Boolean Factor Analysis of Swift GRB Data

Zs. Bagoly[3], L.G. Balázs[1,3] , I. Horváth[2], A. Mészáros[4], J. Kóbori[3], D. Szécsi[3]

[1]MTA CsFK Konkoly Observatory, Budapest, Hungary; [2]Bolyai Military University, Budapest, Hungary;
[3]Eötvös University, Budapest, Hungary; [4]Charles University, Prague, Czech Republic;

E-mail: balazs@konkoly.hu

Abstract

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

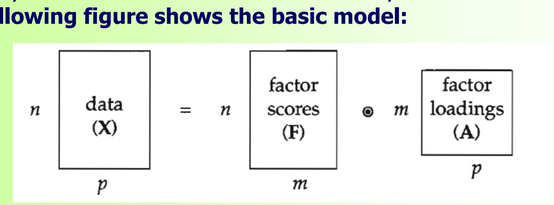
Introduction

The Swift satellite made a major break-through in the simultaneous detection of gamma, X-ray and optical properties of GRBs. 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).

In this work we use the missingness pattern of the γ , X-ray and optical data measured by BAT, XRT, UVOT and ground based measurements of the redshift, collected in the Swift GRB Table which is available at URL location (http://swift.gsfc.nasa.gov./docs/swift/archive/grb_table)

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.* In this kind of factor analysis, the used arithmetic is Boolean, so the scores and loadings are binary. The following figure shows the basic model:



where *A* is the matrix of factor coefficients (loadings) and *n* is the number of observations (cases). 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 (see References).

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, Hydrogen column density, visual magnitude and redshift). The column „Responses" in the following Table 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.

Result of the Boolean factor analysis

Var.	Response		%	Discrepancy		Factor					
	no	yes		neg	pos	F1	F2	F3	F4	F5	F6
T90	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
XFLU	223	326	59	0	0	0	0	0	0	0	1
X24	292	257	47	2	2	0	1	0	0	0	0
XDEC	182	367	67	0	8	0	0	0	0	1	0
XSP	153	396	72	1	1	0	1	1	0	1	0
XNH	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

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. 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) 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.

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). As the following figure demonstrates the optimum number of the cluster is *4*.

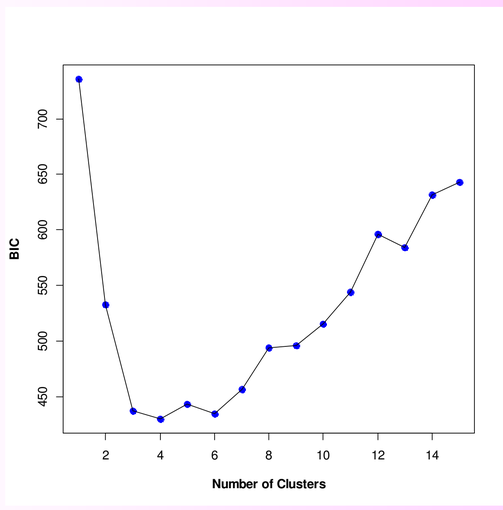


Fig 1. Optimum number of clusters in the k-means clustering.

The following Table gives the means and standard deviation of the quantities measured by BAT, the most complete part of the Swift GRB table:

Group Statistics

Cluster No.		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
1	Pind	1.4418	.51321	125	125.000
	logT90	1.2726	.92093	125	125.000
	logFl	.9372	.75197	125	125.000
	logP	.1060	.77673	125	125.000
2	Pind	1.5336	.50730	86	86.000
	logT90	1.2866	.77534	86	86.000
	logFl	.9803	.61638	86	86.000
	logP	.2114	.50353	86	86.000
3	Pind	1.6060	.45217	179	179.000
	logT90	1.5100	.76190	179	179.000
	logFl	1.0147	.64228	179	179.000
	logP	.1216	.40417	179	179.000
4	Pind	1.5799	.46289	106	106.000
	logT90	1.5599	.68267	106	106.000
	logFl	1.2010	.66157	106	106.000
	logP	.2979	.56534	106	106.000
Total	Pind	1.5465	.48318	496	496.000
	logT90	1.4221	.79936	496	496.000
	logFl	1.0290	.67603	496	496.000
	logP	.1709	.57174	496	496.000

The differences in cluster means is significant according to the following Table:

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Pind	.982	3.090	3	492	.027
logT90	.975	4.131	3	492	.007
logFl	.980	3.275	3	492	.021
logP	.983	2.900	3	492	.035

Summary of the analysis

The following Table 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. The *cl2* and *cl3* are between

Boolean properties of variables within clusters

..Var	. cl1	cl2	cl3	cl4
T90	1	1	1	1
FLU	1	1	1	1
PEAK	1	1	1	1
PIND	1	1	1	1
XFLU	0	0	1	1
X24	0	1	1	0
XDEC	0	1	1	1
XSP	0	1	1	1
XNH	0	1	1	1
V	0	1	0	1
z	0	0	0	1

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.

References

http://www.statistical-solutions-software.com/BMDP-documents/BMDP-8M.pdf

Acknowledgements

This research was supported by OTKA grant K77795