Geosci. Instrum. Method. Data Syst., 4, 139–148, 2015
www.geosci-instrum-method-data-syst.net/4/139/2015/
doi:10.5194/gi-4-139-2015

Geoscientific Instrumentation Methods and Data Systems
Open Access

Analysis of COSIMA spectra: Bayesian approach

H. J. Lehto[1], B. Zaprudin[1], K. M. Lehto[2], T. Lönnberg[3], J. Silén[4], J. Rynö[4], H. Krüger[5], M. Hilchenbach[5], and J. Kissel[5]

[1]Tuorla Observatory, Department of Physics and Astronomy, University of Turku, Väisäläntie 20, 21500 Piikkiö, Finland
[2]Molecular Plant Biology, Department of Biochemistry, University of Turku, 20014 Turku, Finland
[3]Organic Chemistry, Department of Chemistry, University of Turku, 20014 Turku, Finland
[4]Finnish Meteorological Institute, Erik Palmenin aukio 1, PB 503, 00101 Helsinki, Finland
[5]Max Planck Institute for Solar System Research, Justus-von-Liebig-Weg 3, 37077 Göttingen, Germany

Correspondence to: H. J. Lehto (hlehto@utu.fi)

Received: 16 June 2014 – Published in Geosci. Instrum. Method. Data Syst. Discuss.: 11 November 2014
Revised: 26 May 2015 – Accepted: 27 May 2015 – Published: 30 June 2015

Abstract. We describe the use of Bayesian analysis methods applied to time-of-flight secondary ion mass spectrometer (TOF-SIMS) spectra. The method is applied to the COmetary Secondary Ion Mass Analyzer (COSIMA) TOF-SIMS mass spectra where the analysis can be broken into subgroups of lines close to integer mass values. The effects of the instrumental dead time are discussed in a new way. The method finds the joint probability density functions of measured line parameters (number of lines, and their widths, peak amplitudes, integrated amplitudes and positions). In the case of two or more lines, these distributions can take complex forms. The derived line parameters can be used to further calibrate the mass scaling of TOF-SIMS and to feed the results into other analysis methods such as multivariate analyses of spectra. We intend to use the method, first as a comprehensive tool to perform quantitative analysis of spectra, and second as a fast tool for studying interesting targets for obtaining additional TOF-SIMS measurements of the sample, a property unique to COSIMA. Finally, we point out that the Bayesian method can be thought of as a means to solve inverse problems but with forward calculations, only with no iterative corrections or other manipulation of the observed data.

1 Introduction

The COmetary Secondary Ion Mass Analyzer (COSIMA) is a time-of-flight secondary ion mass spectrometer (TOF-SIMS) on board the Horizon 2000 European Space Agency Rosetta mission en route to encounter the 67P/Churyumov–Gerasimenko comet. The space probe consists of an orbiter and a lander. After the in-flight hibernation, the space probe and its instruments were successfully woken up on 20 January 2014. The first orbital maneuvers for the comet approach took place in May 2014. The formal mission end date is 31 December 2015. By that date the comet will have passed perihelion with the Rosetta spacecraft clinging near it all the time.

While the orbiter is traveling at slow speed (meters per second) in the vicinity of the comet (Glassmeier et al., 2007), the COSIMA instrument is collecting dust particles that have been expelled by the comet. A representative set of these particles enter into COSIMA through an open window that extends to the outer surface of Rosetta. The particles are collected on a target plate, which consists of three square 1 cm by 1 cm metal plates and an unexposed 0.3 cm by 3.0 cm reference area. Once exposed, the plate is stored. At a suitable time, the plates are examined one by one with an illuminated optical microscope, the COSISCOPE, which has an optical pixel size resolution of 14 µm (Kissel et al., 2007). By combing several exposures, a super resolution of about 3 µm is possible. Target particles for further analysis are selected, and exposed to an ^{115}In primary ion beam with an ion energy of +8 KeV, a pulse duration of < 3 ns, and a beam width of 50 µm. During each pulse, an unknown number of secondary ions are expelled from the top layer(s) of the target sample. These secondary ions then enter the electric field lens system and end up on the detector, where the flight times of ions are measured. In short, this process is called a shot. Depending on the polarity, positive or negative ions are detected. The

shots are repeated at 500 µs intervals; thus, a 1 s exposure consists of 2000 shots, and during a 3 min exposure, about 360 000 shots are fired. The instrument is described in detail by Kissel et al. (2007).

The outcome of a measurement is a time-of-flight spectrum, which we are directly interested in. In this paper, we discuss the quantitative foundation of understanding the spectrum through statistical analyses of individual spectral lines and touch on some critical issues such as instrument dead-time effects, normalization, and isotope ratio calculation of lines. Multivariate techniques connecting complex chemistry and complete spectra are discussed elsewhere (Silén et al., 2014). Bayesian methods can be extended to the interpretation of these cases, but it is beyond the scope of this paper.

2 Time-of-flight spectrum

We measure the time-of-flight spectrum, i.e., the number of secondary ions as a function of time. This defines the coordinate space we are working in.

The time of flight of an ion scales with the mass m and charge q,

$$t = a + b\sqrt{(m/q)}, \tag{1}$$

where the mass calibration parameters a and b have values of about 4000 and 1600, respectively, in COSIMA. The values of a and b are initially estimated by the onboard software. The charge q is usually $+1$ or -1. The mass of the ion m is expressed in atomic mass units, u. The time of flight is digitized by the time-to-digital converter to bins of 1.953125 ns (Kissel et al., 2007). One COSIMA time-of-flight time bin t corresponds to a mass bin of $0.0013\,u$ at mass $m \sim 1$ and to a mass bin of about $0.04\,u$ at $m \sim 900$.

The resolution of the mass spectrometer is $m/\delta m \sim 1400$ at $m = 100$. At low masses, all the atomic lines are easily separated. Up to a mass of about $m \sim 120\,u$, mineral- and hydrogen-rich organic components can be separated (F. Krüger, personal communication, 1992). The distinction is based on the fact that most minerals due to their internal structure produce elemental masses in their spectra. These have values that usually are below the integer value of the mass. Single mineral ions with $Z > 86$, $m > 222$ have masses above integer values, but they are not expected to have a large contribution to the spectrum. Hydrogen tends to be common in organic molecules and their breakup products. Neutral hydrogen has a mass surplus of $\delta H = 0.0078\,u$ above the integer value of 1. A loss of an electron produces a hydrogen ion, H^+ with an excess of $\delta H^+ = 0.0073\,u$. This implies that organic molecules with ample hydrogen tend to have masses above an integer mass value. It is noteworthy to mention that other elements common in organics have the following deviations: ^{12}C: $0\,\delta$H, ^{14}N: $0.39\,\delta$H, and ^{16}O: $-0.65\,\delta$H, so nitrogen enhances the positive deviation, while the presence of oxygen reduces it. Two relatively common elements, phosphorus and sulfur, often associated with organics, reduce the organic shift by $^{31}\mathrm{P} : -3.36\,\delta\mathrm{H}$ and $^{32}\mathrm{S} : -3.56\,\delta\mathrm{H}$.

The full spectrum consists of 2^{17} or 131 072 time bins and reaches to about $m \sim 6400\,u$. The raw data are in counts per TOF time bin. The lowest mass peak is usually a hydrogen ion at $1.0073\,u$. An electron peak at mass $m_e = (1/1839)\,u$ is present in negative spectra. It is broad due to the significantly larger thermal velocities electrons have compared to ions and due to the high energy of the ion formation and decay processes after the substrate has been irradiated with the indium beam. In principle, ions with the same mass should fall into one time bin. However, the stability of the instrument, the pulse length of the primary beam, finite beam size and thermal distribution of ions all contribute to the resolution so that the time-of-flight arrivals from a single pulse produce a peak with a full width at half of the maximum amplitude (FWHM) of about 2.5 TOF time bins, and close to Gaussian in shape.

Dead-time effects

In weak lines with low secondary ion yield, most of the primary ion beam firings will produce no secondary ions for that mass, and only single ions will be recorded occasionally. For example, a line with a total count of ~ 1000 secondary ions in a 3 min exposure will behave in this way, with about 1 secondary ion on average for every 360 indium shots.

If the ion yield is higher, such that the total counts of a spectral line are on the order of 10 % of the total number of shots, in a 3 min exposure, a few times $\sim 10^4$, then the instrument dead-time effect sets in. After the arrival of a secondary ion, the instrument does not respond to new secondary ions within the next 10 ns, which corresponds to about 5.2 TOF time bins (Kissel et al., 2007). This is important when the number of cases with two or more ions arriving at the instrument becomes significant. Note also that the instrument cannot distinguish between background ions and "good" ions. Both will contribute to the dead-time effect. The contribution of the background is expected to be small, because of the low background levels in COSIMA. It cannot be completely ignored, however. The dead time causes two major distortions to the shape of the spectral line. It reduces the total number of counts detected per time-of-flight bin. Furthermore, a second bias is produced by the asymmetric nature of the dead time. The spectral line becomes skewed by the shifting the peak of the line to shorter flight times than if all ions were recorded or if the dead time was 0.

The single most important parameter for understanding the dead-time effect is the yield of secondary ions to the ratio of counts creating a given line to the number of shots. It gives a measure of how many occurrences of two or more ions in a single shot occur for that particular line. This implies that two spectra with the same absolute line counts for a given mass will have different dead-time effects if they have a dif-

ferent number of primary ion shots. On the other hand, the shape of the same line in a sample from short exposures and long exposures will not change due to dead-time effects if the secondary ion yield does not change.

Ideally, the time of the flight spectrum would show no background, show sharp discrete line peaks and have an exact TOF mass calibration. In reality, we are limited by measurement statistics, finite resolution, dead-time effects, background, multiple nearby lines and various other issues. We will next address how to analyze our COSIMA spectra from a Bayesian perspective.

3 Mathematics and statistics

The ordinate in the data is the TOF time bin, which has a linear relation to the time of flight, which scales as the square root of ion mass. The data themselves are count data and thus Poisson distributed.

The parameters we are interested in at a given mass are the number of spectral components, the integrated count of each component, the mass corresponding to each line and the confidence limits of all these parameters. We approximate the spectral lines as Gaussian in the time-of-flight coordinate system. Standard methods such as least squares or χ^2 fittings are not applicable, however. The reason for this is the nature of the noise in these spectra.

Our data are particle count data and as such positive definite. They follow Poisson statistics. The Poisson probability density function is defined as

$$p(n,\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}, \tag{2}$$

where for large values of n the factorial is calculated either in a logarithmic form

$$\ln(p(n,\lambda)) = n\ln\lambda - \lambda - \ln(n!), \tag{3}$$

or, e.g., by the Batir (2010) equation that is good for $n > 1$ to within a relative accuracy of $< 10^{-6}$, which is sufficient for our calculations.

The Poisson nature of the data implies that the probability density function is not symmetric. The mean has a value different from the median and the mode, the most likely value. As the distribution is not symmetric, the standard square root of variance, "sigma", should not be used to calculate confidence limits or "error limits". Note also that strong peaks have the largest noise in absolute terms, whereas low peaks have a more significant noise contribution in relative terms.

The instrumental dead time brings an additional special complication. The observed data that are affected by the dead time still have a Poisson probability distribution in secondary ion counts per time bin. The "correction" of the dead time used in conventional methods applied to the observed data points effectively distorts the statistical properties of the data by increasing the real noise in the corrected data to a level larger than what is expected from Poisson data, the corrected data being essentially too noisy. This is a potential problem for strong lines. Our approach avoids these problems because we apply the dead-time corrections to the model and not to the observed peaks.

In statistical analyses, one tends to habitually assume Gaussian noise and the propagation of errors through addition of variances. These assumptions are not valid in our case. Their use could cause negative values in "error" limits, which is mathematically and physically an impossible situation, as they would imply negative counts. Furthermore, the way propagation is used contains the hidden assumption of symmetric errors, which is not the case in these data.

We will address the analysis of COSIMA spectra through Bayesian analysis, which will avoid all the problems mentioned above.

3.1 Bayesian analysis

The Bayesian analysis is a universal means of understanding and interpreting measured data. In principle, we could consider our spectrum as one measured entity with several hundred lines and interpret the full spectrum by Bayesian means. This would require working in a data space of a dimensionality of several thousands squared. In practice, it is more convenient to reduce the analysis into an analysis of hundreds of lines, each consisting of a combination of one or several nearby lines. We can do this as, at low masses, there is no overlap between lines of different integer masses and, at high masses, the lines tend to be sparse and still well separated.

The conventional analysis of resolved time-of-flight mass spectrum starts with the observed spectrum, applies a correction term in order to correct for (i.e., remove) the dead-time effects, and then possibly remove a background and then treat the remainder as the real line. Careful conventional analysis does not ignore the contribution of background noise, however. Stephan et al. (2001) presents an elegant way to analyze the classical challenge of separating ^{13}C from the nearby ^{12}CH by deconvolving either the ^{13}C or ^{12}CH with the line shape measured from ^{12}C. From the statistical point, the introduction of the errors from the ^{12}C line and the indeterminacy of the posteriori probability distributions of the fitted lines are something that is not addressed properly without Bayesian methods. All the steps in this procedure distort the original Poissonian nature of the probability distribution of the data.

Our analysis is nearly reverse in many aspects. We start by selecting a model from a large set of models of a beam shape, amplitude and background, after which we apply the effects of the dead time and obtain a model for an observed spectrum. Assuming a Poisson distribution for the model, we then calculate the likelihood that this model will explain the observations. We then iterate towards a cloud of solutions. There are two details we should emphasize here. Our calcu-

lation is a forward calculation. For this reason, we take the dead-time effects into account in a reverse order than what is conventional; thus, we *apply* the effects of the dead time to our model instead of trying to "remove" the effects from the observed data points. Note that at no point do we manipulate or change the values of the real observed data. This has profound implications that we address next in more detail.

Assume that you have an observed peak shape $Y_0(t)$, where t represents a time bin. It is a sum of the true unknown peak shape $\theta(t)$ and a noise term $n_0(t)$. If we have prior knowledge of a likely beam shape, we may assume this shape. It does not mean that we fix the beam shape for good, as we can later apply other models and compare them objectively. This is one of the benefits of Bayesian analysis. It is however good to have a reasonable starting model. The simplest model is that there is no signal in the data and that, at a mass interval, the $y(t)$ is constant. Using a single Gaussian added to a constant background will require three additional parameters, the amplitude, the width and the position of line position. For each additional Gaussian, we need three more parameters. In our model, we need furthermore to take into account the dead-time effects that affect several bins at a time. Our model is thus of the form

$$\theta(t) = D(y(t, x_n)); \tag{4}$$

here, $\theta(t)$ is the calculated model, $D(.)$ is the dead-time effect, y is the model that is a function of time, and n is the number of parameters, one for background and three additional parameters for each Gaussian.

We will search for a solution from the values of model parameters $\theta = \theta(i, x_n)$ that best describes the observed spectrum $Y_0(i)$. As the time is discrete, we use now i instead of the t for continuous time. To within a normalization constant, we can directly calculate for each point i the probability p that our observed data $Y(i)$ are explained by a given model θ. Multiplying all these individually calculated probabilities, we get the conditional probability of our data given the model $p(Y|\theta)$. Note that this is a point where we differ, e.g., from a χ^2 minimization, as we do not square deviations but rather calculate probabilities. Next we take into account any prior information we have of the parameters and their distributions. This probability independent of the data is called the prior probability density $p(\theta)$. It can be considered as the sampling joint probability space of the data model. By multiplying these two probability densities, we get the probability distribution of the values of θ that explain our data.

Mathematically, we are interested in knowing which is the best model θ that describes our data.

$$p(\theta|Y) \propto p(\theta)\, p(Y|\theta) \tag{5}$$

The posteriori density $p(\theta|Y)$ describing the probability density function of the model parameters is thus proportional to the product of the prior density of the model parameters $p(\theta)$ and the probability of the sampling distribution, the data based on the model $p(Y|\theta)$. This is a simplified version of the Bayesian inference (Gelman et al., 1995).

The prior densities $p(\theta)$ are selected such that the position has a uniform density in a mass interval $(m - 0.5, m + 0.5]$: the amplitude of the peak is not well determined in advance, so we have given a prior distribution that is non-informative, i.e., flat in $10\log(\text{amplitude} + 1)$. The prior density for the peak width is somewhat cumbersome. We know that the value cannot be negative, and is not likely to be very wide. The FWHM of the peak of the COSIMA is expected to be close to about 2.5 or a sigma of 1.1 TOF time bins. To take this into account, we apply a prior density distribution $\propto (\lg(\text{FWHM}/2.5))^{-2}$. Note that this does not rule out solutions with single wide peaks. It also biases against identifying single high bins as very unrealistic narrow peaks and against modeling a constant background as an extremely wide Gaussian.

Our result will provide the probability density distribution of various parameters on the left of the equation above. By finding the mode of this distribution, we get the most likely value in the model parameter space describing the data. Confidence limits to the model parameters can be calculated from the posteriori probability distributions.

3.2 MCMC algorithms

Markov chain Monte Carlo (MCMC) algorithms are very useful in determining the posteriori probability space. A random walk in the parameter space is created. This chain converges to the target distribution that, multiplied by the prior distribution, will give the posteriori distribution. In creating the random walk sequence, the next draw from the parameter space depends on the position and the value of the previous sample.

One widely used family of MCMC chains is the Metropolis algorithm. The core in these algorithms is the decision of whether to accept the next move. Say that one has calculated the probability $p(\theta_i|Y)$, we make a move from θ_i to θ_{i+1} by selecting randomly a point from the jumping distributions, which have to be symmetric. If the new θ_{i+1} has a higher probability $p(\theta_{i+1}|y)$, then we accept the move. If the probability is poorer, it will be accepted if the ratio $p(\theta_{i+1}|y)/p(\theta_i|y) > a$, where a is a random number drawn from a uniform distribution [0,1). The selection of points of lower likelihood allows for an effective sampling of the posteriori distribution. This central part of the decision is shared by many Markov chain algorithms that carry different kinds of names.

We use the adaptive Metropolis algorithm (Haario et al., 2001). It has the same Metropolis jumping criterion as above, but it has different means for selecting an optimized step size and the jumping direction in the d-dimensional parameter space. The covariance matrix of the model parameters is calculated from the ever refining posterior distribution in the parameter space. The step size is obtained from the Cholesky

decomposition of the covariance matrix multiplied by a normalization factor $s_d = 2.4^2/d$, where d is the dimension of the free parameters.

$$\boldsymbol{X}_k = \boldsymbol{X}_{k-1} + \sqrt{s_d}\,\mathrm{Chol}(\mathbf{C}_k)\boldsymbol{G} \quad (6)$$

Here, $\boldsymbol{X}_k$ is the parameter vector at step k, $\boldsymbol{G}$ is a random vector from a normal distribution $\mathbf{N}(0, 1)$, $\mathbf{C}_k$ is the covariance matrix of the model parameters calculated from a suitable number of points, and d is the number of parameters in the model (Tamminen, 1994).

For simulated and real spectra, we have used typically 200 iterations in the burn-in phase and 20 000 in the main iteration phase. The upper limit of iterations is determined by convergence to the posteriori distribution and the confidence levels needed.

4 Results of calculations

4.1 Effects of the dead time

The dead-time effects cause the spectral peaks to become weaker, shift the peak maximum to smaller masses and distort significantly small peaks that have a mass slightly larger than a strong peak. To characterize the effects of dead time on COSIMA spectra and to understand the characteristics in detail, we performed simulation. Furthermore, the simulations were run to validate that the equations we derived for our Bayesian method from first principles are valid.

We assume a dead time of 10 ns with a total blockage between the first ion and the subsequent ions. After this dead time, a new ion can be measured initiating a new dead time. From different total secondary ion counts and indium shots, we calculate the probability of various numbers of shots from this line. We simulate each shot separately. The number of secondary ions is obtained by drawing a random number from [0, 1), which is mapped to the cumulative probability distribution of the number of Poisson shots. So, a random number ranging from 0 to a certain value represents the interval of $P(n = 0\,\mathrm{ions})$ giving the mean yield of shots. The next interval represents the range for one secondary ion, etc. This gives us a number that can be interpreted as the probability in the Poisson sample space. Thus, the value tells us how many secondary ions result for each shot. For these ions, we draw the analog of a flight time from a Gaussian distribution with a FWHM of 5 ns (or 2.5 time bins). If two or more secondary ions occur, we determine the ions for which the dead time applies. Finally, we calculate from these simulations two separate line shapes: first, a line with no dead-time correction applied, and then a second one where it has been applied (Fig. 1).

From these simulations, we derive some relevant statistical properties of the dead-time effects. We can confirm the equations derived purely on statistical grounds by Stephan et al. (1994),

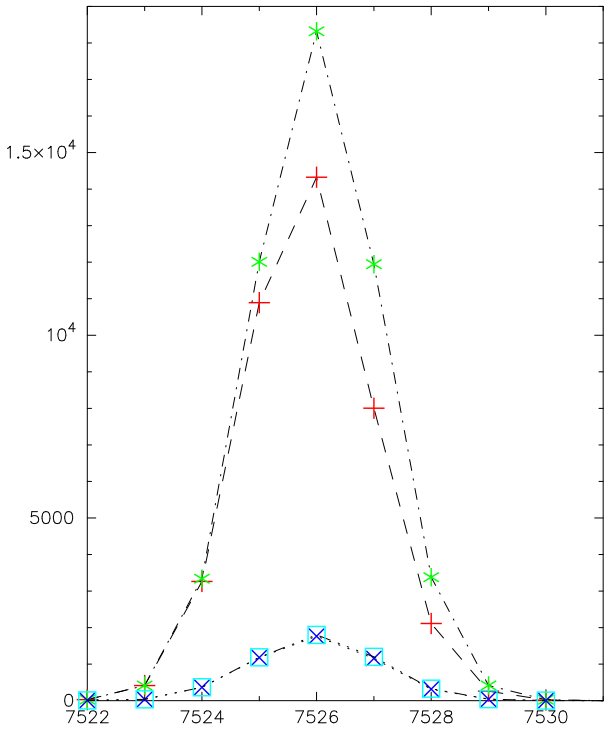

Figure 1. Effects of dead time in COSIMA. The effect is shown for two artificial Gaussian lines. The strongest line has a total yield of 50 %. The fainter line has a yield of 5 %. The top-most curve (green stars and dot-dash line) shows the original 50 % curve; the second curve (red crosses, dash line) shows how the dead-time effect has changed the curve. This is in principle the observed line. Note how the maximum and correspondingly the total line counts have decreased. Also note how the line center and peak have shifted to the left. This is particularly noteworthy on the right side of the line. The lower two curves show similar cases for the 5 % line with blue squares and purple crosses, respectively.

$$I_{\mathrm{cor}} = N \ln(1 - I_{\mathrm{exp}}/N), \quad (7)$$

where N is the number of shots in the spectrum and I_{cor} is the original count with no dead-time effects and I_{exp} is the observed line with dead-time effects in place. If we have a small yield, i.e., if the ratio of the number of secondary ions integrated over to shots is less than about $1/4$, which corresponds to a peak value of $1/15$ of the number of shots, then the intensity of the peak is reduced by the dead time by about

$$\delta I = \left(-\frac{I}{2N}\right). \quad (8)$$

This is the case for most lines observed in COSIMA. For example, if a spectrum results from 360 000 shots, and has an integrated secondary ion count of 24 000, it will lose about 800 counts or 3 % due to the dead-time effects.

The dead time shifts the peak position by about

$$\Delta = -0.3 \cdot I_{\mathrm{exp}}/N, \quad (9)$$

where Δ is in units of TOF time bins. Alternatively, this can be expressed as $\Delta = -0.12 \cdot \mathrm{FWHM} \cdot \ I_{\exp}/N$. The shift has a statistical standard deviation of about $\sigma_\Delta = 1/\sqrt{N_D}$ within 10 %. For large yields and at large count values, Δ has an important contribution for the determination of the line position.

Our Bayesian approach is not affected by the two problems described above as our approach can be considered as an inverse solution calculated by fully forward sub-solutions. In our simulation, it is better that we use a model for the dead time. The magnitude of the dead time depends on the number of counts in the previous bins as follows. A count will be recorded if there are no counts earlier in the same bin or in the previous 10 ns. This is covered by taking into consideration half of the counts in the bin in question, and the previous five full bins and 0.65 of the bin five time steps earlier. Effectively, this is calculating the conjugate of the probability of no ions in the previous bins, and is thus independent of previous derivations. Using this simple formula, we can model the dead-time effects in our simulations:

$$I = I_0 \cdot \left(1 - \exp\left(-\frac{0.5I_0 + \sum_{i=-5}^{-1} I_i + 0.65I_{-6}}{N}\right)\right). \quad (10)$$

4.2 One weak line – analytic estimates, special case

First, we make a simple example. We follow the Bayesian approach, but because of the simplicity of the problem, we need no simulations. We assume no background counts and a line with a total integrated number of counts as A and with internal Poisson noise only. Here we calculate the posteriori distribution directly from a family of Poisson distributions by evaluating the likelihood of the original measured value. We give in Table 1 the median value of the distribution with confidence limits. In the case of a very low background, these values are approximately correct for a given peak. This posteriori distribution is to a high degree identical to a full posteriori distribution with position and beam width parameters marginalized, with a non-informative prior, and with no dead-time effects taken into account.

The differences in the mode, median and mean are important, but fortunately for our case they are not of big concern as the differences are at the most 1 count. More important for us is the asymmetry of the distributions in the case of small peaks.

If background counts are present, e.g., 10 counts in addition to a line of 20 counts, then the distribution above will be the joint distribution.

4.3 Grid simulations – validation

We calculated a set of artificial one- and two-peak cases to validate our method. The exact shape of the line is not critical. For simplicity we chose a Gaussian. We selected an array of 20 time bins and drew a random position for the peak randomly from $[-5, 5]$. In the case of two peaks, we placed them both within that interval. We drew the amplitudes randomly from a uniform distribution in $\lg(N+1)$ space from $N = 0$ to $N = 9999$. The position of the peaks were drawn randomly from a uniform distribution between -10 and 10. The FWHM was fixed to 2.5 time bins.

Table 1. Posteriori distributions of a Poisson peak with a given amplitude A_{obs}, and no background noise.

A_{obs}	2	3	5	7	10	15	20	30
$A_{\mathrm{low\,99}}$	0.1	0.6	1.5	2.5	4.3	7.5	11.0	18.5
$A_{\mathrm{low\,90}}$	0.3	1.3	2.6	3.9	6.1	10.0	14.1	22.4
$A_{\mathrm{low\,68}}$	1.7	2.0	3.6	5.2	7.7	12.0	16.4	24.4
$A_{\mathrm{low\,50}}$	1.7	2.6	4.2	5.8	8.6	13.1	17.7	27.0
A_{mode}	2.0	3.0	5.0	7.0	10.0	15.0	20.0	30.0
A_{median}	2.5	3.5	5.5	7.5	10.5	15.5	20.5	30.5
$\langle A \rangle$	3.0	4.0	6.0	8.0	11.0	16.0	21.0	31.0
$A_{\mathrm{high\,50}}$	3.8	5.1	7.4	9.6	13.0	18.4	23.8	34.5
$A_{\mathrm{high\,68}}$	4.6	5.9	8.3	10.7	14.2	19.9	25.5	36.5
$A_{\mathrm{high\,90}}$	5.3	7.7	10.5	13.1	16.9	23.0	29.0	40.6
$A_{\mathrm{high\,99}}$	12.0	10.9	14.0	17.1	21.3	28.1	34.6	47.2
$\sqrt{A_{\mathrm{obs}}}$	1.4	1.7	2.2	2.7	3.2	3.9	4.5	5.5

We should note that, although the single highest probability A_{mode} is the same as the observed value, the median value has a bias of +0.5 and the mean of the distribution has an even larger positive bias of 1.0. The total width of the 68 % confidence limits agrees within roundoff errors with $\sqrt{A}$. The median point is not centered on the limits. Other low and high confidence limits are shown. Note that, as they are asymmetric, both lower and higher limits are shown for the important confidence limits. The distributions are clearly asymmetric with a positive skew.

Before our full Bayesian tests, we performed a brute force calculation. The best fit parameters were calculated at 0.1 bins in time and 20 intervals per dex in log space, i.e., a resolution of a factor $10^{1/20}$ or 1.122. In total, we calculated the likelihood in $(100 \times 60)^2 = 36\,000\,000$ separate data points per cases. One double peak model takes about 9.7 effective minutes to calculate on a desktop computer (AMD Athlon (tm) IIX(4) 630 Processor 2.79 GHz).

We calculated a total of 3300 cases with a variable background noise and different separations. The major systematic source of error is the discretization of the solutions of the data into the sub-bins. This is an effect that shows the weakness of the direct grid calculation of the probabilities.

4.4 Bayesian case of one line

A faster and more accurate estimate of the line parameters is obtained by the Bayesian method. The additional benefit is that we obtain automatically a distribution for the various parameters of the solution without having to resort to a grid calculation.

We show an example of a real line from the COSIMA full spectrum. The example shown in Fig. 2 is a relatively weak line with a total number count of about 100. The line mass is derived from the in-flight measurements of constants a and b and is caused by $^{19}F^{+}$, which originates from the fuel of the spacecraft.

The solution for a single Gaussian that gave us a mass of 19.0056 u is within 0.0077 u or about 1.1 TOF time bin of F^{+} 18.9979 u. This is slightly more than expected. This could

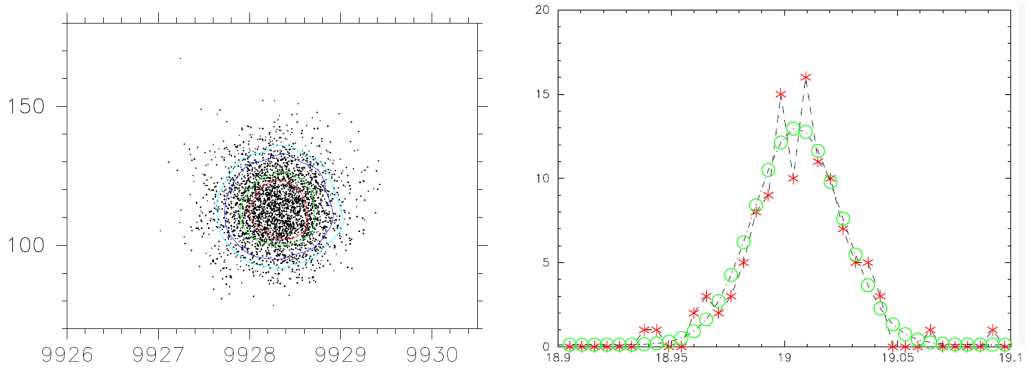

Figure 2. An example of the posteriori distribution of a weak line at mass 19. The background around the line is very low. The weak line has an observed maximum of 16 counts. The top panel shows the posteriori distributions in total count vs. time flight bin. The red curve contains 50 % posteriori confidence limits, the green curve 68 %, the dark blue 90 %, and the light blue 95 % limits. Note that the most likely value has a rather symmetric distribution with a 68 % confidence width of about 0.34 TOF time bins or 0.002 u in mass, and an integrated mean count of 113 ± 11 counts. The mass 19.0056 u is within 0.0077 u or F^+ 18.9979 u. The line does not agree quite as well with heavy water HDO^+ of mass 19.0162 u or the hydrogenated water ion (hydronium) H_3O^+ with a mass of 19.0178 u. These are off by 0.0106 and 0.0122 u, respectively, from the calculated line position. The expected hydronium line would have a time-of-flight bin of 9929.53, which is not in agreement with the posteriori distribution of the upper panel. The spectrum used here is from flight model CS_2D8_20100509T194035_SP_P.TAB.

be explained by systematic errors or small fluctuation in the acceleration voltage.

4.5 Bayesian case of two lines

A simulated two-line case is shown in Fig. 3. The two simulated Gaussian peaks have Poisson noise added to each point. The Gaussians in this simulation have FWHM 2.5 time-of-flight bins or 0.031 u at mass 100 u. Both test cases have a peak with an amplitude of 1000 and a second peak with an amplitude of 100. The total line counts are 5250 and 525, respectively. The general peak finding algorithm detects one peak, but a two-peak fit gives a better result. The x axes in Fig. 3 show the time-of-flight bin around mass 100, with 100 u equal to bin number 40 in these plots. In our simulation, one time bin corresponds to 0.0125 u. The main component has tails extending over 10 time-of-flight bins and the secondary peak is not obvious from the line shape as a strong maximum. The y axes show the total count of the line. In Fig. 3a, the peaks are centered on mass 100.000 and 100.070, respectively. This corresponds to a separation of six bins in the peak locations. The total count is 5775. The figure shows the simulation with total counts given as a function of bins. The calculated centers of the lines are 100.001 and 100.076, and the amplitudes are 988 and 107. The total counts are 5286 and 508, the sum of which, 5794, is very close to the original value. In Fig. 3b, the peaks are centered on masses 100.000 and 100.050, respectively. This corresponds to a separation of four bins in the peak locations. The Bayesian solution shows lines centered on 100.001 and 100.059; the amplitudes are 1024 and 77.4. The total counts are 5521 and 397, the sum of which is 5918, close to the original value. The posteriori probability density distributions with their complex shapes are the very property obtainable with Bayesian methods. This is a feature that could be used to complement the other methods in the seminal paper (Stephan et al., 2001).

We ran 10 000 two-line simulations and modeled them with one and two peaks and investigated which of the models were correctly identified by our Bayesian analysis. The results were quite clear cut. Two nearby peaks are not identified correctly in the presence of Poisson noise if the following limitations are met: the smaller peak has an amplitude of < 7 (or a total count of about 30), the separation is < 4.5 time bins, or the ratio of the counts of the two lines is > 1000. These limits are for general guidance only, and need to be solved separately in each case. In our present algorithm, we have a freely variable line width. With these conditions, we tested a few specific interesting pairs of lines. We are able to separate ^{26}Mg from $^{12}C_2H_2$. Other nearby pairs such as ^{13}C vs. ^{12}CH, ^{14}N vs. $^{12}CH_2$, ^{25}Mg vs. $^{12}C_2H$, and ^{24}Mg vs. $^{12}C_2$ are not separated properly at present, agreeing with the limits from a larger set of simulations. We will investigate this in subsequent papers by fixing the position, the width or the shape of the individual spectral lines. Furthermore, if the b term were larger, then the resolution would improve slightly, rendering better results, but unfortunately the mass scaling parameters are rather fixed for a given instrument.

5 Full COSIMA spectrum

5.1 Analysis of real spectra

To analyze real COSIMA spectra, we make an assumption of the line shape. We have chosen as options a Gaussian shape,

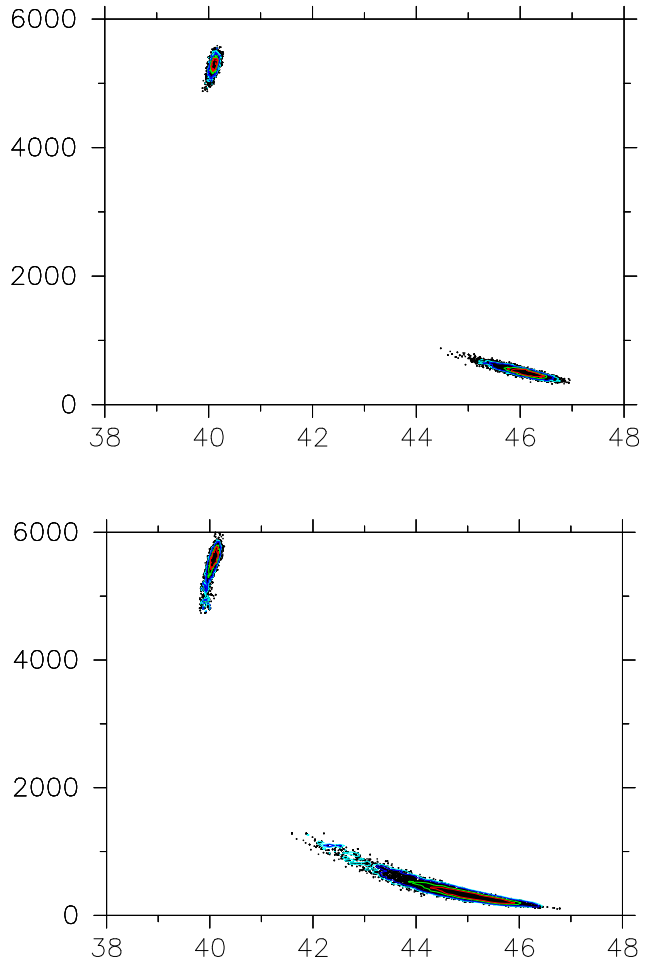

Figure 3. Two examples of the posteriori distributions of the amplitudes and positions. The two simulated Gaussian peaks have Poisson noise added to each point. The Gaussians have a FWHM of 2.5 time-of-flight bins or 0.031 u at mass 100 u. Both test cases have a peak with an amplitude of 1000 and a second an amplitude of 100. Total bin counts are 5250 and 525, respectively. In the first case, the peaks are separated by six bins and in the second simulation by four bins, respectively. The red curve contains 50 % posteriori confidence limits, the green curve 68 %, the dark blue 90 % and the light blue 95 % limits. Note that the distribution maxima are well defined and close to the initial values. Note that the distributions have low density tails that reflect the fact that there is mild degeneracy in the solution. These kinds of distributions are not found trivially by conventional analysis methods.

but on occasions a 80 % Gaussian and 20 % Lorentzian combination is an option that is suitable for modeling lines in positive COSIMA spectra. If an asymmetry of the peak develops in COSIMA for any reason, we will be able to take this into account. Negative ion spectra are more complicated as an additional signal before the main peak is created by the electrons sputtered off the grids inside the reflectron. We will not discuss negative spectra in this paper in detail.

We first estimate the line amplitude and width from the observed line. To the estimated line, we then apply the analytic dead-time correction and obtain a model line that we can compare to the observed line. Note that here we can use the information that the probability distribution of the counts follows a Poisson distribution. With the Bayesian adaptive metropolis algorithm described earlier, we can obtain the posteriori distributions of both the parameters of the original line and of the observed dead-time affected line. These will include automatically the proper line positions and amplitudes. The total counts are obtained by summing discrete counts from the continuous model curves, so the fitted amplitudes of the continuous Gaussian do not represent a real quantity, but just a mathematical aid for measuring the total count from discrete abscissa values.

If we are able to give good guesses for the initial starting points, the algorithm tends to converge faster to a good solution. This is not necessary for the method but aids in reducing the computing time considerably, particularly when estimating multiple spectral lines simultaneously.

The analysis of the lines provides a complicated challenge. Some lines are clear and isolated; often, two separate lines occur together. If they are sufficiently far apart, they can be treated as single isolated lines. Occasionally, a section occurs in the spectrum where several lines appear to be present and mixed in. Sometimes the background levels are somewhat elevated, mimicking multiple merged lines. Our approach is the following: we create a running 5 pixel boxcar sum of the spectrum of the original spectrum and find the local maximum by comparing the adjacent smoothed pixel sums. We then accept as good guesses the points where this maximum has a value that is larger than the background. The background is defined as the smallest of two background measurements. One background estimate is obtained from the 5 pixel sum 20 pixels earlier and the second background 20 pixels on the other side of the maximum. If this difference of the boxcar sum and the background sum is over a certain limit, we accept this point as a guess for a component. We have used an ad hoc lower limit of five counts. This is not a critical limit as it is only a first guess for our Bayesian analysis.

5.2 Normalization issues, and line ratios

A general normalization is often performed by dividing the count of the spectral lines by a certain constant or line, e.g., Si^+ or In^+. This is sometimes preceded by a removal of a variable background. This is fine for rough line identification, but poses a problem when confidence limits are derived for the measured values. In removing and subsequently ignoring the background, one obtains better looking spectra, but in this process, one is introducing a poorly behaving error term on top of the noise created by the "Poisson noise". Poisson noise is additive but not subtractive. Furthermore, in calculating the ratios of lines, the determination of the con-

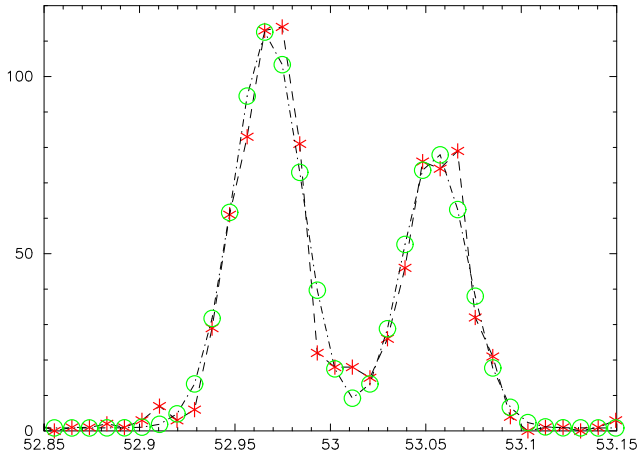

Figure 4. An example of the mass 53 spectral lines in a RM spectrum CS_45D_20110309T074148_SP_P.TAB. The model fit here is a two-Gaussian model and a constant background. The observed line is shown with the red stars and the fit with green circles. The masses derived are 52.967 and 53.055 u. The masses suggest a systematic error of $+0.022\,u$ in mass and an identification of $^{53}Cr^{+}$ at 52.9401 u and $C_4H_5^+$ at mass 53.0386 u.

fidence limits becomes formally ill behaved, as the divider, the reference line, has in principle a probability distribution that includes 0. These are serious issues when any one of the lines is a weak one. If the lines in question are strong, and in particular if the reference line is a strong one, say at least 1000 counts, and if the background level is at the same time low, say less than 50 counts, then the distortion may not be serious.

The proper way to normalize is to build a model where the line ratio is solved for. Take a guess of the stronger integrated line count, make a good guess of the background, and apply a good guess for the line ratio. You have now calculated two integrated line counts. Using the Poisson distribution, calculate what the likelihood is that the observed lines are explained by the given model. Continue with the Bayesian principles of searching for the posteriori probability distribution. Finally, marginalize (integrate) over background and amplitudes to get the likelihood of the line ratio distribution.

5.3 Line identifications and examples of specific isotope lines

The spectra are provided with an initial estimate of the mass calibration parameters, a and b. We have built a simple line identification scheme with the elemental lines and a small set of simple organic lines. These can be applied to the observed values and further improvement with a larger set of lines is possible by removing systematic trends evident in the original spectra as shown in Figs. 2 or 4.

In this study, we have considered so far all lines as independent in the sense that the background level, the line width, position and the maximum amplitude of the peak have been free parameters. However, if we wish to ask a specific question such as whether a certain mass contains lines at predefined exact masses, we can employ different variations to the analysis. For example, if we see ^{24}Mg and $^{12}C_2$ separately, we may want to ask whether mass 25 contains $^{25}Mg^{+}$, $^{24}MgH^{+}$ and $^{12}C_2H^{+}$. We can then fix the interval of the lines in mass and solve for the background level, a single mass offset, and the amplitudes (and the widths) of the three peaks. This reduces the adjustable parameter space from 11 to 9 or 6.

An additional setup can be created between the above multiline kind example and isotope ratios. Consider, e.g., the lines $^{12}C^{+}$, $^{13}C^{+}$, and $^{12}CH^{+}$. We can use a model where the ratio of $^{13}C^{+}/^{12}C^{+}$ has a cosmic value, so that it is not a free parameter. Two free parameters are the position and the amplitude of $^{12}C^{+}$. One free parameter is the amplitude of $^{12}CH^{+}$. The isotope ratio fixes the amplitude of $^{13}C^{+}$, and the mass is fixed by mass difference, so this line would have no additional free parameters. If we consider a common line shape, then we have only six free parameters for the whole model. We can build more elaborate models by finding solutions to the positions and amplitudes. If we keep a fixed carbon isotope ratio, we will have an eight-parameter model.

Investigating the full parameter space of all possible models is beyond the scope of this paper, but we wish to point out the generality of the Bayesian method. These kinds of analyses are not easy to do with conventional means, and the posteriori probability distributions in those cases are at best only guesses. We thus provide posteriori distributions and confidence limits for all measured parameters.

6 Conclusions

We have discussed the basic principles of applying a Bayesian approach to the analysis of COSIMA spectra. We address the accuracy, the fundamental principles of Bayesian analysis. We show that one is able to obtain posteriori distributions for integrated line counts, line positions, and line widths in systems of one or several lines. Even if some of the parameters may turn out to produce strong correlation or degeneracy in the solution, its severity can be characterized by Bayesian analysis.

The instrumental properties of COSIMA that simplify our analysis are the long time interval between the shots so that the secondary ion formation and flight time of the ions can be considered usually statistically independent from shot to shot. Second, the shortness of the pulse and the well-calibrated instrument means that not only each mass line but often also the organic and mineral components can be analyzed separately. Third, the dead time is relatively short and quite nicely matched with the line width, so the dead-time effects will not leak to neighboring lines. The narrow

line shape means that the spectra cannot be well modeled by a line shape derived from the spectrum.

Our analysis methods can be generalized to data with other sorts of noise properties, and nearly any kind of line shapes.

Acknowledgements. COSIMA was built by a consortium led by the Max-Planck-Institut für Extraterrestrische Physik, Garching, Germany, in collaboration with Laboratoire de Physique et Chimie de l'Environnement, Orléans, France, Institut d'Astrophysique Spatiale, CNRS/INSU and Université Paris Sud, Orsay, France, the Finnish Meteorological Institute, Helsinki, Finland, Universität Wuppertal, Wuppertal, Germany, von Hoerner und Sulger GmbH, Schwetzingen, Germany, Universität der Bundeswehr, Neubiberg, Germany, Institut für Physik, Forschungszentrum Seibersdorf, Seibersdorf, Austria, and Institut für Weltraumforschung, Österreichische Akademie der Wissenschaften, Graz, Austria, and is lead by the Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany. The support of the national funding agencies of Germany (DLR), France (CNES), Austria and Finland and the ESA Technical Directorate is gratefully acknowledged. We thank the Rosetta Science Ground Segment at ESAC, the Rosetta Mission Operations Centre at ESOC and the Rosetta Project at ESTEC for their outstanding work enabling the science return of the Rosetta Mission.

H. J. Lehto and B. Zaprudin acknowledge the support of the Academy of Finland (grant number 277375).

Edited by: M. Paton

References

Batir, N.: Very Accurate approximation for the factorial function, J. Math. Inequal., 4, 335–344, 2010.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: Bayesian Data Analysis, Chapman & Hall, London, 1995.

Glassmeier, K. H., Boehnhardt, H., Koschny, D., Kührt, E., and Richter, I.: The Rosetta mission: flying towards the origin of the Solar System, Space Sci. Rev., 128, 1–21, 2007.

Haario, H., Saksman, E., and Tamminen, J.: An adaptive Metropolis algorithm, Bernoulli, 7, 223–242, 2001.

Kissel, J., Altwegg, K., Clark, B. C., Colangeli, L., Cottin, H., Czempiel, S., Eibl, J., Engrand, C., Fehringer, H. M., Feuerbacher, B., Fomenkova, M., Glasmachers, A., Greenberg, J. M., Grün, E., Haerendel, G., Henkel, H., Hilchenbach, M., von Hoerner, H., Höfner, H., Hornung, K., Jessberger, E. K., Koch, A., Krüger, H., Langevin, Y., Parigger, P., Raulin, F., Rüdenauer, F., Rynö, J., Schmid, E. R., Schulz, R., Silén, J., Steiger, W., Stephan, T., Thirkell, L., Thomas, R., Torkar K., Utterback, N. G., Varmuza, K., Wanczek, K. P., Werther, W., and Zscheeg, H.: COSIMA – high resolution time-of-flight secondary ion mass spectrometer for the analysis of cometary dust particles onboard Rosetta, Space Sci. Rev., 128, 823–867, 2007.

Silén, J., Cottin, H., Hilchenbach, M., Kissel, J., Lehto, H., Siljeström, S., and Varmuza, K.: COSIMA data analysis using multivariate techniques, Geosci. Instrum. Method. Data Syst., 4, 45–56, doi:10.5194/gi-4-45-2015, 2015.

Stephan, T., Zehnpfenning, J., and Benninghoven, A.: Correction of dead time effects in time-of-flight mass spectroscopy, J. Vac. Sci. Technol. A, 12, 405–410, 1994.

Stephan, T.: TOF-SIMS in cosmochemistry, Planet. Space Sci., 49, 859–906, 2001.

Tamminen, J.: Adaptive Markov Chain Monte Carlo Algorithms with Geophysical Applications, Finnish Meteorological Institute Contributions No. 47, PhD Thesis, University of Helsinki, Dept. of Mathematics and Statistics, Helsinki, 2004.