

All spectra were collected in the diffuse reflectance mode and obtained by averaging 32 scans using an Antaris II Fourier transform NIR spectrometer (Nicolet, USA). To minimize environment errors, all samples were equilibrated to room temperature (20–25°C) prior to NIR spectra collection. The humidity was kept at ambient levels (60–70% RH) in the laboratory. NIR spectra of leeches and earthworms were collected from 4000 cm^{-1} to 10000 cm^{-1} with a resolution of 8 cm^{-1} and the data are shown in Fig. 1. This study collected 42 average spectra for leech samples (Fig. 1(a)) and 49 average spectra for earthworm samples (Fig. 1(b)). All spectral pretreatments and chemometrics analyses were performed using TQ software.

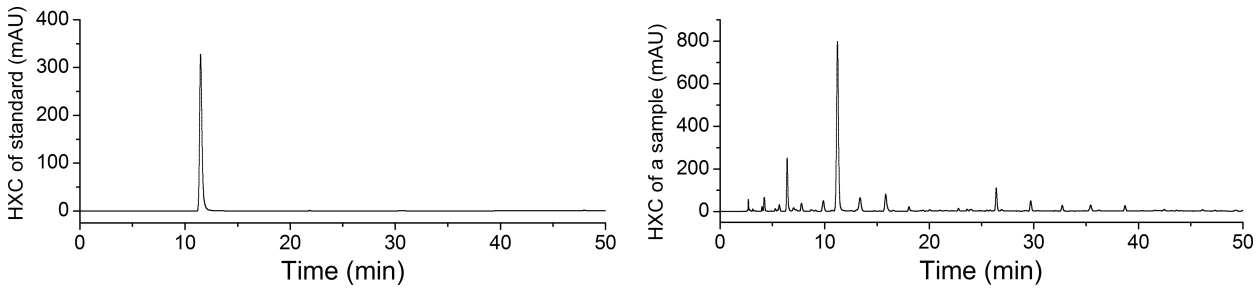
2.3. *HPLC analysis method for hypoxanthine*

HPLC was used as the reference method to quantify the HXC.[21]



(a) (b)

Fig. 1. The raw NIR spectra of leeches (a) and earthworms (b).



(a) (b)

Fig. 2. Chromatogram of hypoxanthine for the standard (a) and one sample (b).

2.3.1. *HPLC conditions*

All samples were separated on an Agilent Eclipse XDB-C18 column (4.6 mm × 250 mm, 5 μm particle size). The mobile phase was a mixture of (A) aqueous solution containing 0.01 mol·L^{-1} potassium dihydrogen phosphate and (B) 50% methanol. The gradient procedure was as follows: initial 100% (A); hold at 100% (A) for 0–5 min; 5–10 min, linear change from 100% to 99% (A); hold at 99% (A) for 10–50 min. The re-equilibration duration between individual runs was 10 min. The mobile phase flow rate was fixed to 1.0 mL/min. The column temperature was kept at 25°C. The detection wavelength was 254 nm.

The extracts were centrifuged at 1500 rpm for 10 min, the supernatant was filtered with 0.45 μm microfiltration membrane, and 10 μL of the filtrate was injected into the HPLC system for analysis. Chromatograms of the hypoxanthine standard (Fig. 2(a)) and one sample of leech (Fig. 2(b)) are shown in Fig. 2. The retention time was 11.2 min for the hypoxanthine. The resolution reached 19.02 for the hypoxanthine.

2.3.2. *HPLC method validation*

The developed HPLC method was validated based on linearity, precision, stability and accuracy. The standard curve was obtained from the linear regression for the peak area versus the respective concentrations for the hypoxanthine standard. The regression equation was $y = 30657x - 42.580$, and the correlation coefficient was 1.0000. A good linear relationship was achieved for the range from 0.06 mg/mL to 0.62 mg/mL for hypoxanthine.

To determine the method repeatability of the method, a randomly selected sample of leech (No. 25) was analyzed by consecutively injecting six needles under the above HPLC conditions. The relative standard deviation (RSD) in the peak areas for hypoxanthine of the instrument precision was 0.11%. This result suggests acceptable instrument precision. Additionally, the stability was tested by analyzing the sample No. 25 every 4 h for 24 h at room temperature. The RSD in the peak areas for hypoxanthine of the sample stability was 1.20%, which indicates the samples were stable for 24 h. The accuracy was evaluated using a recovery test via the standard addition method at three concentrations. Nine samples of No. 25 were spiked with hypoxanthine to 0.5, 1.0 and 1.5 times its amount in the sample. The average recovery was 101.2% for hypoxanthine with RSD values below 2.0%. The above validation data indicates the developed HPLC method was acceptable for determining hypoxanthine.

HXC of all samples were detected at the described conditions. HXC values ranged from 1.1224 mg/g to 2.4903 mg/g for leeches and 0.1526 mg/g to 1.8555 mg/g for earthworms.

2.4. *Outlier multiple diagnoses*

Performance of a multivariate calibration model could be improved significantly by eliminating outliers. Martens and Naes[22] devoted an entire chapter to outliers and discussed conventional outlier diagnoses such as MD and LD in their book "Multivariate Calibration". In the case of multiple outliers, because of masking and swamping,[23] MD and LD may fail to detect true outliers and even mistakenly identify good samples as outliers.[24] To eliminate outliers exactly and get robust models, OMD was proposed and utilized.

2.4.1. *MD*

MD is applied to identify whether samples are outliers based on the spectral information. The MD

measures the distance between a sample spectrum and the mean spectrum of all samples. This diagnosis finds the spectra which are most unlike the other spectra and uses either the Dixon or the Chauvenet test for outliers. If the number of samples is less than 30, the Dixon test is selected. If there are 30 or more samples, the Chauvenet test is used. Because the number of both leech and earthworm samples exceed 30, hence, the Chauvenet test was used for MD. If a sample fails the test, the sample is considered as an outlier.

2.4.2. *LD*

LD shows the relationship between the sample leverage and studentized concentration residual value. The LD can identify those samples that may be outliers. In the LD plot, the data points (one point represents one sample) are expected to be evenly distributed. A data point that is isolated from the others indicate that the corresponding sample is different from the other samples. If a sample with a leverage value or studentized concentration residual value is noticeably different from the leverage values or studentized concentration residual values for the other samples, the sample is excluded as an outlier.

2.4.3. *PCSD*

PCSD is used to diagnose not only the principal components but also the outliers. Normally, the data points representing the samples should evenly distribute in the PCSD plot. The PCSD is applied via calculating score values. A score value represents the multidimensional distance of a sample projected onto a principal component. All of the relevant spectral information in the analysis region or regions of the calibration spectra is condensed into a set of principal components. Each principal component represents an independent source of spectral variation. Principal components are ranked by the amount of variance. Therefore, the first principal component (PC1) and the second principal component (PC2) mainly contain common information in the data. If a data point is isolated from the others, the corresponding sample is different from the others and may be excluded as an outlier.

2.4.4. *FLD*

FLD shows the relationship between the variations of the spectra and contents with certain factors in the analysis region or regions of the calibration samples. The certain factors were confirmed by leave-one-out cross-validation. Each factor represents an independent source of variation. Factors were ranked according to the amount of variation. The first factor describes the principal variation in the calibration samples. Each additional factor describes most of the remaining variations. When the used factor is confirmed, an FLD plot is generated. In the generated plot, one or two samples may be diagnosed as outliers if they are obviously different from most of the samples. The FLD can provide information to help identify samples that may be outliers and decide whether or not to remove them during modeling. If most of the data points are with small variances in an FLD plot, one or two data points with larger concentration variances should be excluded.

2.5. *Chemometrics and data analysis*

Spectral preprocess methods can reduce the effects of systematic noise, baseline variation, light scattering, and path length differences.[25] Multiplicative scatter correction (MSC), standard normal variate (SNV), derivatives, Savitzky–Golay smoothing filter (SGF), Norris derivative smoothing filter (NDF) and combination of them were used for the pretreatment of raw spectra in this work. In general, MSC and SNV are considered as scatter correction methods,[26] and MSC is commonly used to eliminate irrelevant information in the spectra from unknown sources such as surface irregularities, distance variation of sample and detector.[27] Specifically, the MSC corrects any multiplicative effects due to scattering via the linear transformation of each spectrum. The SNV has quite a few differences compared with the MSC. The MSC calculates an ideal spectrum from the calibration standards and uses it to correct the data, while the SNV correction removes the effects of scattering by normalizing the spectra individually.[24] The SNV is recommended to use instead of MSC when the spectra of the unknown samples may have different scattering characteristics than the calibration spectra. Derivatives includes the first derivative (1st Der) and the second derivative (2nd Der). 1st Der was used for removing the baseline, and 2nd Der was introduced to remove both baseline and any spectral baseline drift.[28] The SGF and NDF are used to improve the appearance of peaks that are obscured by random

noise. Besides, the NDF is often used to enhance a sharp band that is overlapped by another broad band. The factors (A set of principal components that contain spectral and concentration information. Factors are used to describe the variation in a PLS method model.) for the calibration model was greater or less than the optimum one, the phenomenon of "overfitting" or "underfitting" would happen, both of which would weaken the performance of the calibration models. To avoid under or over fitting, in this study, leave-one-out cross-validation and the predicted residual error sums of squares (PRESS) value were used to select the optimum factor.[29] The PRESS value decreased obviously with increasing factors. The PRESS value tending to remain almost unchanged or increasing slightly indicated that factor was optimum. The partial least squares (PLS) helped correlate the pretreated spectral data to the indicator contents for constructing the calibration models.[30] The PLS computation were performed by TQ Analyst software (version 8.0).

The predictive capabilities of the developed NIR models were estimated via the correlation coefficient (R), root mean squares error of cross-validation (RMSECV), root mean square errors of calibration and prediction (RMSEC and RMSEP, respectively), and relative standard error of prediction (RSEP). RSEP was calculated for a validation set to assess the quality of the results.[31] The RSEP was calculated as below

$$\text{RSEP} = \sqrt{\sum \frac{(C_1 - C_2)^2}{\sum C_1^2}} \times 100\%.$$

Here, C_1 is the indicator concentration measured by the reference method and C_2 is the indicator concentration predicted by NIR.

An excellent model generally has low RMSEC, RMSECV, RMSEP, RSEP; high R, and a small difference between the RMSEC and RMSECV. Moreover, the RMSEP value should be close to the RMSEC value.

3. Results and Discussion

3.1. *Determination of MC, SSC and HXC*

MC values were calculated according to the drying method of moisture determination in the Chinese Pharmacopoeia (ChP., 2010 version).[32] The MC of leeches and earthworms ranged from 2.1% to 2.7% and 0.5% to 2.7%, respectively.

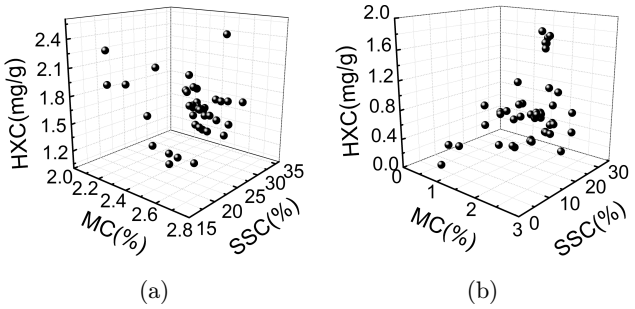


Fig. 3. Diagram of MC, SSC and HXC in leeches (a) and earthworms (b).

SSC was measured using the ChP. Method.[33] The extracts were dried to constant weight at 105°C. SSC ranged from 15.15% to 37.02% and 12.46% to 36.11% for leeches and earthworms, respectively.

HXC were analyzed using the HPLC method described in Sec. 2.3. Figure 3 shows the MC, SSC and HXC of 42 leech samples (Fig. 3(a)) and 49 earthworm samples (Fig. 3(b)).

3.2. *NIR model development*

3.2.1. *Division of calibration set and validation set*

The spectra were randomly divided into a calibration set and a validation set. The calibration set was used to establish a quantitative calibration model, while the validation set was used for testing performance of the established model. The ranges of contents of MC, SSC and HXC in the calibration set covered the ranges in the validation set. Distribution of MC, SSC and HXC of the calibration set and the validation set for the determination of leeches and earthworms by NIR are presented in Tables 1 and 2, respectively. Eight samples named 1, 3, 7, 13, 18, 22, 27 and 33 from leeches were selected into the validation set and the remaining samples were used as the calibration set. Nine samples named 1, 3, 7, 11, 18, 22, 27, 33 and 39 from earthworms were placed into the validation set and the remaining samples were used as the calibration set. The uniform distribution of the validation set in the calibration set were analyzed and testified by principal component scores. A graph of principal component

Table 1. Distribution of MC, SSC and HXC of calibration set and validation set for the determination of leech by NIR.

	Sample sets	Sample number	Content (Min)	Content (Max)	Content ($\overline{x} \pm s$)
MC	Calibration set	34	2.16%	2.64%	2.35 ± 0.11 %
	Validation set	8	2.20%	2.55%	2.37 ± 0.14 %
SSC	Calibration set	34	16.15%	37.02%	29.92 ± 6.43 %
	Validation set	8	22.14%	34.37%	31.50 ± 4.24 %
HXC	Calibration set	34	1.25 mg/g	2.49 mg/g	1.58 ± 0.30 mg/g
	Validation set	8	1.29 mg/g	1.65 mg/g	1.42 ± 0.09 mg/g

scores is presented in Fig. Sup. 1 for leeches (1) and earthworms (2). This graph can demonstrate the rationality of the random splitting for dividing the calibration and validation set.

3.2.2. *Selection of optimal conditions and parameters for NIR model establishment*

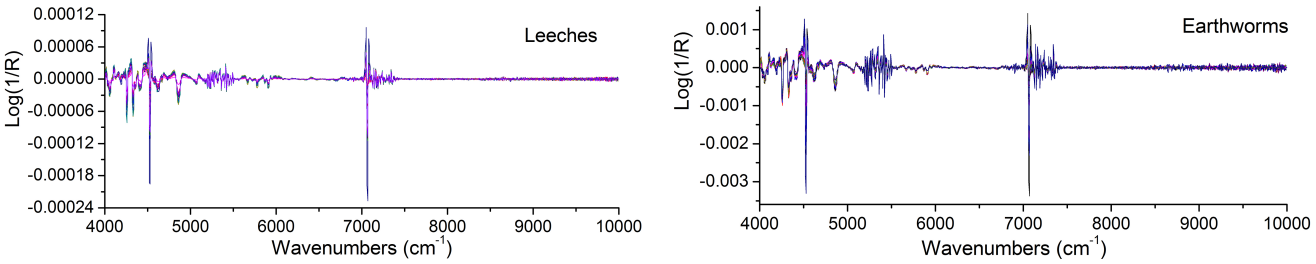
Appropriate selection of specific spectral regions increases usable information and speeds up calibration model computation. The raw NIR spectra of leeches and earthworms are shown in Fig. 1. The 7500–10,000 cm^{-1} region with little characteristic absorption and low signal-to-noise ratio is not recommended for calibration model establishment.[34] There is only a slight variation over the 4000–7500 cm^{-1} region. Therefore, several spectral data pretreatments were used to preprocess NIR spectra for optimizing the calibration performance. These included MSC, SNV, derivatives, SGF, NDF and their combinations. The SNV was selected to remove the effects of scattering, and the 2nd Der was used to eliminate the spectral differences from baseline shifts. To avoid enhancing the noise, the derivative spectra were smoothed with SGF. The pretreated spectra are depicted in Fig. 4. According to the correlation between pretreated spectra and reference values, the optimal conditions and parameters for calibration models of the three indicators are presented in Table 3. The optimum wavebands and factors for MC, SSC and HXC models were suggested by TQ software.

3.2.3. *Establishment of quantitative calibration models*

After identifying the specific spectral regions, selecting appropriate pretreatment methods, choosing the optimum factors, and removing outliers, the manipulated spectral information was correlated with the values measured by the reference assays. The performance of the established models was evaluated in terms of R (the correlation coefficient for the calibration model), R_{CV} (correlation coefficient for leave-one-out cross-validation in calibration), root mean square error of calibration (RMSEC) and root mean square error of cross-validation (RMSECV). A calibration model with high R and R_{CV} as well as low RMSEC and RMSECV with small difference from each other is considered satisfactory. In addition, the predictive ability of the established models was assessed in terms of root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP). The performance parameters of the established models are

Table 2. Distribution of MC, SSC and HXC of calibration set and validation set for the determination of earthworm by NIR.

	Sample sets	Sample number	Content (Min)	Content (Max)	Content ($\overline{x} \pm s$)
MC	Calibration set	40	0.56%	2.69%	1.77 ± 0.47 %
	Validation set	9	1.16%	1.96%	1.57 ± 0.32 %
SSC	Calibration set	40	12.46%	36.11%	24.26 ± 4.82 %
	Validation set	9	18.16%	32.82%	24.65 ± 5.15 %
HXC	Calibration set	40	0.15 mg/g	1.85 mg/g	0.70 ± 0.44 mg/g
	Validation set	9	0.21 mg/g	1.81 mg/g	0.68 ± 0.49 mg/g



(a) (b)

Fig. 4. Chromatogram of pretreated NIR spectra for leech (A) and earthworm (B).

listed in Table 4, which indicate these models have good performance. For example, using the MC of leeches, the regression plots between the reference values and NIR predicted values in the calibration set and the histogram of predictive results for the validation set are depicted in Fig. 5. The two groups of values are highly correlated.

3.3. *Improvement of NIR models using OMD*

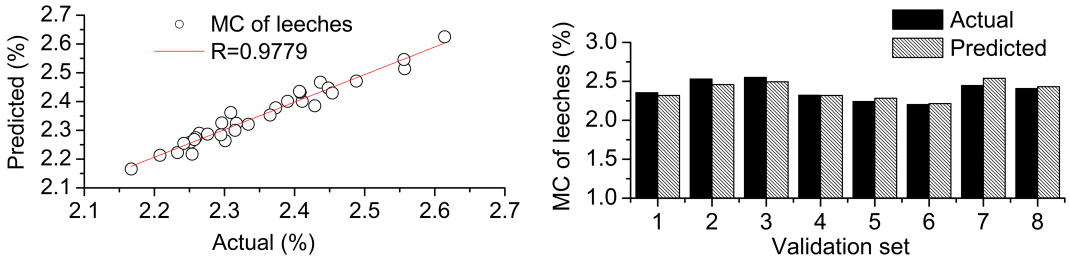
Prior to developing calibration models, it is necessary to identify outlier samples. Outliers in the data can negatively affect the results, and it is important to accurately identify and reject the abnormal data. The performance of a model can be

Table 3. Pretreatments of spectra for MC, SSC and HXC models.

Components	Pretreatment methods	waveband (cm^{-1})	Factors
MC of leeches	SNV + 2nd Der + SGF	5214.57–4855.88 7038.90–6807.49	7
MC of earthworms	SNV + 2nd Der + SGF	7193.18–7112.19 5353.42–5168.29	5
SSC of leeches	SNV + 2nd Der + SGF	7386.03–7316.60 7197.04–7000.33	1
SSC of earthworms	SNV + 2nd Der + SGF	6067.55–5561.40 7395.82–7072.84	8
HXC of leeches	SNV + 2nd Der + SGF	5877.97–5789.26 7081.33–7000.33	4
HXC of earthworms	SNV + 2nd Der + SGF	5955.10–5820.11 5318.71–5303.28	4

Table 4. Parameters of NIR models using OMD.

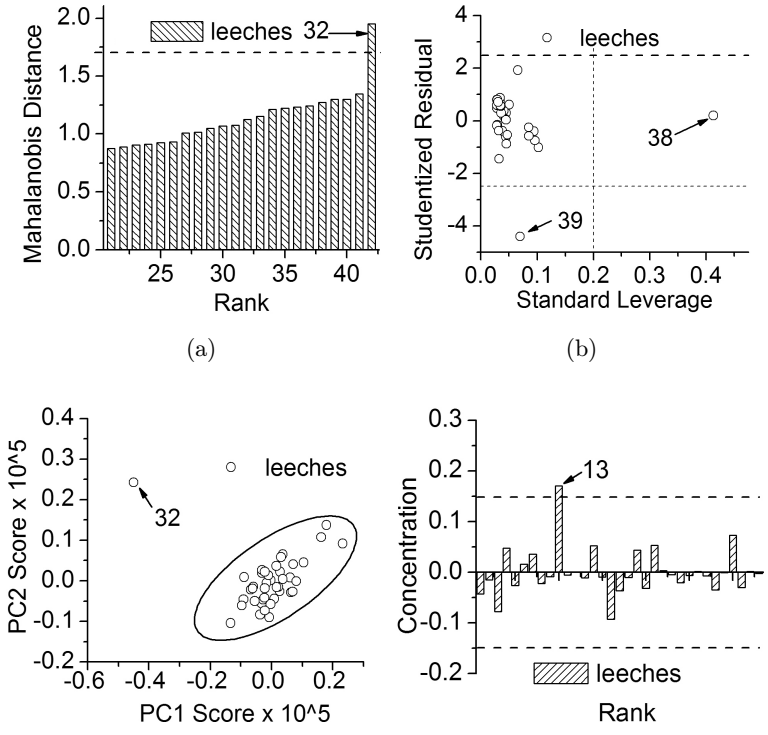
	OMD						Conventional outlier diagnoses					
Model parameters	R	R_{CV}	RMSEC %	RMSECV %	RMSEP %	RSEP %	R	R_{CV}	RMSEC %	RMSECV %	RMSEP %	RSEP %
MC of leeches	0.9779	0.9366	0.0226	0.0629	0.0544	2.3	0.8704	0.8363	0.0525	0.138	0.126	7.8
MC of earthworms	0.9478	0.9116	0.148	0.193	0.141	8.8	0.9203	0.8967	0.190	0.216	0.181	9.1
SSC of leeches	0.9616	0.9542	1.74	1.89	1.64	5.1	0.8991	0.7795	2.84	4.32	3.36	6.2
SSC of earthworms	0.9991	0.9810	0.203	0.918	0.267	2.4	0.9987	0.9771	0.220	0.934	0.275	2.5
	OMD						Conventional outlier diagnoses					
	R	R_{CV}	RMSEC mg/g	RMSECV mg/g	RMSEP mg/g	RSEP %	R	R_{CV}	RMSEC mg/g	RMSECV mg/g	RMSEP mg/g	RSEP %
HXC of leeches	0.9406	0.9050	0.103	0.164	0.129	9.0	0.9164	0.8023	0.123	0.206	0.152	9.3
HXC of earthworms	0.9605	0.9182	0.121	0.172	0.150	12	0.9507	0.9025	0.143	0.197	0.174	18



(a) (b)

Fig. 5. Example calibration plot and validation for MC of leeches.

significantly enhanced if outliers are discarded. Six NIR models were established and improved by employing OMD. The results of OMD for the improved models are shown in from Figs. 6–11. Each of those figures has two histograms (a) and (d) and two plots (b) and (c). In histograms (a), the MD values of the samples are ranked from the smallest to the largest, with a black dashed line representing the threshold of the Chauvenet test. The histograms (a) are aimed at diagnosing spectrum outliers. In plots (b), two horizontal dashed lines represent the upper and lower boundaries of studentized residuals, and vertical dashed lines separate outliers far away from the other data. The plots (c) reflect the PCSD results. Circles were used for visual expression of clustered samples and sparse samples. Generally, those sparse samples outside the circles are treated as outliers. The histograms (d) show the FLD results. Two horizontal dashed lines were added to visualize the samples which have absolute values that were much larger than the others. All diagnosed outliers were flagged with their own names (= numbers).



(a) (b) (c) (d)

Fig. 6. Chromatogram of OMD for MC model of leeches; (a) spectrum outlier diagnostic; (b) leverage diagnostic; (c) principal component scores diagnostic; (d) factor loading diagnostic.

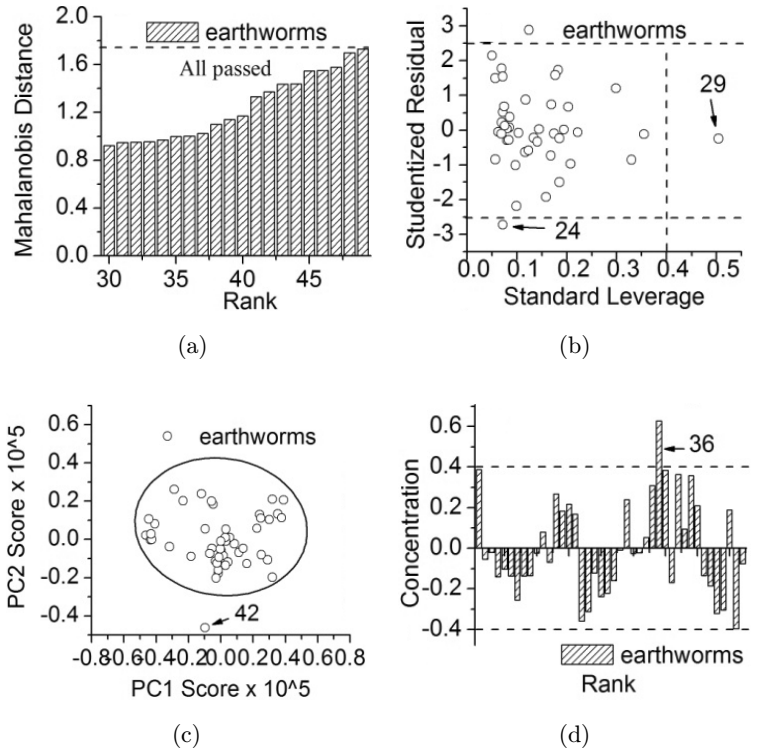


Fig. 7. Chromatogram of OMD for MC model of earthworms; (a) spectrum outlier diagnostic; (b) leverage diagnostic; (c) principal component scores diagnostic; (d) factor loading diagnostic.

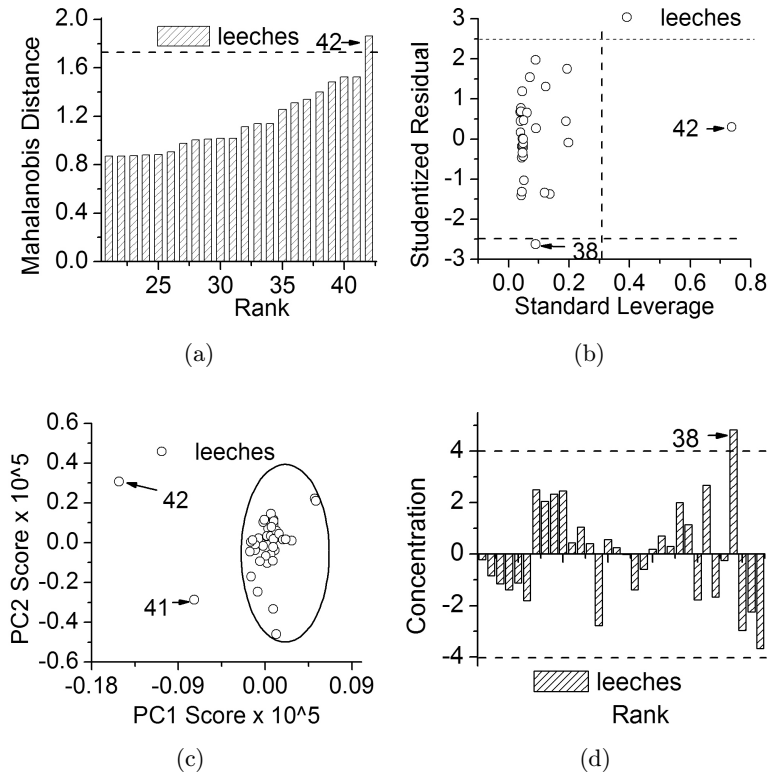


Fig. 8. Chromatogram of OMD for SSC model of leeches; (a) spectrum outlier diagnostic; (b) leverage diagnostic; (c) principal component scores diagnostic; (d) factor loading diagnostic.

3.3.1. *MC models with OMD*

The analytical results of OMD for MC models are displayed in Figs. 6 (leeches) and 7 (earthworms). Figure 6 shows that the OMD identified samples 13, 32, 38 and 39 as outliers, while conventional outlier diagnoses did not recognize the sample 13 as an outlier. If the sample 13 is eliminated, the R value of model is improved from 0.8704 to 0.9779. The RMSEC, RMSEP, RMSECV and RSEP parameters declined from 0.0525% to 0.0226%, from 0.126% to 0.0544%, from 0.138% to 0.0629% and from 3.0% to 2.3%, respectively. In Fig. 7, conventional outlier diagnoses did not recognize samples 36 and 42 as outliers. If the samples 36 and 42 are removed together with samples 24 and 29, the R value increased from 0.9203 to 0.9478, and the other parameters declined from 0.190% to 0.148%, from 0.181% to 0.141%, from 0.216% to 0.193% and from 9.1% to 8.8% for RMSEC, RMSEP, RMSECV and RSEP, respectively. The MC models were all improved by using OMD, by increasing the accuracy of the predicted results.

3.3.2. *SSC models with OMD*

The analytical results of OMD for SS models are displayed in Figs. 8 (leeches) and 9 (earthworms). Figure 8 shows that OMD identified samples 38, 41 and 42 as outliers, while conventional outlier diagnoses did not recognize the sample 41 as an outlier. If the sample 41 is eliminated, the parameter R value was improved from 0.8991 to 0.9616. The parameters RMSEC, RMSEP, RMSECV and RSEP were declined from 2.84% to 1.74%, from 3.36% to 1.64%, from 4.32% to 1.89% and from 6.2% to 5.1%, respectively. In Fig. 9, conventional outlier diagnoses recognized samples 36, 37, 38 and 44 as outliers. However, the sample 44 was not regarded as an outlier according to the PCSD and FLD. If the sample 44 was reserved for modeling, the parameter R value of the SSC model for earthworms increased from 0.9987 to 0.9991, and the other parameters declined from 0.220% to 0.203%, from 0.275% to 0.267%, from 0.934% to 0.918% and 2.5% to 2.4% for RMSEC, RMSEP, RMSECV and RSEP, respectively. Therefore, SSC models with OMD seemed to be superior and more robust.

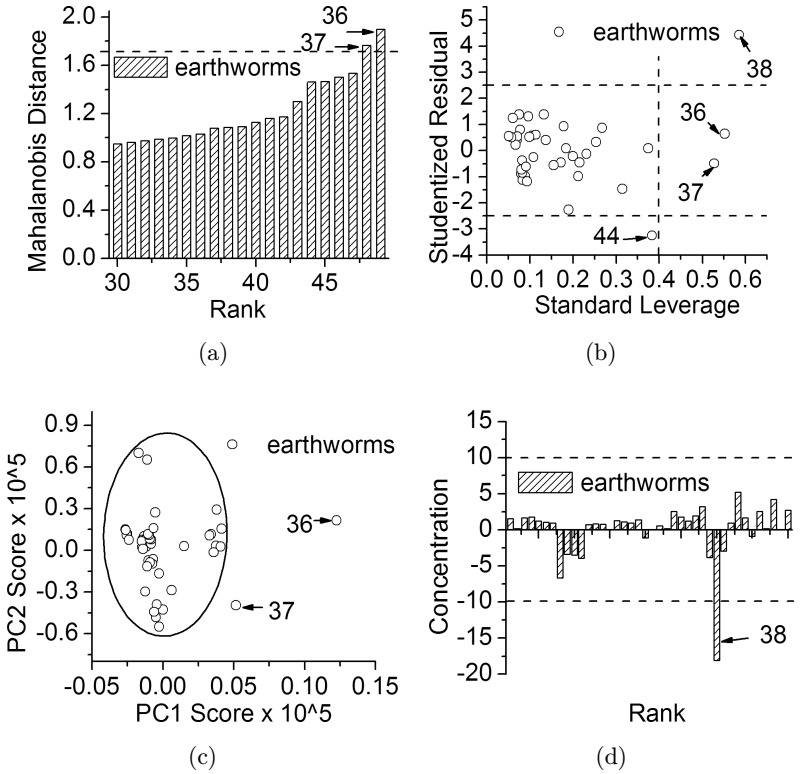


Fig. 9. Chromatogram of OMD for SSC model of earthworms; (a) spectrum outlier diagnostic; (b) leverage diagnostic; (c) principal component scores diagnostic; (d) factor loading diagnostic.

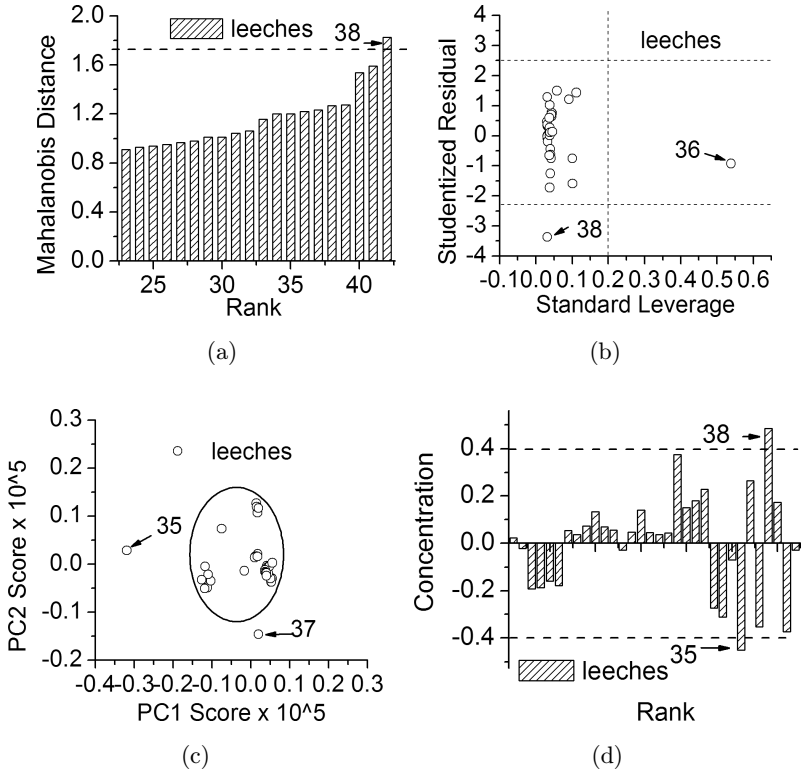


Fig. 10. Chromatogram of OMD for HXC model of leeches; (a) spectrum outlier diagnostic; (b) leverage diagnostic; (c) principal component scores diagnostic; (d) factor loading diagnostic.

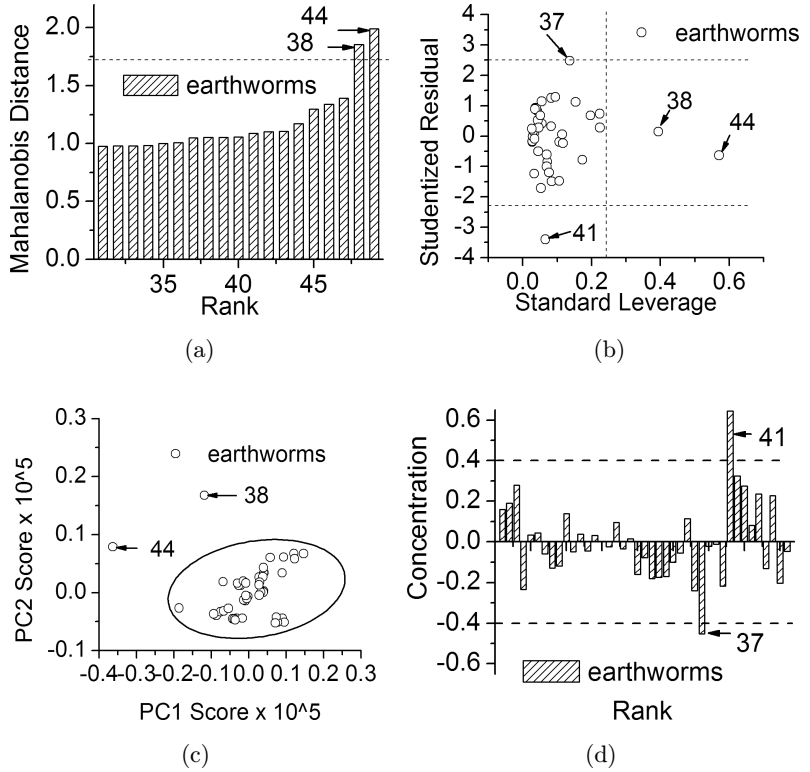


Fig. 11. Chromatogram of OMD for HXC model of earthworms; (a) spectrum outlier diagnostic; (b) leverage diagnostic; (c) principal component scores diagnostic; (d) factor loading diagnostic.

3.3.3. *HXC models with OMD*

The analytical results of OMD for HXC models are presented in Figs. 10 (leeches) and 11 (earthworms). Figire 10 shows that OMD identified samples 35, 36 and 38 as outliers, while conventional outlier diagnoses did not recognize sample 35 as an outlier. If the sample 35 is eliminated, the parameter R value of the HXC model for leeches is improved from 0.9050 to 0.9406. The parameters RMSEC, RMSEP, RMSECV and RSEP declined from 0.123 mg/g to 0.103 mg/g, from 0.152 mg/g to 0.129 mg/g, from 0.206 mg/g to 0.164 mg/g, and from 9.3% to 9.0%, respectively. Figure 11 shows conventional outlier diagnoses did not recognize samples 37 and 41 as outliers. If these two samples are removed, in addition to samples 38 and 44, the parameter R value of the HXC model for earthworms increased from 0.9182 to 0.9605, and the other parameters declined from 0.143 mg/g to 0.121 mg/g, from 0.174 mg/g to 0.150 mg/g, from 0.197 mg/g to 0.172 mg/g, and from 18% to 12% for RMSEC, RMSEP, RMSECV and RSEP, respectively. We conclude that OMD was superior to conventional outlier diagnoses for building HXC models.

The statistics of NIR models based on OMD and conventional outlier diagnoses are listed in Table 4. Based on Table 4 data, it is clear that the NIR models based on OMD had superior calibration results, such as higher correlation coefficients and lower RSEP. All of these results suggest that OMD achieved more satisfactory fitting results and smaller prediction errors. All NIR models provided good prediction results, although the RSEP for HXC model of earthworm was slightly greater.

The SSC models clearly provided the best performance and had the minimum prediction error. This phenomenon was mainly due to the high content of soluble solids, ranging from 15.15% to 37.02% and 2.46% to 36.11% for leeches and earthworms, respectively. In contrast, the MC of earthworms and the HXC of leeches and earthworms were closer to the accepted detection limit of NIR spectroscopy for natural products (0.1%).[35]

On the basis of all the results, OMD was more accurate than conventional outlier diagnoses for establishing satisfactory models.

4. Conclusions

In this paper, MD, LD, PCSD and FLD were mutually complementary. This study established and improved NIR models for accurate prediction of leech and earthworm quality parameters using OMD. The MC, SSC and HXC of leeches and earthworms were quantitatively analyzed and the relevant NIR models were successfully attained based on OMD. OMD generally enhanced the performance and accuracy of NIR models. All R values in NIR models based on OMD were > 0.9 as were the R_{CV} values. In addition, RMSEC, RMSEP, RMSECV and RSEP values of NIR models based on OMD were smaller than of NIR models based on conventional outlier diagnoses. In brief, OMD enabled NIR models to have improved fitting results and smaller prediction errors.

Acknowledgments

We gratefully acknowledge the cooperation and support of Mudanjiang Youbo Pharmaceutical Co. Ltd. (Heilongjiang, China). We also thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

References

1. K. J. Muller, J. G. Nicholls, G. S. Stent (Eds.). *Neurobiology of the Leech*, pp. 27–34, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, ISBN 0-87969-146-8, (1981).
2. J. G. Gasic, D. E. Viner, Z. A. Budzynski, P. G. Gasic. "Inhibition of lung tumor colonization by leech salivary gland extracts from Haementeria ghilianii," *Cancer Res.* **43**(4), 1633–1635 (1983).
3. Y. Munshi, I. Ara, H. Rafique, Z. Ahmad, "Leeching in the history — A review," *Pak. J. Biol. Sci.* **11** (13), 1650–1653 (2008).
4. W. L. Li, C. Wang, Z. J. Sun, "Vermipharmaceuticals and active proteins isolated from earthworms," *Pedobiol.- Int. J. Soil Biol.* **54** (Suppl.), 49–56 (2011).
5. G. Anfossi, I. Russo, P. Massucco, L. Mattiello, F. Cavalot, A. Balbo, M. Trovati, "Adenosine increases human platelet levels of cGMP through nitric oxide: Possible role in its antiaggregating effect," *Thromb. Res.* **105**(1), 71–78 (2002).
6. J. Wang, Z. G. Huang, H. Cao, Y. T. Wang, P. Hui, C. Hoo, S. P. Li, "Screening of anti-platelet aggregation agents from Panax notoginseng using human platelet extraction and HPLC-DAD-ESI-MS/MS," *J. Sep. Sci.* **31**(6–7), 1173–1180 (2008).
7. J. B. Conti, L. Belardinelli, D. B. Utterback, A. B. Curtis, "Endogenous adenosine is an antiarrhythmic agent," *Circulation* **91**(6), 1761–1767 (1995).

8. A. P. Schmidt, D. R. Lara, J. F. Maraschin, A. S. Perla, D. O. Souza, "Guanosine and GMP prevent seizures induced by quinolinic acid in mice," *Brain Res.* **864**(1), 40–43 (2008).
9. P. Chen, W. Li, Q. Li, Y. Wang, Z. Li, Y. Ni, K. Koike. "Identification and quantification of nucleosides and nucleobases in Geosaurus and Leech by hydrophilic-interaction chromatography, *Talanta* **85**(3), 1634–1641 (2011).
10. S. Wang, F. Q. Yang, K. Feng, D. Q. Li, J. Zhao, S. P. Li, "Simultaneous determination of nucleosides, myriocin, and carbohydrates in Cordyceps by HPLC coupled with diode array detection and evaporative light scattering detection," *J. Sep. Sci.* **32**(23–24), 4069–4076 (2009).
11. H. Fan, S. P. Li, J. J. Xiang, C. M. Lai, F. Q. Yang, J. L. Gao, Y. T. Wang, "Qualitative and quantitative determination of nucleosides, bases and their analogues in natural and cultured Cordyceps by pressurized liquid extraction and high performance liquid chromatography–electrospray ionization tandem mass spectrometry (HPLC-ESI-MS/MS)," *Anal. Chim. Acta* **567**(2), 218–228 (2006).
12. F. Q. Yang, J. Guan, S. P. Li, "Fast simultaneous determination of 14 nucleosides and nucleobases in cultured Cordyceps using ultra-performance liquid chromatography," *Talanta* **73**(2), 269–273 (2007).
13. F. Q. Yang, L. Ge, J. W. H. Yong, S. N. Tan, S. P. Li, "Determination of nucleosides and nucleobases in different species of Cordyceps by capillary electrophoresis-mass spectrometry," *J. Pharm. Biomed. Anal.* **50**(3), 307–314 (2009).
14. F. Q. Yang, Sh. P. Li, P. Li, Y. T. Wang, "Optimization of CEC for simultaneous determination of eleven nucleosides and nucleobases in Cordyceps using central composite design," *Electrophoresis.* **28**(11), 1681–1688 (2007).
15. V. M. Fernández-Cabanás, A. Garrido-Varo, J. García Olmo, E. De Pedro, P. Dardenne, "Optimisation of the spectral pre-treatments used for Iberian pig fat NIR calibrations," *Chemometr. Intel. Lab. Sys.* **87**(1), 104–112 (2007).
16. H. M. Mohamed, "Green, environment-friendly, analytical tools give insights in pharmaceuticals and cosmetics analysis," *TrAC Trends Anal. Chem.* **66**, 176–192 (2015).
17. P. J. Tong, Y. P. Du, K. Y. Zheng, T. Wu, J. J. Wang, "Improvement of NIR model by fractional order Savitzky–Golay derivation (FOSGD) coupled with wavelength selection," *Chemometr. Intell. Lab. Syst.* **143**, 40–48 (2015).
18. R. A. Johnson, D. W. Wichern. *Applied Multivariate Statistical Analysis*, 6th Edition Pearson Prentice Hall, USA (2007).
19. X. Y. Chen, Y. Chen, L. H. Wang, C. H. Sun, X. S. Liu, "Fast determination of multiple quality control indexes for concentrating process of Carthamus tinctorius L. Alcohol sedimentation solution by NIRS," *Chin. J. Pharma. Anal.* **30**(11), 2086–2092 (2010).
20. L. J. Luan, N. Chen, X. S. Liu, Y. J. Wu, "Rapid analysis of purification process of grape seed extracts using near infrared spectroscopy," *Chin. J. Anal. Chem.* **40**(4), 626–629 (2012).
21. X. H. Yuan, C. T. Wang, H. U. Jing, "Determination of hypoxanthine in Shuxuetong injection by HPLC," *Chin. J. Biochem. Pharm.* **29**(3), 192–194 (2008).
22. H. Martens, T. Naes, *Multivariate Calibration*, Wiley, New York (1989).
23. C. Andrea, F. Alessio, "General foundations for studying masking and swamping robustness of outlier identifiers," **20**, 79–90 (2014).
24. S. S. Wang, S. Robert, "On masking and swamping robustness of leading nonparametric outlier identifiers for univariate data," *J. Stat. Plan. Inference* **162**, 62–74 (2015).
25. T. Naes, T. Isaksson, T. Fearn, T. Davies, NIR Publications, Chichester: 105 (2002).
26. A. Kohler, M. Zimonja, V. Segtnan, H. Martens. 2.09 - Standard Normal Variate, Multiplicative Signal Correction and Extended Multiplicative Signal Correction Preprocessing in Biospectroscopy, Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, from Comprehensive Chemometrics, 139–162 (2009).
27. J. P. Conzen, *Multivariate Calibration: A Practical Guide for Developing Methods in the Quantitative Analytical Chemistry*, Bruker Optik, Germany, (2003).
28. R. G. Brereton, *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons Ltd., Chichester (2003).
29. G. Gong. Cross-validation, the jackknife and the bootstrap: Excess error estimation in forward regression logistic regression, *J. Am. Stat. Assoc.* **81** (393), 108–113 (1986).
30. S. Wold, M. Sjöström, L. Eriksson. PLS-regression: A basic tool of chemometrics, *Chemometr. Intell. Lab. Syst.* **58**(2), 109–130 (2001).
31. M. Otto, W. Wegscheider, Spectrophotometric multicomponent analysis applied to trace metal determinations, *Anal. Chem.* **57**(1), 63–69 (1985).
32. Chinese Pharmacopoeia Committee, Pharmacopoeia of the People's Republic of China. Beijing: China medical science press, Appendix: 52 (2010).
33. Chinese Pharmacopoeia Committee, Pharmacopoeia of the People's Republic of China. Beijing: China medical science press, Appendix: 62 (2010).

34. N. Gierlinger, M. Schwanninger, R. Wimmer, "Characteristics and classification of Fourier-transform near infrared spectra of the heartwood of different larch species (*Larix sp.*)," *J. Near Infrared Spectrosc.* **12**(2), 113–119 (2004).

35. E. Stark, K. Luchter, M. Margoshes, "Near-infrared analysis (NIRA) a technology for quantitative and quantitative analysis," *Appl. Spectrosc.* **22**(4), 335–399 (1986).