doi: 10.1255/mrfs.2

Application of recursive partial least square regression for prediction of apple juice sensory attributes from NMR spectra

N. Iaccarino[a], C. Varming[b], M.A. Petersen[b], F. Savorani[b,c], A. Randazzo[a], B. Schütz[d], T.B. Toldam-Andersen[e] and S.B. Engelsen[b]*
[a]Department of Pharmacy, University of Naples "Federico II", Via D. Montesano 49, 80131 Naples, Italy.
E-mail: nunzia.iaccarino@unina.it
[b]Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark.
Corresponding Author: se@food.ku.dk
[c]Department of Applied Science and Technology, Polytechnic University of Turin, Corso Duca degli Abruzzi 24, 10129 Turin, Italy.
[d]Bruker BioSpin, Silberstreifen 4, 76287 Rheinstetten, Germany.
[e]Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark.

This study demonstrates the application of a novel variable selection method here employed for the prediction of sweet and sour taste of apple juice from Nuclear Magnetic Resonance (NMR) spectra. The method is called recursive weighted Partial Least Square (rPLS). It operates by iteratively re-weighting the spectral variables using the regression coefficients calculated by PLS. The only parameter to be estimated by the operator is the number of latent factors to be used in the model. This approach provides an easier model interpretation than a regular PLS model, since it converges towards a very limited number of variables and therefore the assignment effort is drastically reduced. These properties suggest a profitable use of the rPLS for the prediction of even more complex sensory features from different types of spectroscopic data.

Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy has been widely applied to food systems in order to obtain a 'holistic view' of the metabolome (foodome) of various kinds of beverages and foods, such as fruit juice[1,2] milk[3,4], wine[5] and olive oil[6]. In recent years, some studies have focused on the correlation between the NMR metabolomic fingerprint of the food samples and the sensory features evaluated by a panel test. These studies include sour cherry juice[7], tomatoes[8], olive oil[9] and coffee beans extracts[10] and some even suggest that NMR spectroscopy can be considered as a "magnetic tongue" for a better characterisation as well as prediction of the taste of food products. In this context Multivariate Data Analysis plays a fundamental role in the understanding of the correlation between the spectral dataset (**X**) and the response parameters from the sensory evaluation (**y**). For this purpose the Partial Least Square regression[11] is the most widely used algorithm. It first calculates a set of loading weights, **W**, which finds the combination between **X** and **y** and then calculates the regression coefficients, **b**, that provide an estimation of **y** when it is multiplied by the **X** matrix. In this study an advanced version of PLS is performed for the prediction of sweet and sour taste of apple juice from NMR spectra (Figure 1). These attributes are considered important drivers of the market preferences[12], therefore their evaluation is crucial and trained sensory panels are employed for this purpose. Finding a method that is able to predict these features avoiding the employment of the panellists and/or reducing chemical analysis to the minimum would help both companies and researchers in selecting only the best cultivars. As far as apples are concerned, titratable acidity and °Brix values were found to be quite good descriptors respectively for the acid and sweet taste[13]. In this study we propose an approach that could be useful when only spectroscopic data are available and a prediction of sensory attributes is needed. This approach is based on the so-called recursive PLS, or just rPLS[14], which is a recently developed variable selection method where the regression coefficients are recursively used as weights on the original data matrix. This concept is based on the fact that the regression vector reflects the importance of the variables: weights around 0 indicate variables not correlated with **y**, and weights with large absolute values indicate important variables. Repeating the weighting, the rPLS model has the property to converge to a limited number of variables (equal to the number

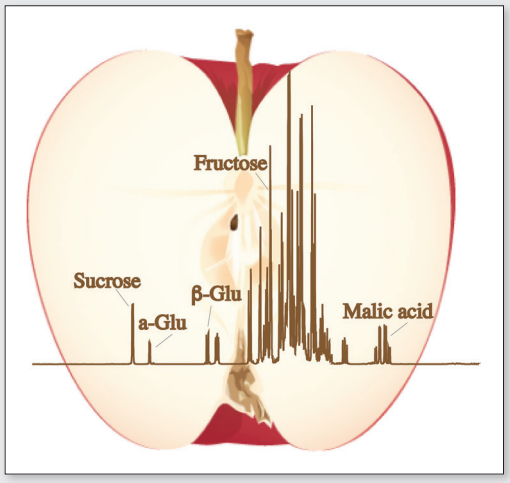

Figure 1. The average NMR spectrum of apple juice.

of Latent Variables/PLS components), facilitating the interpretation and reducing the time consuming step of a thorough signal assignment.

Materials and methods

Samples

The dataset consists of ninety-two apple juice samples obtained from ancient Danish apple cultivars. Each apple juice was then submitted both to NMR analysis and sensory evaluation. Six different descriptors were evaluated: colour, overall odour, apple flavour, overall flavour, sweet taste and sour taste. In order to test the efficacy of the rPLS algorithm, sweetness and acidity, considered easier to interpret, were taken into account in this preliminary study. The sensory panelists were trained with a reference juice as well as with sucrose (11%) and malic acid (0.5%) water solutions for being able to properly recognize all the descriptors. The samples were evaluated using a continuous 0 (none) - 14 (very much) intensity scale and the scores of each sample were averaged over 5 assessors.

Nuclear magnetic resonance data and processing

Bruker Spin Generated Fingerprint (SGF) profiling[15] was employed for the NMR analysis. Each sample required minimal preparation effort consisting of 90% juice with 10% buffer containing 0.1% of TSP (sodium salt of 3-trimethylsilyl-propionate acid-d4) and 0.013% of sodium azide to suppress microorganism activity. This NMR-based screening method is based on an Avance 400 NMR spectrometer with a 9.4-T Ultrashield™ Plus magnet and utilizes flow-injection NMR (BEST™ NMR) with a 4-mm flow-cell probe with Z-gradient and a Gilson liquids handler for sample storage, preparation and transfer. Samples are provided in bar-coded cryovials placed in a Gilson cooling rack that keeps the temperature low (about 4 °C) prior to injection. Then a heated transfer line from the Gilson unit to the probe allows the pre-equilibration of the sample to the desired temperature (300 K) during the transfer. The overall experimental procedure is fully controlled by Bruker's SampleTrack software including temperature adjustment, tuning and matching, locking, shimming and the optimization of the pulses and presaturation power for each sample. The resulting spectra do not need any manual processing step, as they are automatically phase corrected and referenced by the Bruker procedure. Thus, they are ready to be imported in MATLAB (The Mathworks Inc., Massachusetts, USA) where they are at first aligned in the horizontal direction using the icoshift tool developed by Savorani et al.[16] and subsequently mean centered prior to any further chemometrics calculation.

Chemometrics

The rPLS procedure starts by performing a classical PLS regression model between **X** and **y**, while in the following steps, it recursively re-weights the **X** by multiplying it by the regression vector **b** calculated during the previous iteration. This re-weighting is iteratively repeated until no further progress in the regression coefficients occurs. The idea of using the regression vector variables as weights is based on the fact that they reflect the importance of the original spectral variables. Regression vector weights near 0 indicate variables not involved in the correlation with **y**, and weights with large absolute values indicate important variables. The Root Mean Square Error of Cross Validation (RMSECV) is calculated by using venetian blind cross validation (5 groups). The rPLS algorithm is implemented in MATLAB (The MathWorks, Inc.) and made freely available for non-commercial use at www.models.life.ku.dk.

Results and discussion

This apple juice NMR dataset was selected to show the ability of the rPLS algorithm to find the useful information, correlated with the sourness and sweetness in this case, among thousands of spectral variables. The result for the prediction of the sour taste is shown in Figure 2. This plot contains most of the information calculated by the algorithm and it is a direct output of the rPLS script. As observed, the number of variables considered "important" become lower and lower when the iterations increase, until they converge to

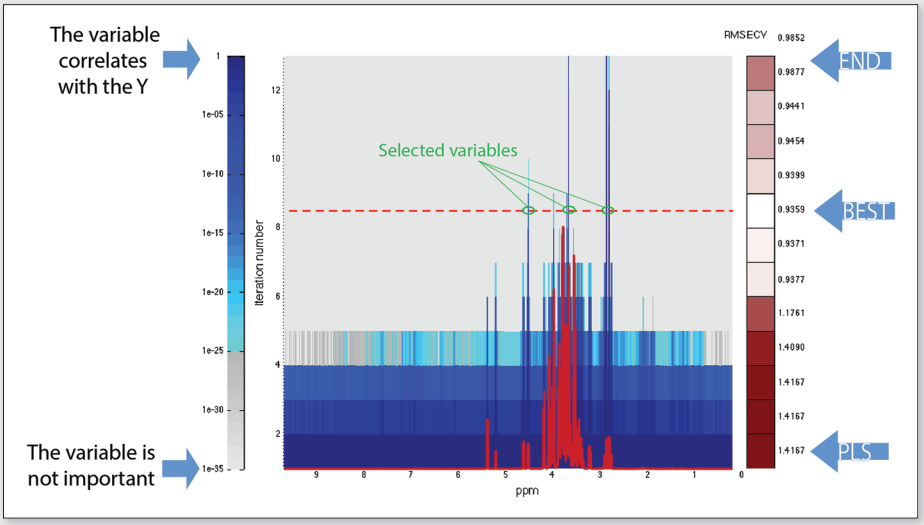

Figure 2. The rPLS result for the prediction of sensory evaluated acidity. The model uses two latent variables. In each row the development of weights according to iterations is shown. The coloured scale on the right represents the RMSECV values, the white box indicates the row where the best model was built and its relative RMSECV value; the bar on the left shows the value of the weights. The value 1 means that the variable has a large weight and thus importance; 1e-35 means that it has not. The red dashed line shows the optimal rPLS model and the green circles indicate the variables selected by the algorithm. The thick red spectrum superimposed to the figure is the average of all the spectra in the dataset.

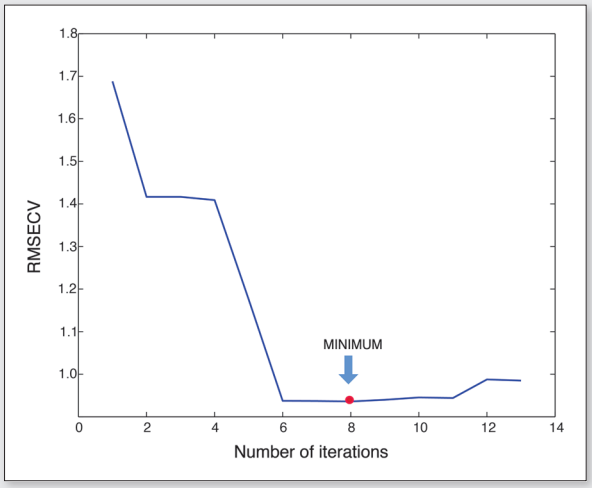

Figure 3. The development in sourness prediction performance (RMSECV) during the rPLS iterations.

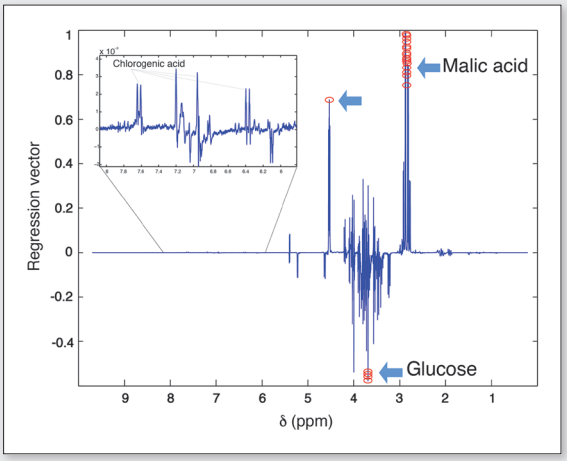

Figure 4. The rPLS result for the prediction of sensory evaluated acidity. The regression coefficients for the full range PLS model (in blue) and for the rPLS reduced model (red circles).

the same number of Latent Variables chosen for the optimal model of the very first step (two in this case). On the left part of the figure, a coloured bar represents the weights used for the variables. The dark blue colour means that the variable has weight equals to 1 and thus it is very important for the regression model. In contrast, the variables situated in the grey region can be considered useless in terms of predictivity since their weight is very close to 0. On the right hand side of the rPLS plot, the RMSECV of each iteration step is shown. The best iterative performance was obtained after eight iterations, as indicated also in Figure 3.

The global PLS model shows a predictive performance of RMSECV=1.70 while the rPLS global minimum shows a predictive performance of RMSECV=0.96. This result has two main advantages, (i) it performs clearly better than the global PLS model and (ii) it is three orders of magnitude more simple, as it contains only 25 variables instead of the 29149 spectral variables included in the global one, allowing the careful inspection of the single variables. Figure 4 shows the regression vector of the global PLS model as well as the variables, counting for the NMR signals at 2.85 ppm (counting for twenty variables), 4.53 ppm (counting for only one variable), 3.69 ppm (counting for four variables), that have been identified by the rPLS as mainly responsible for the sour taste. The peaks can easily be identified as the malic acid methylene (2.85 ppm) and methine (4.53 ppm) protons, while the peak at 3.69 ppm pertains to the glucose pyranose ring protons. These observations are in perfect agreement with the fact that the malic acid is the main acid in apple juice and therefore the main responsible for the sour taste of the samples. Moreover, the fact that glucose is also taken into account by the model, albeit with a numerically lower and negative regression coefficient, indicates an inverse correlation between malic acid content and glucose concentration.

It's also interesting to notice that in the aromatic region, where polyphenols signals arise, the chlorogenic acid shows positive regression coefficients (Figure 4). It is known that polyphenols can give bitterness and astringency to the apple juice[17], however here the main polyphenol found in apple juice seems to have also a positive correlation with the sour taste.

As far as the sweetness is concerned, the best rPLS result occurs after seven iterations (Figure 5). Also in this case the recursive approach brings a clear improvement when compared to the global PLS model, not only in terms of RMSECV, but also in terms of the number of variables to be inspected. Only five peaks have been selected by the rPLS. The signals around 3.81 and 3.67 ppm belong to sucrose and they are positively correlated to the sweetness, while the three glucose peaks (3.99, 3.79 and 3.68 ppm) are negatively correlated to this attribute. The inverse relation between sucrose and glucose content is already known from literature and it is likely due to their interconversion[18]. Surprisingly, the sucrose turned out to be the main responsible for the sweet taste even though the fructose is known to be the main sugar and thus sweetener in apple juice[19]. This confirms the complexity in assigning the sweet taste to a specific chemical compound[13], since it should better be considered as the global result of the combination of several components. One of the advantages of the rPLS approach is that it does not only reduce the variable space and simplify the interpretation of the result, but it also includes the relevant covariation around the selected peaks. The latter information can be extremely useful for assignment purposes.

Conclusion

In this work, we have shown the utility of the rPLS method for the prediction of sensory attributes from Nuclear Magnetic Resonance data of apple juices. Two of the six sensory descriptors available were used for this purpose, namely sweet and sour taste.

In both cases the rPLS was able to develop a good regression model providing just a very limited set of variables that correlate with the y vector. The advantage of this technique is that no parameter must be set by the operator, apart from the optimal number of latent variables required for the initial PLS model. Indeed, the strength of this approach lies in the strong variable selection iteratively performed by the algorithm which not only improves the prediction performance, but it also makes the interpretation of the result tremendously easier for the operator. Thus the aim of this work is to demonstrate the validity of the rPLS method for predicting simple sensory parameters and suggest that it can be a useful tool for the prediction of even more complex sensory features from NMR data avoiding the need of any further chemical analysis. However, it is important to note that when a correlation is obtained from the data,

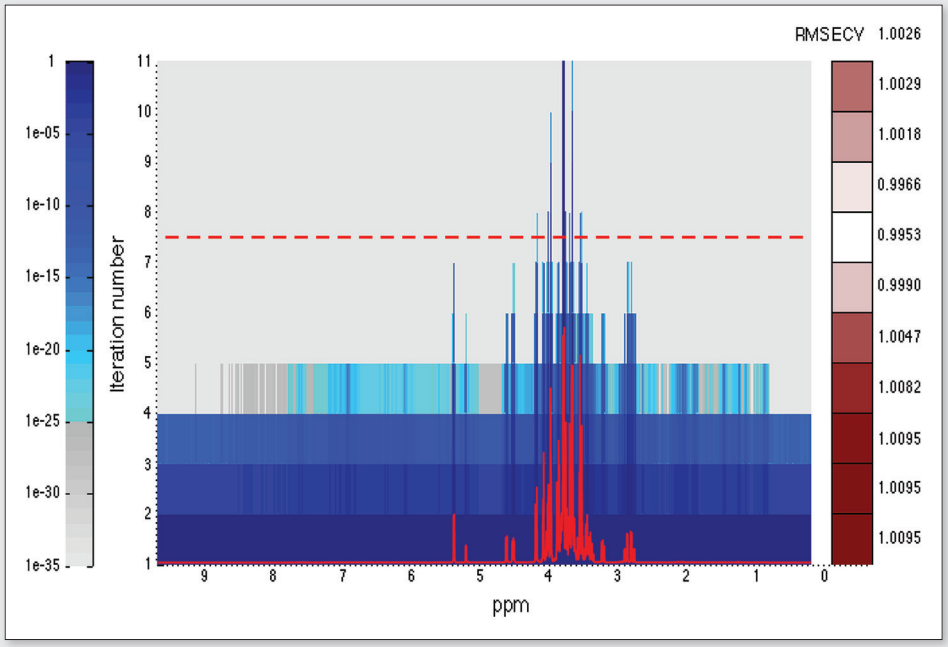

Figure 5. The rPLS result for the prediction of sensory evaluated sweetness.

this does not necessarily mean that there is a direct cause-effect between the correlated parts. An indirect causality is often present in any kind of dataset and this can potentially be misleading for the interpretation. Thus, the results of the model always need to be carefully verified.

The use of the rPLS is also promising for applications to the metabolomics field, as showed by Rinnan et al.[14] since it is able to extract only the useful information from a highly complex metabolomics dataset. Finally, it should be emphasised that in this preliminary research we have been using rPLS to augment the interpretation of the data. If improved prediction models instead are the target, then model validation with an independent test set is obligatory when regression and variable selection is mixed as it is in the rPLS model.

Acknowledgements

The apples were provided by the experimental orchard "Pometum" (Høje Taastrup, Denmark) in the frame of the YDUN project, in collaboration with the Nordic Genetic Resource Center (NordGen). The Ministry of Food, Agriculture and Fisheries financially supported the project through the Danish Food Industry Agency.

References

1. P.S. Belton, I. Delgadillo, A.M. Gil, P. Roma, F. Casuscelli, I.J. Colquhoun, M.J. Dennis and M. Spraul, "High-field proton NMR studies of apple juices", *Magn Reson Chem*. **35**(13), S52-S60 (1997). doi: 10.1002/(SICI)1097-458X(199712)35:13<S52::AID-OMR212>3.0.CO;2-D
2. A.P. Sobolev, L. Mannina, N. Proietti, S. Carradori, M. Daglia, A.M. Giusti, R. Antiochia and D. Capitani, "Untargeted NMR-based methodology in the study of fruit metabolites", *Molecules*. **20**(3), 4088-4108 (2015). doi: 10.3390/molecules20034088
3. J. Belloque and M. Ramos, "Application of NMR spectroscopy to milk and dairy products", *Trends Food Sci Technol*. **10**(10), 313-320 (1999). doi: 10.1016/S0924-2244(00)00012-1
4. F. Hu, K. Furihata, Y. Kato and M. Tanokura, "Nondestructive quantification of organic compounds in whole milk without pretreatment by two-dimensional NMR spectroscopy", *J. Agric. Food. Chem*. **55**, 4307-4311(2007). doi: 10.1021/jf062803x
5. R. Godelmann, F. Fang, E. Humpfer, B. Schütz, M.Bansbach, H. Schäfer and M. Spraul, "Targeted and nontargeted wine analysis by 1H NMR spectroscopy combined with multivariate statistical analysis. Differentiation of important parameters: Grape variety, geographical origin, year of vintage", *J Agric Food Chem*. **61**(23), 5610-5619 (2013). doi: 10.1021/jf400800d
6. R. Sacchi, F. Addeo and L. Paolillo, "1H and 13C NMR of virgin olive oil. An overview", *Magn. Reson. Chem.* **35**, S133-145 (1997). doi: 10.1002/(SICI)1097-458X(199712)35:13<S133::AID-OMR213>3.0.CO;2-K
7. M.R. Clausen, B.H. Pedersen, H.C. Bertram and U. Kidmose, "Quality of sour cherry juice of different clones and cultivars (Prunus cerasus L.) determined by a combined sensory and NMR spectroscopic approach", *J Agric Food Chem*. **59**(22), 12124-12130 (2011). doi: 10.1021/jf202813r
8. A. Malmendal, C. Amoresano, R. Trotta, I. Lauri, S. De Tito, E. Novellino and A. Randazzo, "NMR spectrometers as "magnetic tongues": Prediction of sensory descriptors in canned tomatoes", *J Agric Food Chem*. **59**(20),10831-10838 (2011). doi: 10.1021/jf203803q
9. I. Lauri, B. Pagano, A. Malmendal, R. Sacchi, E. Novellino and A. Randazzo, "Application of "magnetic tongue" to the sensory evaluation of extra virgin olive oil", *Food Chem*. **140**(4), 692-699 (2013). doi: 10.1016/j.foodchem.2012.10.135
10. F. Wei, K. Furihata, T. Miyakawa and M. Tanokura, "A pilot study of NMR-based sensory prediction of roasted coffee bean extracts", *Food Chem*. **152**, 363-369 (2014). doi: 10.1016/j.foodchem.2013.11.161
11. S. Wold, M. Sjöström and L. Eriksson, "PLS-regression: A basic tool of chemometrics", *Chemom Intell Lab Syst*. **58**(2), 109-130 (2001). doi: 10.1016/S0169-7439(01)00155-1
12. S.R. Jaeger, Z. Andani, I.N Wakeling and H.J. MacFie, "Consumer preferences for fresh and aged apples: a cross-cultural compari-

son", *Food Qual Prefer*. **9**(5), 355-366 (1998). doi: 10.1016/S0950-3293(98)00031-7

13. F.R. Harker, K.B. Marsh, H. Young, S.H. Murray, F.A. Gunson, S.B. Walker, "Sensory interpretation of instrumental measurements 2: Sweet and acid taste of apple fruit", *Postharvest Biol Technol*. **24**(3), 241-250 (2002). doi: 10.1016/S0925-5214(01)00157-0
14. Å. Rinnan, M. Andersson, C. Ridder and S.B. Engelsen, "Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS", *J Chemom*. **28**(5), 439-447 (2014). doi: 10.1002/cem.2582
15. M. Spraul, B. Schütz, P. Rinke, S. Koswig, E. Humpfer, H. Schäfer, M. Mörtter, F. Fang, U.C. Marx and A. Minoja, "NMR-based multi parametric quality control of fruit juices: SGF profiling", *Nutrients*. **1**(2),148-155 (2009). doi: 10.3390/nu1020148
16. F. Savorani, G. Tomasi, S.B. Engelsen, "icoshift: A versatile tool for the rapid alignment of 1D NMR spectra", *J Magn Reson*. **202**(2), 190-202 (2010). doi: 10.1016/j.jmr.2009.11.012
17. I. Berregi, J.I. Santos, G. Del Campo, J.I. Miranda, J.M. Aizpurua, "Quantitation determination of chlorogenic acid in cider apple juices by 1H NMR spectrometry", *Anal Chim Acta*. **486**(2), 269-274 (2003). doi: 10.1016/S0003-2670(03)00496-3
18. M. Vermathen, M. Marzorati, D. Baumgartner, C. Good, P. Vermathen, "Investigation of Different Apple Cultivars by High Resolution Magic Angle Spinning NMR. A Feasibility Study", *J Agric food Chem*. **59**, 12784-12793 (2011). doi: 10.1021/jf203733u
19. F. Karadeniz and A. Ekşi, "Sugar composition of apple juices", *Eur Food Res Technol*. **215**(2), 145-148 (2002). doi: 10.1007/s00217-002-0505-2