PROCEEDINGS OF SCIENCE

Tools for analyzing large data-set and handling intensity variations of sources with *INTEGRAL/SPI*

L Bouchet[*1,2], **M. Chauvin**[1,2,5], **P R. Amestoy**[3] , **F.-H. Rouet**[3,6] **and A. Buttari**[4]

[1]*Université de Toulouse, UPS-OMP, IRAP, Toulouse, France*
[2]*CNRS, IRAP, 9 Av. colonel Roche, BP 44346, F-31028 Toulouse cedex 4, France*
[3]*Université de Toulouse, INPT-ENSEEIHT-IRIT, France*
[4]*CNRS-IRIT, France*
[5]*KTH Royal Institute of Technology SE-100 44, Stockholm, Sweden*
[6]*Lawrence Berkeley National Laboratory, Berkeley CA94720, USA*
E-mail: lbouchet@irap.omp.eu

The *INTEGRAL/SPI* X/gamma-ray spectrometer (20 keV-8 MeV) is an instrument for which it is essential to process many exposures at the same time to increase the low signal-to-noise ratio of the weakest sources and/or low-surface brightness extended emission. The processing of several years of data simultaneously (10 years actually) with traditional methods of data reduction is ineffective and sometimes not possible at all. Thanks to the newly developed tools, processing large data-sets from SPI is possible with both a reasonable turnaround time and low memory usage. We present also techniques that we have developed to overcome difficulties related to the intensity variation of sources/background between sources between consecutive exposures.

10th INTEGRAL Workshop: A Synergistic View of the High-Energy Sky
15-19 September 2014
Annapolis, MD, USA

*Speaker.

PoS(Integral2014)070

1. Introduction

Sky imaging, with SPI ([1, 2]), is not direct and relies on a coded-mask aperture associated to a specific observation strategy based on a dithering procedure ([3]). The dithering is needed since a single exposure does not always provide enough information or data to reconstruct the sky region viewed through the instrument $\sim 30^\circ$ field-of-View (FoV). The grouping of these exposures allows to increase the amount of available information on a given sky target through a growing set of independent data. However, sources intensity varies between exposures. Thus, a reliable modeling of sources variability, of at least the most intense ones, is needed to obtain a proper modeling of the data and an accurate measurement of the source's intensity.

2. Handling the intensity variations of sources and background

The variation in intensity of a source (or background) is modeled as a succession of piecewise constant segments of time. In each of the segments ("time-bins"), the intensity of the source is supposed stable. We consider the time-series $x \equiv x_{1:L} = (x_1, \ldots, x_L)$, comprising L sequential elements, following the model,

$$x_i \equiv f(t_i) + \epsilon_i \quad i = 1, 2, \ldots, L \tag{2.1}$$

x_i are the measured data and ϵ_i their measurement errors. The data are assumed to be ordered in time (may be evenly spaced otherwise), meaning that each x_i is associated with a time t_i, and contained in a time interval $T = (t_1, \ldots, t_L)$. $f(t_i)$ is the model to be determined. We choose to model this time-series (source light-curve) as a combination of constant piecewise time segments.

$$f = \sum_{k=1}^{m+1} s_k \mathcal{I}_k \text{ with } \begin{cases} \mathcal{I}_k = 1 & \text{if } t \in [\tau_{k-1}, \tau_k[\\ \mathcal{I}_k = 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Here $\tau_0 = \min(T)$ and $\tau_{m+1} = \max(T)$ or, equivalently, in point number units, $\tau_0 = 1$ and $\tau_{m+1} = L+1$ ($\tau_0 < \tau_1 < \ldots < \tau_{m+1}$).

2.1 Incorporating the sources/background variability into the system of equations

The relation between the data and the sky model can be expressed, schematically, as

$$y = Hx + \epsilon \tag{2.3}$$

Physically, H corresponds to the transfer function or matrix, y to the data and x to the unknown intensity (source plus background) to be determined (a vector of length N).
Let us say that the system of equations with sources of constant intensities is characterized by the matrix H_0 (matrix with N_0 columns) and the solution x_0 (vector of length N_0). Equation 2.4 illustrates schematically how to construct the matrix H when the intensities of sources vary. Each column of the matrix H_0, corresponding to the response of a given source J, is expanded into a sub-matrix with K_J columns. There are K_J intensities or parameters to determine for source J

instead of 1. The final system H is sparser with more unknowns.

$$H_0 = \begin{pmatrix} h_{11} & h_{12} & h_{13} & ... & h_{1N} \\ h_{21} & h_{22} & h_{23} & \ddots & h_{2N} \\ h_{31} & h_{32} & h_{33} & \ddots & h_{3N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ h_{M1} & h_{M1} & h_{M3} & ... & h_{MN} \end{pmatrix} \tag{2.4}$$

$$\longmapsto H = \begin{pmatrix} h_{11} & 0 & 0 & 0 & h_{12} & 0 & 0 & h_{13} & 0 & h_{1N} & ... & 0 \\ 0 & h_{21} & 0 & 0 & h_{22} & 0 & 0 & 0 & h_{23} & h_{2N} & ... & 0 \\ 0 & 0 & h_{31} & 0 & 0 & h_{32} & 0 & 0 & h_{33} & 0 & ... & h_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & h_{M1} & 0 & 0 & h_{M2} & 0 & h_{M3} & 0 & ... & h_{MN} \end{pmatrix}$$

Yet, to allow all the sources to vary on the exposure time-scale is not the appropriate strategy because the problems to solve, for crowded regions of the sky, is again in most cases undetermined. Generally, to deduce coarsely the variability time-scale of sources, a crude and straightforward technique consists in testing several time-scale values until the reduced chi-square, of the associated least-square problem, is around 1 or stops to decrease. However, when defining manually the "time-bins", one might be rapidly overwhelmed with the many time-scales to test and the number of sources. Furthermore, the of modeling the intensity variation of sources turns out to be rather subjective and irksome. To make this step more rational, we propose two methods based on a partition of the data.

3. "Image-space" method

The "image-space" method relies on some already available light-curves (or time-series) that could come either from SPI itself or from another instrument; in our application mainly *INTEGRAL/IBIS* ([4]), but also Swift/BAT ([5]). The purpose is to simplify an original time-series by minimizing the number of constant segments necessary to its description, hence the number of "time-bins". Those "time-bins" will be used to setup the SPI system of equations. This partitioning is done, individually, for all the sources in the FoV. Hence, the basic process to set up "time-bins" characteristics (start, end) is the time series segmentation or partition.

3.1 Partition of a time-series basics

The partition of an interval into segments is closely related to the topic of change-points, widely discussed in the literature. Hence, there are m change-points defining $m+1$ segments, such that the function $f(t)$ is constant between two successive changes-points.
[6] and [7] have proposed a dynamic programming algorithm to explore all the 2^{L-1} possible partitions (See Eq. 2.2). These authors proposed a search method that aims at minimizing the following function (see also [8].

$$\min_{\tau} \left\{ \sum_{i=1}^{m+1} [C(x_{(\tau_{i-1}+1):\tau_i}) + \beta] \right\}. \tag{3.1}$$

The i-th segment contains the data subset $x_{(\tau_{i-1}+1):\tau_i} = (x_{\tau_{i-1}+1}, \ldots, x_{\tau_i})$ (in point number units) and the cost function or effective fitness for this data block is $C(x_{(\tau_{i-1}+1):\tau_i})$. The negative log-likelihood (the chi-square) is a commonly used cost function in the change-point literature. β is a penalty to prevent over-fitting. These authors propose a convenient recursive expression to build the partition in L passes,

$$F(n) = \min_{\tau^*} \{F(\tau^*) + C(x_{(\tau^*+1):n}) + \beta\} \text{ n} = 1, \ldots, \text{L} \tag{3.2}$$

This expression enables to calculate the global optimal segmentation using optimal segmentations on subsets of the data. Once the optimal segmentation for the data subset $x_{1:\tau^*}$ has been identified, it is used to infer the optimal segmentation for data $x_{1:\tau^*+1}$.
The optimal partition of the data is found in $O(L^2)$ evaluations of the cost function.

3.2 Application

We construct few training data-sets. A SPI data-set consists of all the exposures whose angular distance, between the telescope pointing axis and the source of interest direction (central source) is less than 15°. This procedure gathers the maximum number of exposures containing the signal from the source of interest, but at the same the data-set span a $\sim 30°$ radius FoV sky region containing numerous sources. As example, we apply the "image-space" algorithm to the 27-36 keV data-set related to GX 339-4 source. For this study, we rely mainly on IBIS existing light curves in the form of time-series. To illustrate our purpose, we use only the exposures common to SPI and IBIS. The data-set contains 1183 exposures and the sky model has 120 sources. The source 4U 1700-377 is assumed to vary on the exposure timescale (~2700 s) and the instrumental background on ~6 hours. Assuming that all the other sources are constants in intensity gives a final χ^2_r of 2.46 for 19 308 degrees of freedom (dof). which is not acceptable.

3.2.1 Step 1: Segmentation of an existing time-series

The time-series related to all the sources of the sky model are not available, but at least those for strong sources. Each available IBIS time-series is segmented to define the "time-bins"characteristics. To have roughly similar signal-to-noise ratio per sources between IBIS and SPI random Gaussian statistical fluctuations are added to the time series below 100 keV since IBIS is more sensitive than SPI at these energies. Figure 1 shows the application to GX 339-4 time-series.

3.2.2 Step 2: Application to SPI using pre-defined "time-bins"

The SPI related system equations is formed using pre-defined "time-bins" (Equation 2.4) and solved. The χ^2_r, between the data and its model, is 1.186 (18 880 dof). This is clear improvement compared to the previous value of 2.46.

4. "Data-space" method

SPI data contain the variable signal of several sources. The contribution, through a transfer function, of each of the sources to the data, is to be retrieved. For each of the sources, the number

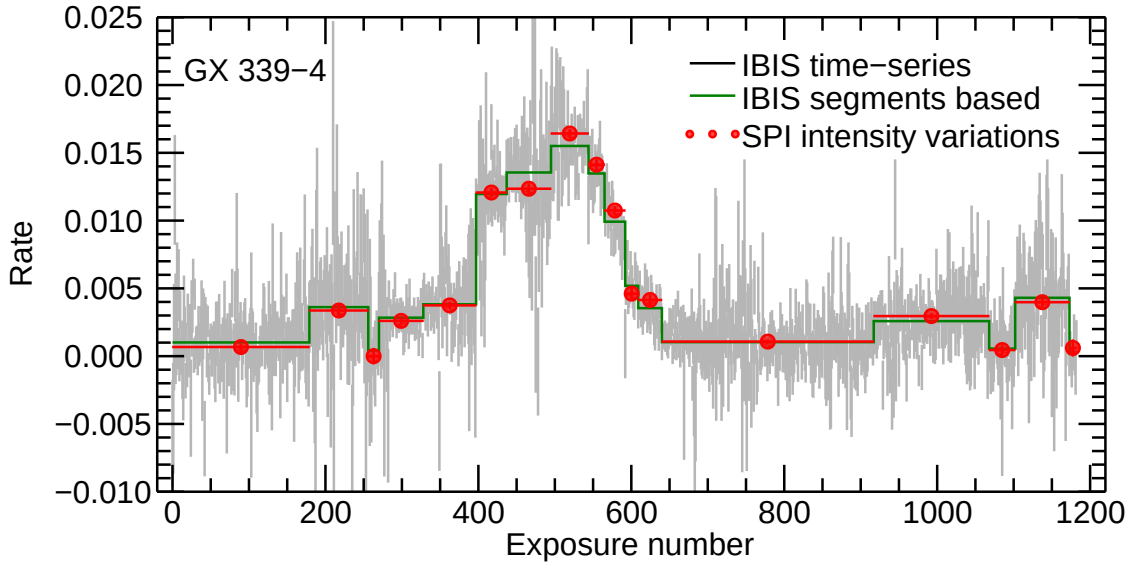

Figure 1: The 26-40 keV IBIS time-series (after scaling of the signal-to-noise ratio) is shown in gray. It contains 1183 data points (one measurement per exposure). This time-series is segmented into 17 constant segments or "time-bins" (green). The reduced-chi-square between the time-series and its segmented version is 1.0006 for 1166 dof. The SPI flux in these segments is shown in red.

and position of segments are parameters to estimate, but the estimates are interdependent because the nature the instrument coding. In short, the system of equations 2.4 is to be formed directly using only the data and the transfer function of the instrument. For this purpose, we developed a specific algorithm. While being more complex, it has the great advantage to be based solely on SPI data.

We formulate the problem to follow as closely as possible the scheme described by eqs. 3.1 and 3.2. Then, we make some simplifications and/or approximations to reduce the computation time and to permit important optimizations of the code, hence to render the problem tractable ([10]). The most important ones concern the search path and the computation of the cost function. Rather than exploring the space for all the sources simultaneously, we explore the reduced space associated to a single source at once and sequentially. Hence, it is straightforward to parallelize the code and processing several sources simultaneously.

In addition, the cost function must be computed many times, each time that a new partition is tested. This is by far, the most time-consuming part of the algorithm since it requires to solve a system of equations to obtain the least-square solution. Fortunately, these calculations can be optimized, at a given iteration, say (n (eq. 2.4), since the different partitions involve matrices, which can be deduced one from each other by suppressing and adding new columns. Therefore, only one decomposition of the matrix ([10]) is needed and the solution is updated.

4.1 "Image-space" versus "data-space" methods

The "time-bins" obtained with the "data-space" are compared to those obtained with the

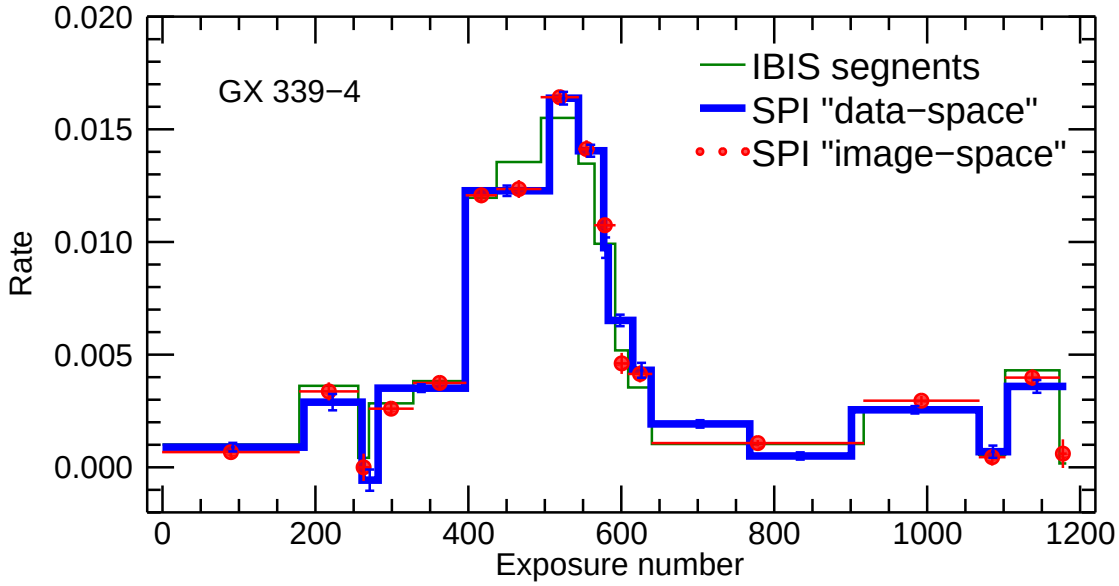

Figure 2: Comparison of GX 339-4 (27-36 keV) intensity variations obtained with the "image-space" and the "data-space" algorithms. The common SPI/IBIS database contains 1183 exposures. The "image-space" method describes GX 339-4 intensity variations with 17 segments (Red) for a χ^2_r of 1.19. The "data-space" method uses 15 segments (Blue) and achieves χ^2_r of 1.20. The GX 339-4 segmented version of the IBIS (26-40 keV) time-series is shown in green.

"image-space" algorithm, using GX 339-4 data-set (Fig. 2). The comparison is done on the SPI and IBIS common 1183 exposures and the sky model consists of 120 sources. This demonstrates the effectiveness of the "data-space" algorithm which is as efficient as the "image-space" algorithm, and better takes into account the SPI sensitivity. Note that, the background is modeled with only 87 segments with the "data-space" algorithm instead of the 286 segments needed with a ~ 6 hours time scale.

4.2 Application

An application to the IGR J17464-3213 data-set is shown on figure 3). The IGR J17464-3213 data-set corresponds to the central, crowded, region of the Galaxy. The sky model consists of 132 sources and the background timescale is fixed to ~6 hours. The data-set is relatively large (7147 exposures) and is artificially split into three subsets to reduce the computation time. The intensity variations of the central source, IGR 17464-3213, is modeled with 29 segments. We have also extracted the intensity evolutions of GRS 1758-258, GX 1+4. GS 1826-24 and GX 354-0 which are derived simultaneously (Fig. 3).

5. Handling of large *INTEGRAL/SPI* data-sets

Processing several years of data simultaneously requires computing not only the solution of a large system of equations, but also the associated uncertainties. We aim at reducing the compu-

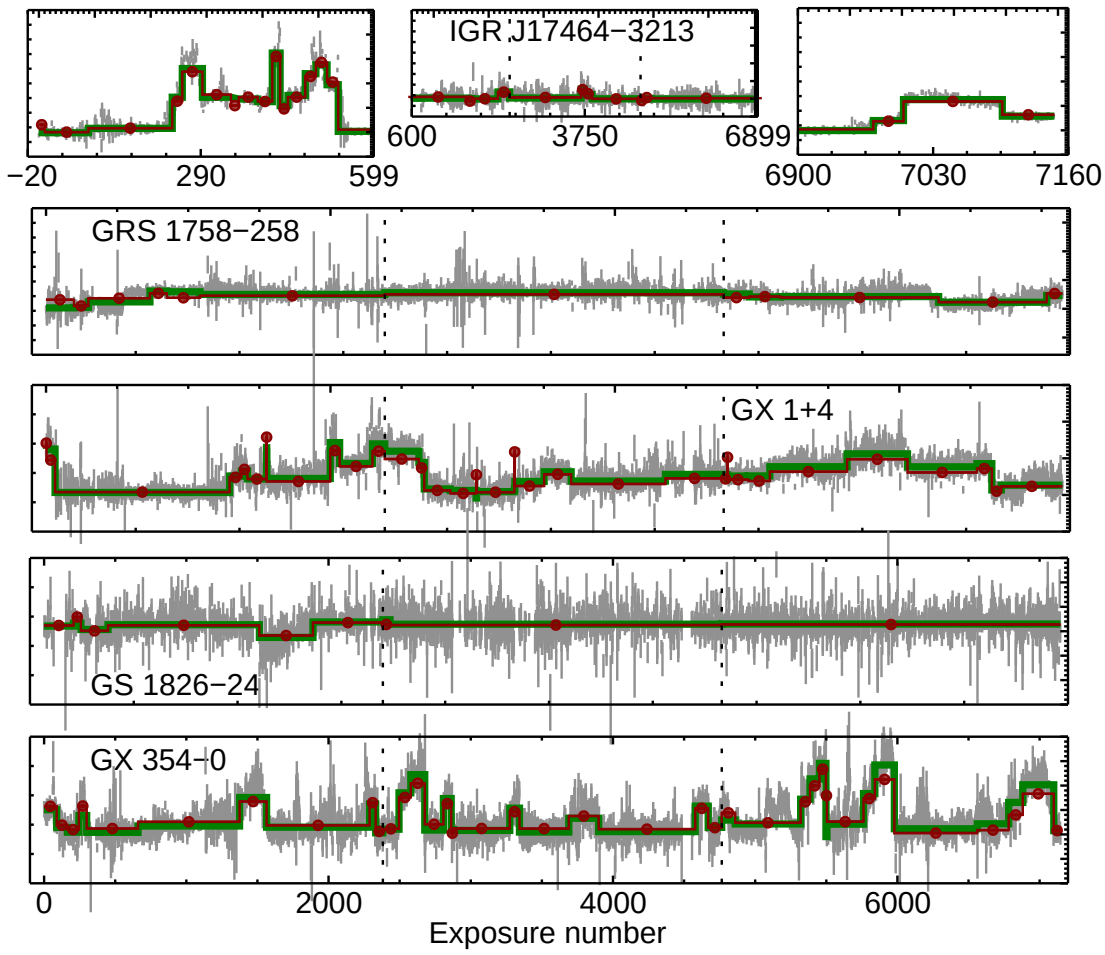

Figure 3: Intensity evolutions (Red) of IGR J17464-3213, GRS 1758-258, GX 1+4, GS 1826-24 and GX 354-0 in the 27-36 keV band. The IGR J17464-3213 data-set is divided in 3 subsets (dashed vertical lines show the limits) during the computation. The averaged fluxes in SPI "time-bins" are in green. The IBIS time-series (30-40 keV) is shown in gray and the green curve corresponds to IBIS flux averaged on SPI computed segments. The distance of GRS 1758-258, GX 1+4, GS 1826-24 and GX 354-0 to IGR J17464-3213 are respectively 7.3, 8.1, 12.7 and 3.4°.

tation time and the memory usage. Since the SPI transfer function is sparse, we have used some popular methods for the solution of large sparse linear systems. The problem can be reduced to solving a linear system with a square matrix : A linear least-squares problem $\min_x \|Hx - y\|$ can be transformed into a square system $Ax = b$ by use of normal equations (Here, ($A = H^T H$ and $b = H^T y$). In the following the term matrix referred to ($A = H^T H$. The characteristics of some of the sample matrices are described in Table 1. We use the MUltifrontal Massively Parallel Solver (MUMPS[1]) to compute the solution of the system of equations. The experiments were carried out

[1]Up-to-date copies of the MUMPS (MUltifrontal Massively Parallel Solver) package can be obtained from the Web pages: http://mumps.enseeiht.fr/ or http://graal.ens-lyon.fr/MUMPS). The software provides a stable and reliable factors

Table 1: Sparsity of the test matrices.

N	Sparsity (%)	Usage
3578	2.96	Central Galaxy (27-36 keV) - 6 years
9437	1.05	
22503	0.28	Diffuse emission (25-50 keV) - 6 years
55333	0.09	
149526	0.04	Simulation (25-50 keV)

The sparsity is the ratio between the number of non-zero elements in the matrix and the total number of elements in the matrix N^2 for the square matrix A.

on Intel I7-3517U processor with 8 GB main memory machine. For the sake of this comparison, all these methods are executed in sequential mode although the codes are parallel.
The results, in Table 2, confirm that sparse, direct solvers achieved a good scalability on the problems of our target application whereas dense linear algebra kernels quickly exceed the limit of modern computing platforms. For the largest problems in Table 2, the dense algorithm cannot be used as the memory requirements are roughly 23 GB and 167 GB respectively. We can extrapolate that on this system, the run time would be around 22 h for the largest problem (instead of 6.7 s using a sparse algorithm). We also need to compute the variance of the solution, which amounts to

Table 2: Times (in seconds) for the computation of the solution.

Matrix size	3578	9437	22503	55333	149526
Sparse	0.2	0.7	1.6	8.0	6.7
Dense	1.2	20.1	169.9	/	/

computing selected entries of the inverse of the sparse matrix corresponding to our linear system. This can be achieved through one of the latest features of the MUMPS software that has been partly motivated by this work. We present experimental results related to the computation of error bars or, equivalently, of the diagonal entries of the inverse matrix A^{-1} (Table 3). As a second term of comparison we also provide experimental results for a brute force approach with no exploitation of sparsity of the right-side and solution vectors. For this purpose, we use directly the MUMPS package and solve several systems of equations in order to compute the inverse matrix. In addition, we analyze the influence of grouping the computation of the diagonal entries (1 right-hand-side (RHS) or 128 at a time). More details can be found in [9].

6. Summary

The imaging properties of SPI rely on the coded mask aperture, but also on a dithering

and can process indefinite symmetric matrix

Table 3: Time (in seconds) to compute the diagonal elements of the inverse of a symmetric matrix.

Matrix size	3578	9437	22503	55333	149526
Left-looking	28.2	376.1	2567.9	489.1	/
MUMPS (1 RHS)	3.77	38.4	204.1	1324.9	8230.5
MUMPS (128 RHS)	1.32	7.34	45.5	245.6	2833.5
MUMPS A^{-1}	0.28	0.9	4.9	36.0	9.5

observation strategy. With only 19 pixels, the SPI detector does not provide enough data to correctly construct and sample the sky image viewed through the aperture of ∼30° FoV. The dithering technique solves this critical imaging problem, by permitting the accumulation of independent data on a given sky region, but at the same time, raises important issues of data reduction and image/data combination through variability of sources. We propose two algorithms to model the intensity variation of sources in the form of combination of piecewise segments of time during which a given source exhibits a constant intensity.
A first method ("image-space") uses existing time series to build segments of time during which a given source exhibits a constant intensity. This auxiliary information is incorporated into the "image-space"system of equations to be solved. The main weakness of this method is that it requires, in most cases, information from other instruments and hence depends both on these instruments characteristics (FoV, sensitivity, ...) and on the level of processing performed on these available external data.
A second, called "data-space" method, determines these segments from SPI data directly and does not suffer from dependence on external data. The dependence across segments, through the transfer function, greatly increases the complexity the problem. We have developed a novel algorithm to handle this problem and made optimizations that accelerate the computations.
Both algorithms allow to introduce more objective parameters, here the "time-bins", in the problem to be solved. They permit to construct an improved sky model which better fit the data and to optimize the signal-to-noise ratio of the sources. For our purposes, these algorithms solve a specific difficulty of SPI data processing, which is the variability of sources during observations.
We have shown that, for processing efficiently and accurately years of data, it is critical to use algorithms that take advantage of the sparse structure of the transfer function (matrix), such as those implemented in the MUMPS software. It was also demonstrated that error bars can be obtained at a relatively inexpensive cost (the same order of magnitude as a simple problem solution) thanks to a recently developed algorithmic feature that efficiently computes selected entries of the inverse of a matrix.

References

[1] Vedrenne, G., Roques, J. P., Schonfelder, V., et al., A&A, 411, L63, 2003

[2] Roques, J. P., Schanne, S., Von Kienlin, A., et al., A&A, 411, L91, 2003

[3] Jensen, P. L., Clausen, K., Cassi, C., et al., A&A, 411, L7, 2003

[4] Ubertini, P., Lebrun, F., Di Cocco, G., et al., A&A, 411, L131, 2003

[5] Barthelmy, S. D., Barbier, L. M., Cummings, , J. R., et al., Space Sci. Rev., 120, 143–164, 2005

[6] Yao, Y., The Annals of Statistics, 12, 1434–1447, 1984

[7] Jackson, B., Sargle, J. D., Barnes et al., IEEE, Signal Processing Letters, 12 (2), 105, arXiv:math/0309285, 2005

[8] Killick, R., Fearnhead, P. & Eckley, I. A., 2011arXiv1101.1438K, 2011

[9] Bouchet, L., Amestoy, P., Buttari et al., Astronomy and Computing, Volume 1, 59, 2013a

[10] Bouchet, L., Amestoy, P., Buttari et al., A&A, 555, 52, 2013B