International Journal of Mass Spectrometry 295 (2010) 13–20

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

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

Multidimensional statistical analysis of PTR-MS breath samples: A test study on irradiation detection

Mattia Fedrigo *, Christoph Hoeschen, Uwe Oeh

Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany[1]

ARTICLE INFO

Article history:
Received 3 November 2009
Received in revised form 7 June 2010
Accepted 14 June 2010
Available online 19 June 2010

Keywords:
Multidimensional analysis
PTR-MS
Breath gas VOC
Detection of exposure

ABSTRACT

A multidimensional statistical analysis of data obtained from breath gas measurements with Proton Transfer Reaction-Mass Spectrometry (PTR-MS) is proposed, based on a chemical-diffusion equilibrium approach. The proposed methodology is developed and demonstrated on the problem of detecting exposure of human beings to ionizing radiation. It could be applied to a general family of non-invasive, high-throughput, breath gas based detection strategies, like for instance a breath gas test for early diagnosis of lung cancer.

1. Introduction

Breath gas analysis is a non-invasive investigation method which can hopefully be used in medical applications [1]. Human breath gas contains volatile organic compounds (VOCs) which are mainly blood born and therefore allow monitoring physiological processes in the body. This technique has already been explored for a number of diseases [2,3].

Proton Transfer Reaction-Mass Spectrometry (PTR-MS) [4–6] is a technology which has become a well-established method to measure a number of VOCs in human breath gas [3,7–15]. It allows highly sensitive, rapid and on-line measurements and is therefore well suited for high-throughput applications. Furthermore, it could be fitted to an ambulance or transported and powered by small vehicles, and it is able to cover a large range of masses, which is not possible with an electronic nose for instance.

When applied to breath gas analysis, PTR-MS can produce easily a large amount of quite variable (noisy) high dimensional data. Consider for instance a mass spectrum from 20 to 200 atomic mass units (AMUs), where for each AMU the signal is the number of protonated molecules with the desired mass, detected in a given amount of time.

One problem with such data is that classic low-dimensional statistical techniques require a large amount of samples to reduce the risk of overfitting. In other words, when devising a test from training data there is a chance that a particular subset of dimensions exhibits the very statistical separation we are looking for not because of a physical link between these measurements and the target phenomenon, but simply because of randomness.

In view of this a multidimensional statistical approach is proposed, implemented as a linear combination of an opportunely preconditioned data subset. The preconditioning is a denoising step, while the linearity keeps the analysis simple and fast. The multidimensionality and the subset selection are two aspects of a trade-off approach between sensitivity, robustness and noise.

The aim is to make use of the extra information provided by the high dimensionality while overcoming its problems by means of a dimension selection scheme parameterized by the signals' noisiness or numerical stability. As a consequence a relatively limited number of samples with respect to the number of dimensions is required.

The proposed breath analysis paradigm (PTR-MS plus multidimensional statistical analysis) could be deployed for a wide range of applications. As a test case, its feasibility for detection of exposure of human beings to ionising radiation is explored in this paper.

After the present introduction, this paper articulates into a presentation of the radiation detection problem, including the sampling methodology and the difficulties in handling the acquired data, followed by the analysis approach and the results.

* Corresponding author. Tel.: +49 89 3187 2405; fax: +49 89 3187 19 2405.
E-mail address: mattia.fedrigo@helmholtz-muenchen.de (M. Fedrigo).

[1] http://www.helmholtz-muenchen.de. Chairperson of Supervisory Board: MinDir'in Bärbel Brumme-Bothe; Board of Directors: Prof. Dr. Günther Wess, Dr. Nikolaus Blum; Register of Societies: Amtsgericht München HRB 6466.

1387-3806/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2010.06.011

Table 1
Breakdown of collected samples.

Sample type	Number of samples	Lung therapy	Full body therapy
Controls' samples	84	0	0
Patients samples' before radiotherapy	15	2	13
Total from not irradiated people	99		
Patients' samples during radiotherapy *directly before* a session	9	3	6
Patients' samples during radiotherapy *directly after* a session	22	8	14
Patients' samples during radiotherapy, no session info	19	15	4
Total from irradiated people	50		

2. The test case: detection of radiation exposure by PTR-MS breath gas samples

Nuclear power plant leaks, terrorist attacks with so-called dirty bombs: one can easily imagine dire scenarios where one would quickly need to ascertain the exposure of a large population to ionising radiation, whose effect can be quite complex [16–18]. Up to now, the extent of the ionizing radiation exposure to individuals can be estimated by biological dosimetry methods based on, e.g., chromosome aberrations, lymphocyte depletions or cytogenetic analysis. Unfortunately, these procedures are invasive, time consuming and may not be feasible for large-scale scenarios. For such incidents, one would need to have a high-throughput technique. Additionally it should be easy to handle, cheap and minimally invasive. PTR-MS analysis of breath gas samples fits these requirements.

Living tissues absorb incident radiation differently depending on the type of radiation and the energy. The radiation may ionize or excite critical targets of the cell, i.e., the DNA molecule (direct effect of radiation) or produce ion radicals which may damage the critical targets (e.g., also membrane lipids and proteins) in subsequent chemical reactions (indirect action of the radiation).

We assume the theory that both direct and indirect actions of radiation result in chemical changes from the breakage of bonds, the chemical products of such processes being released into the blood, reaching the lung and the volatile compounds of them being exhaled via breath [19].

According to this hypothesis, specific VOCs might be detectable in the breath as a consequence of irradiation and might therefore be used as biomarkers to detect human exposure to ionising radiation. There is already a hint due to a pilot study undertaken by Skeldon et al. [20], who found that the ethane level in human breath is enhanced directly after radiation therapy.

The following sections describe the PTR-MS measurements of breath from patients of full or partial body radiotherapy, in the framework of a comparison of the VOC spectra before and after irradiation. The proposed multidimensional statistical analysis uses this information to look for specific VOC biomarkers to be used for radiation exposure assessment.

2.1. Patients and breath sampling

The radiotherapy of tumours is based on the observation that cells of certain tumour types are more radiosensitive than the cells of normal tissue. The patients who donated breath samples were treated with conventional γ-ray radiation therapy. This involved either the thorax region or the entire body, and was performed mostly using multifraction regimes (in a multifraction regime the radiation is not applied in a single dose, but in a number of sessions).

The expired breath gas was collected in commercial Teflon-FEP (fluorinated ethylene propylene) bags of 3 l volume. At the same time, a sample of the room air was collected.

Between 1 and 8 samples were collected at different times from the same patient, before, during and after a therapy cycle. Patients were tested before the beginning of the first session of the irradiation cycle (when no irradiation is administered yet), then at the time of a successive session (immediately before and/or after the session of the day), and eventually after the last session of the cycle.

Samples from controls, who are non-irradiated healthy people, were also collected, preferably in the same room air as the patients (see Table 1).

Patients subjected to full body irradiation received between 4 and 12 Gray of cumulative exposition during a complete cycle, while cumulative organ doses for lung irradiation ranged between 12 and 75 Gray. Most of the patients sampled directly before a session were subjected to a previous session within the last fortnight: 3 of them the day before, one of them 2 days before, one of them 8 days before, one of them 13 days before, while for 3 of them no further information was available.

2.2. PTR-MS analysis

The PTR-MS system employed for this study is described in detail in [21].

The collected bags are stored in an oven at 40 °C for at least 30 min to evaporate the condensed humidity. Then, while still in the oven, they are connected with the inlet of the PTR-MS by a heated Teflon tube. Measurements start after a running-in period of one minute. VOC mass spectra are screened over the range from 20 to 200 atomic mass unit (AMU).

Normally 7 cycles are measured and during each individual cycle each mass is measured with a dwell time of 0.5 s. Average values from the 7 cycles are calculated. This is performed to reduce statistical counting noise. Although one could raise the overall fixed observation time to do so, the noise effect is of course much stronger for small count signals, where counting noise is stronger with respect to signal noise (mean count). In other words, adaptive scanning times are possible. The measurements represent mass counts and are always non-negative, but they do not need to be integers because of averaging.

The choice to focus between masses 20 and 200 is based on the relatively low frequency of very large VOCs (beyond atomic mass 200), on the assumption of irrelevance for our task of protonated molecules with an atomic mass below 19, and on a technical problem with mass 19.

Protonated water ($H30^+$) corresponds to mass signal 19 but this signal is normally so high that it induces detector saturation. For this reason the measurement of mass 19 is not used in current practice. An indirect way to measure water concentration is to consider mass signal 21, which is also a water signal due to natural isotopes. The isotopic ratio is roughly 1 over 500, avoiding saturation.

Moreover, we can also estimate the sample's humidity from the collected data by mean of the various water clusters which are well within our observation range. A water cluster is a cluster of molecules bound together by van der Waals forces and consisting of a protonated water molecule plus any number of further water molecules. The stability of a cluster diminishes with the number of extra water molecules. Water clusters affect mass signals $19 + 18 \times n = 37, 55, 73$, etc.

2.3. Collected data

A sample's data consists of a vector with 181 entries corresponding to protonated molecules' masses from 20 to 200 AMU. The entries are non-negative, can be equal to zero, and provide us with a measure of the relative concentration in the air sample.

This paper's radiation detection test is built upon a collection of 149 breath samples vectors, each paired with a room air sample taken in the same time and place, and processed analogously (the necessity of the room air samples will be discussed later, in the analysis section). Fifty samples were donated by people recently exposed to ionising radiation (immediately after a session or a few days from one, within a therapy cycle), while the remaining 99 correspond to people who were not subjected to it.

3. Breath gas PTR-MS data problems

In this section some of the problems of the collected PTR-MS breath sample data are discussed. It is by no means an exhaustive catalogue, but it helps defining the problem of statistically studying such data and it motivates some of the operations which will be proposed in the analysis section.

Breath PTR-MS data have two structural problems. The first and foremost is the PTR-MS inability to distinguish between molecules with the same nominal mass. There are other kinds of mass spectrometers with comparable resolution, but their cost and size rule them out of the fast-throughput application we are considering.

The second problem is that only molecules with a proton affinity higher than that of water can be detected.

These two problems mean that only the easily protonable subset of breath gas VOCs can be detected, and that groups of same-weight molecules are indistinguishable from one another and consequently lumped together in the spectrum.

There is also an evolution in the behaviour of the PTR-MS device itself, which has to do with the decay of the proton source with time as well as with changing operative conditions. In standard PTR-MS practice this phenomenon is addressed by primary signal normalization.

Moreover, a breath sample is strongly affected by changes in the environmental air, and by conditions of the donor which are not correlated to his exposure to ionising radiation. The weather, the location, the season, the hour of the day or the presence of other people are all examples of external phenomena affecting the air inhaled by a person and consequently any breath sample donated by the same. To smoke, to eat, to drink or to lay still for some time before donating a breath sample also affect it [7,9,11,22].

The 149 samples considered for developing the radiation detection test were collected throughout a time span of roughly one and a half year, in different locations including three hospitals and our research centre. The breath donors did not conform to any standard sampling procedure: some ate or drank shortly beforehand, some performed some physical activity, and the sampling bags were filled with various numbers of respiratory actions—most people would need only one exhalation act, while frail patients might need more. Some people were sampled more than once: the collectors themselves often took a sample of their own breath while in the same room as the patients, and some patients were followed through the radiotherapy arc and breath-sampled accordingly.

For these reasons PTR-MS breath data present a large variability. From the point of view of the development of a statistic correlated with one particular phenomenon among many, this variability is seen as noisiness. The noisiness is such that no single data entry (i.e., molecular mass) appears to be a reliable test for such an indirect phenomenon as the exposure of the donor to ionising radiation, which does not change the VOCs in the breath directly but affects the biological processes producing them.

Strategies to identify and possibly control the sources of this variability [11], while certainly appealing and extensively investigated in the literature, will not be pursued here. The reason is twofold: the focus of the present paper is more on the statistical analysis than the data collection, and the possible high-throughput applications envisioned before would make such a strategy very challenging.

Finally, PTR-MS breath gas data are high dimensional: a single measurement is a set of 181 mass signals. The high dimensionality raises the computational cost of standard strategies like principal component analysis (PCA) or linear discriminant analysis (LDA). Furthermore, the relatively small number of collected samples with respect to dimensionality (149 samples for 181 dimensions) carries the risk of overfitting when devising a test based on this data. This particular aspect leads to forego direct PCA or LDA analysis and to look for new strategies.

4. Multidimensional statistical analysis

In this section the key ideas of the statistical analysis of the breath gas samples from PTR-MS are presented. The analysis is explained on the basis of the radiation detection data, but can be generally applied to any PTR-MS breath gas based diagnosis and it tackles the problem of a relatively limited sample number compared to the sample dimension.

As sketched before, PTR-MS breath sample data present different sorts of unwelcome restrictions and noise. These problems are tackled in three steps. First, the data are preconditioned in order to cancel or at least reduce some of these effects. Secondly, a family of multidimensional tests is produced, indexed by a parameter set and associated to a score. Finally, some parameter values are sampled looking for a good test, that is, a parameter set associated to a high score.

4.1. Preconditioning

The mathematical aspects of simple chemical equilibrium and diffusion processes [23] are a fruitful starting point to devise new analysis strategies. The formulas do not completely apply to real problems, but rather offer an insight into a useful manipulation whose feasibility can be checked by means of simple tests and observations on the available experimental data.

Recall the formula of the chemical equilibrium for a reversible reaction:

$$\frac{K_a^{M_a} K_b^{M_b}}{K_c^{M_c} K_d^{M_d}} = Kost \qquad (1)$$

where K_i are the concentrations of the molecules involved in the reaction with the respective chemical weights M_i, while *Kost* is the equilibrium constant which is independent from the concentrations. If we take logarithms we obtain the linear equation:

$$M_a \ln K_a + M_b \ln K_b + M_c \ln K_c + M_d \ln K_d = \ln Kost \qquad (2)$$

Now consider that a chemical process is the sum of the contribution of all the elementary reactions involved. This suggests invoking the law of large numbers and looking for Gaussian distributions. However these are stable with respect to linear operations, not multiplications.

In probability theory, the sum of two or more Gaussian random variables is a Gaussian random variable as well (possibly degenerate to a Dirac delta). The same is not true for the product of Gaussian random variables.

So one gain the intuition that the logarithm of a concentration might have a Gaussian distribution (see Fig. 1a and b).

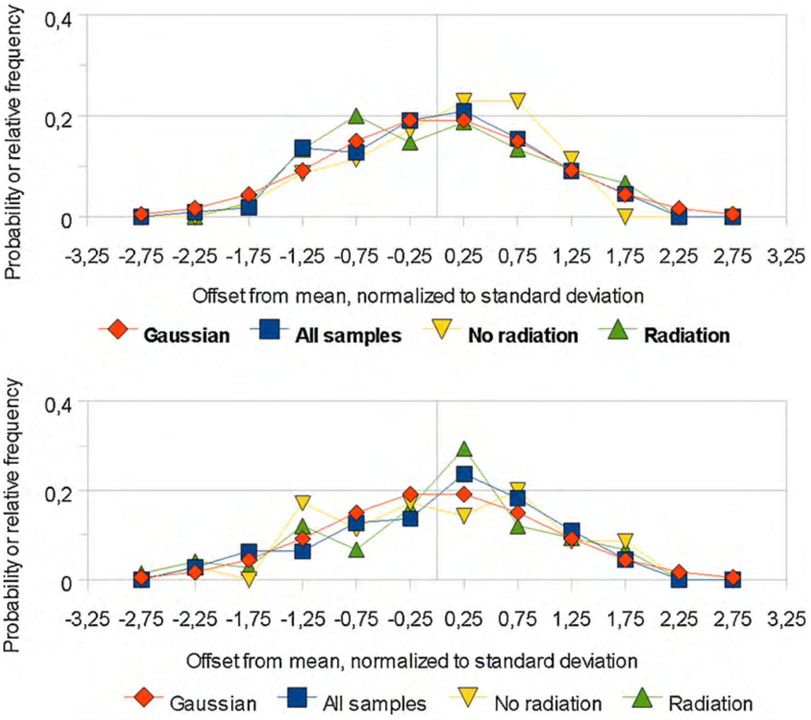

Fig. 1. Distributions of log-concentrations of (a) mass 45 (acetaldehyde) and (b) mass 59 (acetone).

Working with Gaussians is of course attractive because of the low number of parameters required to identify the distribution (mean and variance), as well as the abundance of literature [24] on the subject, both theoretical and operational.

Recall now the most basic diffusion process equation:

$$K_t = K_0\ e^{-a_t} + (1 - e^{-a_t})K_\infty \tag{3}$$

where t is the time, a is a diffusion coefficient and K_t, K_0 and K_∞ are the concentrations at time t, zero and at saturation, respectively. One can roughly associate K_0 to the inhaled room air and K_t to the expired air sample after a time t. In this model the quantity of interest to us would be K_∞, the asymptotic concentration which is independent on K_0 and most closely associated with the state of the lung cells according to the mentioned assumption of blood-born molecules affected by radiation. If one assumes that a_t is small or in other words that a respiratory act is short with respect to the relaxation time of the lungs' diffusion processes, one notice that K_t is a poor estimator of K_∞. With some manipulations the diffusion equation can be reformulated as follows:

$$\ln\frac{K_t}{K_0} = \ln\left[1 + (1 - e^{-a_t})\frac{K_\infty}{K_0}\right] - a_t \cong (1 - e^{-a_t})\frac{K_\infty}{K_0} - a_t \tag{4}$$

which shows that the logarithm of the ratio between breath sample and room air sample might be a better indicator for K_∞, if the relative variation of K_0 is small. But is a_t really small? If that would be the case, one should see a strong correlation between K_t and K_0, that is between the room air and breath samples (see Fig. 2).

According to this approach normalising the breath samples with respect to the room air samples would reduce the noise introduced by the environmental conditions. But it would also reduce the effects of the PTR-MS device fluctuations with time: these affect room air and breath gas samples in the same way, and the considered normalisation would hopefully cancel out such effects.

In brief, the proposed preconditioning strategy will be to take logarithms and perform room air normalisation of the breath samples, while also adding a small constant β to all breath and room air sample counts before normalisation:

$$K_{preconditioned} = \ln\frac{K_{breath} + \beta}{K_{room\ air} + \beta} \tag{5}$$

This constant has a twofold aim. Firstly, to avoid divisions by zero in those cases where no molecules were detected in the room air. Secondly, to use this constant as one of the parameters of the test family, with the understanding that a high value for it would operate as a soft threshold for low count masses, effectively reducing their impact in the test: the logarithm of the ratio concentrates around zero as β increases. In other words, β will be the first parameter of the test family, a *noise parameter*.

4.2. Dimensionality reduction

Let us now focus on the heuristics for producing a parameterised family of tests.

Let us partition the sample data into two non-overlapping sets, a training set and a validation set. A test will be constructed on the basis of the information conveyed by the training set, while the score of the test will be measured on the validation set.

Such a sample data partition can, of course, influence the performance of the test family built on it, for example by selecting most of noisy outliers for the test subset. For this reason, one might repeat the construction and scoring procedures on a small number of partitions, each of them selected independently from the sample data and the other partitions, and then look for a typical result.

In the radiation detection test case independent and identically distributed Bernoulli binary random variables with parameter 0.8 will be used to select 7 independent training sets, each with on average 149 × 0.8 = 112.8 elements. One thus obtains training sets consisting of 119, 120, 118, 105, 110, 110 and 101 samples respectively. The value of 0.8 for the parameter was chosen in order to achieve a good compromise between the statistical robustness of test construction and scoring, the former requiring more information to build upon and hence more sample data.

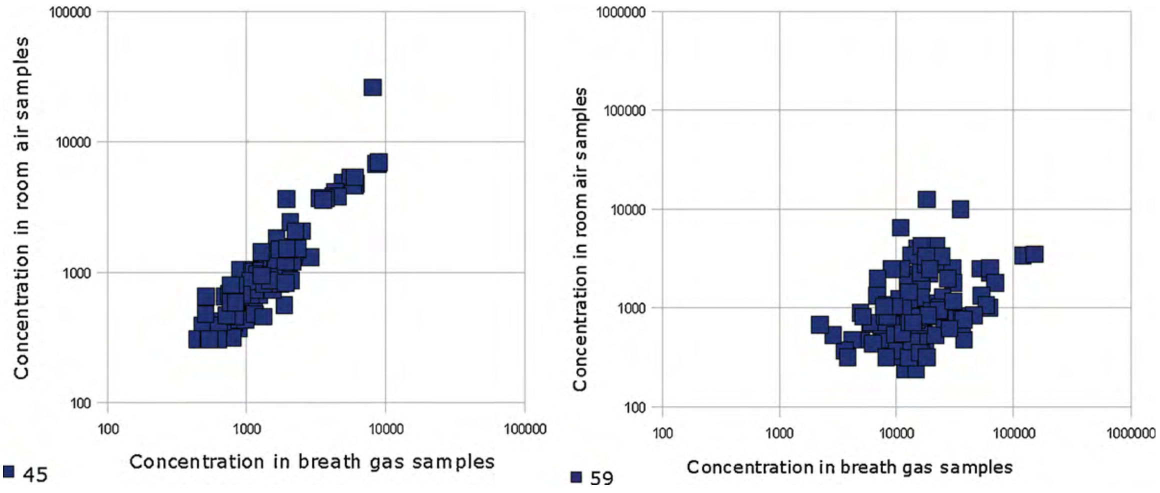

Fig. 2. Room air and breath gas correlation.

The scoring part of the analysis will be simple and informal: the resulting ROC (Receiver Operating Characteristic) curves will be inspected.

As for the test construction part, the target is a relatively small mass signal subset which still conveys as much statistical information as possible about the separation of the two populations, irradiated and not irradiated people.

This is again a trade-off on available resources, the samples, whose nature is limited and noisy. It is a compromise between the test score, which increases with the amount of statistical data used to build the test, and the test cost, since the higher the dimension, the more complex the test and the higher the number of training samples to build it.

Let us compute a separation score for each mass, which is an indicator on how well a single mass separates between the two sampled populations. Let us use the gaussianity assumption for log-signals and assume that the empirical variances of the two populations are the same. In this case a good indicator for separation is the following quantity:

$$D_i = \frac{\mu_{irr,i} - \mu_{not\ irr,i}}{\sigma_i} \tag{6}$$

where

$$\mu_{irr,i} = \frac{\sum_{x \in irradiated} K_i(x)}{S_{irradiated}} \tag{7}$$

$$\mu_{not\ irr,i} = \frac{\sum_{x \in not\ irradiated} K_i(x)}{S_{not\ irradiated}} \tag{8}$$

$$\sigma_i = \sqrt{\frac{\sum_{x \in irradiated}[K_i(x) - \mu_{irr,i}]^2 + \sum_{x \in not\ irradiated}[K_i(x) - \mu_{not\ irr,i}]^2}{S_{irradiated} + S_{not\ irradiated}}} \tag{9}$$

x are the samples, $K_i(x)$ are the preconditioned values for the ith mass, S is the number of irradiated or not irradiated samples in the test construction sample subset, and D_i is the distance between the empirical means of the two population distributions and the common empirical standard deviation.

To obtain our desired subset one just computes these D_i scores for each mass i, order them in decreasing score order and pick the first N of them. N is thus the second test parameter, a *dimension parameter*.

The independent scoring is clearly a sub-optimal approach, because it does not consider correlation between mass signals. Consider for instance the possibility that the two most separating masses are strongly correlated to each other, in the sense that both their informative aspects (the "signals") and the disturbing variations (the "noises") are in an approximately linear relation pairwise with each other. In this case using both the masses might be worse than using just one, because it would fill up an extra mass slot which could have been devoted to a *differently* informative signal, meaning a signal with an independent noise which would have increased the self-averaging effect of the linear combination.

Still, it is a very simple idea whose implementation complexity scales only linearly with the amount of mass signals measured. In the considered case of 181 masses, this is a significant advantage.

The third and last parameter will be a *distortion parameter Z* closely related to the noise parameter β. For each mass i one computes the fraction of times Z_i that the empirical count is lower than the noise parameter β.

$$Z_i = \frac{||x : K_i(x) < \beta||}{S_{irradiated} + S_{not\ irradiated}} \tag{10}$$

This provides an indication on how much the preconditioning distorts the original data. By discarding any mass whose Z_i is greater than the parameter Z, one is reducing the global distortion introduced by the preconditioning at the hopefully small price of discarding relatively small and noisy signals.

4.3. Test construction

After preconditioning and dimensionality reduction the problem is reduced to a small set of opportunely preconditioned signals: a low-dimensional, quasi-Gaussian problem. Simple strategies like the Linear Discriminant Analysis (LDA) are now a viable solution, which will be parameterized by β, N and Z.

Alternatively one can use a much simpler weighted scalar product approach (WSP):

$$\mathrm{WSP} = \sum_{i \in selected\ masses} Q_i \times K_i\{test\ \ sample\} \tag{11}$$

$$Is\ \ Irradiated = Boolean\{WSP > 0\} \tag{12}$$

where

$$Q_i = \frac{\mu_{irr,i} - \mu_{not\ irr,i}}{\sigma_i^2} \tag{13}$$

The WSP can be seen as a radically simplified version of the LDA in which the empirical correlation matrix has been set to zero with the exception of the diagonal entries.

To summarize this relatively long section, a new approach for the analysis of high dimensional breath sample data was described, based on assumptions coming from the theory of statistical mechanics and chemical equilibrium, as well as suggestions about test complexity coming from information theory. The approach consist in preconditioning the data by extracting logarithms and normalising with respect to the room air, then extract a significant mass subset and finally construct a simple linear test on the subset. Three parameters representing noise level, dimension and distortion allow to tune the test for a better performance. The next section illustrates the outcome of the proposed approach when applied to the radiation detection problem.

5. Results of the multidimensional analysis in the radiation detection test case

Before presenting the results obtained by the above approach in the radiation detection case, one shall take care of some extra technical tweaks. In detail, certain mass signals should be masked out because of being known markers for other unrelated phenomena [7,9,22]:

- Mass 42 is typically associated to acetonitrile, a marker for smokers. Since smoking correlates strongly to people developing lung cancer and being consequently subjected to radiotherapy, one cannot associate mass 42 directly to exposure to radiation.
- Masses 69, 72 and 41 are typically associated to isoprene [25,26], a gas which appear to increase with cellular damage and repair, but decrease during exercise [2]. The disputed nature of the decrease (ventilation or metabolism) suggests a cautionary exclusion of this signal.
- Masses 22–28 appear to be background signal, no interesting molecules here.
- Masses 88, 89 and 95 are significantly contaminated by bag impurities.

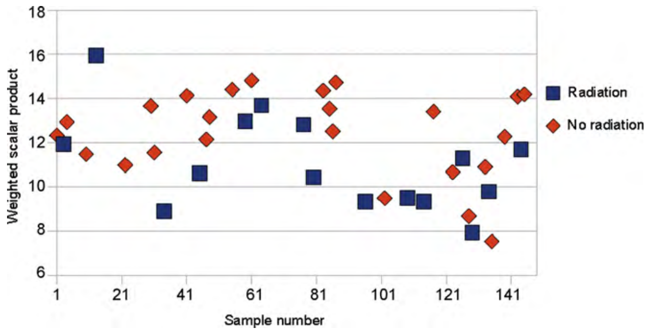

Fig. 3. WSP test.

No quantitative analysis and optimization has yet been done on the test parameters. This will be the topic of further research. Instead, some heuristic exploration of different parameter combinations was performed on seven different sample subset partitions, leading to the following observations:

- A parameter choice producing more coherent, less noisy results between partitions is observed for $\beta = 30$, $N = 7$ and $Z = 0.01$.
- Masses 45, 46, 59, 60, 63 and 73 pop up for most of the partitions, the first four typically associated to Acetaldehyde and Acetone respectively.
- Acetone appears to be critical for the test quality, even if it is a molecule that changes wildly with eating—the results are significantly worse when it is masked out.
- LDA and WSP do not seem to perform too differently. Fig. 3 shows the scores of the WSP test for the test sample subset 6.
- No measurable effect is noticed when confronting pairs of patients' samples before the treatment and after the first session,

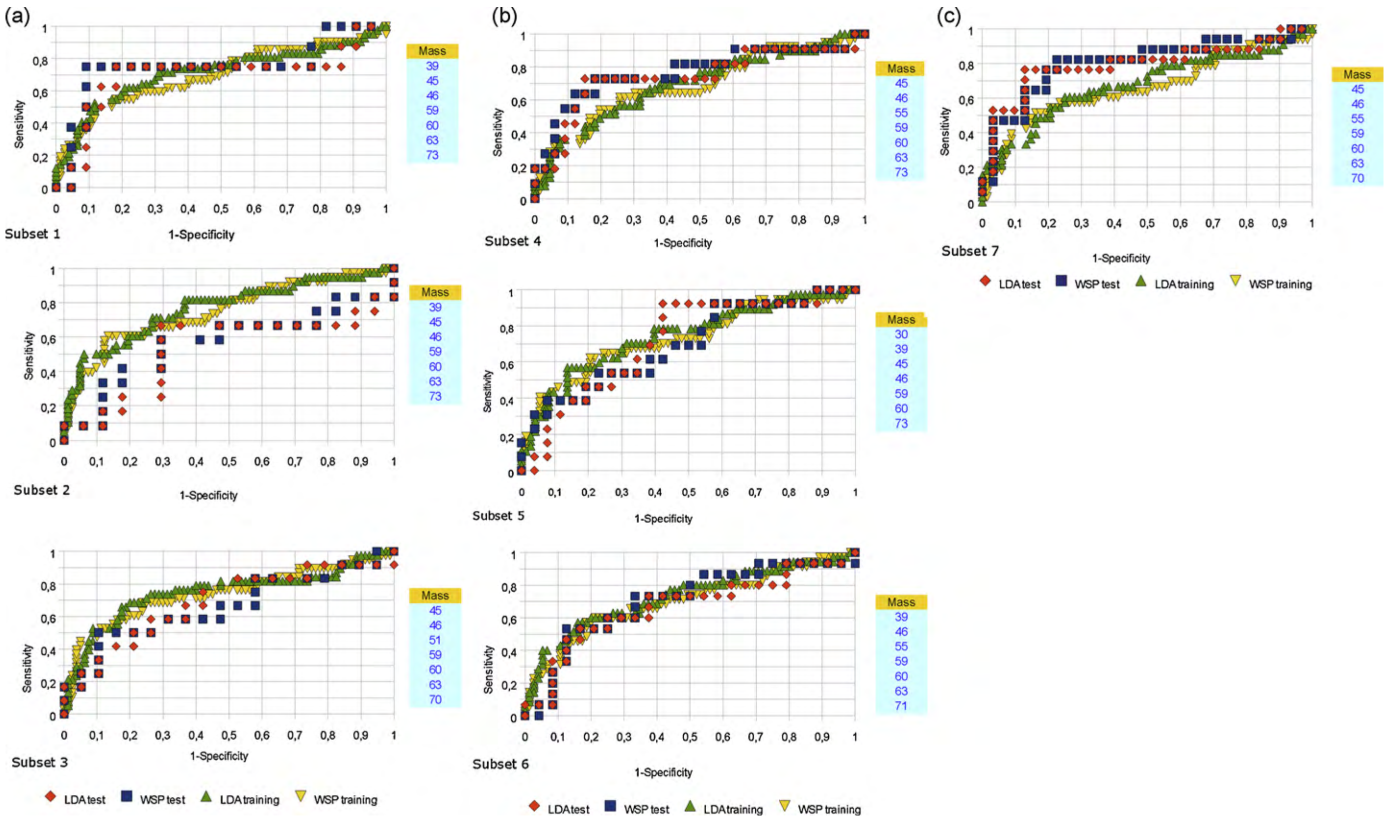

Fig. 4. ROC curves.

the small sample number available (15) is here a statistical limit to decisive observations.

Fig. 3 shows the scoring of the test based on the training partition subset 6.

The ROC curves achieved by applying the proposed statistical analysis are not yet good enough for screening, but they consistently show a significant correlation between the test and the condition of interest, in this case the exposure to ionising irradiation, although no study on dose dependence was performed at this stage. Fig. 4a–c shows seven groups of ROC curves, one group for each data subset partition considered. Each group contains four ROC curves, two for the training data (i.e., the "learning" curve) and two for the test data (the actual "score" or quality measure of the test). The reason for having two pairs of curves each is that both the LDA and WSP strategies were applied. They perform similar to each other. There are some fluctuation in the test performances with respect to the training, namely there is a poor performance for subset two and a good one for subset seven. This is likely due to outliers being concentrated in the test subset two and in the training subset seven, and justifies the approach of considering a small plurality of sample partitions instead of just one.

Fig. 4a–c shows the separation between irradiated and non-irradiated populations (according to tests built with the same parameters but on different data set partitions) between training subset and scoring subset. On the right side are the lists of the selected masses.

6. Conclusions

A multidimensional statistical analysis of PTR-MS data obtained from breath gas samples was described and successfully applied, as a test case, to the problem of detecting exposure of humans to ionising radiation.

The analysis is based on the theory of chemical-diffusion equilibrium and on complexity concepts based on information theory. It consists of a data preconditioning step by means of a normalisation with respect to room air samples, followed by a dimensionality reduction strategy. It works with a relatively limited amount of data, it shows good results on the problem considered and it can be applied to general PTR-MS based detection problems.

Because of the promising ROC curves obtained, the application of this analysis to the PTR-MS based detection of human irradiation might offer a non-invasive, low-cost, high-throughput test. The analysis hints that acetaldehyde and acetone might play an important role in the detection.

The innovative aspects of the present paper lie mostly in the domain of methodology, but they also extend to operative results. In the first place, the introduction of data preconditioning based on room air normalisation and gaussianity, and a dimensionality reduction approach inspired by information theory which allows operating with relatively few samples with respect to dimensionality are the main contributions. Early operative results are a new test for detecting exposure to ionising radiation by means of PTR-MS analysis of breath samples.

Many aspects of both theory and operative protocol need to be further developed.

First of all, on the side of statistical analysis, there is a lot of space for development. One should try to refine the normalisation assumptions, as the correlation between breath and room air shown in Fig. 2 is not ideal and dependent on the mass observed. Plenty of work remains to be done in the choice of the number and the identities of the test masses, since in the test case it was just done empirically by looking at some ROC curves. Other critical points include the hitherto ignored correlation between masses and the use of non-linear tests.

Secondly, the proposed approach is agnostic to any biochemical explanation. Understanding the link between the selected masses and the problem at hand (in the test case, the exposure to ionising radiation) is required, *in primis* to ascertain that the test does not correlate with conditions which characterise the sampled population but have no direct connection to the phenomenon of interest (for instance age or cancer affliction in hospital patients subjected to radiotherapy).

On the operative side, better sample collection protocols are needed in order to reduce dependence from some of the many sources of unwanted sample variation, and consequently noise. This involves specifying patient conditions like sleep and diet, for instance. This would be useful in hospital environments for PTR-MS breath gas based detection of pathological conditions, but not in some dire scenario of a large number of measurements after radiation contamination in the open. Experimenting with different breathing techniques (for instance re-breathing) or storage systems might lead to better results independent of being in a medical facility, as could the selection of narrower bands for the operation of the PTR-MS measurement, like for instance the temperature of the sample. More trivially, collecting more samples to produce better tests would be a very reasonable thing to do.

Finally, the application of the analysis to other PTR-MS breath detection problems, different from irradiation detection, will be needed to truly ascertain the validity of the proposed paradigm.

References

[1] T.H. Risby, S.F. Solga, Current status of clinical breath analysis, Applied Physics B: Lasers and Optics 85 (2006) 421–426.

[2] W. Miekisch, J.K. Schubert, G.F.E. Noeldge-Schomburg, Diagnostic potential of breath analysis—focus on volatile organic compounds, Clinica Chimica Acta 347 (2004) 25–39.

[3] A. Amann, G. Poupart, S. Telser, M. Ledochowski, A. Schmid, S. Mechtcheriakov, Applications of breath gas analysis in medicine, International Journal of Mass Spectrometry 239 (2004) 227–233.

[4] A. Hansel, A. Jordan, R. Holzinger, P. Prazeller, W. Vogel, W. Lindinger, Proton transfer reaction mass spectrometry: on-line trace gas analysis at ppb level, International Journal of Mass Spectrometry and Ion Processes 149/150 (1995) 609–619.

[5] W. Lindinger, A. Hansel, A. Jordan, Proton-transfer reaction mass spectrometry (PTR-MS): on-line monitoring of volatile organic compounds at pptv levels, Chemical Society Reviews 27 (1998) 347–354.

[6] A. Hansel, T.D. Märk (Eds.), International Journal of Mass Spectrometry, Special Issue on Proton Transfer Reaction Mass Spectrometry Volume 239, Issues 2–3 (2004).

[7] A. Jordan, A. Hansel, R. Holzinger, W. Lindinger, Acetonitrile and benzene in the breath of smokers and non-smokers investigated by proton transfer reaction mass spectrometry (PTR-MS), International Journal of Mass Spectrometry and Ion Processes 148 (1995) L1–L3.

[8] J. Taucher, Investigation of Volatile Organic Compounds in Human Breath Using Proton Transfer Reaction Mass Spectrometry, Institute of Ion Physics and Applied Physics Innsbruck, Innsbruck, 1996 (in German).

[9] J. Taucher, A. Hansel, A. Jordan, R. Fall, J.H. Futrell, W. Lindinger, Detection of isoprene in expired air from human subjects using proton-transfer-reaction mass spectrometry, Rapid Communications in Mass Spectrometry 11 (1997) 1230–1234.

[10] T. Karl, P. Prazeller, D. Mayr, A. Jordan, J. Rieder, R. Fall, W. Lindinger, Human breath isoprene and its relation to blood cholesterol levels: new measurements and modeling, Journal of Applied Physiology 91 (2001) 762.

[11] W. Lindinger, J. Taucher, A. Jordan, A. Hansel, W. Vogel, Endogenous production of methanol after the consumption of fruit, Alcoholism: Clinical and Experimental Research 21 (1997) 939.

[12] B. Moser, F. Bodrogi, G. Eibl, M. Lechner, J. Rieder, P. Lirk, Mass spectrometric profile of exhaled breath—field study by PTR-MS, Respiratory Physiology and Neurobiology 145 (2005) 295.

[13] P. Lirk, F. Bodrogi, J. Rieder, Medical applications of proton transfer reaction-mass spectrometry: ambient air monitoring and breath analysis, International Journal of Mass Spectrometry 239 (2004) 221.

[14] A. Wehinger, A. Schmid, S. Mechtcheriakov, M. Ledochowski, C. Grabmer, G.A. Gastl, A. Amann, Lung cancer detection by proton transfer reaction mass-spectrometric analysis of human breath gas, International Journal of Mass Spectrometry 265 (2007) 49.

[15] Greiter Matthias: Atemluftanalyse mittels Protonen Transfer Reaktions Massenspektrometrie (PTR-MS), TU München, 2003.

[16] H.F.-W. Sadrozinski, Radiation effects in life sciences, Nuclear Instruments and Methods in Physics Research A514 (2003) 224–229.
[17] J.B. Little, Non-targeted effects of ionizing radiation: implications for radiation protection, in: Proceedings of 11th International Congress of the International Radiation Protection Association, IRPA 11, Madrid, Spain, May 23–28, 2004.
[18] K.M. Prise, G. Schettino, M. Folkard, K.D. Held, New insights on cell death from radiation exposure, Lancet Oncology 6 (July (7)) (2005) 520–528.
[19] J.C. Anderson, A.L. Babb, M.P. Hlastala, Modeling soluble gas exchange in the airways and alveoli, Annals of Biomedical Engineering 31 (2003) 1402–1422.
[20] K.D. Skeldon, C. Patterson, C.A. Wyse, G.M. Gibson, M.J. Padgett, C. Longbottom, L.C. McMillan, The potential offered by real-time, high-sensitivity monitoring of ethane in breath and some pilot studies using optical spectroscopy, Journal of Optics A: Pure and Applied Optics 7 (2005) 376–384.
[21] L. Keck, U. Oeh, C. Hoeschen, Effects of carbon dioxide in breath gas on proton transfer reaction-mass spectrometry (PTR-MS) measurements, International Journal of Mass Spectrometry 270 (2008) 156–165.
[22] J. Taucher, A. Lagg, A. Hansel, W. Vogel, W. Lindinger, Methanol in human breath, Alcoholism: Clinical and Experimental Research 19 (1995) 1147–1150.
[23] P. Atkins, J. de Paula, Physical Chemistry, 8th edition, Oxford University Press (2006).
[24] S. Ross, Introduction to Probability and Statistics for Engineers and Scientists, 3rd edition, Elsevier (2004).
[25] A. Amann, S. Telser, L. Hofer, A. Schmod, H. Hinterhuner, Breath gas as biochemical probe in sleeping individuals, in: A. Hansel, T.D. Märk (Eds.), 2nd International Conference on Proton transfer Reaction Mass Spectrometry and its Applications, Innsbruck University Press, Obergurgl, 2005.
[26] C. Turner, P. Spanel, D. Smith, A longitudinal study of breath isoprene in healthy volunteers using selected ion flow tube mass spectrometry (SIFT-MS), Physiological Measurement 27 (2006) 13–22.