Multiple testing issues in discriminating compound-related peaks and chromatograms from high frequency noise, spikes and solvent-based noise in LC-MS data sets

Nyangoma, Stephen O.; van Kampen, Antoine A. H. C.; Reijmers, Theo H.; Govorukhina, Natalia; van der Zee, Ate; Billingham, Lucinda J.; Bischoff, Rainer; Jansen, Ritsert

Multiple Testing Issues in Discriminating Compound-Related Peaks and Chromatograms from High Frequency Noise, Spikes and Solvent-Based Noise in LC – MS Data Sets

Nyangoma, Stephen; Kampen, Antoine A.H.C. van; Reijmers, Theo H.; Govorukhina, Natalia; van der Zee, Ate; Billingham, Lucinda J.; Bischoff, Rainer; Jansen, Ritsert

Published in:
Default journal

Publication date:
2007

Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 23

Multiple Testing Issues in Discriminating Compound-Related Peaks and Chromatograms from High Frequency Noise, Spikes and Solvent-Based Noise in LC – MS Data Sets

Stephen O. Nyangoma[*] Antoine A. H. C. van Kampen[†] Theo H. Reijmers[‡]
Natalia I. Govorukhina[**] Ate G. J. van der Zee[††] Lucinda J. Billingham[‡‡]
Rainer Bischoff[§] Ritsert C. Jansen[¶]

[*] University of Birmingham, s.o.nyangoma@amc.uva.nl
[†] Academic Medical Centre Amsterdam, a.h.vankampen@amc.uva.nl
[‡] University of Leiden, t.reijmers@chem.leidenuniv.nl
[**] University of Groningen, N.Govorukhina@rug.nl
[††] University Medical Centre Groningen, A.G.J.van.der.Zee@og.azg.nl
[‡‡] University of Birmingham, l.j.billingham@bham.ac.uk
[§] University of Groningen, r.p.h.bischoff@rug.nl
[¶] University of Groningen, r.c.jansen@rug.nl

Multiple Testing Issues in Discriminating Compound-Related Peaks and Chromatograms from High Frequency Noise, Spikes and Solvent-Based Noise in LC – MS Data Sets*

Stephen O. Nyangoma, Antoine A. H. C. van Kampen, Theo H. Reijmers, Natalia I. Govorukhina, Ate G. J. van der Zee, Lucinda J. Billingham, Rainer Bischoff, and Ritsert C. Jansen

Abstract

Liquid Chromatography - Mass Spectrometry (LC-MS) is a powerful method for sensitive detection and quantification of proteins and peptides in complex biological fluids like serum. LC-MS produces complex data sets, consisting of some hundreds of millions of data points per sample at a resolution of 0.1 amu in the m/z domain and 7000 data points in the time domain. However, the detection of the lower abundance proteins from this data is hampered by the presence of artefacts, such as high frequency noise and spikes. Moreover, not all of the tens of thousands of the chromatograms produced per sample are relevant for the pursuit of the biomarkers. Thus in analysing the LC-MS data, two critical pre-processing issues arise. Which of the thousands of the: 1. chromatograms per sample are relevant for the detection of the biomarkers?, and 2. signals per chromatogram are truly compound-related? Each of these issues involves assessing the significance (deviation from noise) of multiple observations and the issue of multiple comparisons arises. Current methods disregard the multiplicity and provide no concrete threshold for significance. However, with such procedures, the probability of one or more false-positives is high as the number of tests to be performed is large, and must be controlled. Realizing that the cut-offs for declaring a chromatogram (or a signal) to be compound-related can hugely influence which proteins are detected, it seems natural to define thresholds that are neither arbitrary nor subjective. We suggest the choice of thresholds guided by the critical aim of controlling the False Discovery Rate (FDR) in multiple hypotheses testing for significance over a large set of features produced per sample. This involves the use of the regression diagnostics to characterize the signals of a chromatogram (e.g. as outliers or influential) and to suggest suitable tests statistics for the multiple testing procedures (MTP) for discriminating noise and spikes from true signals. The role of

*This work was funded by the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the NGI. We gratefully acknowledge support from two Groningen Bioinformatics Centre funds, the Bolgewasen project (grant number SENTER-TSGE3043) and the Tropisch Fruit project (grant number SENTER-TSOM3005), and also the funding by MRC UK (grant No. RRAK11686). RB was funded by the Dutch Cancer Fund (KWF grant number RUG 2004-3165).

the Generalized Linear Models (GLM) in this MTP is investigated. The method is applied to LC-MS datasets from trypsin-digested serum spiked with varying levels of horse heart cytochrome C (cytoc).

KEYWORDS: liquid chromatography, mass spectrometry, chromatogram, diagnostics, generalized linear models, proteins, biomarkers, noise, spikes, compound-related peaks, outliers, influential observations, false discovery rate, family-wise error rate, multiple testing

1 Introduction

1.1 LC-MS data and pre-processing methods

LC-MS is one of the widely used analytical methods for the analysis and comparison of complex protein and peptide mixtures. It has the advantage of combining the high separation efficiency of high performance liquid chromatography (HPLC) with the selectivity of mass spectrometry. Chromatography is a process in which a chemical mixture carried by a liquid or gas is separated into components as a result of differential distribution of the solutes as they flow around or over a stationary liquid or solid phase. A chromatogram is a time-based graphic record (as of concentration of eluted materials) of a chromatographic separation. Mass spectrometer is an instrument that separates ions according to their molecular masses, also called the mass to charge ratio and denoted by m/z.

The aim of the LC-MS experiment that produced the datasets studied in this manuscript was to detect and quantify proteins and peptides in the serum of cervical cancer patients. This will be achieved by exploring differences in proteomic composition of serum from these patients before and after treatment for cervical cancer. However, the presence of highly abundant proteins e.g. Human Serum Albumin (HSA) and Immunoglobulin G (IgG) often masks those of lower abundance, hindering their identification and quantification. An analytical method for enhancing the detection of those proteins is the depletion of high-abundance proteins from a sample followed by trypsin (an enzyme that catalyzes the hydrolysis of proteins to form smaller polypeptide units, put simply, it is an enzyme that acts to degrade proteins) digestion and LC-MS (Govorukhina, *et al*., 2006). Still, this method does not eliminate the artefacts such as high frequency noise and spikes in chromatograms that also make it difficult to detect the proteins, especially those of lower abundance. This raises the question of which of the thousands (≈ 7000) of the observed signals are truly compound-related, which highlights the first *multiplicity* issue.

The LC-MS of a sample (by sample we mean a biological sample i.e. a specimen) yields a collection of (time, m/z, intensity) measurements, each indicating that at a particular (retention) time, an ion with a particular m/z was detected with a particular intensity. In this space, the observations over a fixed m/z form the chromatogram. The mass is given in *atomic mass units* (amu, also known as Daltons). For small molecules and most ion sources, z $= \pm 1$. For proteins z can be large, up to +50 or higher. At a resolution of 0.1 amu in the m/z domain, an LC-MS run of a sample produces about 14000 *mass spectra* (the pattern of the relative abundances of ions of different atomic masses within a sample) and a chromatographic run time of 2 hours ($\approx$7000 *retention times*).

However, in different runs, measurements will not necessarily occur at the same time or measure exactly the same m/z. To make measurements directly comparable, the common practice is to round (bin) the m/z and time measurements to user-specified levels. The m/z is often rounded to the nearest unit (Windig, *et al.*, 1996; Wiener, *et al.*, 2004) and the time to the nearest 0.05 minutes (3 seconds) (Wiener, *et al.*, 2004), which puts the intensities on a $\text{time} \times \text{m/z}$ plane, which is a subset of $\mathbb{R}^2$.

Chromatograms (sometimes called the *m/z traces*) give information about the elution of compounds, represented by peaks across retention time. Some of these peaks represent possible peptides or proteins (or eluting compounds or analytes) characterizing a phenomenon, while others are spikes and noise (categorized in this manuscript into two groups as high solvent-based (e.g. Fig. 1 (b)) and frequency (e.g. Fig. 2 (a)) noise) that do not contain compound-related information. The compound-related peaks are characterized with broader bases (e.g. Fig. 1 (a)), while the spikes have narrower bases (e.g. Fig. 1 (c)). In this paper, the word peak will be used to refer to a compound-related peak. High quality chromatograms have distinct peaks and bases that contain minimal background noise. The noise level may also be used to categorize chromatograms. A chromatogram that is dominated by solvent-based noise has signals of “constant” magnitudes over the entire retention time range and exhibits no peaks, and thus contains no compound-related information. The solvent-based noise is generally due to mobile phase components that give a signal at each time point over the entire retention time range, while random noise is due to the electrospray ionisation (ESI) interface connecting HPLC and MS. A chromatogram may be contaminated by a complex combination of these artefacts. While many denoising procedures reduce this complexity they do not totally eliminate it. Thus a number of the chromatograms may still exhibit no informative peaks or may contain uncertain compound information even after noise filtering. Examples of such m/z traces include, those contaminated with the solvent-based noise (e.g. Fig. 1 (b)) and those with isolated random noise, expressed at low levels (e.g. Fig. 3 (a), (b) and (d)). Consequently, there is a need to detect and omit the least informative m/z traces from the ultimate data analyses and the question arises of which of the thousands (≈ 1400) of these chromatograms truly contain compounds that may be used in the biomarkers discovery. This issue yet again highlights another multiplicity problem associated with the pre-processing of the LC-MS data, that of detecting the relevant chromatograms.

In this report, we address the two critical pre-processing issues that are vital for reliable peak detection. Which of the thousands of the:

1. chromatograms per sample are relevant for the detection of the biomarkers?, and
2. signals per chromatogram are truly compound-related?

Each of these issues involves assessing the significance (deviation from noise) of multiple observations and the issue of *multiple comparisons* arise. Current methods (e.g. Windig, *et al*., 1996; Windig, *et al*., 2001; Gaspari, *et al*., 2001) disregard the multiplicity issue and provide no concrete threshold for significance. However, with such procedures, the probability of one or more false-positives is high as the number of features to be assessed is large. This means that the probability that some chromatograms will be declared as compound-related by chance alone cannot be neglected and needs to be controlled. Moreover, the preferred methods of eliminating spikes such as the moving median involve fictitious and arbitrary use of the user defined windows sizes. We deduced that depending on the window size chosen; this method could also significantly alter the intensities of the compound-related signals, which hugely lowers its worth. Realizing that the cut-offs for declaring a chromatogram (or a signal) to be compound-related can hugely influence which proteins are detected, it seems natural to define thresholds that are neither arbitrary nor subjective. We suggest the choice of thresholds guided by the critical aim of controlling the False Discovery Rate (FDR) in multiple hypotheses testing for significance over a large set of features produced per sample. This involves the use of the GLM regression diagnostic methods such as the leave-one-out and the case-perturbation (Cook & Weisberg, 1982; Cook, 1986; Williams, 1987) to characterize the signals of a chromatogram (e.g. as outliers (noise) or influential (compound-related)) and chromatograms (as noisy (outliers) or compound-related), which then suggests suitable test statistics for discriminating noise and spikes from true signals using the MTP. MTPs allow one to assess simultaneously, the significance of the results of a family of hypotheses tests. They focus on the specificity by controlling type I (false positive) error rates such as family-wise error rates (Dudoit, *et al*., 2004), false discovery rates (FDR) (Benjamini & Hochberg, 1995) and the false positive proportion (Lehmann & Romano, 2005). The method is applied to LC-MS datasets from trypsin-digested serum spiked with varying levels of horse heart cytochrome C (cytoc).

The MTPs are routinely applied to control the number of false positive results in the microarray studies. However, the application of these methods to detect compound contents of chromatograms is complex, unique and poses further challenges. Assessing the compound contents of a chromatogram involves comparing the raw signals with their smoothed counterparts using simple linear regression (two-sample t-test) and entails using thousands of observed signals (≈ 7000) over the entire retention time range (equivalent to using sample of size ≈ 7000 in microarrays) compared to the usually small sample sizes (e.g. 20 but often even less) in microarrays. Thus whereas the use of p-values (as a basis for FDR control) make perfect sense in microarrays and the MTPs are applied to these values directly, they may be highly misleading in our LC-MS application because of the large sample sizes involved. When sample size is large, the values of test statistics can be extremely large, giving rise to very small p-values, which have to be standardized so that they can be used in the usual way (Good, 1992). Thus we avoid using the conventional FDR controlling procedures, but adopt an alternative approach, the empirical Bayes method, in which each chromatogram is assigned a posterior probability that it is non-compound-related using the distribution of appropriate test statistics (Efron & Tibshirani, 2002; Liao, *et al*., 2004).

There are many features exhibited by mass chromatograms that motivate the use of regression diagnostics in detecting them. A chromatogram with high background noise (high background is used here to mean dominance by solvent-based noise) has a signal (of roughly uniform intensity) over the entire time range leading to it having a higher mean signal value relative to the observed signals than the high-quality ones. This means that the signal to mean-signal ratio is lower for all signals in those chromatograms, but is high for peaks in high-quality chromatograms. With respect to this criterion, a chromatogram with high background noise may be considered as an “outlier” in m-dimensional space. At individual chromatogram level, the intensities of the signals of the spikes and compound-related peaks are large compared to those of the main cluster of signals, noise, that are expressed at background levels. This suggests that these features could be outlier (and/or influential) signals of a chromatogram. In this work, it is shown that these chromatographic features may be detected using diagnostic tools derived from the GLM (with possibly identity or gamma links) in which the signals of a raw chromatogram play the role of independent variables and those of their smoothed and perturbed versions, the predictor variables. The noisy signals turn out to be those observations that are not well fitted by the model.

Finally, the issue of run-to-run retention time variation has been identified as a significant impediment of LC-MS data analysis. Therefore, time warping is an essential pre-processing step in the analysis of LC-MS datasets, undertaken to align the compound-related features of chromatograms with same m/z values drawn from different LC-MS runs (see e.g. Nielsen, *et al*., 1998; Bylund, *et al*., 2002; Johnson, *et al*., 2003). These algorithms assume that the corresponding chromatograms have similar landmark profiles across the retention time with possibly small shifts in positions that may be corrected by time warping. However, artefacts such as high frequency noise and spikes often mask some landmark features, making the matching process complex. We develop methods for discriminating peaks from noise that may lead to more effective matching of the chromatographic features.

In Section 2, we review chemometric methods for ranking chromatograms according to their compound contents that partially motivated our work and spell out their failures. We then propose a GLM that describes the relationship between the raw chromatogram and its smoothed version and set up the null hypotheses for the MTPs in Section 3. In Section 4, we derive the test statistics for assessing the compound contents of chromatograms. In Section 5, the signals of a chromatogram are characterized and test statistics for detecting them proposed. We derive a link function for the GLM in Section 6 and introduce the corresponding test statistics. The multiple testing procedures for controlling the false positives are described in Section 7. Some computation results are presented in Section 8 and concluding remarks are discussed in Section 9.

2 Chemometric methods for detecting qualities of chromatograms

The noise inherent in the LC-MS data makes it difficult to identify the components present. Each LC-MS run produces thousands of chromatograms with varying noise contents. The high quality chromatograms have dominant distinctive peaks and have bases that are less contaminated with background noise. The other category of chromatograms are dominated by high frequency noise and spikes that mask the compound-related information, especially those of the lower abundance compounds. Another cohort of chromatograms is dominated by the solvent-based noise (e.g. Fig. 1 (b)). They exhibit no compound-related peaks and are thus irrelevant to the biomarkers discovery effort. In some cases, the random noise can be detected and discriminated from the compound-related signals. Because of the complexity in the composition of a chromatogram and the fact that not all chromatograms produced from an LC-MS run are compound-related, some attention has been devoted to the developing of the methods for selecting which chromatograms may be used in the biomarkers discovery (see e.g. Windig, *et al*., 1996; Windig, *et al*., 2001; Gaspari, *et al*., 2001). But unfortunately, none of these authors have considered the multiplicity nature of this problem. In this section, we review some popularly used chemometric methods for discriminating the compound-related chromatograms from those dominated by noise.

2.1 The Component Detection Algorithm (CODA)

To obtain data with improved signal-to-noise ratio the raw data is often filtered using smoothing algorithms, e.g. the moving average. The high quality chromatograms have low noise and background and are not so affected by the filtering process. The converse is true of the noisy chromatograms. Consequently, a first step in developing algorithms for detecting the quality (the noise level) of a chromatogram is the choice of a similarity measure $d(,)$ that quantifies the deviations of the raw chromatogram from its smoothed version.

Let c_k^* be the smoothed version of the raw chromatogram c_k of the kth m/z channel using a moving average method with window size *w*. Suppose u_k is a vector with elements $u_{jk} = (c_{jk}^* - \bar{c}_k^*)$, where $\bar{c}_k^* = \sum_j c_{jk}^* /(n-w+1)$ is the mean of the smoothed ion currents, then u_k is a smoothed and mean-subtracted version of c_k. CODA (Windig, *et al.*, 1996) utilizes a similarity index defined by

$$d(c_k,u_k) = \sum_j c_{jk}\, u_{jk} / \|c_k\| \|u_k\|,$$

where $\|u_k\| = \sqrt{\sum u_{jk}^2}$ is the Euclidean length of u_k, for example. Thus $d(c_k,u_k)$ is essentially the cosine of the angle between c_k and u_k, i.e. $0 \le d(c_k,u_k) \le 1$.

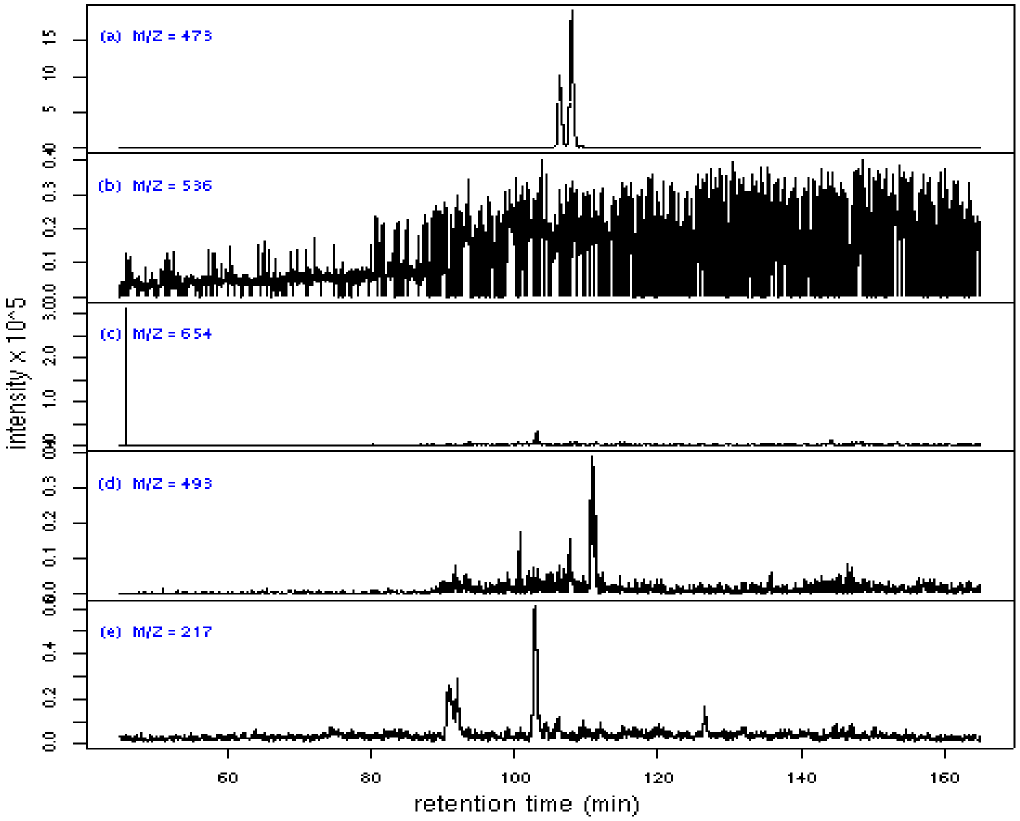

Figure 1. Chromatogram: (a) of high quality (b) dominated with solvent-based noise (c) with dominant spike (d) that is compound-related with high noise levels (e) that is compound-related with a baseline contaminated with solvent-based noise.

The value $d(c_k, u_k) = 0$ corresponds to no common peaks between the raw chromatogram and its smoothed version, while $d(c_k, u_k) = 1$ represents total similarity between the two spectra. The similarity thresholds are usually user defined, thus their choices can be quite arbitrary. It may be recognized that u_k is in a sense just a perturbation of the raw chromatogram c_k. Thus, simply put, CODA compares a raw chromatogram with its perturbed version, with the perturbation done to reduce the noise level (smoothing) and to pinpoint chromatograms dominated by solvent-based noise (mean subtraction).

2.2 Other algorithms for assessing quality of chromatograms

Another popularly used algorithm for m/z traces selection is the so-called Impress Quality (IQ) algorithm (Gaspari, *et al*., 2001). It is an employs an entropy-based measure of similarity. Taguchi's signal-to-noise ration method (Taguchi, 1986; Massart, *et al*., 1997; p. 799) is also popularly used. For LC-MS data, we established that Taguchi's method has poor performance, while it is not clear how the solvent-based noise is dealt with in the IQ algorithm.

Although these methods provide experimentalists with some criteria for ranking chromatograms according to their compound content, they are inadequate in three ways:

1. The quality thresholds are user defined, thus the chromatogram choices can be quite arbitrary.
2. These algorithms do not adequately provide means of discriminating the lower abundance proteins from spikes and noise.
3. None of these methods account for the *multiplicity* nature of the chromatogram selection procedure or of peak detection and the results are thus prone to include uncontrolled numbers of false positives.

We suggest methods to account for these oversights guided by the critical aim of controlling the FDRs in multiple hypotheses testing for significance over a large set of features produced per sample. To facilitate the quality detection process, noise and spikes are first detected and screened from the chromatograms using the Family Wise Error Rate (FWER) MTPs such as the Bonferoni, and the Holm's methods (Dudoit, *et al*., 2004) to control for the false detections of the spikes and noise from the chromatograms. The empirical Bayes criterion (Efron & Tibshirani, 2002) is then used to control for the number of chromatograms that are falsely declared to be compound-related. The main advantage of using the new method is that standard statistical quantities for setting up thresholds such as F-statistic (and/or their corresponding p-values) and coefficients of determination, *et* *cetra*, may be used as chromatogram quality selection criterion, which enables one to take the multiplicity nature of the selection process into account.

3 Detecting features on chromatograms

Several features of the m/z traces (chromatograms) that result from an LC-MS run make it difficult to identify the components present and hence to evaluate the differences in proteomic composition of samples describing different conditions. Fig. 2 displays a chromatogram with m/z = 303 from different runs of LC-MS of serum sample from a patient with cervical cancer, spiked with different levels of cytoc (so that each of the chromatograms (a), (c), (e) (g) and (j) depicts a profile for a different level of cytoc). For this m/z trace the experimenter knows that the cytoc peak elutes at the retention time ≈ *86* minutes. However, we see that at lower concentrations of cytoc, this peak is of lower abundance and is masked by high levels of noise and spikes that makes it indistinguishable from these artifacts. To facilitate the inter-sample comparisons of these m/z traces, it is imperative to first screen out the noise so as to unmask the hidden (lower abundance) peaks, a process that involves the vetting of the thousands of signals for compound-related information. If done on a per signal basis, the probability of one or more false-positives is high as the number of features to be assessed is large and must be controlled. We propose a method for setting up cut-offs for rejections driven by the crucial aim of controlling the number of signals that are falsely declared to be compound-related. An application of the step-down FWER MTPs, resulted in the transformation of the noisy chromatograms in Fig. 2 (a), (c), (e) (g) and (i) into the compound-related m/z traces depicted by fig. 2 (b), (d), (f), (h), (j), and thus unmasked the lower abundance cytoc peaks. This extols the sensitivity of the new method in detecting noise and its specificity in identifying the compound-related signals. The virtue of this method is that it does not involve the use of the fictitious window-sizes or subjective peak widths or arbitrary intensity thresholding, as do the conventional filtering methods. The test statistics for the ensuing MTPs are functions of the diagnostic tools from the GLM with the raw signals as response variables and their smoothed versions as predictor variables. How the GLM arises is explained below.

3.1 The model

Chromatograms contain various forms of noise, both of instrumental and chemical origin in addition to genuine compound-related information. The chemical noise results from a number of sources, such as components in the LC mobile phase or the sample that give rise to a significant background signal. Thus reduction of noise is especially important for biological samples, where matrix components may be much more concentrated than the analytes of interest.

To obtain data with improved signal-to-noise ratio the raw data is often filtered using smoothing algorithms, e.g. moving average. Consequently, a major issue in the development of the algorithms for assessing compound contents of a chromatogram is the choice of a similarity measure that quantifies the deviation of a raw chromatogram from its smoothed counterpart. The popularly used measures include the angle between the vectors of the signals of a raw chromatogram and those of its smoothed counterpart (Windig, *et al*., 1996; Massart, 1997; Windig, *et al*., 2001) and the entropy-based measures (Gaspari, *et al*., 2001). However, some non-informative chromatograms e.g. those dominated by the solvent-based noise are classified as compound-related by these measures. This led to the use of modified version of the angles between two vectors that instead measures the deviations of the raw m/z traces to their smoothed and mean subtracted versions (i.e. a perturbed version of the angle between two vectors), the so-called CODA to filter these artefacts. This was motivated by the simple fact that the signals of a solvent-based noise dominated chromatogram are somewhat uniformly distributed in the entire retention time range, which means that its mean-signal value is large relative to the intensities of its entire signals, unlike those of the other types of chromatograms. Thus a mean subtraction heavily penalizes signals of a solvent-based noise dominated chromatograms leading to poor correlation with its smoothed and perturbed counterpart. Although these measures give criteria for ranking chromatograms according to their compound contents, they do not provide thresholds for selecting these features. We propose novel thresholds for selecting relevant chromatograms guided by the crucial aim of controlling the number of m/z traces incorrectly declared as compound-related using MTPs. This requires the use of test statistics with known distributional properties. Thus instead of using arbitrarily modified statistics, we propose the use of diagnostics derived from a regression of the signals of a raw chromatogram on their perturbed versions.

3.1.1 The chromatogram smoothing method

Because we propose to derive our results from the regression of the raw signals on their smoothed counterparts as a means of detecting noise, the first critical question is which smoothing method is preferable. A suitable smoothing procedure must effectively remove the noise, while preserving the key characteristics of the compound-related peaks. A number of methods have been used including, moving average, moving median and even polynomial smoothers such as Savitzky-Golay method (Windig, *et al*., 1996; Massart, 1997; Windig, *et al*., 2001; Listgarten & Emili, 2005). Moving median appears to be a top preference. However, we deduced that it also significantly alters the intensities of the compound-related signals, which hugely lowers its worth. On the other hand, the moving average method reduces the noise contents of a chromatogram but it does not totally eliminate the spikes and high frequency noise. However, it does not alter the compound-related intensities much, this being the reason for its

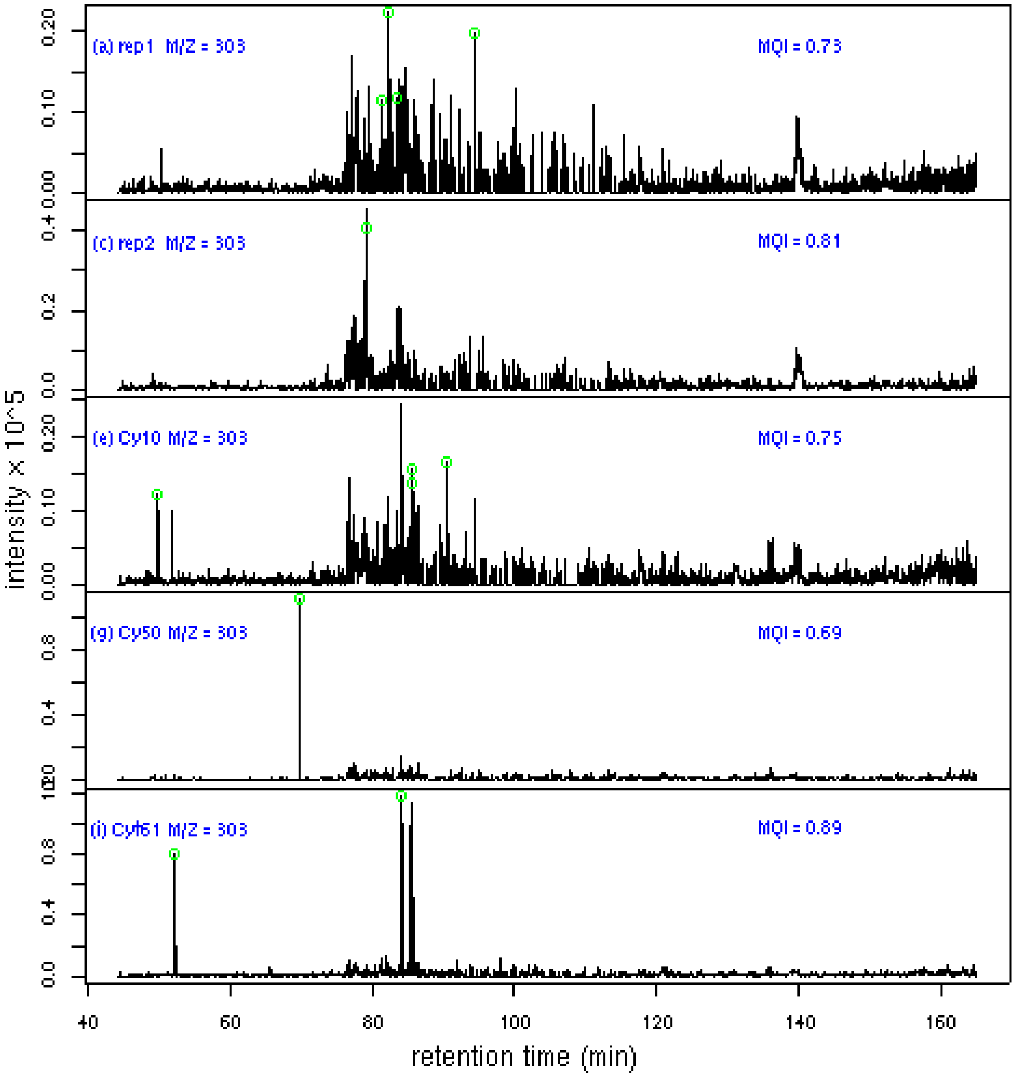

Figure 2: profile plots of signals from m/z = 303. (a) and (c) raw chromatogram from two replicates of the serum. (e), (g) and (i) are replicates of same serum with different levels of spikeins of Cytochrom C (10, 50 & 61 pm). The green circles are various artefacts detected in the first run of outliers detection algorithm using the Bonferroni MTP. Their corresponding versions after pre-processing using the MTP procedures proposed in this manuscript are shown in Figures 2 (b), (d), (f), (h) and (j), respectively.

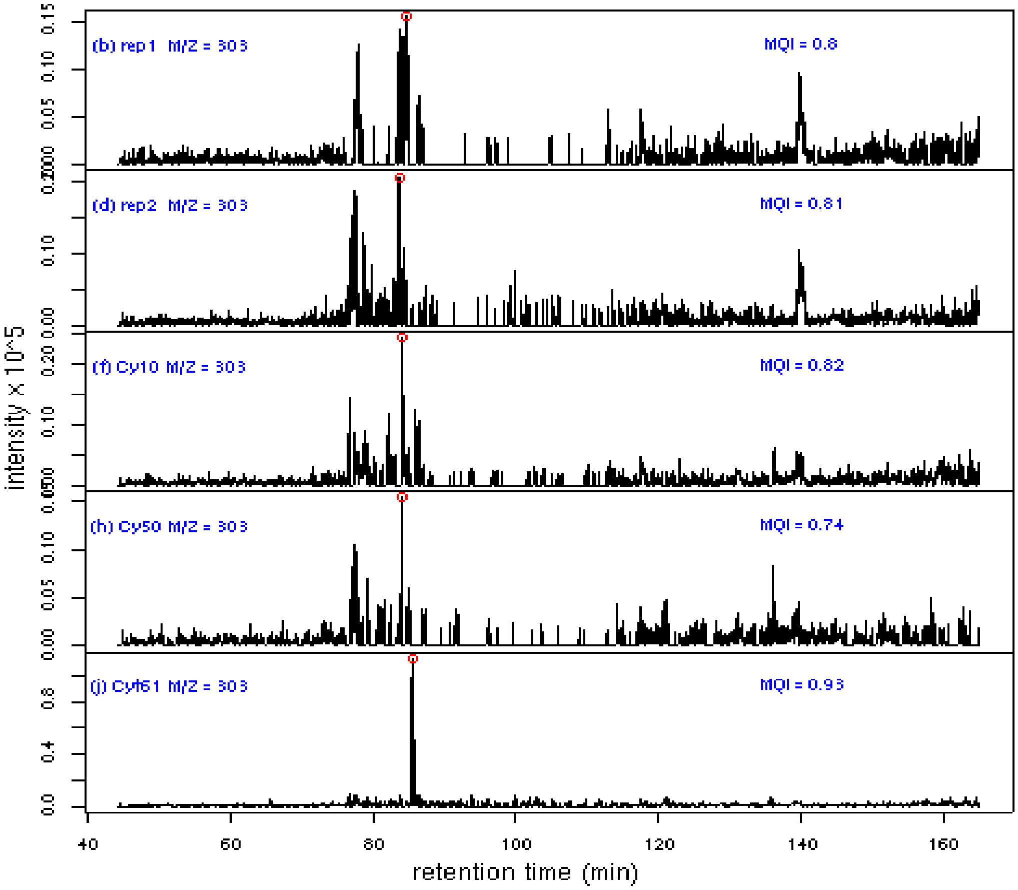

Figure 2: (b), (d), (f), (h) and (j) profile plots of signals from m/z = 303 (corresponding to the plots of raw chromatograms shown in Fig. 2: (a), (c), (e), (g) and (i), respectively) after they are preprocessed using repeated runs of single outlier detection algorithm. The red dots are the compound-related peaks that may be used to compare the samples with spikein and the ones without. Note the time shifts in the location of the corresponding peaks from different runs or conditions (e.g. peaks marked by red circles). The MQIs are the r-squared values from various regressions.

preference by Windig, *et al.* (1996) and Windig, *et al.* (2001). The polynomial smoothers, e.g. the Savitzky-Golay method can be used, but the daunting task is the choice of appropriate window-sizes. Thus in this report the moving average is our method of choice but we reinforce it with the regression diagnostic tools such as residuals and use the MTPs to set the rejection regions for detecting and removing the high frequency noise and spikes that eludes the moving average method. An advantage of our method is that window-size choices that are central to many methods that are solely dependent on smoothing criterion for noise reduction are of modest concern here.

3.1.2 The regression model and hypotheses of interest

Our problem is two-fold: For each chromatogram, we want to detect those signals for which we have high confidence that they are truly compound-related using the MTP. Secondly, for each run of LC-MS, we want to choose those chromatograms for which we have high confidence that they are truly compound-related, i.e. they are not dominated by noise. MTPs consist of a number of steps including, choosing appropriate parameters of interest (e.g. the mean difference in intensities between the raw chromatogram and its smoothed version), specifying the null hypothesis that relates this parameter to the question of interest (e.g. mean difference = 0), specifying the test statistic for which the null distribution is known (e.g. two-sample t-statistic), performing a test on each signal over the entire retention time or performing a test on each chromatogram over the m/z range, choosing an appropriate experimentwise error rate to control (e.g. the number of false positives or Type I errors) and choosing a method to control this rate (e.g. Bonferroni) (Birkner, *et al*., 2006). In this report all these requirements will be derived from the regression of the raw chromatogram on its smoothed version as derived below. This suggests that the candidate test statistics for the MTP are the residuals (or some functions of them) or the estimates of the regression parameters as derived below. But first, we set up the null hypotheses to be tested.

Let c_i^* be the denoised version of the raw chromatogram c_i of the ith m/z channel. Then a measure of noise is defined by the "residuals", $e_i = c_i - c_i^*$. In general e_i may be measured by distances of the type

$$e_{ij} = c_{ij} - \mu_{ij}(\beta), \qquad (2)$$

where $\mu_{ij}(\beta) = h(c_{ij}^*, \beta)$ is a function describing suitable relationship between the raw data and its smoothed version. Thus the model for evaluating the noise level of a chromatogram is

$$c_{ij} = \mu_{ij}(\beta) + e_{ij}, \qquad (3)$$

so that if it is assumed that $E[e_{ij}] = 0$, $0 < Var(e_{ij}) < \infty$ and $Cov(e_{ij}, e_{ik}) = 0$ for $j \neq k$, then c_{ij} are independent with $E[c_{ij}] = \mu_{ij}$. To model this relationship, we adopt the GLM approach (McCullagh & Nelder, 1989) in which it is assumed that $g(\mu_{ij}(\beta)) = \eta(c_{ij}^*, \beta)$, where $\beta = (\beta_{i0}, \beta_{i1}) \in \mathbb{R}^2$ is a 2-dimensional vector of unknown regression parameters that must be estimated, $\eta(c_{ij}^*, \beta) = \beta_{i0} + \beta_{i1} c_{ij}^*$ is a linear predictor and $g(.)$ is a link function describing the relationship between c_{ij}^* and μ_{ij}. Suitable link functions for the current problem are derived in Section 6.

Determining if a chromatogram of m/z channel i is compound-related involves simultaneous tests of hypotheses $H_{i0} : \beta_{i1} = 0$, $i = 1, 2, \ldots , n$, about the regression parameters, while detecting whether the jth ion current c_{ij} of this m/z channel is compound-related is a multiple testing problem whose test statistic is the jth component of that for testing the hypothesis H_{0i}, which is the sum of the test statistics for individual ion currents over the entire retention time. More specific and suitable form of the test statistics for screening individual ion currents are dictated by their characteristics, for example if they are outliers, then the test statistics are functions of the corresponding residuals defined by

$$e_{ij} = c_{ij} - \mu_{ij}(\hat{\beta}), \qquad (4)$$

where $\hat{\beta}$ are estimates of the regression parameters that are necessarily efficient, e.g. the maximum likelihood (ML) estimates. The parameter $\hat{\beta}_{i1}$ measures the association between the raw chromatogram and its smoothed counterpart. Two approaches for obtaining such estimates are the likelihood and the quasi-likelihood approaches.

3.2 The parameter estimation

In the likelihood approach, it is assumed that the intensities (ion currents) of a chromatogram, c_{ij}, are independently and identically distributed with density functions belonging to the exponential family of the form

$$f(c_{ij}; \theta, \phi) = exp\left\{ \frac{c_{ij}\theta_j - b(\theta_j)}{\phi} + h(c_{ij}, \phi) \right\},$$

where as in McCullagh and Nelder (1989), θ_j is a canonical parameter and ϕ is a dispersion parameter and $b(.)$, $h(.)$ are known functions satisfying, $E[c_{ij}] = \mu_{ij} = b'(\theta_j)$ and variance $Var(c_{ij}) = v_{ij} = b''(\theta_j)\phi = V(\mu_{ij})\phi$, and $V(\mu_{ij})$ is some function of μ_{ij}.

Let $\hat{\beta}$ be the ML estimate of the parameter $\beta = (\beta_{i0}, \beta_{i1})^T$. Now suppose that $\{P\}(\beta)$, represents a P computed at β. Then the elements of the vector $\hat{\beta}$ may be obtained through the iterative procedure

$$\beta_{i0}^{(r+1)} = \{\lambda_{i0} / \delta_i\}\left(\beta_{i0}^{(r)}\right) \text{ and } \beta_{i1}^{(r+1)} = \{\lambda_{i1} / \delta_i\}\left(\beta_{i1}^{(r)}\right), \qquad (5)$$

where $\lambda_{i0} = (\alpha_{i1}\alpha_{i2} - \alpha_{i3}\alpha_{i4})$, $\lambda_{i1} = (\alpha_{i4}\alpha_{i5} - \alpha_{i2}\alpha_{i3})$, $\delta_i = \left(\alpha_{i1}\alpha_{i5} - \{\alpha_{i3}\}^2\right)$,

$$\alpha_{i1} = \sum_{j=1}^{m} c_{ij}^{*2}\gamma_{ij}, \ \alpha_{i2} = \sum_{j=1}^{m} z_j\gamma_{ij}, \ \alpha_{i3} = \sum_{j=1}^{m} c_{ij}^{*}\gamma_{ij}, \ \alpha_{i4} = \sum_{j=1}^{m} c_{ij}^{*} z_j \gamma_{ij}, \ \alpha_{i5} = \sum_{j=1}^{m} \gamma_{ij}^{(r)},$$

$\gamma_{ij} = \dfrac{f_{ij}}{e_{ij}} \dfrac{\partial \mu_{ij}}{\partial \eta_{ij}}$, $f_{ij} = \dfrac{e_{ij}}{v_{ij}} \dfrac{\partial \mu_{ij}}{\partial \eta_{ij}}$, and

$$z_{ij} = \beta_{i0} + \beta_{i1} c^{*}_{ij} + e_{ij} \frac{\partial \mu_{ij}}{\partial \eta_{ij}} \tag{5.1}$$

is the so-called working variable in GLM terminology. This algorithm is faster and hence more efficient than the traditional GLM algorithms that invert the matrix $D = X_i^T W_i X_i$ at each iteration, where X_i is an $n \times 2$ matrix whose first column is *1*, a vector of 1's and whose second column is c_i^* and W_i is a diagonal matrix with elements γ_{ij}. The elements of $\hat{\beta}$ are then $\hat{\beta}_{i0} = lim_{r\to\infty} \beta_{i0}^{(r)}$ and $\hat{\beta}_{i1} = lim_{r\to\infty} \beta_{i1}^{(r)}$, respectively.

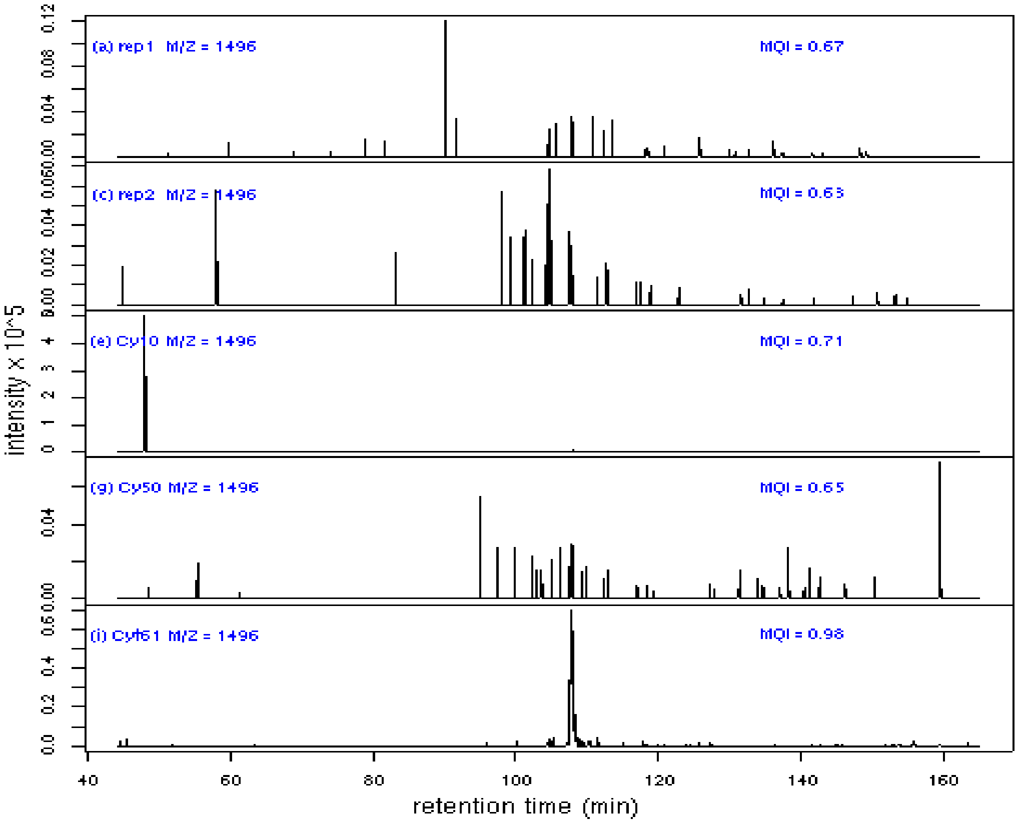

Figure 3. Profile plots of signals from a chromatogram with (m/z = 1496). (a) and (c) are plots of replicates of raw chromatogram from serum of a cervical cancer patient. (e), (g) and (i) are replicates of same serum with different levels of spikeins of cytoc (10, 50 & 61 pmol). In this example, the cytoc peaks are expected to appear at around 108 minutes. Their corresponding versions after pre-processing using the tools proposed in this manuscript are displayed in Figures 2 (b), (d), (f), (h) and (k), respectively.

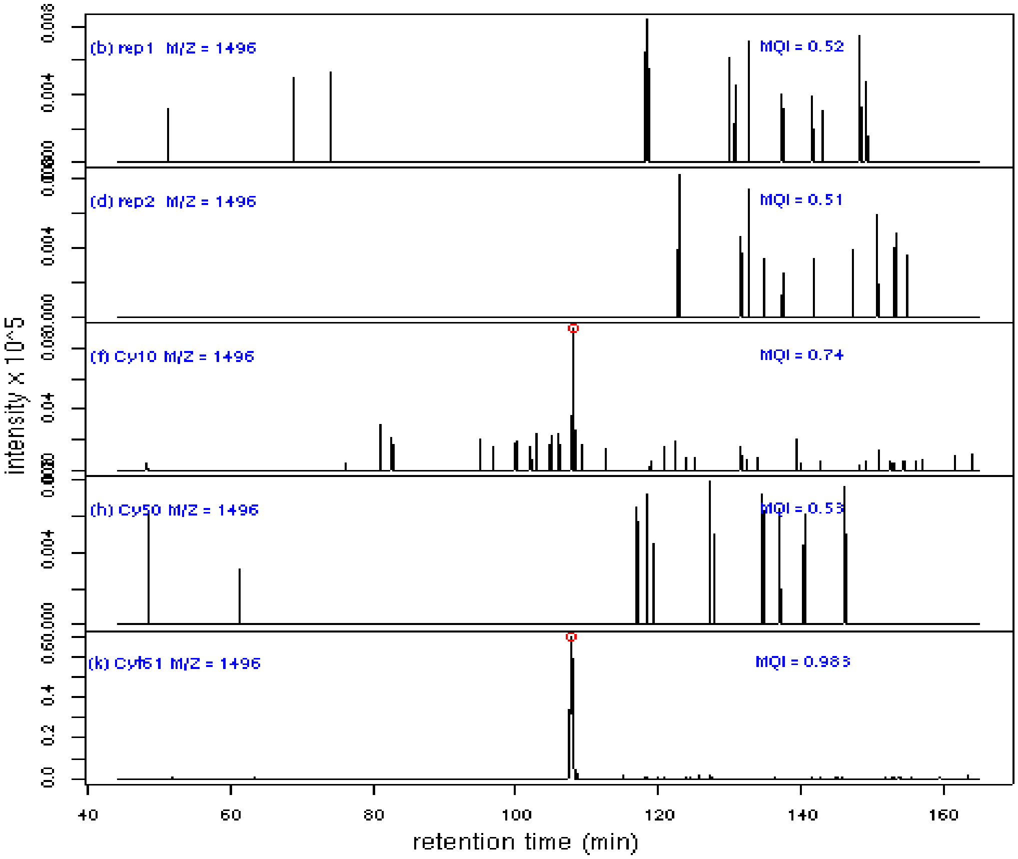

Figure 3. (b), (d), (f), (h) and (k) are the corresponding pre-processed versions of chromatograms displayed in Figures 3 (a), (c), (e), (g) and (i), respectively. The red dots are the cytoc-related peaks detected after screening out noise using MTP. The MQIs are the r-squared values from various regressions.

In the absence of sufficient information to construct the likelihood for the intensities, the theory of quasi-likelihood (McCullagh & Nelder, 1989, p. 323) enables us to draw inference about the associations between the raw chromatogram and its smoothed version. The Quasi-score equations are equivalent to the normal equations in the weighted least squares approach, in which the weights are γ_{ij} (Dobson, 1990; p. 13).

4 Test statistics for assessing compound contents of signals and chromatograms

It is well known that not all the chromatograms obtained from a single run of LC-MS are relevant for the purpose of biomarkers discovery (see e.g. Fig. 1 (b), Fig.

3 (b), (d) and (h)) and consequently, significant attention has been devoted to the development of the methods for selecting the relevant mass traces (also known as the high quality chromatograms) (Windig, *et al*., 1996; Windig, *et al*., 2001 and Gaspari, *et al*. 2001). This process involves vetting thousands of chromatograms for compound-related information and the issue of multiplicity arises. We want to use the MTPs to discriminate compound-related chromatograms from noise. A first step in such procedures is the determination of appropriate test statistics for the individual hypotheses, followed by the choice of the false discoveries or type I error control method that combines them into a simultaneous tests procedure (see e.g. Lehmann and Romano, 2005).

A compound-related chromatogram has low noise levels and is not significantly altered when smoothed. Thus these m/z traces may be detected by evaluating the significance of the regression parameter β_{i1} introduced in the model defined in Equation (3). In general, tests of the hypotheses $H_{i0} : \beta_{i1} = 0$ provide information regarding the quality (noise level) of a mass chromatogram, with a rejection, $\beta_{i1} \neq 0$, indicating the chromatogram of the m/z channel i contain relevant compounds. If there were only a small number of tests to be done, then standard techniques from McCullagh and Nelder (1989) that involves testing each hypothesis individually may be used to detect chromatograms that deviate from noise. However, as there are thousands of m/z traces to be tested, the chance of one or more false diagnosis is large and must be controlled. However, these procedures utilize test statistics for individual hypotheses, that must be derived to pave way for simultaneous testing of H_{i0}, $i=1, 2, ..., n$. Testing each hypothesis H_{i0} is equivalent to comparing two GLMs with link functions $\eta(c^*_{ij}, \beta) = \beta_{i0}$ and $\eta(c^*_{ij}, \beta) = \beta_{i0} + \beta_{i1}c^*_{ij}$, respectively. Suppose that D_{i0} and D_{i1} are the respective deviances corresponding to these functions, i.e. for example, $D_{i0} = \sum_{j=1}^{m} d^2_{ij}$, where d^2_{ij} is the likelihood ratio test statistic for testing a null hypothesis with a corresponding link function $\eta(c^*_{ij}, \beta) = \beta_{ij}$, $j = 1, ... , m$, that is different for every ion current, against an alternative with $\eta(c^*_{ij}, \beta) = \beta_{i0}$ for all j. Then a test statistic for each hypothesis is the change in the deviance, $\Delta D_i = D_{i0} - D_{i1}$. ΔD_i, may be interpreted as a measure of the quality of a chromatogram with mass channel i. If it is assumed that the ion currents (signals) of this chromatogram are independent, then the null distribution ΔD_i is χ^2_1. The values of ΔD_i greater than the upper tail $100 \times \alpha\%$ point of the χ^2_1 distribution would lead to a rejection of H_{i0} and the conclusion that the chromatogram of the m/z channel *i* contain relevant compounds.

Let $h_{ir} = \gamma_{ir}\left(\alpha_{i1} - 2\alpha_{i3}c^{*}_{ir} + c^{*2}_{ir}\alpha_{i5}\right)/\delta_{i}$, then asymptotically, $Var(\hat{e}_{ij}) = \hat{\vartheta}_{ij} = \hat{v}_{ij}(1-\hat{h}_{ij})$, where the hats indicate evaluation of the variables at $\hat{\beta}$. Then h_{ir} is the rth diagonal element of the matrix $H_{i} = W_{i}^{1/2}X_{i}(X_{i}^{T}WX_{i})^{-1}X_{i}^{T}W_{i}^{1/2}$, where $\sum_{r} h_{ir} = trace\,H = 2$. Then the noise level associated with the signal at the *j*th time point of the mass channel r may be evaluated by the standardized Pearson residuals (see e.g. Williams, 1987),

$$\hat{r}_{ij} = \hat{e}_{ij} / \left\{\hat{\vartheta}_{ij}\right\}^{1/2} \qquad (6)$$

which is a scaled version of (2). This statistic will be large for the noisy or spiky chromatograms as demonstrated shortly. Define $h^{'}_{ir} = mh_{ir}$, and then following Hoaglin and Welsch (1978) the rth ion current will be a point of high-leverage if $h^{'}_{ir} > 4$. For LC-MS data, points of high leverage are likely to be spikes.

The studentized residuals may be combined into an overall goodness-of-fit statistic for testing H_{0} resulting into the Pearson statistic defined by

$$X^{2} = \sum_{j=1}^{m} \varphi(1-\hat{h}_{ij})\hat{r}_{ij}^{2}\ , \qquad (7)$$

where in our case, asymptotically, $X^{2} \sim \varphi\chi^{2}_{m-2}$, m being the number of retention times over which the ion currents are sampled. For normal distribution, this is just the residual sum of squares, which is also equivalent to the deviance for this model (McCullagh & Nelder, 1989, p. 34). Obviously X^{2} is another candidate statistic for the MTPs.

Another way to test the hypotheses H_{i0} is to compare the saturated model with the proposed GLM using log-likelihood ratio statistic or the deviance. For every mass channel, a saturated model is a GLM in which each of the m ion currents has a distinct linear component $\eta_{j} = \beta_{i0} + \beta_{i1}c^{*}_{ij}$, say, so that this model has m parameters. Suppose that $f(c_{ij};\mu_{ij},\varphi)$ is the density function of the ion current at the *j*th retention time, c_{ij}, then the deviance is defined by $D_{i} = \sum_{j=1}^{m} d_{ij}^{2}$, where $d_{ij}^{2} = 2\varphi \log\left\{f(c_{ij};c_{ij},\varphi)/f(c_{ij};\hat{\mu}_{ij},\varphi)\right\}$. Large values of D_{i} suggest that the raw chromatogram is compound-related, thus D_{i} is yet another candidate test statistic in the MTP. The significance of the deviance may be evaluated by comparing D_i with the quantiles of χ^2_{m-2}. The scaled deviance residuals (Williams, 1987),

$$\hat{r}^D_{ij} = sign\{\hat{e}_{ij}\}d_{ij} / \sqrt{\varphi(1-\hat{h}_{ij})}, \tag{8}$$

may be used to evaluate the compound content of the ion current at the *j*th time point.

5 Characterizing signals of a chromatogram

Although the residuals such as those defined by Eq. (6) and (8) are in general useful in diagnosing the compound information in a signal, some specific noise types such as outliers require special statistics that involve the transformation of these residuals in order to efficiently be detected. Thus the question of which statistics to use in discriminating compound-related signals from noise will depend on the characteristics of the spurious signals in a chromatogram. Therefore to make informed decision of the appropriate test statistic for the MTPs, we need to carefully characterize the signals of a chromatogram. We need to answer questions such as: are there signals that wildly deviate from their smoothed versions e.g. outliers and high-frequency noise? If there are no deviations, are the signals compound-related or are they solvent-based noise?

For each chromatogram, we see two pools of observations:

(1) High intensities: exhibited by spikes, high frequency noise and compound-related signals.

(2) Low intensities: exhibited by the bulk of the observations that is expressed at background levels.

The fact that spikes, high frequency noise and compound-related signals have large (extreme) intensities relative to the bulk of the signals that are expressed at background levels hints that these signals could be categorized as either outliers or influential observations or simply some observations that may not fit the GLM well. The latter can be detected using residuals of the types defined by Eq. (6) and (8), while the former requires special transformations of these residuals to be detected, as will be seen later. We use scatter plots of the raw signals versus their smoothed versions as an aid in characterizing spikes and peaks.

Fig. 4 (a) is a scatterplot of the high quality chromatogram, with m/z = 1236, versus its smoothed version corrected by “mean subtraction” to segregate the solvent-based noise and 4 (b) is similar plot for the same chromatogram after pre-processing using the MTP. In this chromatogram, the only dominant feature is the compound-related peak. In both cases the raw chromatogram is strongly linearly related with its smoothed version. The tip of the compound-related peak is marked by a green dot that does not show significant deviation from the trend line suggested by the scatterplot. However, the signals of the peak have high intensities that are clustered on the upper part of the trend line, while the signals expressed at the background levels tend to cluster near the origin. This suggests that the inclusion of a compound-related signal can increase the precision of the estimates of the intercepts, β_{i1}, implying that they are definitely not outliers but are possibly influential observations (see Cook & Weisberg, 1982, for the definition of influential observations). Fig. 4 (c) is a scatterplot of the spike-contaminated chromatogram (Fig. 4 (ii)). The green dots are the signals of the spikes that clearly deviate from the trend line suggested by the scatterplot, implicating them to be outliers. Fig. 4 (d) is the scatterplot of the chromatogram displayed in Fig. 4 (iii), obtained by detecting and deleting spikes and high frequency noise from the chromatogram in Fig. 4 (ii) with the aid of MTPs. There is an improved fit between this chromatogram and its smoothed and perturbed counterpart (R^2 value of 0.80, that is now larger than FDR suggested threshold of 0.71, compared to the original value of 0.65). The green dots in Fig. 4 (d) correspond to the signals of the compound-related peaks seen in Fig. 4 (iii) that have been unmasked by deleting the spikes.

5.1 Test statistics for detecting noise, spikes and compound-related signals in a chromatogram

We have seen that spikes and high frequency noise may be characterized as outliers, while compound-related signals could be possibly classified as influential observations. This suggests that candidate test statistics for segregating spikes and high frequency noise from compound-related signals are those for detecting outliers in a GLM, while the tools for detecting influential observations may be used to detect the compound-related signals. The usual strategy is to monitor changes in some aspect of the fit of the model (e.g. the ML estimates) caused by deleting an observation (leave-one-out strategy), i.e. changes in these statistics by refitting the model when the case of interest is omitted. The extreme observations results in the largest changes. Thus the relevant test statistics for the MTP for discriminating noise from compound-related signals are those for detecting outliers. These test statistics are derived as follows.

Let $\hat{\beta}_{(ir)}$ denote the ML estimate of β when fitting the GLM model with the rth ion current excluded. Then $\hat{\beta}_{(ir)}$ may be obtained by fitting the mean shift outlier model (Williams, 1987)

$$\eta(c_{ij}^{*},\beta) = \beta_{i0} + \beta_{i1}c_{ij}^{*} + u_{ir}\lambda_r, \qquad (9)$$

where $u_{ir} = 1$, if $r = j$ and $u_{ir} = 0$, otherwise. Then the possibility that the rth ion current of the ith mass channel is outlier (spike) may be evaluated by testing the null hypothesis $H_{0r} : \lambda_r = 0$ against the alternative hypothesis, $H_{1r} : \lambda_r \neq 0$.

This requires computation of $\hat{\beta}_{(ir)}$, the full iterate estimates of β, which may be approximated by its one-step estimate, $\hat{\beta}^{1}_{(ir)}$.

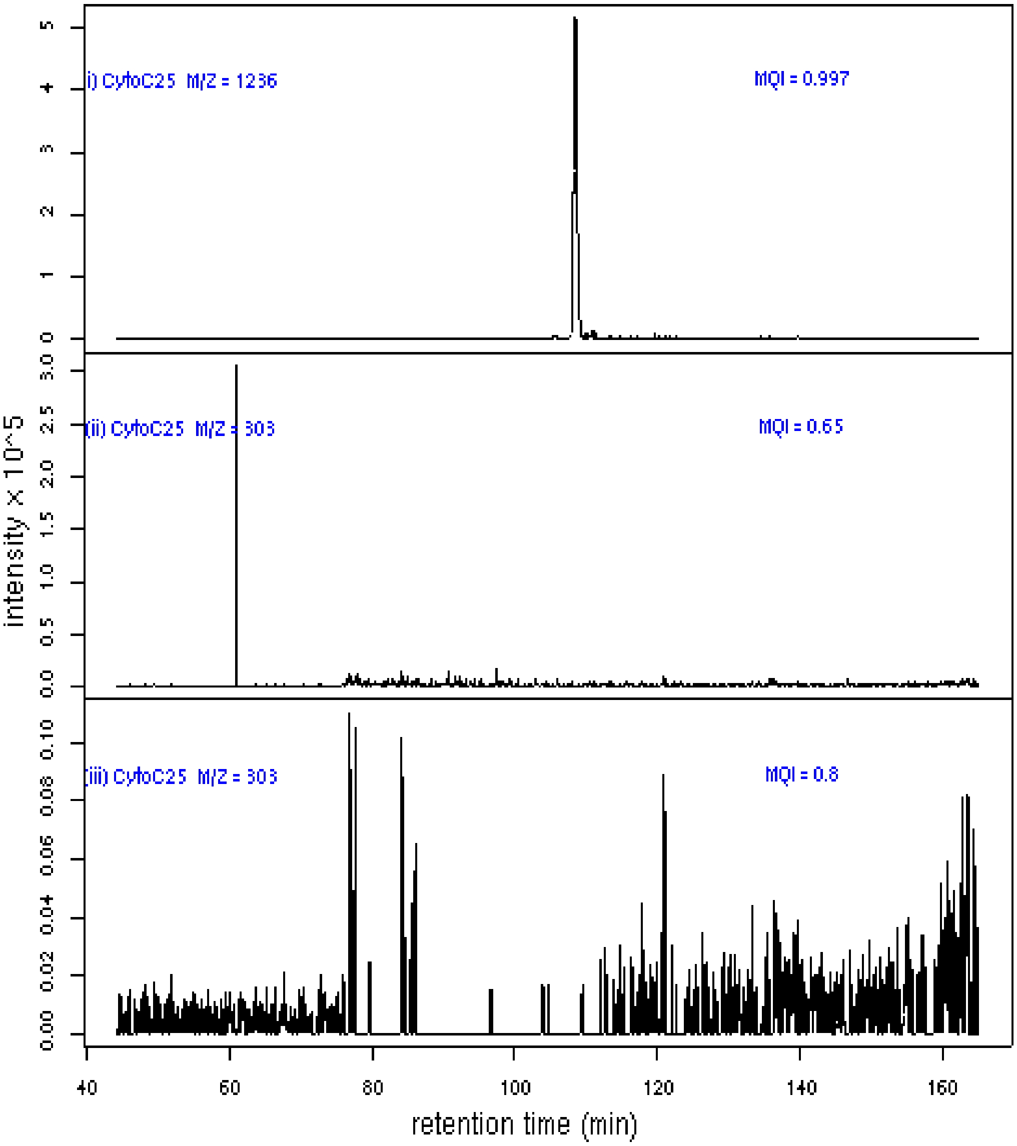

Figure 4. The profile plots of: (i) the high quality m/z trace, (ii) m/z trace characterized by a single dominant spike, (iii) m/z trace obtained from (ii) after screening with diagnostic tools for outlying and influential observations. The MQIs are the r-squared values from various regressions. These chromatograms are used to characterize the signals of chromatograms and to derive the null distributions of chromatograms that are then used to derive test statistics for MTPs. How they are used for these purposes are explained by Figures 4 (a), (b) and (c) below.

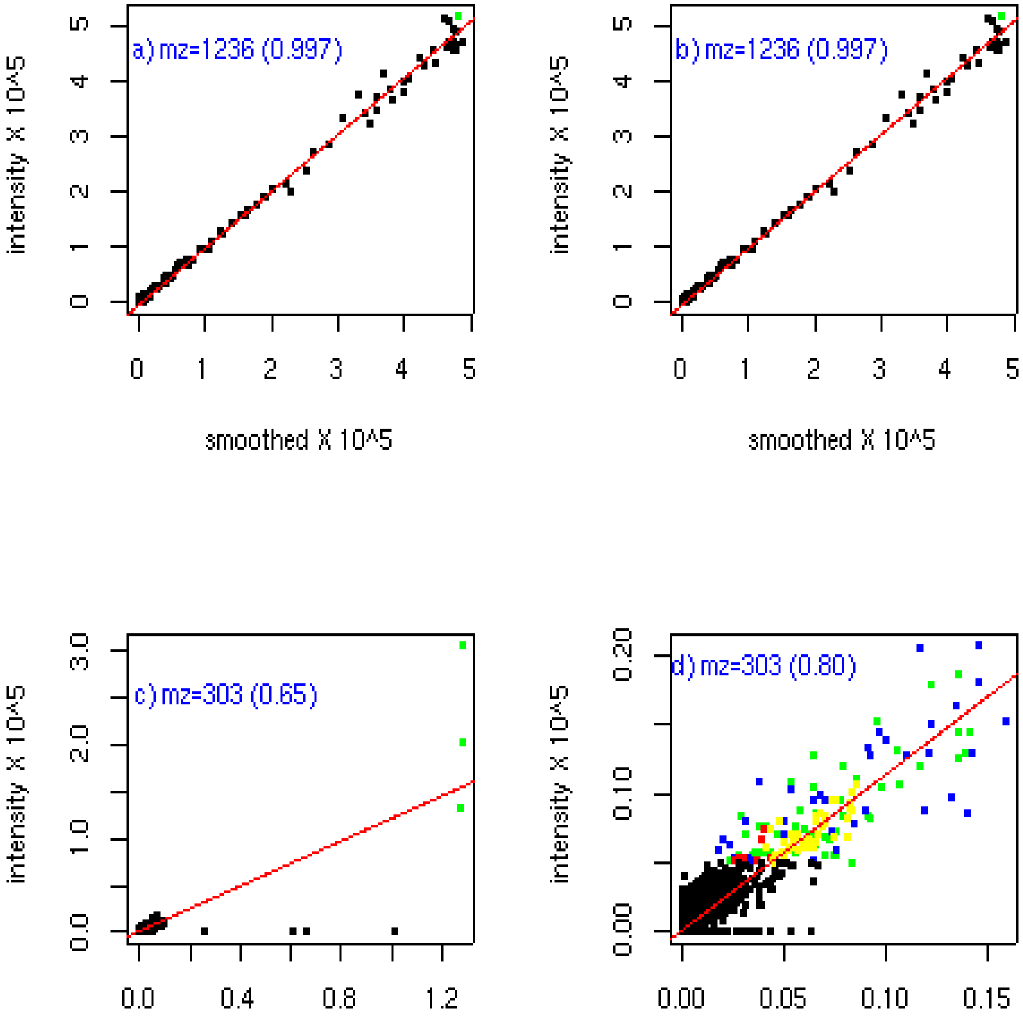

Figure 4. (a) scatterplot of the high quality m/z trace, Fig 4 (i), against its smoothed and perturbed version. The green dot is the maximum intensity of this mass trace. (b) A scatterplot of (i) after removing artefacts using tools for outliers detection. (c) scatterplot of the high quality m/z trace, Fig. 4 (ii), characterized by a single dominant spike, against its smoothed and perturbed version. (d) A scatter plot of m/z trace in Fig 4 (iii), obtained from m/z trace in Fig 4 (ii) after screening with diagnostic tools for outlying and influential observations. The blue, yellow, red and green dots are the intensities of peaks of m/z trace in Fig. 4 (iii). The bracketed numbers are the r-squared values from various regressions.

Lemma 1. *Let* $\tau_{ir} = \hat{\delta}_i^2 (1 - \hat{h}_{ir}) / \hat{\omega}_{ir}$, *then the elements of* $\hat{\beta}^1_{(ir)}$ *are*

$$\hat{\beta}^1_{i0(r)} = \hat{\beta}_{i0} - \left\{ \hat{r}_{ir} (c_{ir} \hat{\alpha}_{i1} - \hat{\alpha}_{i3}) / \sqrt{\hat{\tau}_{ir}} \right\} \text{ and } \hat{\beta}^1_{i1(r)} = \hat{\beta}_{i1} - \left\{ \hat{r}_{ir} (\hat{\alpha}_{i5} - c_{ir} \hat{\alpha}_{i3}) / \sqrt{\hat{\tau}_{ir}} \right\} \qquad (10)$$

5.1.1 Changes in deviance as test statistics to detect spikes

For the linear regression models, many of these expressions simplify, for example $\eta(c_{ij}^{*},\beta)=\mu_{ij}=\beta_{i0}+\beta_{i1}c_{ij}^{*}$, and consequently, $\omega_{ij}=1/Var(c_{ij})$. Using the one-step estimates it can be shown that the statistic $\Delta D_{i(r)}^{1}=(1\text{-}h_{ir})\left(\hat{r}_{ir}^{D}\right)^{2}+h_{ir}\hat{r}_{ir}^{2}$ may be used to rank the signals according to their impact on the estimation and hence may be used to detect the dominant features like spikes or the compound related peaks. Let $D_{i(r)}=\sum_{j\neq r}^{m}\tilde{d}_{ij}^{2}$ be the deviance for the GLM model with the linear predictor $\eta(c_{ij}^{*},\beta)=\beta_{i0}+\beta_{i1}c_{ij}^{*}$, with the rth ion current omitted and evaluated at $\hat{\beta}_{(ir)}^{1}$ instead of $\hat{\beta}$. Then $D_{i(r)}$ is the deviance of the GLM model $\eta(c_{ij}^{*},\beta)=\beta_{i0}+\beta_{i1}c_{ij}^{*}+u_{ir}\lambda_{r}$. It can be shown that $\Delta D_{i(r)}^{1}$, is the one-step approximation to the second-order Taylor series expansion of the change in deviance $\Delta D_{i(r)}=D_{i1}-D_{i(r)}$ (see e.g. Nyangoma, *et al.*, 2006), where D_{i1} is the deviance for the GLM model $\eta(c_{ij}^{*},\beta)=\beta_{i0}+\beta_{i1}c_{ij}^{*}$. $\Delta D_{i(r)}$ is the likelihood ratio statistic for testing H_{0r}. If it is assumed that the ion currents follow a normal distribution and since a chromatogram consists of a large number of ion currents, $\Delta D_{i(r)}^{1}$ is exactly distributed as χ_{1}^{2}. In this case, $\Delta D_{i(r)}^{1}$ is an appropriate test statistic for detecting spikes in a chromatogram using the MTPs.

5.1.2 Changes in parameter estimates as test statistics to detect compounds and Spikes

The changes in parameter estimates when a case is deleted may also be used to characterize the signals of a chromatogram. More importantly, large changes in intercepts of the linear fit of the raw chromatograms to their smoothed versions, i.e. large values of $\hat{\beta}_{0}-\hat{\beta}_{0(ir)}^{1}$, would imply that the *r*th case is an outlier. Deletion of such a case is also accompanied by large change in a goodness-of-fit statistic, e.g. the deviance. Thus $\hat{\beta}_{0}-\hat{\beta}_{0(ir)}^{1}$ is another candidate test statistic for detecting spikes in a MTP. However, the changes in slopes $\hat{\beta}_{1}-\hat{\beta}_{1(ir)}^{1}$ are usually small for such observations. McCullagh and Nelder (1989, p. 403) draw similar conclusions for normal distributions. This fact was established for our data, for example when outliers in Fig. 2 (g) are deleted we obtain results represented by Fig. 2 (h), in which, the deletions resulted in over two-fold changes in intercept, while there was little change in slopes. In addition, there was a substantial improvement in the goodness-of-fit statistic (increase in r-squared of about 8%).

The relationship between the signals of the compound-related peaks and its smoothed counterpart are consistent with the linear trend suggested by the main cluster of the data that are expressed at background (noise) levels. However, the noise and the compound-related signals are still distinct in these linear plots, for example the noise are clustered at the lower end of the line of best fit, while the compound-related are spread out towards the upper end. This suggests that the inclusion of a compound-related peak will improve the accuracy of the slope ($\hat{\beta}_1$) of the line of the best fit to the data. This means that ion currents whose omission results in large changes in slopes, i.e. large values of $\hat{\beta}_1 - \hat{\beta}^1_{1(ir)}$, would most likely be compound-related. This suggests that the change in slope $\hat{\beta}_1 - \hat{\beta}^1_{1(ir)}$ is a candidate test statistic for compound-related signal in a MTP.

5.1.3 Influence curves and Cook's distance as statistics for detecting compounds and spikes

The above ideas lead to consideration of many other statistics for detecting compound-related peaks. Key among them, are Cook's distance defined by $D_{ij} = h_{ij}\hat{r}^2_{ij}/2(1-h_{ij})$ (Cook, 1977; Cook and Weisberg, 1982; p. 117) and the sample influence curves.

Lemma 2. *Let $\hat{\pi}_{ij} = \sqrt{\hat{\omega}_{ij}\hat{\tau}_{ij}}$ and m be the number of retention times. Then the sample influence curve for the "intercept" is an $m \times 1$ vector with the j th element defined by*

$$SIC^I_{ij} = \left\{(m-1)\sqrt{\hat{\gamma}_{ij}}\left(\hat{\alpha}_{i1} - c_{ij}\hat{\alpha}_{i3}\right)\hat{r}_{ij}\right\}/\hat{\pi}_{ij}, \qquad (11)$$

while that for the "slope" is an $m \times 1$ vector with the j th element defined by

$$SIC^G_{ij} = \left\{(m-1)\sqrt{\hat{\gamma}_{ij}}\left(c_{ij}\hat{\alpha}_{i5}\text{-}\hat{\alpha}_{i3}\right)\hat{r}_{ij}\right\}/\hat{\pi}_{ij}. \qquad (12)$$

The distributions of these statistics are discussed in Cook and Weisberg (1982).

5.2 Test statistics for detecting solvent-based noise

We have so far suggested per-signal test statistics for detecting the high frequency noise, spikes and compound-related signals. However, we have not discussed how to detect or deal with artefacts such as the solvent-based noise that is also a common feature of the LC-MS data. A chromatogram that is dominated by the solvent-based noise has a signal (of roughly uniform intensity) over (almost) the entire retention time range and thus it has a larger mean signal value relative to (all) its signals than a high-quality or a spiky chromatogram. The same effect is incurred by other chromatograms whose signals are mainly random noise expressed at low levels (e.g. Fig. 4 (b)). Thus a regression of the signals of such a chromatogram on its smoothed and perturbed (by subtracting (adding) from (to) them a quantity that is some function of their mean or median) versions, results in large residuals for all its observations. Such a perturbation will have insignificant effect on the other types of chromatograms. This suggests that a measure of location of signals of a chromatogram may be used to characterize the type of noise dominating it.

The foregoing implies that the diagnostic tools from a regression of the raw signals of an m/z trace on its smoothed and perturbed version may be used to sensitively detect spikes, random and solvent-based noise in a single step. Thus in this paper we essentially compare the raw chromatograms, not to their smoothed versions (as so far implied), but to their smoothed and perturbed counterparts.

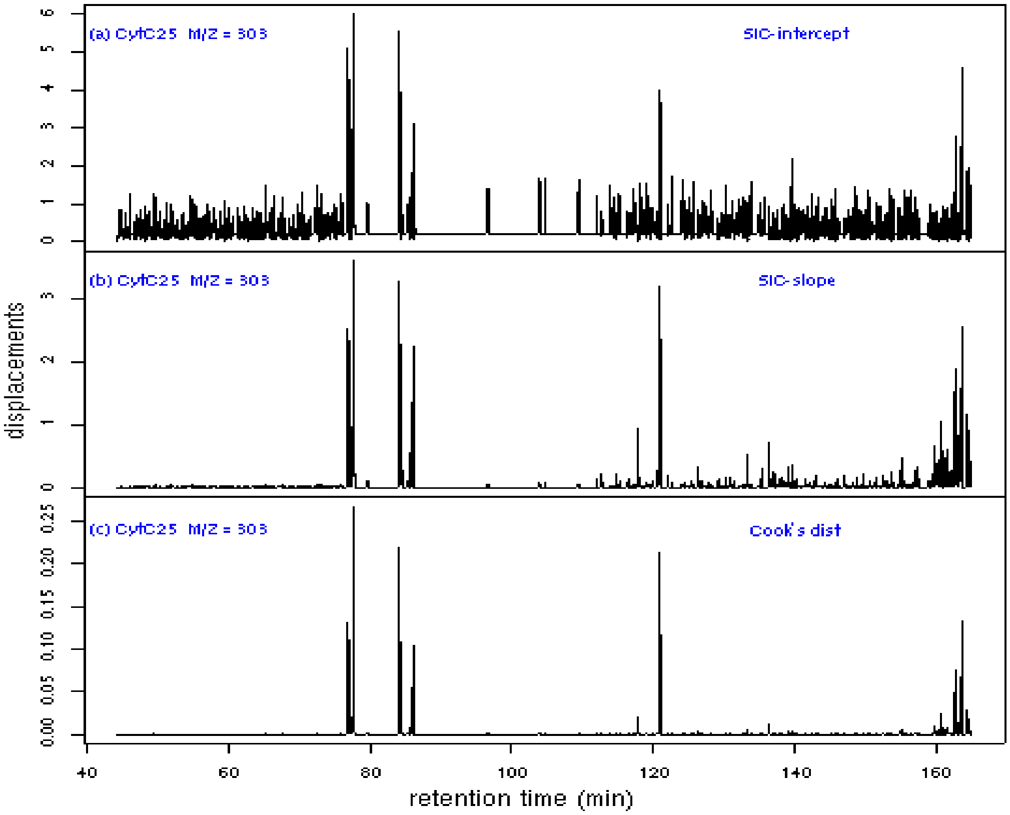

Figure 5. retention time index plots of influence statistics: (a) Sample influence curve (SIC) for change in intercept (b) SIC for change in slope (c) Cook's distance, for (m/z = 303, shown in Fig. 2 (iii)). These may be compared to plots for the high quality m/z = 1236 (Fig. 2 (i)) displayed in Fig. 5 (d), (e) and (f), below.

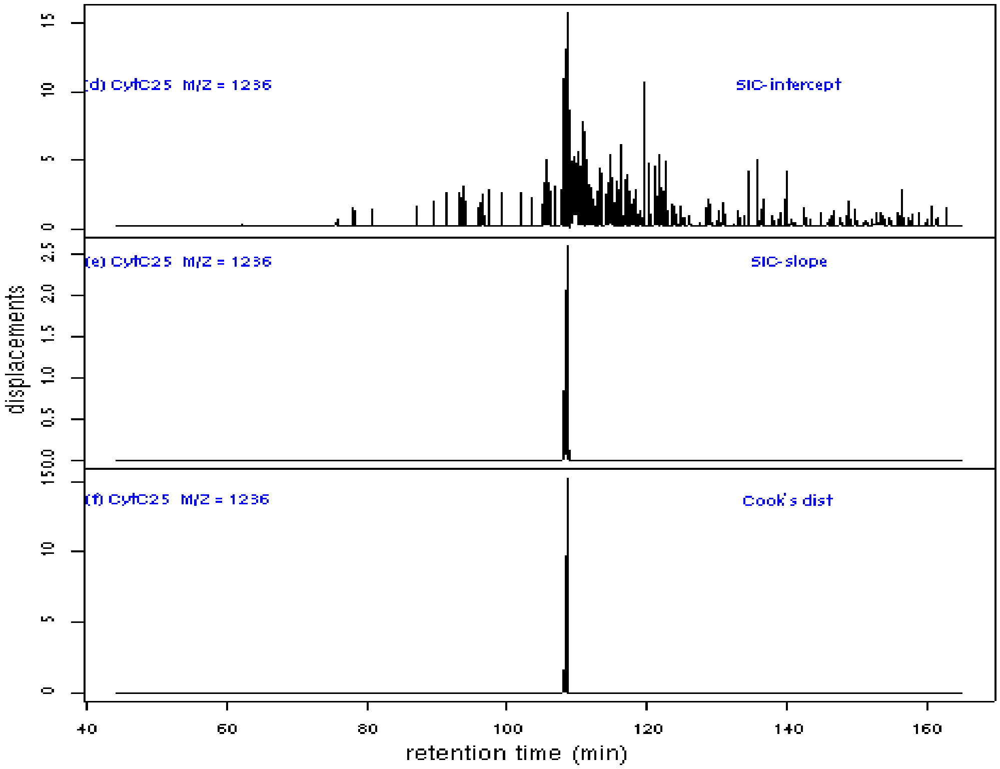

Figure 5. retention time index plots of influence statistics: (d) SIC for change in intercept (e) SIC for change in slope (f) Cook's distance, for the high quality m/z = 1236, shown in Fig. 2 (i)).

This makes sense because the mean-subtraction proposed here does not affect the statistics for detecting high frequency noise, spikes and compound-related signals and likewise, the moving average filtering does not affect the detection of solvent-based noise. However, it is important to point out that as opposed to the CODA method, direct mean or median subtraction is not useful in a regression setup. Thus we suggest modifications of this procedure that are relevant in the regression regime.

With respect to the size of the mean or median relative to the observed intensities, we argue that a solvent-based noise dominated chromatogram may be considered as an "outlier" in m-dimensional space. To distinguish these m/z traces by their solvent-based noise contents, we exploit the local characteristics of individual chromatograms (multivariate observations in m-dimensional space) e.g. their means and medians. We note that the current methods for detecting a single outlier in a multivariate sample (e.g. Wilk's, 1963, criterion) would measure the deviation of each chromatogram from an overall (global) mean over all the chromatograms. With the inherent iter-chromatogram variabilities in intensities, using such global measures to detect outlying chromatograms would mischaracterize the chromatograms through the borrowed incompatible information. Our approach thus extends the use of the usual regression diagnostics to identifying the peculiarities of points in m-dimensional spaces.

5.2.1 The perturbation

To be able to detect the elusive solvent-based noise in m/z channel i, we fit a GLM with the observed ion currents as independent variables but with the matrix of explanatory variables given by

$$X_i(\omega) = X_i + W_i A_i \qquad (13)$$

where X_i is an $m \times 2$ matrix, whose first column is a vector of 1's and its second column is a vector of the smoothed currents of m/z channel i, c^*_{ij}. A_i is a 2×2 diagonal matrix with diagonal elements M_i, a constant that may be proportional to the total ion current for the m/z channel i (e.g. mean ion current), W_i is an $m \times 2$ matrix of weights, whose element in the rth row and first column is $\omega_{r1} = 0$, while the element in the rth row and second column is $\omega_{r2} = \delta_r$, a constant reflecting the contribution of the rth signal to the total ion current, and $-1 \le \delta_r \le 1$. This transformation is equivalent to a simultaneous perturbation of the elements in the second column of X_i to

$$x_{ij} = c^*_{ij} + \omega_j, \qquad (14)$$

where $\omega_j = \delta_j M_i$ and $i = 1, ..., m$. A special case of this scheme in which $\delta_j = -1$ for all j and $M_i = \sum_j c_{ij} / m$, the mean ion current for the m/z channel i, leads to comparing the raw chromatogram with its smoothed and mean subtracted version as used in CODA (Windig, *et al.*, 1996; Windig, *et al.*, 2001). This scheme inflicts a constant penalty on all ion currents irrespective of their contribution to the total ion current and is thus rather conservative. Equation (14) represents the simultaneous perturbation of the explanatory variables, a technique popularly used to detect outlying or influential observations in linear regression (Cook, 1986). The value $\omega_j = 0$ is interpreted as the null perturbation. This method may be used to characterize *m*-dimensional observations in addition to evaluating the influence of individual cases, which makes it different from that of Williams (1987), who studied the effect of deleting individual cases from data that can be modelled as GLM.

To detect solvent-based noise using GLM, we found it suitable to use a scheme in which δ_j is defined by

$$\delta_j = \begin{cases} -1, & c_{ij} > M_i \\ -\frac{c_{ij}}{M_i}, & otherwise \end{cases} \tag{15}$$

where choices of M_i may include the mean, median or mean of the lowest and highest 10% of the ion currents for the m/z channel i. This is equivalent to perturbing a smoothed chromatogram using a mixture scheme involving mean (or median) subtraction from intensities that are greater than the mean and setting the rest of the differences to zero. Our proposed scheme allows a more meaningful penalty (that is somewhat proportional to the contribution to the total ion current) to be investigated.

Because the proposed perturbation scheme has insignificant effect on the detection of other features of a chromatogram other than the solvent-based noise, one could use the diagnostics from the fit of a GLM regression of the raw m/z trace on its smoothed and perturbed counterpart defined by

$$c_{ij} = \mu_{ij}(\omega) + e_{ij}(\omega)\,. \tag{16}$$

where it is assumed that $g(\mu_{ij}(\omega)) = \eta_{ij}(\omega) = \beta_{i0} + \beta_{i1} x_{ij}$ to simultaneously detect all types of artefacts, including spikes, high frequency noise, solvent-based noise and even the compound-related signals. A chromatogram with high level of solvent-based noise will have a large M_i value and thus will be heavily penalized by perturbation leading to poor fit of model (16) and the converse is true of a high quality m/z trace. The change in parameter estimate when the smoothed chromatogram is perturbed may be used to detect the solvent-based noise.

Theorem 1.

Suppose that W_ω *is an* $m \times m$ *diagonal matrix with elements* $w_i(\omega) = v_i^{-1}(\omega)\left(\partial\mu_{ij}(\omega)/\partial\eta_{ij}(\omega)\right)^2$ *and suppose* a_i *is an* $m \times 1$ *vector with jth element* $a_{ij} = \omega_j \beta_{i1} + e_{ij}(\omega)\frac{\partial\mu_{ij}(\omega)}{\partial\eta_{ij}(\omega)} - e_{ij}\frac{\partial\mu_{ij}}{\partial\eta_{ij}}$ *and let* $\hat{b}_i = \left(I - X_\omega (X_i^T \hat{W}_i X_i)^{-1} X_i^T \hat{W}_i\right)\hat{z}_i$ *be an* $m \times 1$ *vector, where* X_i *is an* $m \times 2$ *matrix, whose first column is a vector of 1's and whose second column is a vector of the smoothed currents of m/z channel* i*,* c_{ij}^**,* X_ω *is its perturbed counterpart with perturbed ion currents,* x_{ij} *'s, in its second column and* $\hat{z}_i$ *is a vector of working variables (defined in Eq. (5.1)) and* W_i *is a matrix of weights, associated with the unperturbed GLM model. Let* $\hat{\alpha}_\omega$ *and* $\hat{\theta}_\omega$ *be respectively, the parameter estimates obtained by regressing* a_i *and* b_i *on the columns of* X_ω *with respect to the weights* W_ω*. Then if* $\hat{\beta}_\omega^1$ *is the one-step parameter estimate of* $\tilde{\beta}_\omega$*, the ML estimate under the perturbed model, then the change in parameter estimates due to the perturbation is approximately* $\tilde{\beta}_\omega^1 - \hat{\beta} = \hat{\alpha}_\omega + \hat{\theta}_\omega$*.*

Proof. The proof of this theorem follows by noting that the working variable under the perturbed model may be expressed as $z_{ij}(\omega) = z_{ij} + a_{ij}$ and that $\hat{\beta} = \left(X_i^T \hat{W} X_i\right)^{-1} X_i^T \hat{W}_i \hat{z}_i$. □

It may be seen that if ω is an $m \times 1$ vector with rth element $u_{ir}\lambda_r$ and $u_{ir} = 1$ if $r = j$, and zero otherwise, where λ_r is a parameter not depending on β, then $\hat{\beta}_{\omega}^{1} - \hat{\beta}$ is the statistic for investigating the effect of deleting an observation.

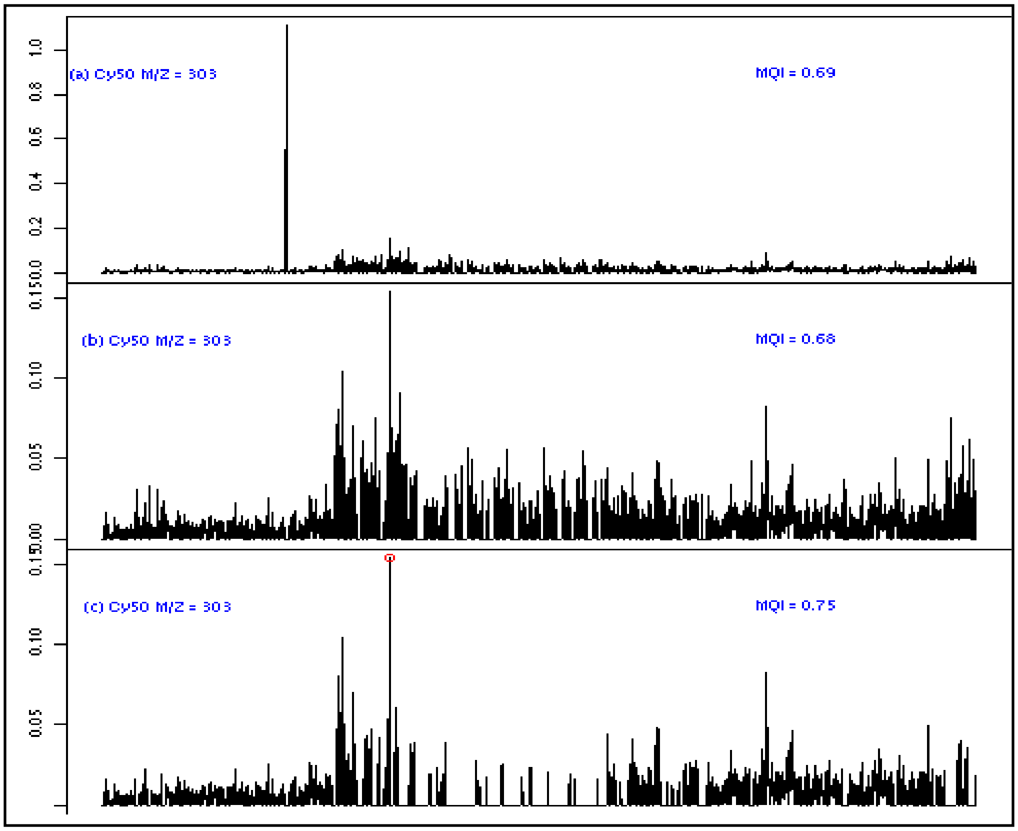

Figure 6. Profile plots of m/z = 303 (a) with dominant spike that masks detection of compound-related peaks (b) shows same mass trace after deleting dominant spike (c) shows chromatogram in (b) after screening high levels of noise and spikes.

6 Choosing a link function and the test statistics

So far we have proposed in a rather general way how to construct the test statistics for the MTPs for our two-dimensional signal profiling. The form of the relevant statistics is actually dictated by the nature of the link function for the

GLM. In this section, we discuss the possible forms of the link function for the problem of discriminating noise from compound-related signals and chromatograms.

6.1 The link functions and related diagnostic tools

We use the method of McCullagh & Nelder (1989; p. 401), who suggest an informal method for defining the link function that involves examining the plot of the estimated adjusted dependent variable $\hat{z}_i$ against the estimated linear predictor, $\hat{\eta}_i$. If we assume a normal model for the LC-MS, then $\hat{z}_i = c_i$, the vector of raw chromatograms, while $\hat{\eta}_i = \hat{\mu}_i$, is the predicted value from the regression of c_i on the column space of its smoothed and perturbed version. In this case, the informal check translates to a scatter plot of c_i versus $\hat{\mu}_i$. Since we have only a single independent variable, x_i, the informal check for linearity is the scatter plot of c_i versus x_i. Figures 4 (a), (b), (c) and (d) are scatterplots of the raw chromatograms against their smoothed and perturbed versions. We see that the chromatograms that contain compound-related peaks (e.g. 4 (a) and (d)) are highly linearly correlated (R^2 values of 0.997 and 0.80) with their smoothed and perturbed versions. On the other hand, chromatograms with high levels of noise or other artefacts (e.g. 4 (c)) have poor linear correlations (e.g. R^2 = 0.65) with their perturbed and smoothed versions. This suggests that the appropriate link function for the GLM that models the relationship between the compound-related raw chromatogram and its smoothed counterpart may be the identity link, which is the null model in this case. Departures from this model describe various forms of noise.

The departure from linearity observed in Fig. 4 (c) is due to the outlying artefacts (spikes) shown by Fig. 4 (ii). The appropriate model is then the simple linear regression model

$$c_{ij} = \mu_{ij}(\theta) + e_{ij} \qquad (17)$$

where $\mu_{ij} = \beta_{i0} + \beta_{i1} x_{ij}$. Thus it is sensible to assume that the errors are identically and independently distributed as normal, i.e. $e_{ij} \sim N(0, \sigma^2)$.

The scatter plots of c_i against x_i do not show a pronounced increase in variance, so there is a remote chance of considering gamma errors or the inverse link (Crawley, 2003; p. 662). A formal method for defining the link functions involves adding $\hat{\eta}_i^2$, an extra covariate and assessing the fall in deviance (Hinkley, 1985).

6.1.1 Test statistics for detecting noise and compounds

The use of identity link function leads to consideration of several statistics for detecting compound-related signals and artefacts. Key among them, are Cook's distance defined by $D_{ij} = h_{ij}\varepsilon_{ij}^{2} / 2(1-h_{ij})$ (Cook, 1977, 1982; p. 117) and the sample influence curves $SIC_{ij} = (m-1)(X_i^T\ X_i)^{-1} x_j \varepsilon_{ij} / \sqrt{(1-h_{ij})}$ (Cook & Weisberg, 1982; p. 110), where we now define X_i as a $m \times 2$ matrix, with the first column having elements 1 and x_{ij}'s in the second column, h_{ij} is the jth diagonal element of the hat matrix $H_i = X_i(X_i^T X_i)^{-1} X_i^T$ and ε_{ir} is the studentized residual from the regression defined by (17). The elements of the sample influence curve for the intercept of the simple linear regression model may be used to assess if the jth ion current of the mass channel i is a spike (an outlier), while that of slope can be used to assess whether jth ion current is compound-related.

Spike or random noise can be detected by single outlier statistic

$$t_r = \varepsilon_{ir}\left(\frac{m-3}{m-2-\varepsilon_{ir}^{2}}\right)^{1/2} \qquad (18)$$

(see Cook & Weisberg, 1982, p. 20), where ε_{ir} is the studentized residual. t_r follows a student *t*-distribution with $m-2$ degrees of freedom.

6.1.2 Test statistics for detecting compound-related chromatograms

Note also that under the identity link, $Var(c_{ij}) = v_{ij} = \sigma^2$, introducing a nuisance parameter in estimation and hence the deviance is not fully determined. In this case, the test statistic for testing the null hypothesis $H_{0i} : \beta_{1i} = 0$ is the ratio of the deviances $F_i = \Delta D_i / D_{i1}^{s}$, where $D_{i1}^{s} = D_{i1} / (m-2)$ is the unbiased estimate for $\varphi = \sigma^2$. Large values of F_i, typically greater than $F_{(1,m\text{-}2)}(\alpha)$, the $100 \times \alpha\%$ point of the F distribution with 1 and $m-2$ degrees of freedom, indicate that the chromatogram of m/z channel i is compound-related. Here $S_T = \sigma^2 D_{i0}$ and $S_R = \sigma^2 D_{i1}$ are the total and the residual sums of squares, respectively and $\rho = S_R / S_T$ is the proportion of the total variation not explained by the linear model (LM). Note that S_T and S_R may be interpreted as the squared Euclidean lengths of the vector of residuals from the GLMs under H_{i0} and H_{i1}, respectively. In particular S_T is just the sum of the squared deviations of the raw observations from their mean. Then another candidate test statistic is $R^2 = 1 - \rho = \Delta D_i / D_{i0}$, the proportion of the total variation explained. It is interpreted as the square of the correlation between the raw chromatogram and its smoothed version. For noisy chromatograms the residuals are large leading to small values of R_i^2. Note that this statistic for testing H_{i0} may be used to measure of the quality of the chromatogram and we call the mass chromatogram quality index (MQI).

6.2 The software

All test statistics derived in this paper are easily implemented in the **R** statistical package (Ihaka & Gentleman, 1996); in fact they are by-products of the *lm* object. The codes are available from the corresponding author on request.

7 Multiple Testing Procedures and thresholds for noise detection

In the previous section, we derived several test statistics for profiling the individual signals of a chromatogram as well as the individual chromatograms for compound-related information. Because each of these issues involves assessing the significance (deviation from noise) of multiple observations, we need a method to combine these individual tests into a simultaneous test procedure that take into account the multiplicity nature of these problems. Disregarding such characteristics increases the probability of one or more false rejections (Type I errors) because of the large numbers of tests to be done in each case. In this section we assess the MTPs that may provide efficient rejection regions for each of the test statistics for individual null hypotheses that control Type I error rates. Type I error is committed by rejecting a true null hypothesis.

There are several Type I error rates: 1) the family-wise error rate (FWER), which controls the probability of rejecting more than one false positive (Holm,1979; Lehmann & Romano, 2005) ; 2) tail probability of the proportion of false positives, which controls the proportion of false positives to total rejections (TPPFP) (van der Laan, *et al.*, 2004); 3) false discovery proportion (FDP) (Lehmann & Romano, 2005), which controls the proportion of false positives to total rejections; 4) false discovery rate (FDR) (Benjamini & Hochberg, 1995), which controls the $E[FDP]$. These MTPs are routinely applied to control the number of false positive results in the microarray studies. However, FWER has been criticised for being too conservative, especially for many biological applications. Despite this criticism, we found the FWER methods, especially the step-down procedures, quite efficient in detecting dominant chromatographic features such as spikes and high frequency noise. Of special interest are the more general $k-$FWER, the probability of rejecting at least k true null hypotheses. By using such error rates with $k > 1$, one is willing to tolerate one or more false rejections, provided the number of false rejections is controlled. The control of the $k-$FWER requires that $k - FWER \leq \alpha$.

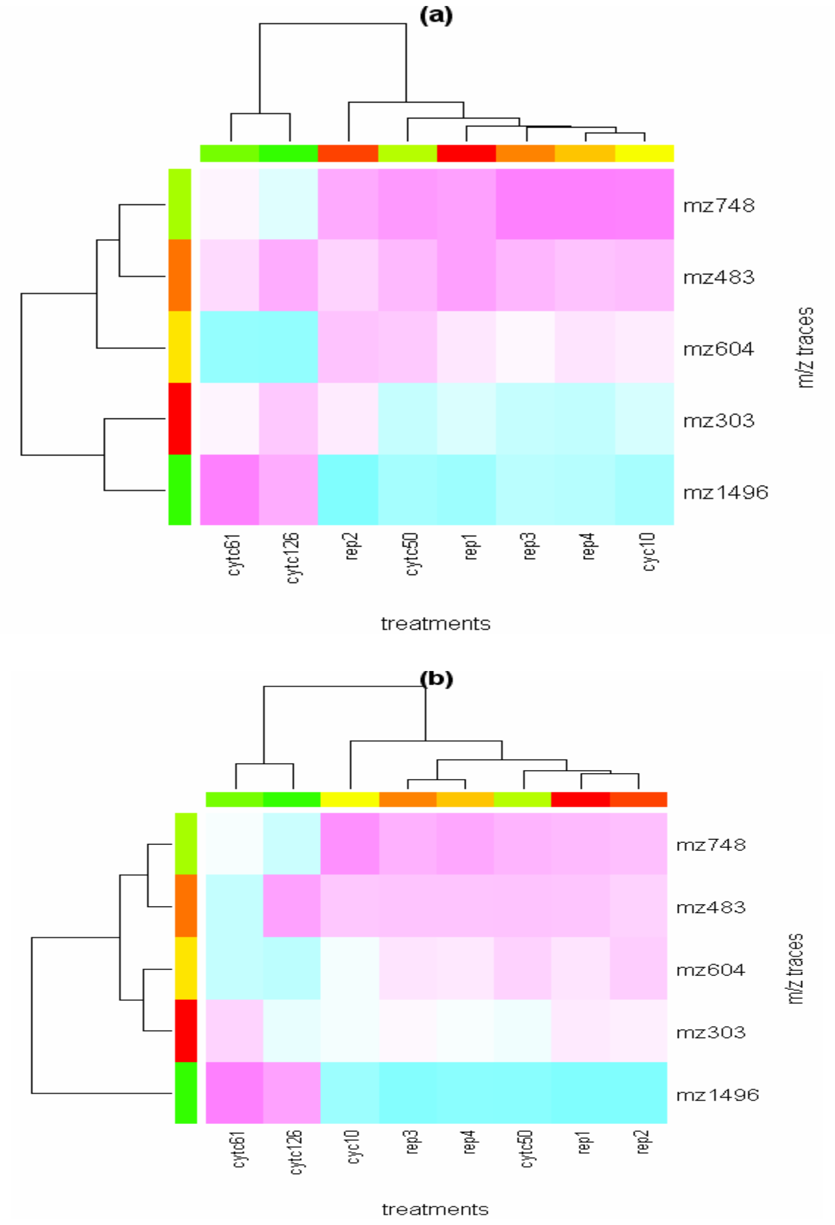

Figure 7. Heatmaps of the goodness-of-fit statistic (r-squared) (a) before and (b) after pre-processing

7.1 Detecting spikes and high frequency noise

To detect spikes and high frequency noise in m/z channel i, we use step-down MTPs, i.e. we consider the hypotheses successively, from the most significant to the least significant, with further tests depending on the outcomes of earlier ones. At step $k = 0, 1, 2, ..., m-2$ we computed the statistics

$$t_r = \varepsilon_{ir}\left(\frac{m-3-k}{m-2-k-\varepsilon_{ir}^2}\right)^{1/2}, \quad r = 1, 2, \ldots, m-2-k \qquad (19)$$

and used the step-down rejection rule $max|t_r| > t(m-2-k, \alpha_i)$, where $\alpha_k = \alpha / 2(m-3-k)$ to declare the largest signal over the range of retention times (still undeclared significant) a spike or random noise. This resulted, for example, in the filtering of the noisy chromatogram in Fig. 2 (a) to produce the compound-related m/z trace shown in Fig. 2 (b). In this chromatogram, a cytoc protein was spiked at around 86 minutes, and this is clearly discriminated from noise by this method (see this peak in Fig. 2 (b)). The other possible criterion for choosing a t-value cut-off is provided by the control of FDP (Lehmann & Romano, 2005). Suppose we want $P[FDP > \gamma] \le \alpha$, where $\gamma \in (0,1)$ then Lehmann and Romano (2005) proposes a step-down procedure with $\alpha_k = (\lfloor \gamma k \rfloor + 1)\alpha / (s + \lfloor \gamma k \rfloor + 1 - k)$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x, and s is the number of tests. For two-sided tests such as those considered here, an appropriate threshold is $\alpha_k^{'} = \alpha_k / 2$.

Table 1: Mass Quality Index: before and after screening noise

m/z = 330	raw m/z	screened m/z
no-spike	0.73	0.80
no-spike	0.81	0.81
CytC10	0.75	0.82
CytC50	0.69	0.74
CytC61	0.89	0.93

7.2 Detecting compound-related chromatograms

After filtering noise from a chromatogram, our second problem is to evaluate the utility of each of the resulting chromatograms in the pursuit of biomarkers, a problem similar to the chromatogram profiling proposed by Windig, *et al.* (1996). The application of MTPs to detect noisy chromatograms is complex, unique and poses further challenges compared to the applications in microarrays. Investigating the compound content of a chromatogram involves comparing thousands of its observed signals (≈ 7000) with their smoothed counterparts using simple linear regression (or two-sample *t*-test), and thus entails the usage of extremely large "sample" sizes compared to the usually small sample sizes (e.g. 10 and often even less) used to detect the significance of genes in microarray data sets. Large sample sizes inflate the values of the test statistics, resulting in extremely small p-values, which must be standardized before being used in the usual way (Good, 1992). Thus whereas the use of p-values make sense in microarrays and the MTPs are directly applicable on these quantities, they may be highly misleading in this LC-MS application and hence the MTPs that utilize p-values are not directly applicable in this case. The use of the rule of thumb of correcting the p-values for the large-sample-size-effect (Good, 1992) did not yield p-values of magnitudes amenable to the analyses by the conventional FDR methods. Hence we adopt an alternative approach that instead utilizes the distribution of the test statistics, an empirical Bayes approach (Efron & Tibshirani, 2002; Liao, *et al.*, 2004), in which each chromatogram is assigned a posterior probability that it is compound-related given that the test statistic takes a specific value, also known as the local FDR. This method is closely related to the Benjamini-Hochberg FDR method (Efron & Tibshirani, 2002).

Let a_i be 1 if the ith chromatogram is compound-related (i.e. if the alternative hypothesis is true or $\beta_{i1} \neq 0$) and 0, when $\beta_{i1} = 0$ (i.e. when $H_{i0} : \beta_{i1} = 0$ is true) and let R_i be the square-root of the coefficient of determination from the regression of the raw chromatogram on its smoothed and perturbed version, $i = 1, 2, \ldots, n$, where n is the number of chromatograms to be profiled (tests to be performed). Let f_0 be the density of R_i given $a_i = 0$ and f_1 be the corresponding density when $a_i = 1$. Then the density of R_i is a finite mixture

$$f(r_i, \theta) = \pi_0 f_0(r_i, \theta_0) + (1 - \pi_0) f_1(r_i, \theta_1), \qquad (20)$$

where $\pi_0 = P[a_i = 0]$ is the expected proportion of true H_{i0}. When the square roots of the co-efficient of determinations from the regression of the raw chromatogram on its smoothed and perturbed counterpart are used as test statistics, it was found that for $R_i = r$, $f_0(r) = \phi(r, \mu_0, \sigma_0)$, the density of a $N(\mu_0, \sigma_0)$ distribution and $f_1(r) = \phi(r, \mu_1, \sigma_1)$, for example, in one run of LC-MS we found that these statistics were fitted by two normal distributions $N(\mu_0 \approx 0.71, \sigma_0 \approx 0.11)$, $N(\mu_1 \approx 0.90, \sigma_1 \approx 0.06)$ with $\pi_0 \approx 0.38$ shown in Fig. 8 (a). These estimates were computed using the **R** statistical package (Ihaka & Gentleman, 1996) function, ***optim***, with the method L-BFGS-B. This procedure requires careful choices of the starting values. The local FDR is the posterior probability of H_{i0} being true given $r_i = r$ and is given by

$$FDR(r) = \frac{\pi_0 \phi(r, \mu_0, \sigma_0)}{\pi_0 \phi(r, \mu_0, \sigma_0) + (1 - \pi_0) \phi(r, \mu_1, \sigma_1)} \qquad (21)$$

This statistic was used to classify chromatograms as being noise or compound-related. For values, $0 \le r \le 1$, the resulting FDRs are depicted by the curve shown in Fig. 8 (b), that suggests a conservative threshold of $R = 0.71$ (or $R^2 = 0.5041$). If however, one is willing to tolerate, for example, only 20% of the false discoveries then $R \approx 0.84$ (or $R^2 \approx 0.71$). This agrees with what we have observed in practice, i.e. $R^2 \ge 0.71$ provides adequate evidence that an m/z trace is compound-related. Thus we recommend an FDR threshold of $R^2 = 0.71$. The definition of this FDR-based threshold is novel as it allays the arbitrariness with which compound content of a chromatogram are declared using current methods such as CODA. For example, using CODA, we used a highly conservative threshold of $r \approx 0.98$ (for the same data analysed in this paper), which resulted in the use of only 45 (see Govorukhina, *et al*., 2006), compared to 1139 out of 1401 m/z traces if an FDR threshold of $R^2 \approx 0.71$ is used.

We also considered mixture modelling of the Fisher normal transformed correlations

$$z_i = 0.5log((1+r)/(1-r)),$$

but we found the model just presented easier to interpret.

8 Applications

In search for biomarkers for a particular condition, one explores differences in proteomic composition of samples from different conditions, for example before and after treatment for cervical cancer. The aim is usually to detect the proteins that are up or down regulated or conserved between the conditions. A major challenge in pursuit of this objective is that chromatograms from different samples are often of different qualities, containing varying levels of noise that mask the detection of common markers. For high quality chromatograms, e.g. Fig. 3 (i), it is easy to obtain accurate markers to be used for comparison. In this case, the application of the so-called one-dimensional peak picking algorithms, e.g. those used to analyze SELDI-TOF data sets (e.g. Li, *et al*., 2005) would give reliable list markers. However, lower quality chromatograms must be pre-processed to screen out the spikes and noise to enable detection of markers with confidence. To demonstrate this problem, we use the profile plots of a chromatogram with m/z = 303 drawn from LC-MS datasets from different samples of trypsin digested human serum from a cervical cancer patient. Figure 2

(a) and (c) show two replicates of the same sample, while (e), (g) and (i) are samples from the same patient spiked with varying levels of cytoc, being 10 picomole (pmol), 50 pmol and 61 pmol, respectively. The purpose of adding cytoc into serum was to have internal standards to evaluate the ability of the proposed methods to reliably detect this added protein. In these datasets, the properties (e.g. the retention times) at which related compounds elute are known. For example, the experimentalist knows that the cytoc peak elutes at around 86 minutes for the chromatogram with m/z = 303, meaning that an important marker for comparison of the spiked and nonspiked samples must be around this time. However, the raw chromatograms for both conditions are riddled with many artefacts making cytoc peaks for lower concentrations indistinguishable from noise. The chromatograms in Fig. 2 (a), (c) and (e) are dominated by spikes and high noise levels, which mask the distinction of cytoc peaks. In Fig. 2 (g), a single spike has completely masked the compound-related peaks, while the spikes in Fig. 2 (i) make it difficult to locate the compound-related peaks. Green circles indicate some selected artefacts in the raw chromatograms. Clearly, there is a need to screen out noise so that compound-related peaks may be located with certainty. Having detected and removed the noise that mask detection of compounds, some of the resulting chromatograms e.g. those dominated with solvent-based noise (Fig. 1 (b)) and low levels of random noise (Fig. 3 (b), (d) and (h)) still exhibit no compound-related information. Such chromatograms are irrelevant for the purposes of biomarkers discovery and must be detected and removed to reduce the dimensionality of the data. We use the MTPs that control the FWER stated in the previous section to set thresholds for selecting the relevant features in a chromatogram, as well as the empirical Bayes-based MTP to detect m/z traces that contain no compound-related information.

We applied the step-down FWER MTPs on the test statistics (diagnostic tools for outlier detection, Eq. (19)) drawn from the linear regression of raw chromatogram on its smoothed version, to set thresholds for detecting the artefacts on a chromatogram. Fig. 4 (ii) is a profile plot of an m/z trace with m/z = 330 from a sample spiked with 25 pmol of cytoc. The scatterplot of this m/z trace against its smoothed and perturbed version, shown in Fig. 4 (c) indicates that the signals of the spike depicted by green dots are outliers. This artefact masks the detection of multiple outliers that characterize this chromatogram, which in turn masks the compound-related peaks. Fig. 6 (a) depicts the raw m/z trace shown in Fig. 2 (a), while Fig. 6 (b) is its version when the dominant spike is deleted. The fall in MQI value from 0.69 to 0.68 demonstrates that deleting the dominant spike in this case unmasks the effects of the other contaminants. The screening of the dominant artefacts results in the quality improved m/z trace (with MQI = 0.74 compared to 0.69), shown by Fig. 6 (c). This value is greater than the FDR-defined threshold, indicating that this chromatogram contains compound-related information. To detect the multiple outliers that characterize this m/z trace, we used a step-down MTP for controlling the FWER that involves the successive use of the single-step procedures for outlier detection (see Cook & Weisberg, 1982, p. 20).

Another important formal tool for detecting outlying signals is the changes in intercepts of the linear fit of the raw chromatograms to their smoothed versions. For example deletion of outliers in m/z traces represented in Fig. 2 (g) (resulting into m/z trace in 2 (h)) resulted in over two-fold changes in the value of the intercept, with only a small change (3%) in the value of slope. Deletion of outlying cases is often accompanied by large change in a goodness-of-fit statistic, e.g. the deviance. For example, there was a substantial improvement in the goodness-of-fit statistic (increase in r-squared of about 10%) for the m/z trace represented in Fig. 2 (b). This means that both the changes in the parameter estimates and the goodness-of-fit are potentially important test statistics that may be used in the MTP for simultaneous detection of artefacts provided their null distributions could be defined.

An important value of the new pre-processing method is that it can detect and remove artefacts and hence unmask compounds in a chromatogram. Table 1 gives the value of r-squared (MQI) of the m/z traces shown in Fig. 2 before and after screening the noise. We see substantial increases in the goodness-of-fit (r-squared) when outliers are deleted, signalling improved evidence of compound-related content. In similar vein, Table 2 shows the MQI values for chromatograms with m/z = 303, 483, 604, 748, 1496. These m/z traces were chosen because they were spiked with proteins of known properties that we wish to discriminate from artefacts. Overall, the new method results in the increases in the MQI values in many m/z traces, irrespective of whether the samples are spiked with cytoc or not. These MQI values strongly exceed the FDR-determined threshold of 0.71, indicating improved evidence of compound-related features in these m/z traces. This demonstrates the efficacy of the new method in discriminating compounds from noise. This means that compounds can be detected with increased confidence after noise filtering using our method. As expected, high concentrations of cytoc (61 pmol and 126 pmol) tend to have higher MQI values because of the well-formed cytoc peaks than at lower concentrations. To further assess the efficacy of the new method, we use the heatmaps to visualize differences in patterns of MQI results in Table 2. Fig. 7 (a) and (b) are heatmaps of MQI values before and after pre-processing. The heatmap in Fig. 7 (a) identifies two similar clusters of treatments, that is the high concentrations of cytoc in one group and its lower concentrations confounded with the unspiked replicates, in the other. There is better separation between the spiked and nonspiked chromatograms after pre-processing by our method. Fig. 7 (b) identifies three clusters, that is, the high concentrations of cytoc, the low concentrations (10 pmol) and the nonspiked replicates. Note that pre-processing of chromatograms leads to correct groupings, lumping together replicates 4 and 3, replicates 2 and 1 and separating cytoc-spiked samples. We note that improved quality leads to efficient detection of compound-related peaks.

A good pre-processing tool must effectively detect and remove artefacts from a chromatogram but must not alter the intensity of the compound-related signals. Fig. 4 (i) depicts a high quality chromatogram. It has a dominant compound-related peak and is noise free. The scatterplot of this chromatogram against its smoothed version (Fig. 4 (a)) reveal that both the compound-related signals and the signals of the main cluster of observations that are expressed at background levels describe a linear trend, indicating that these signals are not significantly affected by smoothing. An application of this algorithm to this high quality chromatogram exhibits similar results, indicating that it does not alter the MQI (see the scatterplot in Fig. 4 (b)) or the compound-related information. However, pre-processing a chromatogram (e.g. Fig. 4 (ii)) that is riddled by artefacts, results in an m/z trace given in Fig. 4 (iii), which exhibits clear compound-related peaks and marked increase in MQI (from 0.65 to 0.80; see Table 1). This indicates that this algorithm effectively detects and removes noise and thus unmasks compound-related information.

After the pre-processing step, an important succeeding step is the determination of the compound-related peaks. Although the relationship between the signals of the compound-related peaks of a raw chromatogram and its smoothed counterpart are consistent with the linear trend suggested by the main cluster of the raw data that are expressed at background (noise) levels versus their smoothed versions, the latter are still distinct from the former in this linear relationship, for example the noise are clustered at the lower end of the line of the best fit, while compound-related peaks are concentrated in the upper end of this line. This suggests that the inclusion of a compound-related peak will improve the accuracy of the slope ($\hat{\beta}_{i1}$) of the line of the best fit to the data. This means that ion currents whose omission results in the large changes in slopes, i.e. large values of $\hat{\beta}_{i1} - \hat{\beta}^{1}_{i1(r)}$, would most likely be compound-related. To test the null hypothesis that *r*th ion current of the *i*th m/z channel is compound-related, one may use the elegant MTP e.g. the bootstrap method (van der Laan, *et al*., 2004) to estimate the test statistic null distribution using appropriate resampled variables $\sqrt{n}(\hat{\beta}_1 - \hat{\beta}^{1}_{i1(r)})$. However, since this step is undertaken after a rigorous MTP pre-processing is performed to eliminate spikes and random noise (expressed at high levels) and thus unmasking the compound-related peaks, we feel that another MTP step may not be necessary and that the use of statistics that rank signals based on the magnitude of the changes in slopes of the line of best fit (influence), e.g. the sample influence curves and the Cook's distance, may just do. The

Cook's distance combines information from both the intercept and the slope into a single measure of influence, while the sample influence curves provide separate information for these parameters. We applied three measures of changes in parameter estimates when a signal is deleted, to chromatograms in Fig. 4 (i) and Fig. 4 (iii). Figure 5 (a), (b) and (c) show the influence curve for the intercept, the influence curve for the slope and the Cook's distance, respectively for chromatogram shown in Fig. 4 (ii). Both the slope and the Cook's distance have high values corresponding to (compound-related) peaks, indicating that the compound-related signals are influential. The intercept also shows clear peaks but with high values, at times not corresponding to the compound-related peaks. This shows that intercept may not be a preferable measure for detecting compound-related peaks. The same conclusions can be drawn for the high quality chromatogram (Fig. 4 (i)). Its influence curve for the intercept, the influence curve for the slope and the Cook's distance, are given by Figure 5 (d), (e) and (f), respectively. Again, both the slope and Cook's distance give clearer separation of compound-related peaks from noise, than does the intercept. All this confirm our conjecture that compound-related peaks are influential. Few top ranked signals represent compound-related peaks. A gain empirical Bayes-based FDR may be used to set a threshold for rejection.

Note the time shifts in the positions of markers (e.g. those shown by red points in Fig. 2 (b), (d), (f), (h), and (j)). This is a well-known instrumental problem with LC-MS data. It can be corrected using time warping algorithms e.g. Correlation Time Warping (COW) (Nielsen, Carstensen & Smedsgaard**,** 1998) may be used as well. For our dataset, we found that the time shift could be corrected by a linear search in the neighbourhoods within 3.2 minutes of points on the target chromatograms corresponding to the peaks on the reference chromatogram.

A chromatogram dominated by solvent-based noise (e.g. a raw chromatogram m/z = 536 shown in Fig. 1 (b)) has no well-defined peaks thus contain no compound-related information. The simultaneous perturbation proposed in Section 5 (Eq. (13)) heavily penalizes this m/z trace, resulting in lack of evidence of compound-related information (R^2 = 0.602, see Table 3) in it. This demonstrates the ability of the new method in discriminating compound-related chromatograms from those with high background noise and may be used as a dimension reduction tool.

9 Discussion

We presented here a novel algorithm to integrate the critical issues of setting thresholds for discriminating compound-related peaks and chromatograms from high frequency noise, spikes and solvent-based noise in LC – MS data sets and

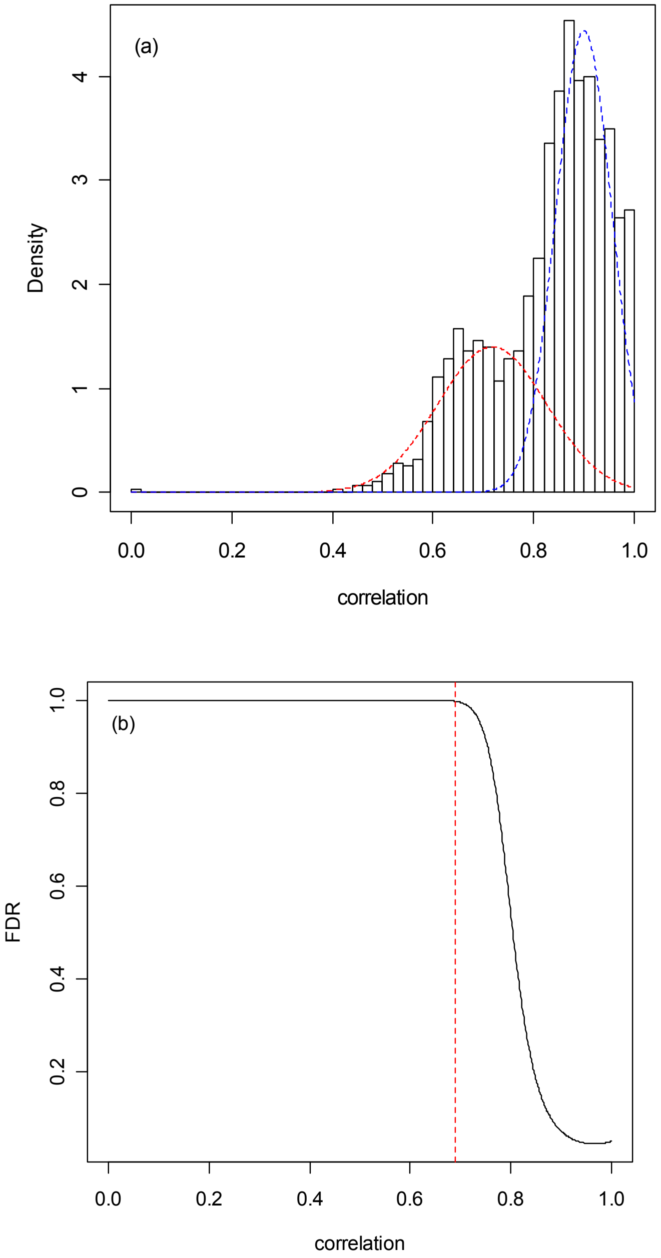

Fig. 8. (a) a mixture of normals fit to the square-roots of the coefficients of determinations of 1401 chromatograms. (b) local FDRs associated with each of these chromatograms.

the control of FDR. This presents a new approach for diagnosing and dealing with artefacts in LC-MS datasets. In the process we introduced new ways of characterizing the signals of a chromatogram, namely (a) that spikes and high levels of noise appear as univariate outliers relative to the linear trend suggested by the bulk of signals that are expressed at background levels and those that are compound-related, (b) the chromatograms dominated by solvent-based noise are multivariate outliers and (c) the compound-related peaks are influential observations. It has been recognised for the first time that the detection of these features involves complex multiple hypotheses testing processes with additional challenges compared to those encountered in microarray studies. In particular, the problem of detecting which of the thousands of chromatograms should be used in biomarkers discovery, involves the use of thousands of data points ("large sample size") compared to the use of the usually smaller sample sizes in microarray studies, and poses the new challenge of how to standardize the p-values so that they may be interpreted and analysed using the conventional FDRs control methods. It is found out that the methods of FDR controls that do not involve the direct use of the p-values such as the empirical Bayes method are easier to adapt and avoids the problem of having to find a suitable standardizing strategy for the p-values.

We have proposed a regression framework for constructing test statistics for the hypotheses about the compound contents of signals and chromatograms. It turns out that a number of the usual linear regression diagnostics may be used to judge the compound contents of these features. Key among them, are the outlier statistic based on the externally studentized residuals (Cook & Weisberg, 1982, p. 20) and the sample influence curve. The simultaneous detection and deletion of the artefacts using the MTP is shown to result in chromatograms with improved qualities and hence compelling evidence of compound-related information. This means that using our method as a pre-processing technique, one is able to detect the compound-related signals with higher degree of confidence and precision. This can highly improve the efficiency of the retention time alignment procedures, which must be done, especially if the intent is to compare samples from different conditions. The compound-related signals have been shown to be influential and a number of formal and informal techniques have been proposed to detect them.

Both the Cook's distance, a measure that combines information from both the intercept and slope of a simple linear regression model, and the slope component of the sample influence curve have been shown to accurately detect these signals by assigning them higher ranks. It has been demonstrated that the local characteristics of chromatograms e.g. their means and medians are indispensable tools for characterizing the compound contents of m/z traces.

The new method accounts for the multiplicity nature of both the peak detection and the chromatogram selection procedures in the analysis of LC-MS data. The use of the MTPs enabled us to give concrete thresholds for detecting compound-related information guided by the crucial aim of controlling FDRs. This allays the arbitrariness with which these thresholds are currently obtained.

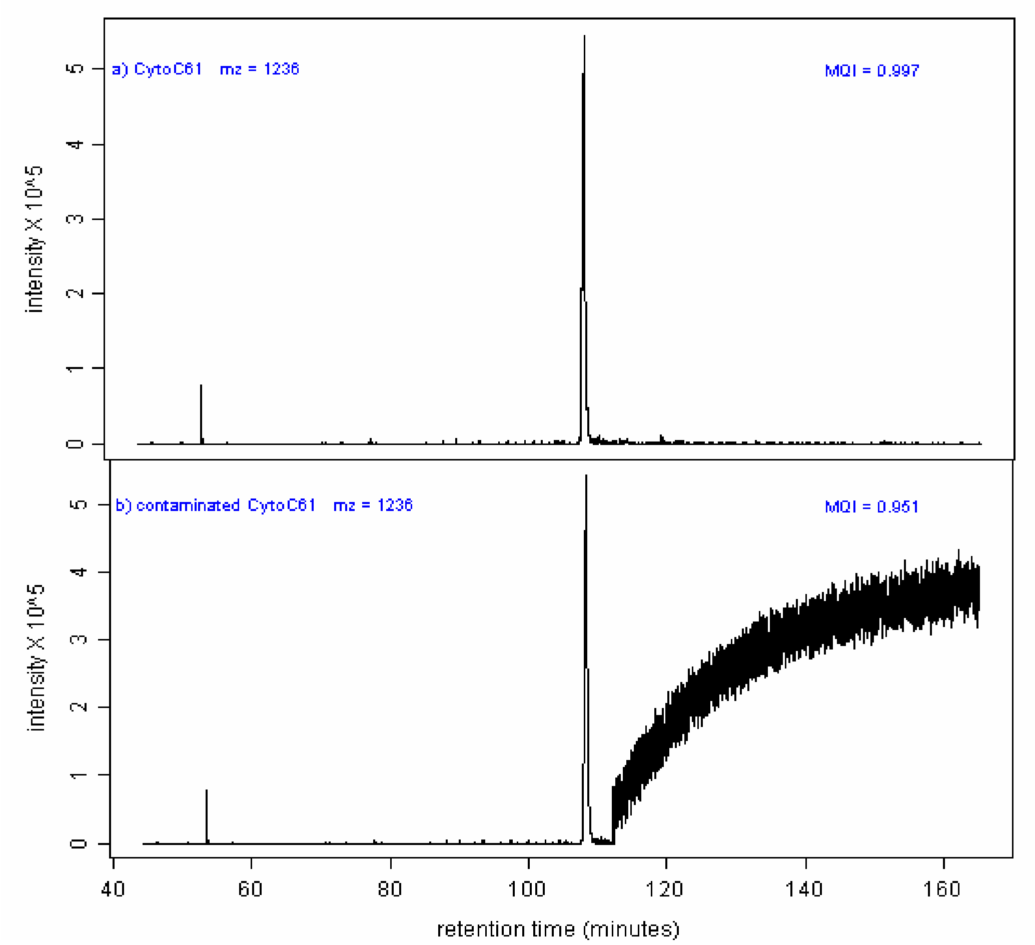

Fig. 9. (a) a high-quality chromatogram from serum spiked with 61 pmol. of cytoc. (b) the same chromatogram contaminated at the upper end with noise that is a mixture of uniform, normal and exponential distributions to create a high background noise.

Perhaps the utmost achievement of our work is the demonstration that regression diagnostics are indispensable tools for unraveling the compound information of chromatograms.

We described our approach in terms of LC-MS technology. However, our method is quite general and applies to many forms of instrumentation such time-of-flight spectrometry.

Finally, it is important to note that when chromatographic peaks are present with a significant contribution of the solvent system, especially during gradient analyses, which results in elevated baselines, the MQI values can be lower (see e.g. Fig. 9 (b) that shows a contaminated version of the high-quality chromatogram in Fig. 9 (a)). In such cases, we recommend that a preprocessing

Table 2: The MQI for chromatograms with same m/z value before and after they are spiked.

m/z	Non spiked (replicates)				Spiked with cytochrome C (pmol)			
	1	2	3	4	10	50	61	126
303	0.73 (0.80)	0.81(0.81)	0.71(0.80)	0.72(0.79)	0.75(0.82)	0.69(0.74)	0.89(0.93)	0.96(0.91)
483	0.86(0.87)	0.84(0.87)	0.83(0.88)	0.83(0.88)	0.84(0.88)	0.85(0.89)	0.91(0.87)	0.99(0.99)
604	0.79(0.81)	0.86(0.88)	0.77(0.83)	0.80(0.83)	0.80(0.82)	0.83(0.86)	0.80(0.87)	0.79(0.88)
748	0.86(0.89)	0.89(0.91)	0.88(0.91)	0.89(0.92)	0.89(0.93)	0.89(0.92)	0.89(0.90)	0.87(0.89)
1496	0.67(0.52)	0.63(0.51)	0.70(0.60)	0.71(0.64)	0.71(0.74)	0.65(0.53)	0.98(0.98)	0.99(0.99)

Table 3: MQI values for chromatograms shown in Fig. 1

m/z	R-squared (rank)
217	0.941 (102)
473	0.998 (2)
493	0.818 (266)
536	0.602 (592)
654	0.704 (494)

step of baseline subtraction be performed before the use of the methods discussed in this paper. There are a number of useful baseline-correction methods including those based on local regression methods such as loess (see e.g. Li *et al*., 2005).

References

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,* Series B, **57**, 289-300.

Birkner, M. D., Hubbard, A. E., van der Laan, M. J., Skibola, C. F., Hegedus, C. M. and Martyn T. Smith, M. T. 2006. Issues of Processing and Multiple Testing of SELDI-TOF MS Proteomic Data, *Statistical Applications in Genetics and Molecular Biology*, **5**, 1, Article 11.

Bylund D. *et al.*, 2002. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modeling of liquid chromatography-mass spectrometry data. *Journal of Chromatography* A, **961**, 237-44.

Cook, R.D., 1977. Detection of influential observations in linear regression. *Technometrics*, **19**, 15-8.

Cook, D., 1986. Assessment of local influence. *Journal of the Royal Statistical Society*, Series B, **48**, 133-69.
Cook, R.D. & Weisberg, S. 1982. *Residuals and influence in regression*. Chapman & Hall.

Crawley, M.J., 2002. *An Introduction to Data Analysis using S-Plus*, John Wiley & Sons, Ltd.

Dobson, A.J., 1990. *An introduction to generalized linear models*. Second Edition. Chapman & Hall.

Dudoit, S., van der Laan, M. J. & Polard, K. S. 2004. Multiple testing Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3 (1): Article 13.

Efron, B. and Tibshirani, R. 2002. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70-86.

Gaspari M; Vogels F; Wulfert F; Tas A; Venema K; Bijlsma S; Vreeken R; van der Greef J. 2001. Novel strategies in mass spectrometric data handling. *Advances in Mass Spectrometry*, **15**, 283-96.

Good, I. J. 1992. The Bayes/Non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, **87**, 597-606.

Govorukhina, N. I., Reijmers, T. H., Nyangoma, S. O., van der Zee, A. G., Jansen R. C., Bischoff , R. 2006. Analysis of human serum by LC-MS: improved sample preparation and data analysis. *Journal of Chromatography* A, **1120**, 142–50.

Hinkley, D.V., 1985. Transformation diagnostics for linear models. *Biometrika*, **72**, 487-96.

Hoaglin, D.C. & Welsch, R.E., 1978. The hat matrix in regression and ANOVA. *The American Statistician*, **32**, 17-22.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.

Ihaka R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 3. 299-314.

Johnson K.J., Wright B.W., Jarman K.H., Synovec R.E. 2003. A high-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *Journal of Chromatography* A, **996**, 141-55.

Lehmann, E. L. and Romano, J. P. 2005. Generalizatons of the family-wise error rate. *Annals of Statistics*, **33**, 3, 1138-54.

Li, X., Gentleman, R., Lu, X., Shi, Q., Iglehart, J. D., Harris, L., Mirion, A. 2005. SELDI-TOF mass spectrometry protein data. In *Bioinformatics and Computational Biology solutions using R and Bioconductor*, Eds. Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S. Springer London, 91-109.

Liao, J. G., Yong Lin, Selvanayagam Z. E., Shih, W. J. 2004. A mixture model for estimating the local false discovery rate. *Bioinformatics*, **20**, 2694-2701.

Listgarten, J. & Emili, A. 2005. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics*. **4**, 419-34.

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., de Jong, S., Lewi, P. J., Smeyers-Verbeke, J., Mann, Charles K., 1997. *Handbook of Chemometrics and Qualimetrics*: Part A. Elselvier.

McCullagh, P. & Nelder, J.A., 1989. *Generalized Linear Models*. Chapman & Hall.

Nielsen N-P.V., Carstensen J. M. & Smedsgaard J, 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping. *Journal of Chromatography* A, **805,** 17–35.

Nyangoma, S.O., Fung, W.K., Jansen, R.C. 2006. Identifying influential multinomial observations. *Computational Statistics and Data Analy*sis, **50**, 10, 2799-821.

Taguchi, G., 1986. *Introduction to Quality Engineering: designing Quality into products and Processes*. Kraus International Publication, White Plains, NY.

van der Laan, M., Dudoit, S. & Pollard, K. 2004. Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives. Working Paper Series, Paper 141, Div. Biostatistics, Univ. California, Berkeley.

Wiener, M. C., Sachs, J. R., Deyanova, E. G. and Yates, N. A. 2004. Differential mass spectrometry: A label-free LC-MS method for finding differences in complex peptide and protein mixtures. *Analytical Chemistry,* **76,** 6085-96.

Wilks, S., 1963. Multivariate statistical outliers. *Sankhyá,* **A**, **25**, 507-26.

Williams, D.A., 1987. Generalize linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, **36**, 181-91.

Windig, W., Phalp J. M. & Payne, A.W., 1996. A noise and background reduction method for component detection in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry*, **68**, 3602-6.

Windig, W., Smith W.F. & Nichols, W.F., 2001. Fast interpretation of complex LC-MS data using chemometrics. *Analytica Chemica,* **446**, 467-76.