PROCEEDINGS OF SCIENCE

Boolean Factor Analysis of Swift GRB Data

Zsolt Bagoly*[ab], Lajos G. Balázs[cd], István Horváth[b], Gábor Tusnády[e], József Kóbori[a], Dorottya Szécsi[a] and Attila Mészáros[f]

[a] *Dept. of Physics of Complex Systems, Eötvös University, H-1117 Budapest, Pázmány P. s. 1/A, Hungary*
[b] *Dept. of Physics, Bolyai Military University, H-1581 Budapest, POB 15, Hungary*
[c] *Dept. of Astronomy, Eötvös University, H-1117 Budapest, Pázmány P. s. 1/A, Hungary*
[d] *MTA CSFK Konkoly Observatory, H-1525 Budapest, POB 67, Hungary*
[e] *Rényi Institute of Mathematics, Hungarian Academy of Sciences, H-1364 Budapest, POB 127, Hungary*
[f] *Faculty of Mathematics and Physics, Astronomical Institute, Charles University, V Holešovičkách 2, CZ 180 00 Prague 8, Czech Republic*
E-mail: zsolt.bagoly@elte.hu

Up to 24 Feb 2012 Swift has triggered on 638 GRBs. In this work we use the pattern of the missing gamma, X-ray and optical data measured by BAT, XRT, UVOT and ground based measurements of the redshift, collected for the Swift GRBs ([1]).

Using the Boolean factor analysis of multivariate statistical methods we studied the missing data patterns of the gamma, X-ray and optical observed quantities of GRBs. We found that the measured gamma properties have some impact on the missing data in the X-ray and optical domains. The missing data pattern depends, however, on random effects as well.

Gamma-Ray Bursts 2012 Conference -GRB2012,
May 07-11, 2012
Munich, Germany

*Speaker.

PoS(GRB 2012)039

1. Introduction

The Swift satellite made a major break-through in the simultaneous detection of gamma, X-ray and optical properties of GRBs [2]. The burst alert, given by the BAT on board of the satellite, is followed by slewing over the target. A significant fraction of GRBs, however, remains undetected by the XRT and UVOT.

The failure of detection in these energy regimes can have different reasons. The obvious reason for the failure is the faintness of the signal in comparison to the detection limit of XRT or UVOT. In a considerable number of cases the slewing is blocked the Moon and/or the Sun and the detection can happen only after a considerable time. Normally, these detections are also treated as missing in the further statistical analysis. The redshift is measured by ground based facilities following the positions given by the Swift, assuming the necessary optical brightness and access to the necessary telescope time (in some cases only the host is measured). The missingness of the data can have also information about the astrophysical nature of the objects. Boolean factor analysis dealing with binary data is a way to use this kind of information (1 - detected, 0 - undetected).

2. Mathematical Summary

Boolean factor analysis is dealing with dichotomous (binary) data. Its goal is similar to the classical factor analysis: to represent p variables ($X = x_1, x_2 \dots x_p$) by m factors ($F = f_1, f_2 \dots f_m$), where m is significantly smaller than p. n is the number of observations (cases). In this kind of factor analysis, the used arithmetic is Boolean, so the scores and loadings are binary. The basic model is $X = F \times A$, where X is the $n \times p$ data matrix, F is the $n \times m$ factor scores matrix and A is the $m \times p$ matrix of factor loadings. One can get negative or positive discrepancies between the observed and predicted values. The positive discrepancy occurs when the observed score is one but the analysis estimates it to be zero, and in the case of the negative discrepancy the observed score is zero but the estimated value is one.

In the present analysis we used the 8M module of the BMDP statistical package[3].

3. Boolean factor analysis of Swift Data

For performing the analysis we used the missing data pattern of the Swift GRB Table. We used 11 variables of this Table (duration, fluence, peak flux, photon index, early X-ray flux, 24 hour X-ray flux, X-ray decay index, X-ray spectral index, X-ray hydrogen column density, visual magnitude and redshift). The column “Response” in Table 1 summarizes the missing data pattern ("yes" - detected, "no" - undetected). The most complete part of the data is the gamma energy domain. It is not surprizing because the triggering of the event is produced by the BAT.

The algorithm attempts to minimize the number of discrepancies between the observed and predicted values of the variables. At the end 6 factor were resulted. The binary pattern of these factors is given in the last six columns of the Table 1. These factors themselves do not represent necessarily observed missing data patterns of individual cases. The missing data pattern of the individual cases (burst events) proceeds from the linear combination of these factors using the factor scores obtained also from the analysis. The cases can be partitioned into groups (clusters)

	Response		%	Discrepancy		Factor					
	no	yes		neg	pos	F1	F2	F3	F4	F5	F6
T_{90}	41	508	93	12	2	1	0	1	1	1	0
Flu	25	524	95	1	7	1	0	1	1	1	0
$Peak$	39	510	93	9	1	1	0	1	1	1	0
$Pind$	24	525	96	0	7	1	0	1	1	1	0
X_{flu}	223	326	59	0	0	0	0	0	0	0	1
X_{24}	292	257	47	2	2	0	1	0	0	0	0
X_{Dec}	182	367	67	0	8	0	0	0	0	1	0
X_{SP}	153	396	72	1	1	0	1	1	0	1	0
X_{NH}	161	388	71	8	0	0	1	1	0	1	0
V	415	134	24	15	4	0	0	1	0	0	0
z	399	150	27	0	0	0	0	0	1	0	0

Table 1: The Swift GRBs' missing data pattern.

	$Pind$		$\log T_{90}$		$\log Flu$		$\log Peak$		N
	Mean	σ	Mean	σ	Mean	σ	Mean	σ	
$cl1$	1.44	0.51	1.27	0.92	0.94	0.75	0.11	0.78	125
$cl2$	1.53	0.51	1.29	0.78	0.98	0.62	0.21	0.50	86
$cl3$	1.61	0.45	1.51	0.76	1.01	0.64	0.12	0.40	179
$cl4$	1.58	0.46	1.56	0.68	1.20	0.66	0.30	0.57	106
Total	1.55	0.48	1.42	0.80	1.03	0.68	0.17	0.57	496

Table 2: The k-means clusters' means and standard deviations of the quantities measured by BAT.

according to the factor scores indicating which of the 6 factors is necessary to describe the missing data pattern of a particular case. The distance between two particular case is given by the total number of discrepancies in their factor scores:

$$d_{ij} = \sum_{k=1}^{m} |f_{ik} - f_{jk}|$$

Since the value of f is 1 or 0 the distance given above is also the squared Euclidean distance.

4. K-means clustering of the data

We used k-means clustering to define centers in the parameter space of the factor sores to partitione the cases into groups where every particular case is assigned to the closest center with respect to the squared Euclidean distance. We obtained the optimum number of clusters by minimizing the Bayesian Information Criterion

BIC = sum of squared distances within groups + number of parameters x log(number of cases)

Fig. 1 demonstrates the BIC value for different number of clusters: the optimum is 4 clusters.

Table 2 gives the means and standard deviations of the quantities measured by BAT. According to the F-test the cluster means are significantly different: for the $Pind$, $\log T_{90}$, $\log Flu$ and $\log Peak$ the significances are $0.027, 0.007, 0.021$ and 0.035 respectively.

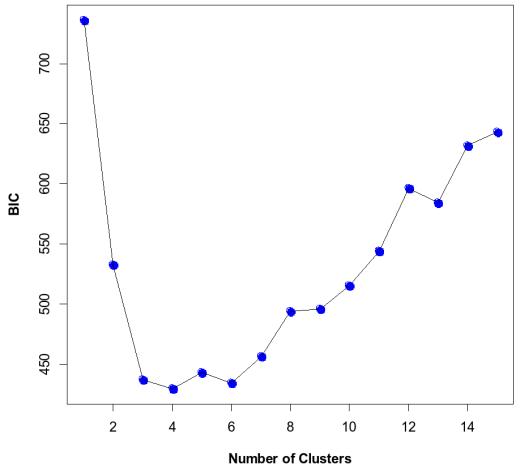

Figure 1: Optimum number of clusters in the k-means clustering.

	T_{90}	Flu	$Peak$	$Pind$	X_{flu}	X_{24}	X_{Dec}	X_{SP}	X_{NH}	V	z
$cl1$	1	1	1	1	0	0	0	0	0	0	0
$cl2$	1	1	1	1	0	1	1	1	1	1	0
$cl3$	1	1	1	1	1	1	1	1	1	0	0
$cl4$	1	1	1	1	1	0	1	1	1	1	1

Table 3: The typical missing data pattern of the different clusters.

5. Summary of the analysis

Table 3 summarizes the typical missing data pattern of the different clusters. In $cl1$ only the gamma data are recorded, it has the smallest mean fluence and peak flux. In the contrary, $cl4$ has recorded gamma, X-ray and optical data and it has the largest mean fluence and peak flux.

One may conclude therefore that the fluence and peak flux are the major factors in defining the missing data patterns of the X-ray and optical data of the Swift GRB Table. Looking for a bias-free database these result emphasize the importance of the analysis of the different observational effects near the treshold.

This work was supported by OTKA grant K077795, by OTKA/NKTH A08-77719 and A08-77815 grants (Z.B.), by the Grant Agency of the Czech Republic grant P209/10/0734 (A.M.), and by the Research Program MSM0021620860 of the Ministry of Education of the Czech Republic (A.M).

References

[1] `http://swift.gsfc.nasa.gov/docs/swift/archive/grb_table`

[2] S. D. Barthelmy *et al.*, *The Burst Alert Telescope (BAT) on the SWIFT Midex Mission*, *Space Science Reviews* **120** (143–164)

[3] `http://www.statistical-solutions-software.com/BMDP-documents/BMDP-8M.pdf`