Analytica Chimica Acta 648 (2009) 207–214

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Analysis of MALDI FT-ICR mass spectrometry data: A time series approach

Donald A. Barkauskas[a,*], Scott R. Kronewitter[b], Carlito B. Lebrilla[b], David M. Rocke[c]

[a] Children's Oncology Group, 440 E. Huntington Drive Suite 402, Arcadia, CA, 91006, USA
[b] Department of Chemistry, University of California, Davis, CA, 95616, USA
[c] Division of Biostatistics, School of Medicine, University of California, Davis, CA, 95616, USA

ARTICLE INFO

Article history:
Received 20 May 2009
Received in revised form 24 June 2009
Accepted 25 June 2009
Available online 5 July 2009

Keywords:
Fourier transform ion cyclotron resonance
Generalized gamma distribution
Matrix-assisted laser desorption/ionization

ABSTRACT

Matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry is a technique for high mass-resolution analysis of substances that is rapidly gaining popularity as an analytic tool. Extracting signal from the background noise, however, poses significant challenges. In this article, we model the noise part of a spectrum as an autoregressive, moving average (ARMA) time series with innovations given by a generalized gamma distribution with varying scale parameter but constant shape parameter and exponent. This enables us to classify peaks found in actual spectra as either noise or signal using a reasonable criterion that outperforms a standard threshold criterion.

1. Introduction

Matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry (MALDI FT-ICR MS) is a technique for high mass-resolution analysis of substances that is rapidly gaining popularity as an analytic tool in proteomics. Typically in MALDI FT-ICR MS, a sample (the *analyte*) is mixed with a chemical that absorbs light at the wavelength of the laser (the *matrix*) in a solution of organic solvent and water. The resulting solution is then spotted on a MALDI plate and the solvent is allowed to evaporate, leaving behind the matrix and the analyte. A laser is fired at the MALDI plate and is absorbed by the matrix. The matrix becomes ionized and transfers charge to the analyte, creating the ions of interest (with fewer fragments than would be created by direct ablation of the analyte with a laser). The ions are guided with a quadrupole ion guide into the ICR cell where the ions cyclotron in a magnetic field. While in the cell, the ions are excited and ion cyclotron frequencies are measured. The angular velocity, and therefore the frequency, of a charged particle is determined solely by its mass-to-charge ratio. Using Fourier analysis, the frequencies can be resolved into a sum of pure sinusoidal curves with given frequencies and amplitudes. The frequencies correspond to the mass-to-charge ratios and the amplitudes correspond to the concentrations of the compounds in the analyte. FT-ICR MS is known for high mass resolution, with separation thresholds on the order of 10^{-3} Daltons (Da) or better [1,2].

The spectra analyzed in this article were recorded on an external source MALDI FT-ICR instrument (HiResMALDI, IonSpec Corporation, Irvine, CA) equipped with a 7.0 T superconducting magnet and a pulsed Nd:YAG laser 355 nm. In addition to hundreds of spectra generated as described above for a cancer study [3] using human blood serum as the analyte, we generated 56 spectra using neither analyte nor matrix. We will refer to the latter category of spectra as "noise spectra" and use them in Sections 2 and 3 to develop our model, then apply the model to a spectrum with known contents in Section 4.

We find that an autoregressive, moving average (ARMA) time series with innovations given by a generalized gamma distribution can closely model the properties of the noise spectra, and that this representation is useful for accurately identifying real substances in spectra produced using analyte. The modeling assumptions developed in this article are implemented in the *R* package `FTICRMS`, available either from the Comprehensive *R* Archive Network (http://www.r-project.org/) or from the first author.

2. Methods

2.1. Description of data

A typical noise spectrum is shown in Fig. 1 with frequency in kilohertz (kHz) plotted on the horizontal axis. (In the mass spectrometry literature, it is more usual to see m/z—the mass-to-charge ratio—on the horizontal axis, but the actual process of measurement uses equally spaced frequencies, and the m/z values are computed using one of several non-linear transformations on the frequencies [4]. Thus, the spectrum pictured in Fig. 1 is how it appears after the

* Corresponding author.
E-mail address: don.barkauskas@curesearch.org (D.A. Barkauskas).

0003-2670/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2009.06.064

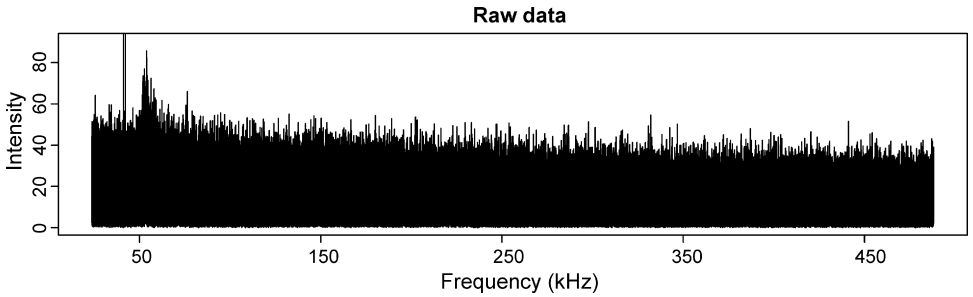

Fig. 1. Typical noise spectrum. A MALDI FT-ICR spectrum produced without matrix or analyte. The spike extending off the top of the picture is actually two peaks at frequencies of 41.21 and 42.21 kHz which extend upward to intensities of approximately 222.7 and 95.4, respectively.

fast Fourier transform is applied to the measured data.) The thick spike at a frequency of roughly 40 kHz is actually two peaks at frequencies 41.21 and 42.21 kHz which extend upward to intensities of approximately 222.7 and 95.4, respectively, and are apparently instrumental noise—they appear in all 56 noise spectra at roughly the same spots and have no isotope peaks. In the analysis that follows, we set the values of the spectra at frequencies corresponding to these two peaks to be missing.

2.2. Properties of noise spectra

We start by considering two striking properties of the noise spectra. The first property is the special forms of the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) of the noise spectra; Fig. 2 displays the graphs of the sample ACF and sample PACF of the noise spectrum from Fig. 1. Starting with lag 7, the sample ACF is nearly constant at roughly 0.0613. The sample PACF, on the other hand, oscillates between positive and negative values before decaying to a small positive value. As we show in Section 2.3, the sample ACF enables us to get information not only about the baseline but also about the coefficients to use in the ARMA representation of the spectrum. The sample PACF will be useful for evaluating the final ARMA model for accuracy. The second property comes from looking at the sample "homogenized" cumulants $\hat{\kappa}'_1, \hat{\kappa}'_2, \ldots$ of the spectrum. (The sample homogenized cumulants of a set of data are related to the mean, variance, skewness, kurtosis, etc., of the data and will be defined precisely in Section 2.4, Eq. (5).) Fig. 3 displays scatterplots of the running sample homogenized cumulants (with bandwidth 4001 points—other bandwidths give similar plots) of the noise spectrum from Fig. 1. It is clear that the first three sample homogenized cumulants have strong relationships. As we show in Section 2.4, this enables us to get information about the proper parameters to use in the generalized gamma distribution for the innovations in the ARMA representation of the spectrum.

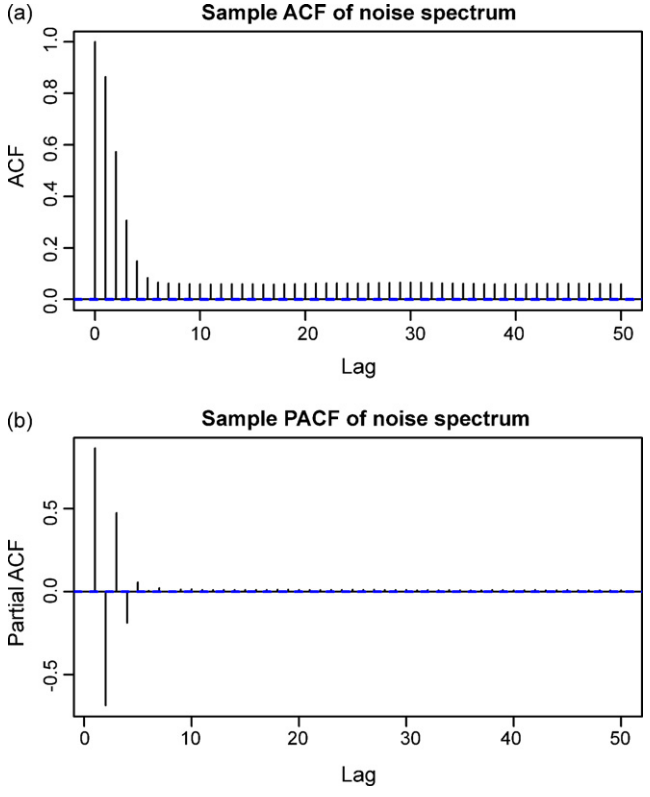

Fig. 2. Sample ACF and sample PACF of typical noise spectrum. The sample autocorrelation function (top) and sample partial autocorrelation function (bottom) through lag 50 of the noise spectrum from Fig. 1.

2.3. Analysis of the ACF

The sample ACF $\hat{r}_k$ at lag k of a realization $\{y_t\}_{t=1}^n$ of a time series $\{Y_t\}_{t=1}^n$ is defined by

$$\hat{r}_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}, \tag{1}$$

where $\bar{y}$ is the sample mean. This is usually defined for stationary time series, in which (among other criteria) the means $\{\mu_t\}$ of the underlying random variables $\{Y_t\}$ are assumed to be constant. However, estimating the underlying means for a noise spectrum by some method (running means, running medians, etc.) clearly shows that they are not constant.

Thus, suppose that $Y_t \sim (\mu_t, \sigma_t^2)$ with known means $\{\mu_t\}_{t=1}^n$ and suppose that the correlation between Y_t and Y_{t-k} is given by $\tilde{\rho}_k$ (independent of t). Then, we have

$$\begin{aligned} \tilde{\rho}_k &= \frac{\mathbb{E}\{(Y_t - \mu_t)(Y_{t-k} - \mu_{t-k})\}}{\sqrt{\mathbb{E}\{(Y_t - \mu_t)^2\}} \cdot \sqrt{\mathbb{E}\{(Y_{t-k} - \mu_{t-k})^2\}}} \\ \tilde{\rho}_k \sum_{t=1}^{n} (y_t - \mu_t)^2 &\approx \sum_{t=k+1}^{n} (y_t - \mu_t)(y_{t-k} - \mu_{t-k}), \end{aligned} \tag{2}$$

where $\mathbb{E}(\cdot)$ is the expected value operator. We subtract the right-hand side of Eq. (2) from the left and add the result to the numerator