J. Anal. Test. (2017) 1:10
DOI 10.1007/s41664-017-0003-y

ORIGINAL PAPER

Spectral Representation of Proton NMR Spectroscopy for the Pattern Recognition of Complex Materials

Peter de B. Harrington[1] · **Xinyi Wang**[1]

Received: 5 December 2016 / Accepted: 3 January 2017 / Published online: 24 February 2017

Abstract Proton nuclear magnetic resonance (NMR) spectroscopy provides a powerful tool for chemical profiling, also known as spectral fingerprinting, because of its inherent reproducibility. NMR is now increasing in use for authentication of complex materials. Typically, the absorbance spectrum is used that is obtained as the phase-corrected real component of the Fourier transform (FT) of the free induction decay (FID). However, the practice discards half the information that is available in the dispersion spectrum obtained as the imaginary component from the FT. For qualitative analysis or quantitative analysis of small sets of absorbance peaks, the symmetric and sharp peaks of the real spectra work well. However, for pattern recognition of entire spectra, trading peak resolution for peak reproducibility is beneficial. The absolute value of the complex spectrum gives the length or magnitude of magnetization vector in the complex plane; therefore, the magnitude relates directly to the signal (i.e., induced magnetization). The magnitude spectrum is obtained as the absolute value from the real and imaginary spectral components after the FT of the FID. By breaking with tradition and using the magnitude spectrum the reproducibility of the spectra and consequent recognition rates can be improved. This study used a 500-MHz ^{1}H NMR instrument to obtain spectra from 4 diverse datasets; 12 tea extracts, 8 liquor samples, 9 hops extracts, and 25 *Cannabis* extracts. Six classifiers were statistically evaluated using 100 bootstrapped Latin partitions. The classifiers were a fuzzy rule-building expert system (FuRES) tree, support vector machine trees (SVMTreeG and SVMTreeH), a regularized linear discriminant analysis (LDA), super partial least squares discriminant analysis (sPLS-DA), and a one against all support vector machine (SVM). All classifiers gave better or equivalent results for the magnitude spectral representation than for the real spectra, except for one case of the 24 evaluations. In addition, the enhanced reproducibility of the absolute value spectra is demonstrated by comparisons of the pooled within sample standard deviations. For pattern recognition of NMR spectra, the magnitude spectrum is advocated.

Keywords *Cannabis* · Tea · Hops · Liquor · *Humulus* · NMR fingerprinting · Magnitude spectrum · Absolute value spectrum · Pattern recognition · Classification

Introduction

Authentication of herbal medicines and nutraceuticals is growing in importance, especially as the global economy grows and products are shipped worldwide. A useful approach is chemical profiling or spectral fingerprinting of plant extracts [1–5]. Although less sensitive than mass spectrometry (MS), nuclear magnetic resonance (NMR) spectroscopy provides a more reproducible complementary technique for the identification and quantification of metabolites in plant extracts [6].

NMR is a key method for metabolomics and the number of papers has been growing exponentially as demonstrated by a nice review [7]. However, much of this growth has been in targeted analysis for which sets of metabolites are identified and quantified in the NMR spectrum. For

✉ Peter de B. Harrington
peter.harrington@ohio.edu

[1] Clippinger Laboratories, Department of Chemistry and Biochemistry, Center for Intelligent Chemical Instrumentation, Ohio University, Athens, OH 45701-2979, USA

authentication and screening, especially in industry, a faster and easier untargeted analysis approach is provided by chemical profiling which is also known as spectral fingerprinting. These approaches avoid the inherent problems in selecting and quantifying peaks in complex NMR spectra. Chemical profiling is an untargeted analysis for which the individual components of the botanical material are not identified or quantified; instead, the spectra are compared point by point using chemometric classifiers. The use of NMR for untargeted profiling coupled to chemometrics is a burgeoning and important application area. Here are some nice reviews on the topic of NMR metabolic profiling [8–12].

Typically for NMR spectroscopy the real spectral component of the Fourier transform of the free induction decay (FID) is used. After phase correction, the real absorbance spectrum has sharp and symmetric peaks. However, additional information in the imaginary dispersion spectrum is only used for phase-correcting the real spectrum. Because the rotating magnetization vector is modeled in the complex plane by using only the real spectrum some of the analytical signal is unused. The use of the magnitude or amplitude spectrum is proposed because this spectrum although less visually appealing will have greater signal-to-noise and reproducibility compared to the real absorbance spectrum. Reproducibility is important for classification or pattern recognition approaches to work effectively. The increase of signal in the magnitude spectrum results from the greater peak areas of the wider peaks than those found in the real spectrum. This finding is not surprising because it is a trading rule between signal and resolution [13].

NMR was used to profile four diverse sets of extracts. The samples were classified using six different classification methods. The average classification rates were statistically compared between the real NMR absorbance spectra and the magnitude spectra obtained from the absolute value of the complex spectrum. All the validations used 100 bootstrapped Latin partitions (BLPs) [14].

Theory

Pooled Standard Deviation

The pooled standard deviation is a useful measure of experimental uncertainty about the sample mean. It also is useful for scaling the variables of sets of spectra, especially for cases when informative peaks have smaller intensities than other peaks in the spectra. The pooled standard deviation ***sp*** is obtained from the equation given below:

$$sp_j = \sqrt{\frac{\sum_{k=1}^{g}\sum_{i=1}^{m_k}\left(x_{ij} - \bar{x}_{kj}\right)^2}{m-g}}, \quad (1)$$

for which x_{ij} is an element of a data matrix for which each row is an NMR spectrum and each column is a chemical shift measurement. Bold italic upper case typeface denotes a matrix and lower case bold italic typeface denotes a vector. The data matrix $\mathbf{X}$ comprises m rows of spectra and n columns of measurements j. The sum of squares is calculated as the difference between the m_k spectra of each sample or group g and their group mean $\bar{x}_{kj}$. The pooled standard deviation is a measure of the pooled error about the samples.

Fuzzy Rule-building Expert System

The fuzzy rule-building expert system (FuRES) builds a classification tree that comprises branches (i.e., rules) of linear discriminants that minimize the fuzzy entropy of classification. The algorithm initiates by projecting the data from a multidimensional space onto a normalized weight vector to yield scalar scores [15] which are used to calculate the fuzzy entropy of classification. The fuzzy logistic values are the consequents of each rule, and the multivariate rules comprise the branches of the classification tree. The divide and conquer algorithm continues until all the data of each node consist of a single class [16], and the final classification tree allows the visualization of the inductive structure of the rules.

Super Partial Least Squares-Discriminant Analysis

Super partial least squares-discriminant analysis (sPLS-DA) is used as reference method for the other classifiers [17, 18]. The response matrix $\mathbf{Y}$ is a set of binary variables describing the class membership of the spectra in rows of the matrix $\mathbf{X}$. An internal BLP is applied to the training data to calculate an average prediction error [19]. The number of latent variables is selected that yields the least prediction error and then this number is used for the entire calibration set to generate the model. Because the response matrix has a binary encoding, PLS estimates greater than unity or less than zero are set to the corresponding limits (e.g., 0 and 1) during the iterative cycles.

Support Vector Machine

A support vector machine (SVM) is a learning algorithm that can recognize subtle patterns in complex datasets [20]. The SVM is a binary linear classifier that optimizes a classification hyperplane between the surface data points of two clusters in the data space [21]. The one against all

Table 1 Description of the 12 tea samples

ID	Tea name	Water temperature	Amount of tea	Steeping time (min)	Color of the extract
A	Golden Dragon	Before boiling	Level tsp	3	Light green
B	Gyokuro	Before boiling	Level tsp	3	Dark green
C	Puerh Imperial	Boiling	Level tsp	3	Light green
D	Puerh Liu An Anhui				Very light green
E	Sessa Assam	Boiling	Level tsp	3	Colorless
F	Silver Needle	Before boiling	Level tsp	3	Colorless
G	Singelli Darjeeling	Boiling	Level tsp	3	Colorless
H	Tieguanyin				Light green
I	Vivid Huoshan Yellow Bud	Before boiling	Level tsp	3	Very light green
J	White Peony	Before boiling	Level tsp	3	Very light green
K	Wild Yeti				Very light green
L	Yi Wu Beencha	Boiling	Level tsp	3	Light green

Table 2 Description of the eight liquor samples

ID	Type
A	Primary fermentation ambrosia
B	Secondary fermentation ambrosia
C	First bottle
D	First carboy
E	First distillate
F	Second distillate
G	Third distillate
H	Fourth distillate

Table 3 Description of the 25 *Cannabis* extracts

ID	Name
A	Grape Stamper
B	F10
C	HOG
D	Agent Orange
E	Blue Dream
F	Jah Kush
G	Golden Goat
H	Big Black
I	Sour D
J	Denver OG
K	Chem 4
L	Moby Chem
M	Chem 91
N	Micado
O	Head Band
P	Super Lemon Haze
Q	Jack Herer
R	Hit Man OG
S	Wreckage
T	Glass Slipper
U	Skunk
V	Purple Kush
W	Power F10
X	Green Crack
Y	Sage N Saw

method builds an SVM model for each class and all the other objects are grouped together into an opposing class. During prediction, the SVM that yields the largest output designates the predicted class of the object. The main advantage of the SVM is its fast construction of the classification models, especially for megavariate data which have many more measurements than objects.

Support Vector Machine Tree

The support vector machine tree (SVMTreeG) builds a classification tree of SVMs whose encodings are achieved by the separation of scores with the least fuzzy entropy [21]. The key advantage of this tree-based classifier is that nonlinearly separable data may be classified, and for SVMs, this advantage avoids the necessity of finding a workable kernel transform. By variance driven [based on principal component analysis (PCA)] or covariance driven (based on PLS), after the SVM models are built, the one that provides the lowest entropy of classification is the most efficient classifier and is selected for the branch of the tree. The SVMTreeH [22] is a modification to the support vector machine tree that uses fuzzy entropy to encode overlapping clusters in the data space.

Regularized Linear Discriminant Analysis

A regularized version of linear discriminant analysis (LDA) was used that uses a pseudo-inverse to invert the pooled within group covariance matrix [23]. The shortest Mahalanobis distances calculated from the scores on the canonical variates are used to designate the best fitting class membership.

Experimental Section

Sample Preparation

Tea, liquor, and hops samples and Cannabis extracts were supplied by Chemical Mapping, Inc. (Golden, CO). Direct $CDCl_3$ extraction instead of extraction drying and reconstitution was used for samples except for the liquor. Twelve varieties of commercial tea leaves of 50.0 mg each were extracted with 2.0 mL of $CDCl_3$ (99.8%, Sigma-Aldrich, St. Louis, MO, USA) in a glass vial with a screw phenolic cap for 18 h at room temperature; then the extract was vortexed and filtered with 0.45 µm polyvinylidene fluoride (PVDF) filter (Bonna-Agela Technologies, Wilmington, DE, USA). A 693-µL filtrate was mixed with 7 µL of a 1% (*v*/*v*) solution of tetramethylsilane (TMS) in $CDCl_3$ (99.8%, Sigma-Aldrich, St. Louis, MO, USA) in the NMR tube to calibrate the NMR spectra. An overview of the tea extracts according to the labeling is given in Table 1.

For the eight liquor samples, 540 µL of each liquor sample was mixed with 60 µL 99.9% D_2O (Cambridge Isotope Laboratories, Andover, MA, USA) in the NMR tube to calibrate the NMR spectra. An overview of the liquor samples is given in Table 2.

For the nine hops samples, 300.0 mg of each, was powdered by sieve and extracted with 10.0 mL of $CDCl_3$ in a glass vial with a screw phenolic cap for 17 h at room temperature; then the extract was vortexed and filtered with 0.45 µm PVDF filter. The filtrate was treated with 100 rods of 12 mesh 3Å molecular sieves (Fluka Analytical, USA) which were added into each of the vials for more than 24 h before analysis. Then a 500-µL aliquot of the filtrate was mixed with 5 µL $CDCl_3$ with 1% TMS in the NMR tube.

For the 25 *Cannabis* samples, plant buds, 300.0 mg of each, were powdered by sieve and extracted with 10.0 mL of $CDCl_3$ in a glass vial with a screw phenolic cap for 17 h at room temperature; then the extract was vortexed and filtered with 0.45 µm PVDF filter. A 495-µL aliquot of the filtrate was mixed with 5 µL of a 1% TMS in $CDCl_3$ in the NMR tube to calibrate the NMR spectra. Samples were stored in their NMR tubes at 4 °C between daily analyses. An overview of the types of all *Cannabis* samples per the labeling is given in Table 3.

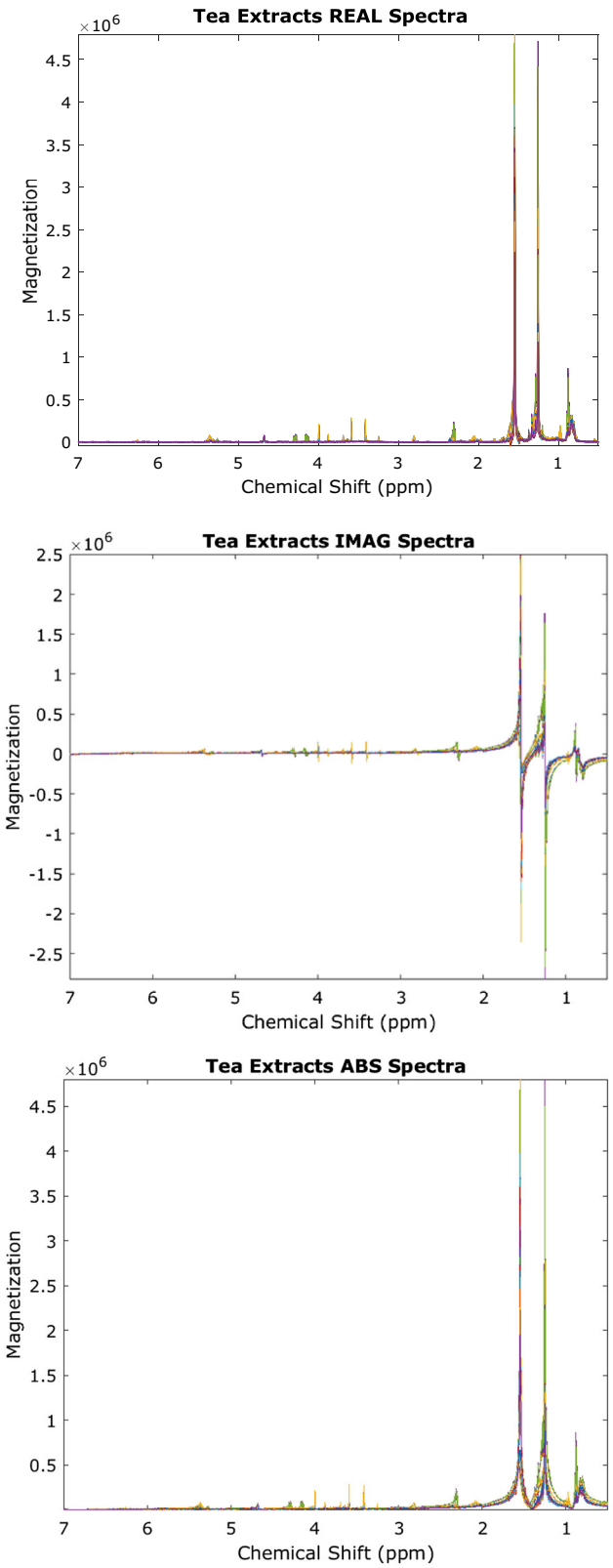

Fig. 1 *Top* REAL absorbance spectra of 60 tea extracts; *middle* IMAG dispersion spectra; and *bottom* ABS magnitude spectra

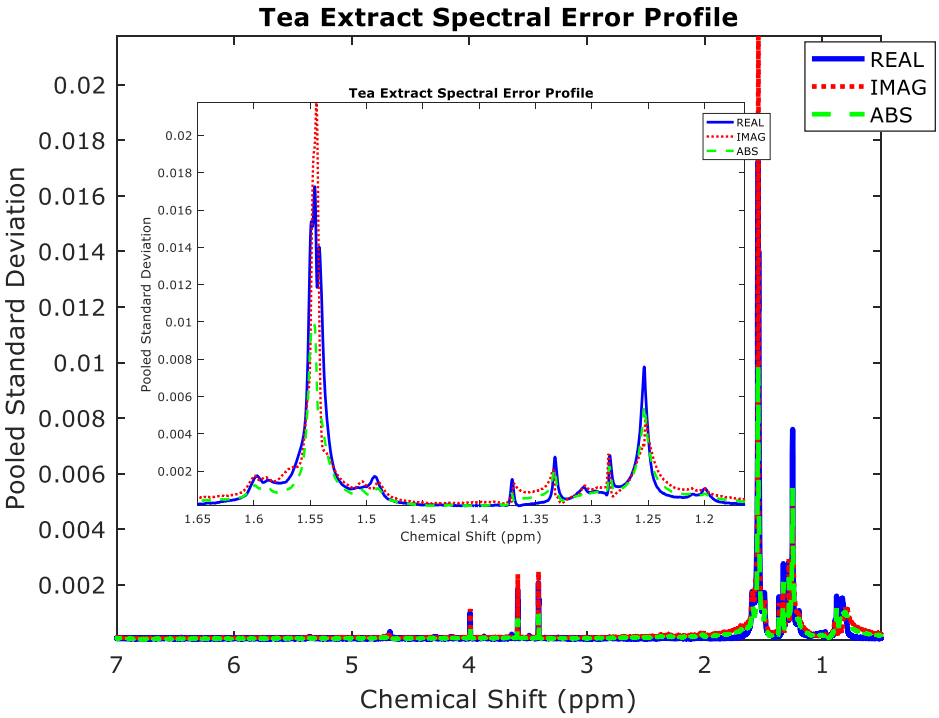

Fig. 2 The pooled standard deviation about the sample means for the REAL, IMAG, and ABS spectral datasets that gives the error with respect to chemical shift

Instrumental Parameters

All the NMR measurements were performed on a Bruker Avance III HD and Bruker Ascend^{TM} 500 nuclear magnetic resonance spectrometer (Bruker BioSpin AG, Fällanden, Switzerland) equipped with a *Ø*5-mm broadband multinuclear (PABBO) probe. Proton NMR spectra were acquired at 298.0 K. Sixteen scans and two prior dummy scans of 65,536 spectra measurements were acquired with a spectral range of 19.9923 ppm. Data were acquired with random block designs with each block collected on a subsequent day to minimize the instrument drifts effect. The IconNMR^{TM} version 4.7 software was used to collect, and TopSpin^{TM} version 3.2 software was used to automatically phase- and baseline-correct the spectra. Chemical shifts were calibrated with the TMS signal at δ 0.00 ppm for all samples except the liquor samples which used the H_2O peak at δ 4.79 ppm [24]. Calibration of the chemical-shifts was accomplished on the instrument using the TopSpin^{TM} software.

Data Processing

All of the raw NMR data were read and converted to the MATLAB mat file format by the *rbnmr* function [25]. All evaluations used the range of [0.5, 7.0] ppm for processing. Each magnitude spectrum was created in MATLAB by using the *complex* function and the phase-corrected imaginary and real spectra from the *rbnmr* function. The absolute value of the complex spectra gave the magnitude spectra. Before multivariate analysis, all the data were normalized to unit vector length. For some datasets, the classification rate was improved by error scaling for which the spectra are divided by the pooled within sample standard deviation. MATLAB R2016b (MathWorks Inc., Natick, MA, USA) was used to process the NMR spectra and calculate statistics from the classification results. The computer was equipped with a Core i7 930 K CPU (Intel Corporation, Santa Clara, CA, USA) operating at 3.2 GHz with six physical and six logical processing units (i.e., hyperthreading turned off). The computer had 64 GBs of quad channel memory. The operating system is MS Windows 8 64-bit Enterprise edition (Microsoft Corp., Redmond, WA, USA).

Discussion of Results

Spectral Representation

Three spectral representations from the Fourier transformed FIDs, the real spectrum (REAL), the imaginary spectrum (IMAG), and the absolute value spectrum (ABS) are given for the set of 60 tea spectra in Fig. 1. The ABS is the absolute value of the complex spectrum (i.e., REAL + IMAG*i*) and represents the magnitude of the magnetization in the complex plane. The peaks of the ABS spectrum are broader and less symmetric than those in the

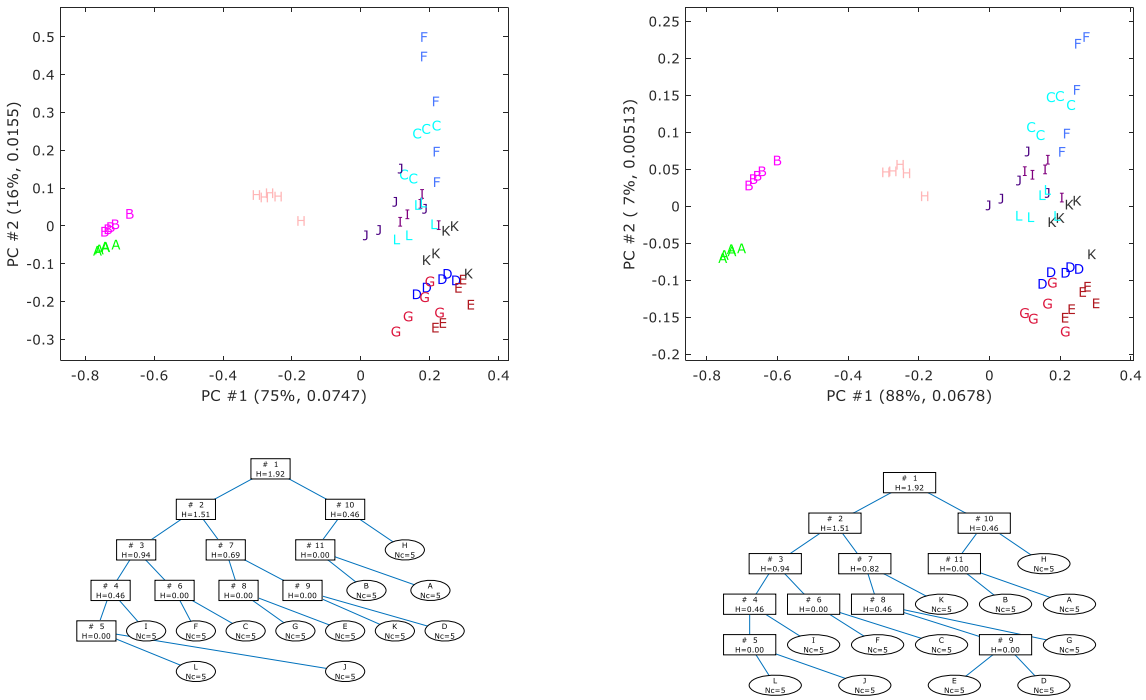

Fig. 3 Tea extracts of 12 samples and 5 replicates. *Top left* principal component scores for the REAL spectra; *top right* principal component scores for the ABS spectra; *bottom left* SVMTreeH for the REAL spectra; and *bottom right* SVMTreeH for the ABS spectra

Table 4 Comparison of spectral representation for 6 classifiers using 100 bootstraps and 5-Latin partitions for 12 tea extracts

	REAL (%)	ABS (%)	T	p value
FuRES	88.4 ± 0.5	92.2 ± 0.4	12.8	«0.001
LDA	96.2 ± 0.3	96.8 ± 0.2	4.1	«0.001
sPLS-DA	99.5 ± 0.2	100.0 ± 0.05	5.7	«0.001
SVM	96.2 ± 0.4	99.2 ± 0.2	17.1	«0.001
SVMTreeG	89.4 ± 0.3	92.9 ± 0.2	20.8	«0.001
SVMTreeH	88.6 ± 0.3	93.9 ± 0.2	29.3	«0.001

Average classification accuracies with 95% confidence intervals

REAL spectrum. For this reason, the REAL spectrum is the preferred choice for spectroscopists who are concerned with qualitative analysis. Note that the IMAG spectrum does not contribute to the amplitude of the ABS spectrum, because it passes through zero at chemical shifts where the REAL peak maxima occur. However, since the contribution occurs at the peak edges, wider peaks will comprise more signal by the larger peak areas. When the peak resolution is unimportant as is the case for spectral pattern recognition and comparison, the ABS spectrum will be beneficial because it uses the entire NMR signal. In theory, the signal-to-noise ratio should improve by a factor of the square root of two.

To evaluate the reproducibility, the pooled standard deviation about the 12 tea sample means was calculated from the normalized spectra. This figure of merit measures the inherent error of the measurement. The pooled standard deviation has two functions for this paper. First, it is used to characterize measurement error of the experiment. Second, it will be used to scale some of the datasets that have high dynamic range (i.e., very large and very small peaks). The benefit will be demonstrated with the liquor study.

The pooled standard deviations for the REAL, IMAG, and ABS spectra are given in Fig. 2. The larger the peak or the intensity of standard deviation, the greater the error. In this figure, the ABS error profile gives the minimum error throughout most of the spectral range, while the REAL and IMAG spectra have greater errors. For pattern recognition, reproducibility is key and the classification results will be consistent with this finding.

All the evaluations of the four datasets used consistent conditions. The spectral range was [0.5, 7.0] ppm to eliminate the solvent peak at δ 7.26 ppm and the TMS peak at δ 0.00 ppm. The number of spectral measurements (i.e., data points per spectrum) was 20,000. Each spectrum was normalized to unit vector length. For two datasets, the liquor and hops, the spectra were scaled by the pooled standard deviation; because those spectra have high dynamic ranges, without scaling poor classification

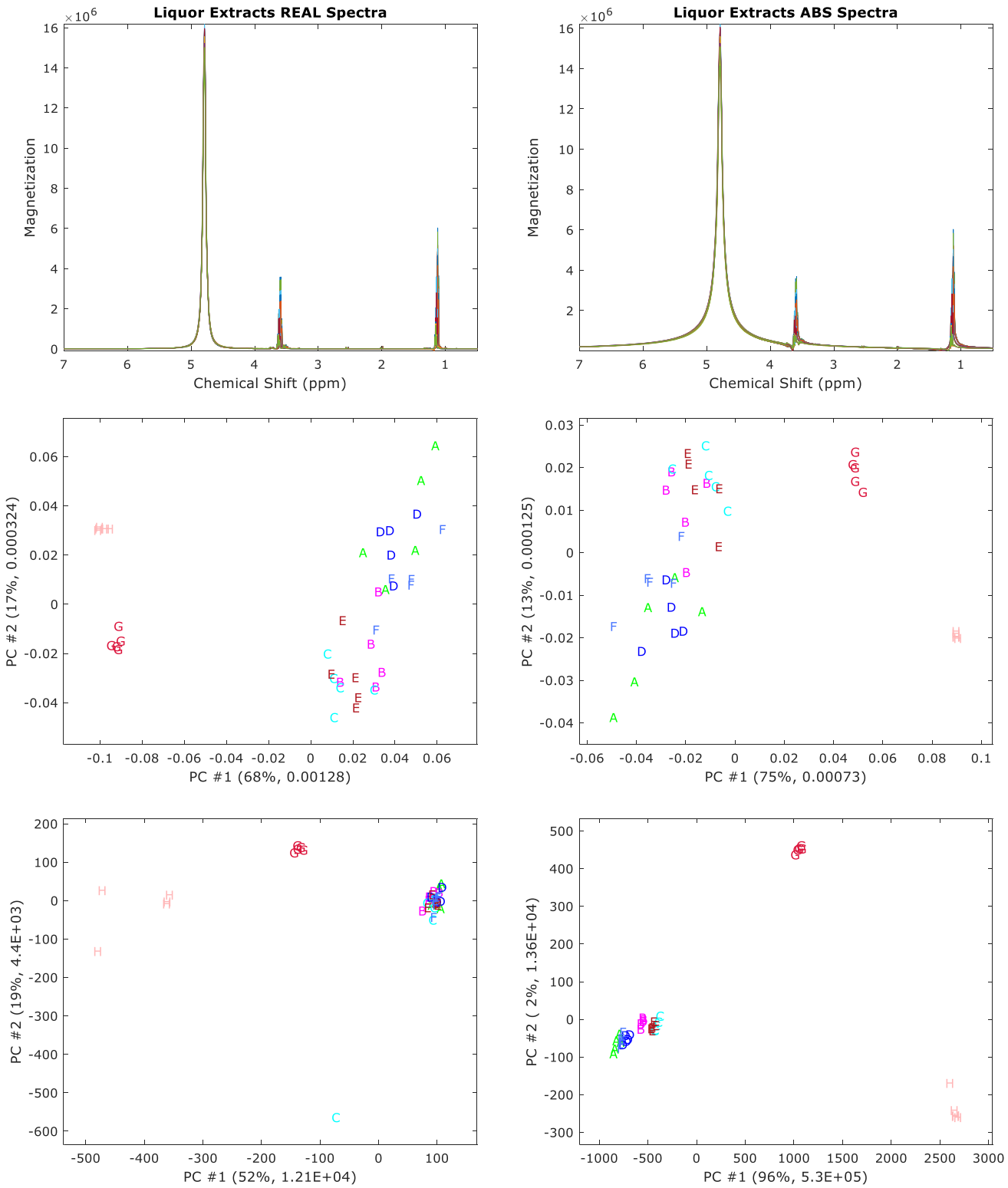

Fig. 4 *Top left* liquor REAL spectra; *Top right* ABS spectra; *middle left* principal component scores of the REAL spectra; *middle right* principal component scores of the ABS spectra; *bottom left* principal component scores of the error-scaled REAL spectra; and *bottom right* principal component scores of the error-scaled ABS spectra

Table 5 Comparison of spectral representation for 6 classifiers using 100 bootstraps and 5-Latin partitions for 8 liquor samples using error scaling

	REAL (%)	ABS (%)	*t*	*p* value
FuRES	83.9 ± 0.7	99.6 ± 0.2	46.0	<0.001
LDA	95.9 ± 0.6	99.5 ± 0.2	13.3	<0.001
sPLS-DA	88.0 ± 0.8	99.6 ± 0.2	26.9	<0.001
SVM	95.5 ± 0.7	99.1 ± 0.3	10.1	<0.001
SVMTreeG	89.0 ± 0.6	99.9 ± 0.1	35.8	<0.001
SVMTreeH	88.8 ± 0.7	99.9 ± 0.1	29.9	<0.001

Average classification accuracies with 95% confidence intervals

Table 6 Description of the nine hops samples

ID	Name
A	Chinook
B	Apollo
C	Mount Hood
D	Centennial
E	Citra
F	Simcoe
G	CTZ
H	Cascade
I	Galaxy

accuracy was obtained (e.g., 60%). This scaling is hence referred to as error-scaling. All comparisons will examine the REAL versus the ABS spectrum because the IMAG spectrum generally gave the worst classification results. BLPs were used to achieve a statistical validation with 100 bootstraps to yield sufficient statistical power. Positive t scores will favor ABS and negative REAL spectral representations. The matched sample *t* test is used to compare the classification results for each bootstrap between the REAL and ABS spectrum.

Most of the classifiers were parameter free, except for the SVM. The SVM had its cost *C* factor arbitrarily set to *inf* which is a MATLAB variable for a very large number. The sPLS-DA was the super PLS implementation which determines the optimal number of latent variables by an internal BLP of the calibration set. FuRES is the softest classifier and tends to be the most sensitive to the representation of the data because it balances variance and bias (i.e., larger peaks are favored over smaller features). The SVMTreeG is the softest of the SVM classifiers and the SVMTreeH trades softness for efficiency in building minimal spanning trees.

A brief description of the teas is given in Table 1. Missing fields in the table correspond to unknown information. The spectra for the tea extracts are given in Fig. 1. The principal component scores and the classification trees are given in Fig. 3. The principal component scores allow for the visualization of the distribution of the spectra. The REAL results are in the left column and the ABS results on the right column of this figure. Both sets of scores appear to be similar; however, the percent total variances (sum of the percentages on each axis) of the ABS scores of 95% is greater than the value for the REAL scores 92%, which indicates that the ABS scores have a better noise distribution. At the bottom are two classification trees obtained from SVMTreeH, a fuzzy entropy-based support vector machine tree. For both trees, all the classes have been resolved. The tree structures are the same except for rules #6, #8, and #9 that characterize groups that are closer together in the dataspace. Table 4 reports the average results of the 100 bootstraps and 5-Latin partitions. The measures of precision presented with the averages are 95% confidence levels. A matched sample *t* test was used to compare the classification rates between the REAL and ABS spectra. Positive *t* scores indicate a higher classification rate for the ABS set of data. For all six classifiers, the ABS spectra gave significantly better classifications.

The next set is a set of eight liquor samples from various phases of production. A description is given in Table 2. Figure 4 demonstrates the usefulness of the error-scaling procedure. The spectra for both the REAL (left) and ABS (right) are dominated by the peaks for ethanol. The characteristic peaks are from the other compounds that are minuscule. The middle of the figure comprises the principal component scores for the normalized spectra and the bottom of the figure comprises principal component scores that were obtained after the error scaling procedure. Two trends are obvious. First, error scaling greatly enhances the resolution of the objects in the different classes by giving appropriate weights to the smaller peaks in the spectra. Second, the ABS spectral scores exhibit much greater resolution of samples than the REAL spectral scores. The classification results using 100 bootstraps and 3-Latin partitions are given in Table 5. The ABS dataset gave significantly improved results for all classifiers.

A set of data were nine samples of hops extracts that had replicate measurements collected on different days. A description of these samples is given in Table 6. The spectra and principal component scores are given in Fig. 5. There are many smaller but characteristic peaks downfield from 2 ppm. For this case, error scaling improved the classification results significantly as well. There are subtle differences between the principal component scores of the REAL and ABS sets. The ABS scores have a greater cumulative variance than the REAL scores. The results are reported in Table 7. For all six classifiers, the results were significantly better for the ABS data.

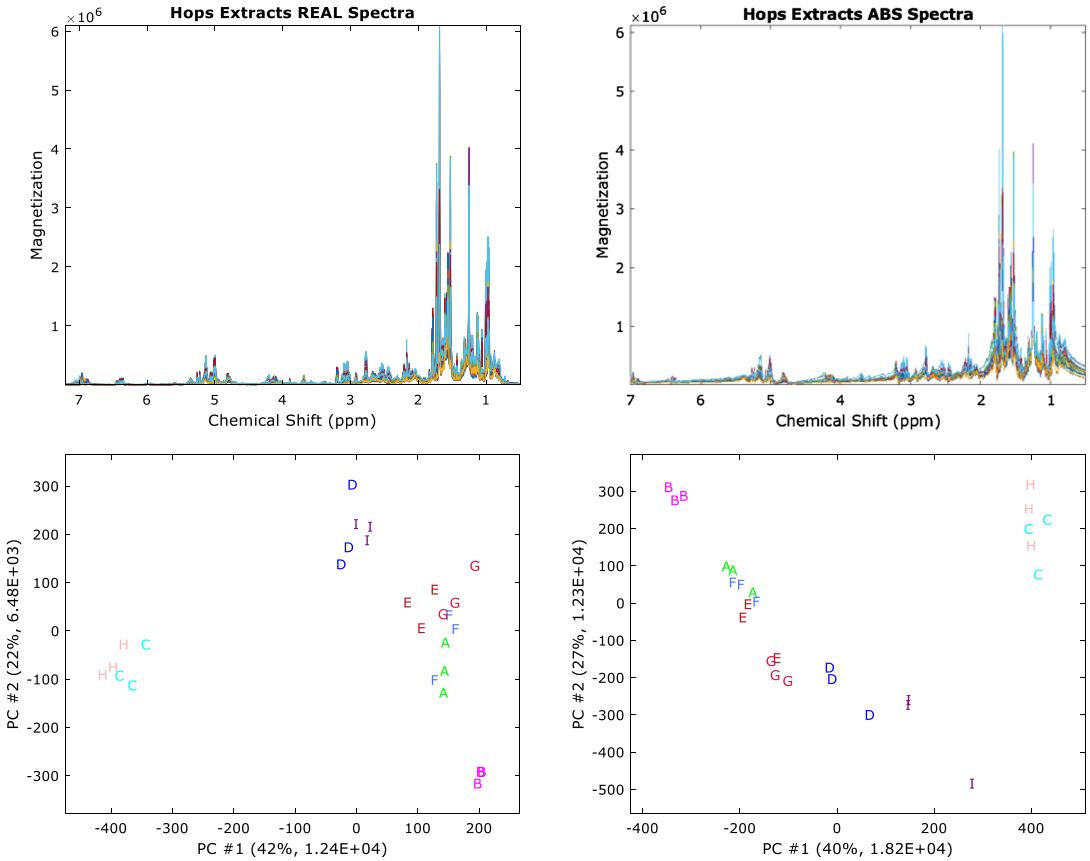

Fig. 5 *Top left* hops REAL spectra; *top right*, ABS spectra; *bottom left* principal component scores of the REAL spectra; and *bottom right* principal component scores of the ABS spectra

Table 7 Comparison of spectral representation for 6 classifiers using 100 bootstraps and 3-Latin partitions for 9 hops extracts using error scaling

	REAL (%)	ABS (%)	*t*	*p* value
FuRES	100.0	100.0	0.0	1
LDA	97.0 ± 0.3	100.0	18.8	<0.001
sPLS-DA	98.4 ± 0.5	99.0 ± 0.4	1.6	0.1
SVM	100.0	100.0	0.0	1
SVMTreeG	98.7 ± 0.3	100.0	7.1	<0.001
SVMTreeH	98.2 ± 0.6	100.0	6.0	<0.001

Average classification accuracies with 95% confidence intervals

The last set was also the largest. It comprised 25 *Cannabis* extracts that each had 5 replicates yielding 125 spectra. Error scaling was not required for this data. Table 3 gives a description of the sample extracts and Fig. 6 contains the spectra and principal component scores. When comparing the principal component scores, REAL has the greater cumulative variance of 80% compared to 79% for the ABS. The classification results are given in Table 8. For all classifiers, except for SVMTreeG, the ABS representation gave significantly better results.

Conclusions

For characterization or authentication of botanical extracts and other complex materials, NMR coupled to pattern recognition is a powerful and robust tool. For pattern recognition spectral reproducibility is important. By adding signal via increased peak width will improve the reproducibility. This requirement may be a departure from conventional NMR spectroscopy for qualitative analysis for which peak resolution is more important. The ABS spectral representation measures the magnitude of the NMR magnetization. It combines the information obtained from the real absorption and imaginary dispersion spectra. The magnitude spectra obtained from the absolute value of the complex spectrum is less visually appealing because the peaks are broader and lack symmetry. However, for pattern recognition of NMR spectra, the increase in reproducibility and signal-to-noise ratio as exhibited by the pooled standard deviation spectrum yields better classification accuracy. This behavior typically occurs as a trading-rule [13] between spectral resolution and signal-to-noise ratio. It also is typical in chemometrics that data beautification by an assortment of methods, e.g.,

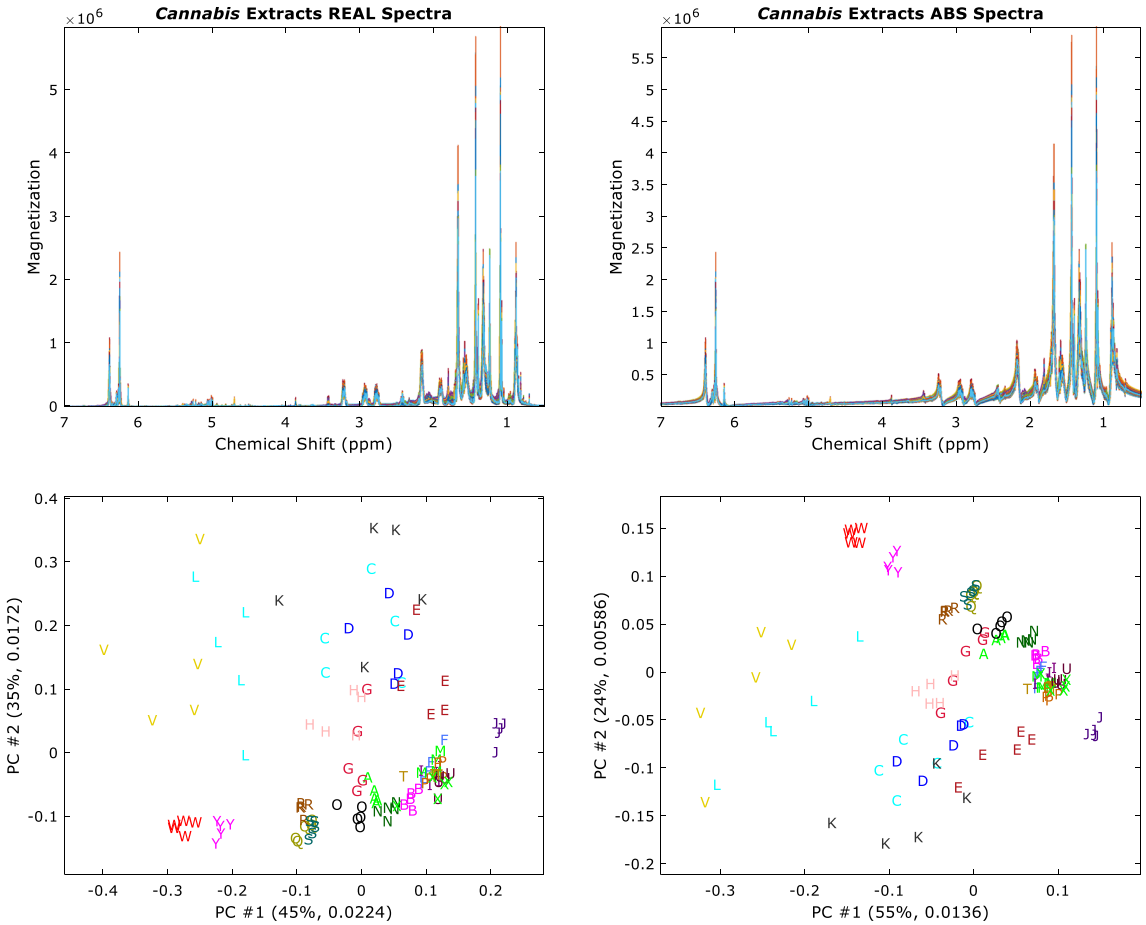

Fig. 6 *Top left Cannabis* REAL spectra; *top right* ABS spectra; *bottom left* principal component scores of the REAL spectra; and *bottom right* principal component scores of the ABS spectra

Table 8 Comparison of spectral representation for 6 classifiers using 100 bootstraps and 5-Latin partitions for 25 *Cannabis* extracts

	REAL (%)	ABS (%)	*t*	*p* value
FuRES	92.4 ± 0.3	94.0 ± 0.3	8.3	«0.001
LDA	98.0 ± 0.2	98.9 ± 0.1	10.4	«0.001
sPLS-DA	99.1 ± 0.1	99.5 ± 0.1	4.3	«0.001
SVM	96.8 ± 0.1	97.6 ± 0.1	9.6	«0.001
SVMTreeG	96.7 ± 0.2	96.2 ± 0.2	−4.8	«0.001
SVMTreeH	94.8 ± 0.3	95.4 ± 0.2	3.4	0.0009

Average classification accuracies with 95% confidence intervals

deconvolution and peak fitting, may make the data visually appealing but at the cost of reducing the inherent reproducibility.

Furthermore, error scaling by using the pooled standard deviation about the sample means provides a measure of the experimental error. It also is beneficial for scaling spectra that have a large dynamic range and a mix of large and small characteristic peaks.

Spectral representations from four diverse sets of data were statistically evaluated with six classifiers. For all 24 classifier comparisons except for one, the ABS spectral dataset yielded improved or equal performance. Therefore, the use of the magnitude or ABS spectrum is advocated for pattern recognition and classification of NMR spectra.

Acknowledgements Steve Baugh at Chemical Mapping, Inc. is thanked for supplying the botanical samples and extracts. Dr. Andrew Tangonan is thanked for his helpful comments in the NMR experiments. The OHIO Center for Intelligent Chemical Instrumentation is thanked for support of this project. We would like also to thank our reviewers for their hard work and helpful suggestions.

References

1. Chen P, Harnly JM, Harrington PD. Flow injection mass spectroscopic fingerprinting and multivariate analysis for differentiation of three *Panax* species. J AOAC Int. 2011;94(1):90–9.
2. Chen P, Luthria D, Harrington PD, Harnly JM. Discrimination among *Panax* species using spectral fingerprinting. J AOAC Int. 2011;94(5):1411–21.
3. Harnly J, Chen P, Harrington PD. Probability of identification: adulteration of American ginseng with Asian ginseng. J AOAC Int. 2013;96(6):1258–65.
4. Sun XB, Chen P, Cook SL, Jackson GP, Harnly JM, Harrington PB. Classification of cultivation locations of *Panax quinquefolius* L samples using high performance liquid chromatography–electrospray ionization mass spectrometry and chemometric analysis. Anal Chem. 2012;84(8):3628–34.
5. Harrington PD, Voorhees KJ, Basile F, Hendricker AD. Validation using sensitivity and target transform factor analyses of neural network models for classifying bacteria from mass spectra. J Am Soc Mass Spectrom. 2002;13(1):10–21.
6. Mahrous EA, Farag MA. Two dimensional NMR spectroscopic approaches for exploring plant metabolome: a review. J Adv Res. 2015;6(1):3–15.
7. Larive CK, Barding GA, Dinges MM. NMR spectroscopy for metabolomics and metabolic profiling. Anal Chem. 2015;87(1):133–46.
8. Monakhova, Y. B.; Kuballa, T.; Lachenmeier, D. W., Chemometric methods in NMR spectroscopic analysis of food products. *J Anal Chem* + **2013,** *68* (9), 755-766.
9. Rolin D, Deborde C, Maucourt M, Cabasson C, Fauvelle F, Jacob D, Canlet C, Moing A. High-resolution H-1-NMR spectroscopy and beyond to explore plant metabolome. In: Rolin D, editor. Adv Bot Res, vol. 67. San Diego: Elsevier Academic Press Inc; 2013. p. 1–66.
10. Lamanna R. Proton NMR profiling of food samples. In: Webb GA, editor. Annu Rep Nmr Spectro, vol. 80. San Diego: Elsevier Academic Press Inc; 2013. p. 239–91.
11. Smolinska A, Blanchet L, Buydens LMC, Wijmenga SS. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. Anal Chim Acta. 2012;750:82–97.
12. McKenzie JS, Donarski JA, Wilson JC, Charlton AJ. Analysis of complex mixtures using high-resolution nuclear magnetic resonance spectroscopy and chemometrics. Prog Nucl Mag Res Sp. 2011;59(4):336–59.
13. Griffiths PR. "Trading rules" in infrared Fourier-transform spectroscopy. Anal Chem. 1972;44(11):1909–13.
14. Harrington PDB. Statistical validation of classification and calibration models using bootstrapped Latin partitions. Trac Trends Anal Chem. 2006;25(11):1112–24.
15. Wang ZF, Chen P, Yu LL, Harrington PD. Authentication of organically and conventionally grown basils by gas chromatography/mass spectrometry chemical profiles. Anal Chem. 2013;85(5):2945–53.
16. Harrington PB. Fuzzy multivariate rule-building expert systems—minimal neural networks. J Chemom. 1991;5(5):467–86.
17. Aloglu AK, de Boves Harrington P, Sahin S, Demir C. Prediction of total antioxidant activity of *Prunella* L. species by automatic partial least square regression applied to 2-way liquid chromatographic UV spectral images. Talanta. 2016;161:503–10.
18. Harrington PD, Kister J, Artaud J, Dupuy N. Automated principal component-based orthogonal signal correction applied to fused near infrared-mid-infrared spectra of French olive oils. Anal Chem. 2009;81(17):7160–9.
19. Selander E, Heuschele J, Nylund GM, Pohnert G, Pavia H, Bjærke O, Pender-Healy LA, Tiselius P, Kiørboe T. Solid phase extraction and metabolic profiling of exudates from living copepods. PeerJ. 2016;4:e1529.
20. Xu ZF, Bunker CE, Harrington PD. Classification of jet fuel properties by near-infrared spectroscopy using fuzzy rule-building expert systems and support vector machines. Appl Spectrosc. 2010;64(11):1251–8.
21. Harrington PD. Support vector machine classification trees. Anal Chem. 2015;87(21):11065–71.
22. Harrington, PB. Support vector machine classification trees based on fuzzy entropy of classification. Anal Chim Acta 2017;954:14–21.
23. Mehay AW, Cai CS, Harrington PD. Regularized linear discriminant analysis of wavelet compressed ion mobility spectra. Appl Spectrosc. 2002;56(2):223–31.
24. Fulmer GR, Miller AJM, Sherden NH, Gottlieb HE, Nudelman A, Stoltz BM, Bercaw JE, Goldberg KI. NMR chemical shifts of trace impurities: common laboratory solvents, organics, and gases in deuterated solvents relevant to the organometallic chemist. Organometallics. 2010;29(9):2176–9.
25. Nyberg N. https://www.mathworks.com/matlabcentral/fileexchange/40332-rbnmr. Accessed 19 Feb 2017.