Chemometrics and Intelligent Laboratory Systems xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Linking GC-MS and PTR-TOF-MS fingerprints of food samples

Luca Cappellin [a,b], Eugenio Aprea [a], Pablo Granitto [c], Ron Wehrens [a], Christos Soukoulis [a], Roberto Viola [a], Tilmann D. Märk [b], Flavia Gasperi [a], Franco Biasioli [a,*]

[a] IASMA Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach, 1, 38010, S. Michele a/A, Italy
[b] Institut für Ionenphysik und Angewandte Physik, Leopold-Franzens Universität Innsbruck, Technikerstr. 25, A-6020, Innsbruck, Austria
[c] CIFASIS, French Argentina International Center for Information and Systems Sciences, UPCAM (France)/UNR-CONICET (Argentina), Bv 27 de Febrero 210 Bis, 2000, Rosario, Argentina

ARTICLE INFO

Article history:
Received 2 December 2011
Received in revised form 28 March 2012
Accepted 11 May 2012
Available online xxxx

Keywords:
PLS
LASSO
Proton transfer reaction-mass spectrometry
Time-of-flight
Prediction
Multivariate correlation

ABSTRACT

Recently the first applications in food science and technology of the newly available volatile organic compound (VOC) detection technique proton transfer reaction-mass spectrometry, coupled with a time of flight mass analyzer (PTR-TOF-MS), have been published. In comparison with standard techniques such as GC-MS, PTR-TOF-MS has the remarkable advantage of being extremely fast but has the drawback that compound identification is more challenging and often not possible without further information. In order to better exploit and understand the analytical information entangled in the PTR-TOF-MS fingerprint and to link it with SPME/GC-MS analyses we employed two multivariate calibration methods, PLS and the more recent LASSO. We show that, while in some cases it is sufficient to consider a single PTR-TOF-MS peak in order to predict the intensity of a SPME/GC-MS peak, in general a multivariate approach is needed. We compare the performances of PLS and LASSO in terms of prediction capabilities and interpretability of the model coefficients and conclude that LASSO is more suitable for this problem. As case study, we compared GC and PTR-MS data for different matrices, namely olive oil and grana cheese.

1. Introduction

Proton transfer reaction-mass spectrometry (PTR-MS) allows the on-line monitoring of volatile organic compounds (VOCs) with low detection limit and fast response time [1]. It is considered an essential tool for environmental chemistry and environmental sciences [2], which are probably the fields wherein PTR-MS is mostly applied [1]. It has, however, also been applied successfully in medical science [3] and food science and technology [4], agronomy [5] and genetics [6]. The rapidity of PTR-MS fingerprints makes it possible to analyse a great number of samples in a much shorter time than more established techniques such as GC-MS. In fact, the time required to characterize a single sample can be reduced of about one hundred times with the use of PTR-MS in place of GC-MS. The questions arise whether the two approaches provide comparable or complementary information and which is the better way to exploit the information entangled in these two different approaches. On the one hand, it is well established that compound identification is usually possible with GC-MS, helped by the availability of commercial mass spectra libraries (i.e. NIST Mass Spectral Database, Wiley Mass Spectral Libraries) containing reference electron-impact (EI) spectra for a large number of compounds. The link between GC-MS data and VOC headspace concentration can then be obtained via a calibration procedure. On the other hand, with the first realisations of PTR-MS apparatuses, usually equipped with a quadrupole mass analyser, compound identification is normally very difficult [7], since usually only the nominal (protonated) VOC m/z value is measured. Recently a new version of the PTR-MS based on time-of-flight mass spectrometer has been commercialised: the PTR-TOF-MS [8]. It is characterized by a larger mass range (up to 400 Th in our settings), a faster acquisition time (0.1 s), and a high mass resolution (nominally $\Delta m/m$ up to 8000) [8]. This last improvement, together with an achievable high mass accuracy [7], allows in many cases the separation of isobaric compounds and strongly enhanced the possibility of compound identification or, at least, provides the sum formula of the observed peaks. The further step of identifying the actual VOCs, although facilitated by the above mentioned ameliorations remains often still only tentative and relies very much on the knowledge of the nature of the considered samples to rule out improbable isobars. Moreover, the presence of fragments, which are common to several compounds, often complicates PTR-TOF-MS spectra. In fact, thanks to its limited collision energy PTR induced fragmentation is often reduced but it remains an issue especially when complex mixtures have to be measured, as it is the case of food samples.

Therefore, the link between PTR-TOF-MS peaks and GC-MS data of the same sample is generally not obvious, meaning that a one to one relation between GC-MS and PTR-MS peaks is, in general, not expected. Sometimes it is possible to connect a single compound identified by GC to a single PTR-MS peak as it is often the case in

* Corresponding author. Tel.: +39 0461 615187; fax: +39 0461 650956.
E-mail address: franco.biasioli@iasma.it (F. Biasioli).

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2012.05.008

environmental chemistry [1] and has been also used by food scientists [9–12]. For instance Nestlè laboratories showed that it is, in some cases, possible to demonstrate by off line analysis that the peak at $m/z = 72$ (corresponding to the protonated compound) can be used as a monitor for acrylamide concentration [9]. In a similar way the technique has been used to quantify furan and methylfuran in different matrices [13] or benzene formation in drinking model systems [10]. However, this is not the case in most situations because of the presence of residual fragmentation and isobaric compounds. Thus, the problem of correlating PTR-MS fingerprints with GC-MS indication cannot be addressed by univariate correlation (one-to-one) but is, in general, a complex problem of multivariate nature.

PLS (Partial Least Squares) [14] is one of the most popular multivariate regression methods, especially in the field of chemistry that also formed its origin. It has shown good performance in a wide variety of fields and is available in many different software packages (e.g. [15]). Moreover, it has successfully been applied to problems similar to ours, e.g. [16,17]. It can be shown that PLS effectively applies a shrinkage penalty to the regression coefficients in order to regularize the otherwise underdetermined system of equations — PLS is most popular in areas where the number of variables regularly exceeds the number of samples [18,19]. Also, alternatives like Principal Component Regression [20] and Ridge Regression [21] can be shown to shrink the regression coefficients [18]. The LASSO (Least Absolute Shrinkage and Selection Operator) [22] is a relatively new alternative. It, too, employs shrinkage, but rather than a quadratic penalty on the coefficient size, it uses the absolute size of the coefficients. An interesting side effect is that for any given size of the penalty, only a limited number of coefficients is non-zero. This allows to utilize LASSO as a variable-selection tool and presents significant advantages when not only prediction quality is important but also model interpretation.

In the present work, we employ two efficient multivariate correlation methods, PLS and LASSO, to tackle the problem of the correlation between PTR-MS and GC-MS assessment of the headspace of agroindustrial samples. In particular, we concentrate on the ability of these methods of predicting VOC concentrations, as measured by GC-MS, starting from PTR-TOF-MS spectra. We compare the performance of PLS and LASSO and discuss the interpretation of the model coefficients on the basis of the PTR-MS fundamentals. As a first case study we consider a sample set of 72 Trentingrana cheese that have been produced under controlled cheese-making procedure starting from milk stored in different conditions. The original goal of the experiment, described in Fabris et al. [23], was to evaluate the effect of milk storage temperature and collection modality on the final quality of ripened cheese. As a second case study, the headspace of 56 extravirgin olive oils produced in a pilot scale plant under controlled conditions is considered. The aim of this study was the characterization of a large number of monocultivar olive oils obtained from an olive tree cultivars collection in Tuscany (Italy).

2. Materials and methods

2.1. GC/SPME-MS

Volatile organic compounds, present in the headspace of food samples, equilibrated at 40 °C for 30 min before the analysis (10 min in the case of olive oils), were extracted and preconcentrated (30 min) by means SPME (Solid Phase Microextraction) according to the procedure reported in Endrizzi et al. [24] and were analysed in a GC interfaced with a quadrupole mass detector which operates in electron ionisation mode (EI, internal ionisation source; 70 eV) with a scan range from m/z 30 to 300 (GC Clarus 500, PerkinElmer, Norwalk CT, USA). Separation was achieved on a HP-Innowax fused-silica capillary column (30 m, 0.32 mm ID, 0.5 μm film thickness; Agilent Technologies, Palo Alto, CA, USA). Compound identification was based on mass spectra matching with the standard NIST-98/Wiley library and linear retention indices (LRI) of authentic reference compounds. Further details can be found in [24]. In the case of olive oils, we refined our results using the R package "xcms" [25] which allows to analyse the filtered chromatogram for each nominal mass. Via selecting a suitable fragment in order to estimate the concentration of a particular compound, we tried to avoid contaminations from other compounds of similar retention time but different fragmentation pattern and problems related to detector saturation. Moreover, our approach allowed to extract the intensity of some peaks which in the total ion chromatogram were hidden under the peak of another compound present in larger amount, while displaying a clear signal in the filtered chromatograms. For olive oils VOC concentrations were expressed in mg/kg equivalent of the internal standard 4-methyl-2-pentanol. In the case of grana cheese the internal standards were 4-methyl-2-pentanone, ethyl heptanoate and isobutyric acid [24].

2.3. PTR-TOF-MS

Rapid PTR-TOF-MS measurements were performed with a commercial PTR-TOF 8000 instrument supplied by Ionicon Analytik GmbH, Innsbruck (Austria) [8]. The TOF was operated in V mode. The sampling time of the TOF spectra was 0.1 ns and the ionisation conditions were controlled by drift voltage (600 V), drift temperature (110 °C) and drift pressure (2.11 mbar). The mass resolution was about 4000 ($m/\Delta m_{50\%}$). Samples were equilibrated at 40 °C for 30 min in a water bath before the analysis to mimic the GC procedure; they were then measured by direct injection of the headspace mixture into the PTR-TOF-MS drift tube via a heated (110 °C) peek inlet for 20 s, allowing to take 20 averaged spectra. All spectra were corrected for count losses due to the detector dead time [26] and calibrated in the m/z domain according to [7]. Peak extraction was performed according to the methodology described in [27]. Normalization of ion counts was performed via the approximated formula proposed in [28], using a constant reaction rate coefficient of $2 \cdot 10^9\ cm^3/s$ [29]. We refer to [23] for further details. Notice that in the case of VOCs that do not fragment after being ionised, the method gives an estimation of the VOC concentrations (in parts per billion by volume or ppbv), provided isomeric compounds are not interfering [30].

2.4. Statistical analysis

The data set related to the grana cheese experiment consists of two matrices with 72 rows corresponding to the 72 samples. The first matrix contains the GC data and has 32 columns corresponding to the concentrations of the 32 identified compounds, while the second matrix has 401 columns corresponding to the normalized intensities of the identified PTR-TOF-MS peaks.

Analogously, the second data set, which refers to the 56 olive oil measurements, consists of two matrices corresponding to the results of the GC analyses ($56 \text{ rows} \times 59 \text{ columns}$) and the one corresponding to PTR-TOF-MS (56×1053).

In the following, we will denote with X and Y the datasets related to PTR-TOF-MS and SPME/GC-MS analysis, respectively. The columns of X will be referred to as PTR-TOF-MS variables while those of Y as GC-MS variables. The rows of both X and Y will be referred to as samples. Data pre-processing and multivariate statistical analysis have been performed employing R packages [15,25,31]. Pre-processing included taking a log transformation of X and Y in order to get a more homogeneous distribution of the concentrations values, thus limiting spurious correlations caused by the presence of few samples of outlying intensities, that would hinder the subsequent multivariate analysis. Both matrices were then standardized by setting the mean to 0 and the standard deviation to 1 for all columns.

Please cite this article as: L. Cappellin, et al., Linking GC-MS and PTR-TOF-MS fingerprints of food samples, Chemometr. Intell. Lab. Syst. (2012), doi:10.1016/j.chemolab.2012.05.008

Preliminary insight on the correlation between X and Y variables was provided by employing standard Pearson correlation. More sophisticated analyses were carried on by multivariate methods such as PLS and LASSO. All Y variables were considered separately in the analyses.

A general linear multivariate model can be written as

$$Y = XB + E$$

where Y is the matrix of dependent variables (the properties to be predicted), X is the matrix of independent variables (the measurements to be used in the prediction), B is the matrix of regression coefficients to be estimated in the modelling procedure, and E the matrix of residuals. PLS decomposes both X and Y into latent variables that not only show high correlations between the two blocks (so that prediction is possible) but also cover large parts of the variances of X and Y (so that predictions are stable as well). Note that the shrinking of the coefficients is not explicitly enforced but results as a property of the algorithm. More details can be found in the literature [14].

The LASSO does use explicit penalization, and the model can be written as

$$Y = XB + \text{lambda}|B| + E$$

where the second term is the penalization term using the absolute values of the coefficients B. |B| signifies the L1 norm of the coefficients, i.e., the length of the coefficient vector based on the sum of the absolute coefficient sizes, rather than the sum of the squared coefficients. Here, the size of the penalty coefficient lambda needs to be optimized. Again, cross validation is used for this. Note that it is possible to calculate all models for all possible values of lambda simultaneously, so that the whole procedure takes almost the same time as ordinary linear regression.

The complexity optimization of PLS and LASSO models as well as prediction error estimations were performed via the repeated double cross validation (rdCV) procedure proposed by Filzmoser and coworkers [32]. We set the number of segments to 10 in both the inner (SEG_{CALIB}) and outer (SEG_{TEST}) loop. The rationale behind this choice is the relatively small number of samples we have for both grana cheese and olive oil. The number of repetitions (n_{REP}) was set to 100. We reported the root mean square error (RMSE) for the preprocessed GC-MS variables as prediction error for both PLS and LASSO. Such RMSE gives an estimate of the prediction error relatively to the standard deviation of the compound concentration, providing a reliable picture of the model prediction performance. For instance, a RMSE of 1 suggests that the model is unsuitable to predict the concentration of the selected compound from the corresponding PTR-TOF-MS fingerprints, the prediction error being about one standard deviation.

3. Results and discussion

3.1. Model predictions

Table 1 reports the prediction errors, namely root mean square error (RMSE), provided by rdCV for both PLS and LASSO applied to the Trentingrana cheese dataset. Although the estimated models often differ, in general, the two methods provide similar and consistent results. On other datasets PLS and LASSO have been shown to have different prediction capabilities [33]. In Table 1 compounds such as 2-octanone, 2-heptanol, ethyl octanoate and most acids display the lowest prediction errors. This means that, for these compounds, there exists a close relation between GC peaks and PTR-MS spectra that the models are able to catch and convert into prediction capability. For other compounds the RMSE shows intermediate values, indicating that the prediction is still possible but it is not as accurate as in the case of the above-mentioned compounds. In the case of acetic acid the RMSEs are 0.93 and 0.82 for LASSO and PLS, respectively, suggesting that the concentration of this compound, for the grana cheese matrix, is almost unpredictable from the PTR-TOF-MS peaks. This holds also true for the olive oil matrix (Table 2). It is worth pointing out that there are some other compounds whose prediction error is very close to 1, meaning that the models are unable to predict these GC variables from X. These compounds are 2-methylbutanal, ethyl isobutanoate, ethyl hexanoate and δ-octalactone. The dominant signal for 2-methybutanal is at m/z 69.067 Th corresponding to a generic fragment $C5H9^+$ that is common to many compounds (aldehydes, C5 alcohols, 1-octen-3-ol and several terpenes) [34]. Ethyl isobutanoate shares the same fragmentation profile with ethyl butanoate both present in cheese at similar concentration and shares a fragment with isobutanoic acid added as internal standard. The dominant signal of ethyl hexanoate, after protonation, is at mass 145.122 Th, the same obtained from the protonation of octanoic acid. The latter one is present in a concentration 20–80 times higher thus interfering with the possibility to correlate the ethyl hexanoate with PTR-MS profiling. Therefore, in general, poorly predicted compounds may be molecules that produce overlapping peaks.

Furthermore, 4-methyl-2-pentanone, ethyl heptanoate and isobutanoic acid are included among compounds with a prediction error very close to 1 as well. These compounds are standards that are introduced into the GC column for calibration purposes but not added for PTR-MS measurements and therefore no relation is expected with the PTR-TOF-MS variables. This observation may thus be seen as mere check of consistency.

Table 1
Case study 1: Trentingrana cheeses. Root mean square prediction error for the multivariate calibration models PLS and LASSO. In brackets the optimal model parameter (λ for LASSO and number of components for PLS) is reported.

Compound	RMSE	
	LASSO	PLS
Ethyl acetate	0.68 (0.2)	0.64 (2)
2-Methylbutanal	0.99 (0.42)	0.96 (1)
3-Methylbutanal	0.88 (0.35)	0.83 (1)
Ethyl isobutanoate	1.02 (0.46)	0.92 (1)
2-Pentanone	0.51 (0.14)	0.48 (2)
4-Methyl-2-pentanone[a]	1.02 (0.38)	1.08 (1)
Ethyl butanoate	0.53 (0.18)	0.55 (3)
2-Hexanon	0.47 (0.05)	0.34 (5)
2-Heptanon	0.41 (0.17)	0.42 (3)
3-Methylbutanol	0.46 (0.04)	0.47 (3)
Ethyl hexanoate	0.97 (0.7)	0.84 (1)
2-Octanon	0.33 (0.05)	0.42 (5)
Acetoine	0.46 (0.11)	0.44 (4)
2-Heptanol	0.31 (0.12)	0.42 (4)
2,6-Dimethyl pyrazine	0.6 (0.22)	0.52 (3)
Ethyl heptanoate[a]	1.01 (0.4)	1.08 (1)
1-Hexanol	0.8 (0.28)	0.74 (4)
2-Nonanone	0.5 (0.11)	0.52 (2)
Ethyl octanoate	0.35 (0.07)	0.33 (4)
Acetic acid	0.93 (0.46)	0.82 (2)
Isobutanoic acid[a]	1.01 (0.39)	1.08 (1)
Ethyl decanoate	0.53 (0.07)	0.42 (4)
Butanoic acid	0.52 (0.17)	0.41 (3)
Isovaleric acid	0.44 (0.16)	0.35 (3)
Valerianic acid	0.28 (0.04)	0.28 (4)
Hexanoic acid	0.43 (0.07)	0.34 (4)
δ-Octalactone	0.96 (0.37)	0.83 (1)
Heptanoic acid	0.33 (0.03)	0.33 (4)
Octanoic acid	0.36 (0.04)	0.34 (4)
δ-Decalactone	0.83 (0.25)	0.79 (1)
Nonanoic acid	0.7 (0.25)	0.62 (2)
Decanoic acid	0.46 (0.07)	0.44 (4)

[a] Standards for calibration of GC-MS, not added for PTR-TOF-MS measurements.

Please cite this article as: L. Cappellin, et al., Linking GC-MS and PTR-TOF-MS fingerprints of food samples, Chemometr. Intell. Lab. Syst. (2012), doi:10.1016/j.chemolab.2012.05.008

Table 2
Case study 2: extravirgin olive oils. Root mean square prediction error for the multivariate calibration models PLS and LASSO. In brackets the optimal model parameter (λ for LASSO and number of components for PLS) is reported.

Compound	RMSE	
	LASSO	PLS
Acetaldehyde	0.63 (0.24)	0.55 (3)
trans-1,3-pentadiene	0.6 (0.26)	0.54 (3)
Ethanol	0.81 (0.34)	0.72 (3)
2-Methylbutanal	1.06 (0.44)	1 (1)
3-Methylbutanal	0.75 (0.23)	0.64 (3)
Benzene	0.74 (0.31)	0.71 (2)
Hydrocarbon (C10H18)	0.53 (0.15)	0.47 (4)
Hydrocarbon (C10H18)	0.52 (0.15)	0.5 (4)
3-Pentanone	0.72 (0.21)	0.58 (3)
3-Ethyl-1,5-octadiene	0.54 (0.09)	0.57 (3)
α-Pinene	1 (0.5)	1.02 (1)
trans-2-hexene	0.45 (0.16)	0.65 (4)
Toluene	0.38 (0.2)	0.79 (4)
3-Ethyl-1,5-octadiene (cis or trans)	0.39 (0.21)	0.46 (3)
Hexanal	1 (0.61)	0.9 (1)
trans-2-pentenal	0.62 (0.19)	0.67 (3)
cis-3-hexenal	0.66 (0.16)	0.58 (3)
1-Penten-3-ol	0.95 (0.5)	0.93 (1)
Limonene	0.65 (0.26)	0.51 (3)
4-Methyl-3-pentanal	1.01 (0.52)	0.96 (1)
3-Methyl-1-butanol + 2-methyl-1-butanol	0.58 (0.15)	0.55 (3)
trans-2-hexenal	0.32 (0.1)	0.41 (3)
3-Ethyltoluene	0.46 (0.12)	0.5 (3)
β-cis-Ocimene	0.58 (0.21)	0.67 (3)
β-trans-Ocimene	0.61 (0.25)	0.77 (4)
Styrene	0.32 (0.12)	0.56 (3)
p-Cymene	0.94 (0.45)	0.89 (3)
Hexyl acetate	0.7 (0.3)	0.87 (3)
Perillene	0.38 (0.11)	0.57 (4)
trans-2-penten-1-ol	0.7 (0.22)	0.54 (3)
cis-3-hexenyl acetate	0.46 (0.14)	0.53 (3)
cis-2-penten-1-ol	0.63 (0.22)	0.52 (3)
Hexanol	0.81 (0.32)	0.68 (3)
trans-3-hexen-1-ol	0.7 (0.27)	0.67 (3)
trans-Alloocimene	0.27 (0.06)	0.69 (3)
cis-3-hexen-1-ol	0.12 (0.04)	0.36 (3)
trans-2-hexen-1-ol	0.21 (0.07)	0.55 (4)
Unknown (possibly *p*-mentha-1,3,8-triene)	0.74 (0.39)	0.86 (4)
Acetic acid	1.01 (0.46)	1.09 (1)
2-Ethyl-1-hexanol	0.6 (0.16)	0.52 (4)
α-Copaene	1.01 (0.44)	0.97 (2)
Benzaldehyde	0.58 (0.17)	0.55 (4)
1-Octanol	0.79 (0.23)	0.61 (4)
Dimethyl sulfoxide	0.68 (0.22)	0.89 (3)
5-Ethyl-2-(5 H)-furanone	0.65 (0.26)	0.59 (3)
Butyrolactone	0.54 (0.22)	0.47 (3)
Cis-β-farnesene	0.95 (0.54)	0.84 (1)
Acetophenone	0.56 (0.21)	0.48 (3)
Nonanol	0.96 (0.21)	0.77 (3)
Eremophilene	1.02 (0.46)	1.04 (1)
5-methyl-4-Hexen-3-one	0.15 (0.04)	0.31 (4)
trans,trans-α-farnesene	0.97 (0.35)	1.13 (1)
Methyl salicylate	0.61 (0.24)	0.72 (2)
Benzyl alcohol	0.68 (0.24)	0.71 (2)
β-Phenylethyl alcohol	0.83 (0.52)	0.79 (2)
Phenol	0.62 (0.19)	0.55 (3)
Heptanoic acid	0.77 (0.34)	0.71 (1)
Nonanoic acid	0.66 (0.31)	0.59 (2)
Benzophenone	0.57 (0.22)	0.49 (3)

Table 2 reports the RMSE for the prediction of the VOCs identified by SPME/GC-MS in the headspace of the olive oil samples. Analogously to the case of Trentingrana cheese, most compounds show a reasonably good prediction error.

LASSO and PLS provide similar results, with a few exceptions. The prediction of some compounds was not possible with none of the two methods. This is the case of 2-methylbutanal, α-pinene, limonene, 3-methyl-1-butanol, 2-methyl-1-butanol, p-cymene, acetic acid, α-copaene, cis-β-farnesene and eremophilene.

In olive oil, the concentration of β-trans-ocimene is more than 20 times higher than that of other isomeric terpenes such as α-pinene, limonene and β-cis-ocimene and this explains why β-trans-ocimene is predicted reasonably well by the models while, at the contrary, α-pinene and limonene are not predicted. The prediction error found for β-cis-ocimene is probably driven by the correlation that this compound displays with β-trans-ocimene.

It is worth noting that the chromatographic signal for styrene (retention time 687 s) is hidden in the tail of the very intense peak corresponding to trans-2-hexenal at retention time 616 s whose tail extends till retention time 696 s and the tail of β-trans-ocimene (Fig. 1). Nevertheless, our peak extraction approach is able to disentangle the three compounds making use of the different filtered signals: trans-2-hexenal or β-trans-ocimene have no significant fragment at nominal mass 104. Moreover, this hidden peak is well predicted from the PTR-TOF-MS fingerprint (RMSE 0.32 for LASSO).

3.2. Model interpretation (examples)

LASSO and PLS are not only prediction tools but may also provide insight on the relation between GC-MS and PTR-MS measurement, for instance via interpretation of the model coefficients. As an example we will discuss the cases of trans-2-hexen-1-ol and 3-ethyl-1,5-octadiene for the olive oil dataset and those of 3-methylbutanol and butanoic acid for the grana cheese dataset. Our choice is based on their different and peculiar behaviours, providing paradigmatic examples of possible outcomes of the models.

3.2.1. Trans-2-hexen-1-ol ($C_6H_{12}O$)

Trans-2-hexen-1-ol is a product of trans-2-hexenal reduction. Together the other aliphatic C6 component, it contributes significantly to the green odours of olive oils [35].

The protonated mass of this compound is 101.096 Th. In fact correlation analysis shows that trans-2-hexen-1-ol (as measured by SPME/GC-MS) displays the largest correlation with the normalized intensities of the PTR-TOF-MS peaks at m/z 101.095 (Pearson correlation 0.97), 102.098 (0.98), 83.086 (0.95), 84.089 (0.95). The mass accuracy reached in PTR-TOF-MS by proper mass calibration [7] allows thus to identify the 101.095 Th peak as corresponding to $C_6H_{13}O^+$ and the 102.098 Th peak to its first order isotope.

Fig. 2A plots the predicted values for trans-2-hexen-1-ol by LASSO in a Leave-One-Out (LOO) experiment; that is the sample to be predicted is left out and the remaining samples are used to build the LASSO model (the optimal parameter is provided by rdCV) to be employed in the prediction. These unbiased estimations lie very close to the theoretical line of perfect prediction. It is very interesting to note that even oil samples having a very different concentration of trans-2-hexen-1-ol from all other samples are well predicted by the LASSO models. The model coefficients of the optimal LASSO model are depicted in Fig. 2B. It is clear that only the PTR variables at m/z 101.095 and 102.098 play a significant role in the model.

The prediction of trans-2-hexen-1-ol by the PLS method provides worse results (RMSE 0.55, see table 2), as it is shown in Fig. 2C. For this compound the optimum number of PLS components is 4. The coefficients of all PTR-TOF-MS variable for the corresponding optimum model are plotted in Fig. 2D. There appears no clear predominance of two variables as for the LASSO method and the differences between coefficients are narrower. The coefficients of the PTR variables 101.095 Th, 83.086 Th, 55.0542 Th (corresponding to the fragmentation of trans-2-hexen-1-ol [36]) and their isotopes are among the most important variables but many others have coefficients which are non negligible and are not suppressed as in the case of LASSO. The interpretability of the best PLS model is not as straightforward as in the case of the LASSO model.

Please cite this article as: L. Cappellin, et al., Linking GC-MS and PTR-TOF-MS fingerprints of food samples, Chemometr. Intell. Lab. Syst. (2012), doi:10.1016/j.chemolab.2012.05.008

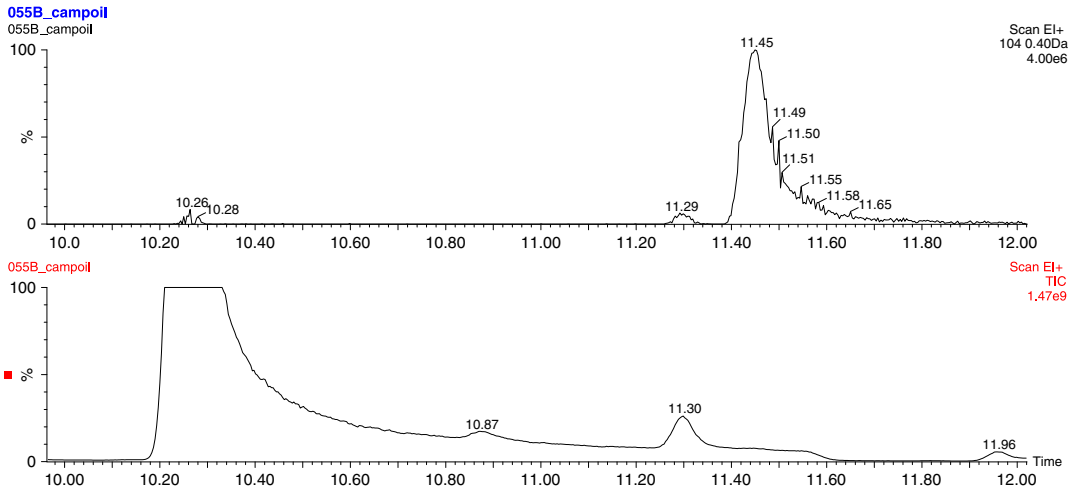

Fig. 1. GC-MS. Chromatogram of an extravirgin olive oil. In the upper panel the single ion scan at m/z 104 is reported showing a peak at 11.45 min (687 s), corresponding to styrene. In the lower panel, where total ion count spectra is reported, the styrene peak is hidden in the tails of compounds presenting very intense signals.

3.2.2. 3-Ethyl-1,5-octadiene ($C_{10}H_{18}$)

A different case is that of 3-ethyl-1,5-octadiene. This is a common compound found in the headspace of extravirgin olive oils. It is a hydrocarbon with molecular form $C_{10}H_{18}$ and it is supposed to be one of the pentene dimmers formed during olive oil production [37].

The mass of the protonated form is 139.148 Th. Fig. 3 shows a mass window of the PTR-TOF-MS spectrum for an olive oil sample at nominal mass 139. Four peaks were identified and one of them was associated with the sum formula $C_{10}H_{19}^+$. The Pearson correlation between the corresponding PTR variable and 3-ethyl-1,5-octadiene is rather low (0.60), and there are many other PTR peaks having larger correlation with 3-ethyl-1,5-octadiene; nevertheless no correlation exceeds 0.8. A look at Fig. 4 shows that in the optimum model the largest positive coefficient is remarkably associated with m/z 139.147, i.e. with $C_{10}H_{19}^+$. In fact in PTR-MS spectra the protonated form of the compound is expected. In the best model, other variables have non-negligible coefficients, for instance m/z 135.012 has the largest negative coefficient.

PLS in this case has similar prediction capabilities to LASSO. The coefficient of the PTR variable corresponding to m/z 139.148 Th is again the largest, but, as in the case of trans-2-hexen-1-ol, the spread with the coefficients belonging to the other PTR variable is not marked as for the LASSO model (Fig. 5). As exemplified by both PLS and LASSO the choice of a multivariate approach leads to significant improvements in the prediction error, while a monovariate approach would not be suitable in this case. For instance the RMSE of the monovariate model employing only the most correlated PTR variable to 3-ethyl-1,5-octadiene would be 0.98.

3.2.3. 3-Methylbutanol ($C_5H_{12}O$)

Isoamyl alcohol, or 3-methylbutanol, is a common compound often found in ripened cheeses. It originates from branched aliphatic amino acid leucine that are degraded during cheese ripening to 3-methylbutanal that is further reduced to 3-methylbutanol [38].

The mass of the protonated form of this compound is 89.0961 Th. Upon correlation analysis, it is found that 3-methylbutanol displays the highest correlations with the PTR-TOF-MS peak at estimated m/z 71.087 Th (Pearson correlation 0.77) and its first order isotope at m/z 72.090 Th. Such peaks are identified as a general fragment corresponding to the sum formula $C_5H_{11}^+$. In fact alcohols typically undergo water loss after protonation in PTR-MS producing a generic fragment $M^+-(H_2O)$. Our analysis (not shown) shows that in optimal LASSO model the largest and most significant coefficient is indeed associated with the peak at m/z 71.087 Th, while a negative contribution comes from the PTR variable at 69.071 Th, identified as $C_5H_9^+$. This fact could be explained by the observation that its second order isotope has a m/z very close to 71.087 Th, thus interfering with the signal of $C_5H_{11}^+$.

The prediction of 3-methylbutanol by the PLS method also provides good results. The coefficients (not shown) of all PTR-TOF-MS variables at this minimum do not indicate a clear predominance of a single variable as for the LASSO method. Moreover, the differences between coefficients are narrower. The contribution of the 71.087 Th variable

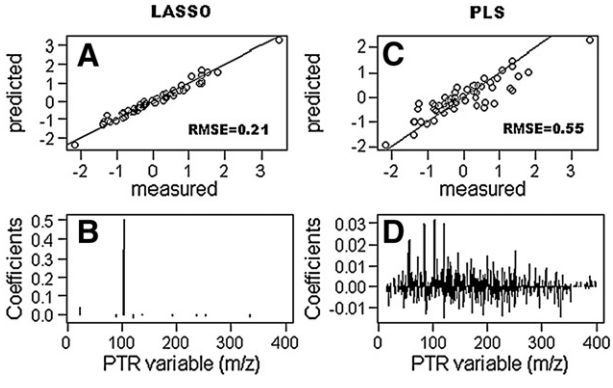

Fig. 2. Trans-2-hexen-1-ol (olive oil headspace). A,C: Prediction of trans-2-Hexen-1-ol (as measured by GC/SPME-MS) with LASSO (A) and PLS (C) from the PTR-TOF-MS fingerprint using a LOO procedure (see text). B: LASSO, coefficients of the PTR variables for the optimal LASSO model. Note that the largest coefficients are associated with the PTR variables at m/z 101.095 Th and 102.098 Th. D: PLS, coefficients of the PTR variables for the best PLS model.

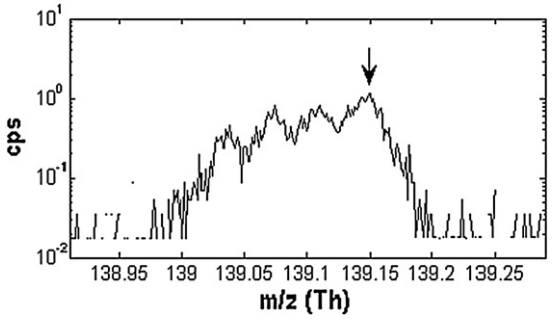

Fig. 3. Extract of a PTR-TOF-MS spectrum for an olive oil sample. The arrow indicates the position of the peak at m/z 139.149, corresponding to $C_{10}H_{19}^+$. The ordinate axis units are counts per second (cps).

Please cite this article as: L. Cappellin, et al., Linking GC-MS and PTR-TOF-MS fingerprints of food samples, Chemometr. Intell. Lab. Syst. (2012), doi:10.1016/j.chemolab.2012.05.008

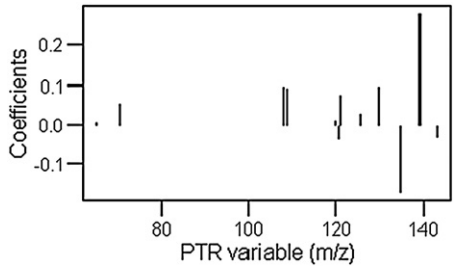

Fig. 4. 3-Ethyl-1,5-octadiene (olive oil headspace). Model coefficients for LASSO. The largest coefficients (in absolute value) are associated with m/z 139.147 and m/z 135.012.

is still among the largest but the interpretability of the coefficients is compromised.

3.2.4. Butanoic acid ($C_4H_8O_2$)

During cheese ripening, lipolysis, due to the action of the indigenous lipases of milk or to the action of microbial lipases, generates many free fatty acids. At PTR-MS normal conditions (120–140 Td) the main signal recorded for volatile fatty acids is the protonated molecular mass MH^+ followed to the fragment generated from water loss $M^+(-H_2O)$. Butanoic acid plays an important role in the flavour of Grana Padano, if present in a balanced amount, conferring a rancid cheese-like odour [39]. The molecular mass of butanoic acid is 88.0524 thus the protonated form of this compound is 89.0603 Th. As expected, butanoic acid correlates with the PTR-TOF-MS peak at 89.060 Th (Pearson correlation 0.77). Other masses, excluding for brevity isotopologues, well correlated with butanoic acid are recorded at 99.082 Th (0.75), 103.075 Th (0.72), 117.091 Th (0.80), 131.106 (0.70), 135.102 Th (0.78), 145.121 Th (0.77), 159.135 Th (0.71), 163.132 Th (0.73) and 173.150 Th (0.81). The mass series $89.06025 + z \cdot 14.01565$ ($z = 1,2,\ldots$) is associated to volatile fatty acids of increasing chain length, the found correlations are interpretable with their common origin from lipase activity. The presence of these correlations is reflected in the results of the LASSO models. In fact, in the optimum model, a largely predominant positive coefficient is associated with m/z 173.150, the reason relying in the above considerations.

In this case, PLS provides a slightly better prediction performance than LASSO, the RMSE being 0.41, but the interpretation of the model coefficients is less clear, because of very large number of variables with similar coefficients (not shown), many of those not having a clear connection to butanoic acid.

4. Conclusions

GC-MS is a hyphenated technique widely used in food science analysis; its success is mainly due to the powerful compound identification and to the reliable quantification. However, it is time consuming and may be not viable or too expensive if a high number of analyses are required as in the case of food control (quality control, geographical origin determination, contaminant screening, label protection). In these contexts, fingerprinting techniques are becoming more and more widespread and in particular direct injection mass spectrometric fingerprinting. In general, being fingerprinting techniques non-selective, analytical information is limited. In the case of PTR-MS fingerprinting has been demonstrated in several works the possibility of retrieving the analytical information entangled in the produced spectra.

We investigated the relationship between the data sets obtained by these two different headspace techniques: on one side the well established but time consuming GC-MS and, on the other, the novel and rapid PTR-TOF-MS. Two multivariate correlation methods (PLS and LASSO) were used to predict the concentration of the volatile organic compounds as measured by SPME/GC-MS on the basis of the rapid PTR-TOF-MS fingerprinting. We tested our methodology on two data sets related to complex but interesting and economically relevant food matrixes: olive oil and ripened, parmesan like, cheese.

In the case of some compounds, a one-to-one relation between GC-MS and a PTR-TOF-MS peaks can be established. Thus, in these cases, the use of multivariate calibration methods would not be necessary, but, in general, the prediction is more complicated, for example, because of the interference of several compounds on the same PTR-TOF-MS peak, and the use of multivariate methods allows a better understanding of the data. They, in fact, consider the analytical information contained in the whole PTR-TOF-MS spectrum (e.g. correlations, fragmentation or interferences of compounds having close m/z values), thus improving the information retrievable from the PTR-TOF-MS fingerprint and bettering the prediction.

We suggest that the calibration methods must be developed for each food matrix. In fact, it is not realistic, given the complexity of food samples, to expect that a single calibration model can work in general. In our investigations, PLS and LASSO provided in general comparable results in terms of prediction capabilities, nevertheless LASSO produced more interpretable models.

Even if it was not possible to develop predictive models for all the GC identified peaks, our results indicate that many relevant compounds can be predicted with sufficient accuracy and the availability of GC data improves the possibility to disentangle the information in PTR-TOF-MS data. On the other side, even this point was not addressed in the present work, it must be pointed out that PTR-TOF-MS detects compounds that are not easily detectable with a single GC-MS analysis [12]. In fact, together with information related to the compounds detected by GC, PTR-TOF-MS data provides quantitative information, for instance, on low molecular mass important metabolites as methanol, acetaldehyde, ethanol, sulphur compounds and many others.

In conclusion, the newly available PTR-TOF-MS technique may complement the results of GC-MS and remarkably reduce of about 100 times the measuring time: our proposal is thus to measure a reduced number of samples with GC-MS for a reliable compound identification and then to extend the analysis to a larger number of samples by PTR-TOF-MS. A metabolomic approach can follow on a large samples-set characterized by PTR-TOF-MS fingerprinting and the possibly relevant markers will then be chemically identified on the basis of GC analysis.

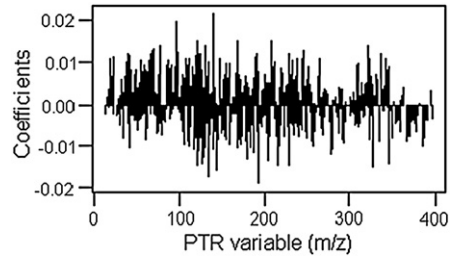

Fig. 5. 3-Ethyl-1,5-octadiene (olive oil headspace). Model coefficients for PLS.

Acknowledgements

This work was supported by the Autonomous Province of Trento, Italy, as part of the project called "Qualità della filiera Grana Trentino". The authors acknowledge Tormod Næs for fruitful suggestions and thank Alessandra Fabris and Emanuela Betta for providing technical support on GC-MS analyses.

References

[1] J. de Gouw, C. Warneke, Measurements of volatile organic compounds in the earth's atmosphere using proton-transfer-reaction mass spectrometry, Mass Spectrometry Reviews 26 (2007) 223–257.

[2] C. Hewitt, S. Hayward, A. Tani, The application of proton transfer reaction-mass spectrometry (PTR-MS) to the monitoring and analysis of volatile organic compounds in the atmosphere, Journal of Environmental Monitoring 5 (2003) 1–7.
[3] A. Critchley, T. Elliott, G. Harrison, C. Mayhew, J. Thompson, T. Worthington, The proton transfer reaction mass spectrometer and its use in medical science: applications to drug assays and the monitoring of bacteria, IInternational Journal of Mass Spectrometry 239 (2004) 235–241.
[4] F. Biasioli, F. Gasperi, C. Yeretzian, T.D. Märk, PTR-MS monitoring of VOCs and BVOCs in food science and technology, TrAC, Trends in Analytical Chemistry 30 (2011) 968–977.
[5] P. Granitto, F. Biasioli, E. Aprea, D. Mott, C. Furlanello, T. Mark, et al., Rapid and non-destructive identification of strawberry cultivars by direct PTR-MS headspace analysis and data mining techniques, Sensors and Actuators B: Chemical 121 (2007) 379–385.
[6] E. Zini, F. Biasioli, F. Gasperi, D. Mott, E. Aprea, T.D. Märk, et al., QTL mapping of volatile compounds in ripe apples detected by proton transfer reaction-mass spectrometry, Euphytica 145 (2005) 269–279.
[7] L. Cappellin, F. Biasioli, A. Fabris, E. Schuhfried, C. Soukoulis, T.D. Märk, et al., Improved mass accuracy in PTR-TOF-MS: another step towards better compound identification in PTR-MS, International Journal of Mass Spectrometry 290 (2010) 60–63.
[8] A. Jordan, S. Haidacher, G. Hanel, E. Hartungen, L. Mark, H. Seehauser, et al., A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS), International Journal of Mass Spectrometry 286 (2009) 122–128.
[9] P. Pollien, C. Lindinger, C. Yeretzian, I. Blank, Proton transfer reaction mass spectrometry, a tool for on-line monitoring of acrylamide formation in the headspace of Maillard reaction systems and processed food, Analytical Chemistry 75 (2003) 5488–5494.
[10] E. Aprea, F. Biasioli, S. Carlin, T. Mark, F. Gasperi, Monitoring benzene formation from benzoate in model systems by proton transfer reaction-mass spectrometry, International Journal of Mass Spectrometry 275 (2008) 117–121.
[11] M.A. Mortenson, G.A. Reineccius, Encapsulation and release of menthol. Part 2: direct monitoring of l-menthol release from spray-dried powders made with OSAn-substituted dextrins and gum acacia, Flavour and Fragance Journal 23 (2008) 407–415.
[12] E. Aprea, F. Biasioli, S. Carlin, I. Endrizzi, F. Gasperi, Investigation of volatile compounds in two raspberry cultivars by two headspace techniques: solid-phase microextraction/gas chromatography-mass spectrometry (SPME/GC−MS) and proton-transfer reaction-mass spectrometry (PTR−MS), Journal of Agricultural and Food Chemistry 57 (2009) 4011–4018.
[13] Julia Märk, P. Pollien, C. Lindinger, I. Blank, T. Märk, Quantitation of furan and methylfuran formed in different precursor systems by proton transfer reaction mass spectrometry, Journal of Agricultural and Food Chemistry 54 (2006) 2786–2793.
[14] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[15] B.-H. Mevik, R. Wehrens, The pls package: principal component and partial least squares regression in R, Journal of Statistical Software 18 (2007).
[16] A.C. Pereira, M.S. Reis, P.M. Saraiva, J.C. Marques, Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC-MS, HPLC-DAD, Chemometrics and Intelligent Laboratory Systems 105 (2011) 43–55.
[17] V. Gaydou, J. Kister, N. Dupuy, Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil, Chemometrics and Intelligent Laboratory Systems 106 (2011) 190–197.
[18] T. Hastie, The Elements Of Statistical Learning: Data Mining, Inference, And Prediction: With 200 Full-Color Illustrations, Springer, New York, 2001.
[19] N. Butler, M. Denham, The peculiar shrinkage properties of partial least squares regression, Journal of the Royal Statistical Society: Series B: Methodological 62 (2000) 585–593.
[20] T. Næs, Harald Martens, principal component regression in NIR analysis: viewpoints, background details and selection of components, Journal of Chemometrics 2 (1988) 155–167.
[21] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 42 (2000) 80.
[22] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society: Series B: Methodological 58 (1996) 267–288.
[23] A. Fabris, F. Biasioli, P. Granitto, E. Aprea, L. Cappellin, E. Schuhfried, et al., PTR-TOF-MS and data-mining methods for rapid characterisation of agro-industrial samples: influence of milk storage conditions on the volatile compounds profile of Trentingrana cheese, Journal of Mass Spectrometry 45 (2010).
[24] I. Endrizzi, A. Fabris, F. Biasioli, E. Aprea, E. Franciosi, E. Poznanski, et al., The effect of milk collection and storage conditions on the final quality of Trentingrana cheese: sensory and instrumental evaluation, International Dairy Journal 23 (2011) 105–114.
[25] R Development Core Team, A Language and Environment for Statistical Computing, Vienna, Austria, 2009.
[26] L. Cappellin, F. Biasioli, E. Schuhfried, C. Soukoulis, T.D. Märk, F. Gasperi, Extending the dynamic range of proton transfer reaction time-of-flight mass spectrometers by a novel dead time correction, Rapid Communications in Mass Spectrometry 25 (2011) 179–183.
[27] L. Cappellin, F. Biasioli, P. Granitto, E. Schuhfried, C. Soukoulis, T.D. Märk, et al., On data analysis in PTR-TOF-MS: from raw spectra to data mining, Sensors and Actuators B: Chemical 155 (2011) 183–190.
[28] W. Lindinger, A. Hansel, A. Jordan, On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) — Medical applications, food control and environmental research, International Journal of Mass Spectrometry 173 (1998) 191–241.
[29] L. Cappellin, M. Probst, J. Limtrakul, F. Biasioli, E. Schuhfried, C. Soukoulis, et al., Proton transfer reaction rate coefficients between H3O+ and some sulphur compounds, International Journal of Mass Spectrometry 295 (2010) 43–48.
[30] L. Cappellin, T. Karl, M. Probst, O. Ismailova, P.M. Winkler, C. Soukoulis, et al., On quantitative determination of volatile organic compound concentrations using proton transfer reaction time-of-flight mass spectrometry, Environmental Science and Technology 46 (2012) 2283–2290.
[31] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (2010) 1–22.
[32] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double cross validation, Journal of Chemometrics 23 (2009) 160–171.
[33] D. Lee, W. Lee, Y. Lee, Y. Pawitan, Sparse partial least-squares regression and its applications to high-throughput data analysis, Chemometrics and Intelligent Laboratory Systems 109 (2011) 1–8.
[34] K. Buhr, S. van Ruth, C. Delahunty, Analysis of volatile flavour compounds by Proton Transfer Reaction-Mass Spectrometry: fragmentation patterns and discrimination between isobaric and isomeric compounds, International Journal of Mass Spectrometry 221 (2002) 1–7.
[35] J.M. Olias, A.G. Perez, J.J. Rios, L.C. Sanz, Aroma of virgin olive oil: biogenesis of the «green» odor notes, Journal of Agricultural and Food Chemistry 41 (1993) 2368–2373.
[36] E. Aprea, Addressing Issues of Sensory Analysis by PTR-MS: Applications in Food and Environmental Science, Doctoral Thesis, Univ, Innsbruck, 2005.
[37] F. Angerosa, L. Camera, N. d' Alessandro, G. Mellerio, Characterization of seven new hydrocarbon compounds present in the aroma of virgin olive oils, Journal of Agricultural and Food Chemistry 46 (1998) 648–653.
[38] J. Adda, The chemistry of flavour and texture generation in cheese, Food Chemistry 9 (1982) 115–129.
[39] L. Moio, F. Addeo, Grana Padano cheese aroma, The Journal of Dairy Research 65 (1998) 317–333.

Please cite this article as: L. Cappellin, et al., Linking GC-MS and PTR-TOF-MS fingerprints of food samples, Chemometr. Intell. Lab. Syst. (2012), doi:10.1016/j.chemolab.2012.05.008