IL NUOVO CIMENTO Vol. 28 C, N. 3 Maggio-Giugno 2005
DOI 10.1393/ncc/i2005-10045-1

Principal-Component Analysis of Gamma-Ray Bursts' Spectra(*)

Z. Bagoly(1), I. Horváth(2), L. G. Balázs(3), L. Borgonovo(4), S. Larsson(4), A. Mészáros(5) and F. Ryde(4)

(1) *Laboratory for Information Technology, Eötvös University - H-1117 Budapest Pázmány P. s. 1./A, Hungary*
(2) *Department of Physics, Bolyai Military University - H-1456 Budapest, POB 12, Hungary*
(3) *Konkoly Observatory - H-1525 Budapest, POB 67, Hungary*
(4) *Stockholm Observatory, AlbaNova - SE-106 91 Stockholm, Sweden*
(5) *Astronomical Institute of the Charles University - V Holešovičkách 2 CZ-180 00 Prague 8, Czech Republic*

(ricevuto il 23 Maggio 2005; pubblicato online il 20 Settembre 2005)

Summary. — The principal-component analysis is a statistical method, which lowers the number of important variables in a data set. The use of this method for the bursts' spectra and afterglows is discussed in this paper. The analysis indicates that three principal components are enough among the eight ones to describe the variablity of the data. The correlation between the spectral index α and the redshift suggests that the thermal emission component becomes more dominant at larger redshifts.

PACS `98.70.-f` – Unidentified sources of radiation outside the Solar System.
PACS `98.70.Rz` – γ-ray sources; γ-ray bursts.
PACS `98.80.-k` – Cosmology.
PACS `01.30.Cc` – Conference proceedings.

1. – Introduction

The principal-component analysis (PCA) is a powerful statistical method, if the observational data contain several variables [1]. If there are some correlations among the variables, then PCA lowers the number of variables needed for the description of the dataset without any essential loss of generality. It introduces the quantities called principal components (PCs). Then one can omit some PCs not leading to essential loss of information about the dataset (for more details about PCA see [1]).

In the topic of gamma-ray bursts (GRBs) the paper [2] used PCA on the dataset of the BATSE Catalog [3]. This Catalog contains nine measured variables in the gamma-band alone (three different peak fluxes, the four fluences in the four different energy channels, and the two durations) for any GRB. Bagoly *et al.* [2] have shown that of these nine

---

(*) Paper presented at the "4th Workshop on Gamma-Ray Burst in the Afterglow Era", Rome, October 18-22, 2004.

variables only three are also enough (one of the durations, a combination of the fluences and the peak fluxes, and the fluence on the highest energy channel alone).

As far as it is known, since 1998 no further PCA was applied on any GRBs' datasets. This is remarkable, because for the dozens of GRBs further observational quantities are also known from the analyses of spectra and from the afterglow (AG) data. Thus, the usefulness of such generalized PCA is a triviality. This contribution presents the preliminary results of a new PCA, in which more variables than in [2] are included.

In Jochen Greiner's list [4], up to September 30, 2004, there are 38 GRBs with known redshift z: for GRB980329 and GRB011030X there are upper limits only, and in one case z is queried (GRB980326). These 41 GRBs define the primary sample here.

For every GRB in our sample we will consider 8 variables: z, α, β, AG_{break}, E_{peak}, $\mathrm{AG}\alpha$, T_{90} and fluence. The last two variables are the usual T_{90} durations and fluences used already in the BATSE Catalog. The indices α and β are the slopes in Band's spectrum, E_{peak} defines the photon energy, where the Band spectrum $EF(E)$ peaks ($F(E)\mathrm{d}E$ is the fluence from the infinitesimal photon energy interval $E, (E+\mathrm{d}E)$). These three variables define the spectrum of GRBs. AG_{break} is the observed time (in days), when the power law decay of AG changes into a faster decay; the slope of the decay before this AG_{break} is denoted as $\mathrm{AG}\alpha$. (This slope is usually denoted also as α, but using this notation there would be a confusion with the spectrum's slope α.)

All this means that we have taken only two variables from BATSE Catalog (the T_{90} duration and fluence), which in essence defined the two PCs in [2]. The third PC of BATSE Catalog, namely the fluence at the highest energy channel, will not be considered here, because for this channel the data are often vanishing and/or noisy. Our main aim here is to obtain information and conclusions about the six remaining new quantities, not about the BATSE values themselves; hence, it is better not to consider here the fluence at the highest energy channel.

There are 16 GRBs from the mentioned 41 GRBs that have all these variables in Greiner's list. These 16 GRBs define the net sample for our PCA. For fluence, T_{90}, AG_{break} (days), $1+z$ and E_{peak} we will take the logarithms of the measured values.

2. – Principal-Component Analysis

PCA needs, first, the calculation of correlations among the variables. The following pairs (with the corresponding probabilities p) give strong correlations: $\log T_{90}$-$\log(1+z)$ ($p = 66.49\%$), $\log(1+z)$-α ($p = 99.75\%$), $\log \mathit{Fluence}$-$\log T_{90}$ ($p = 95.49\%$), $\log E_{peak}$-$\log T_{90}$ ($p = 97.98\%$) and $\log E_{peak}$-$\log \mathit{Fluence}$ ($p = 99.72\%$). All other correlations are weaker.

The second step in PCA is to calculate the eigenvalues and eigenvectors. PCA gives us the eigenvalues and the eigenvectors of the "best" projection (table I).

Then the third step in PCA is to define the important PCs. In accordance with [1] the important PCs are that ones, which have eigenvalues above 1.0. From table I we obtain the result that only 3 PCs are important. This is a highly remarkable conclusion: we have 8 variables, and their reduction to 3 is an essential lowering. In addition, two PCs were expected already from the earlier paper [2], because these two PCs were given by the BATSE data alone. In other words, the 6 new quantities—not used already in [2]—define only one further new important PC.

As is seen from table I, the first three eigenvectors (Comp1-3) are the linear combination of $\log E_{peak}$, $\log T_{90}$, $\log \mathit{Fluence}$, $\log(1+z)$, α, $\log \mathrm{AG}_{break}$ and $\mathrm{AG}\alpha$. No variable alone is dominating in these PCs. All this means that, beyond the two PCs coming

Table I. – *PCA result.*

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
$\log(1+z)$	0.345	−0.359	−0.399	0.301	0.278	0.348	−0.538
α	0.355	−0.394	−0.283	0.336	−0.392	−0.188	0.497
β	−0.342	0.016	0.394	0.798	0.226	0.128	0.098
$\log \mathrm{AG}_{break}$	−0.225	0.473	−0.560	0.023	0.132	0.135	−0.013
$\mathrm{AG}\alpha$	0.381	−0.142	0.484	−0.285	0.217	0.377	0.051
$\log E_{peak}$	−0.427	−0.285	−0.145	−0.173	−0.218	0.716	0.313
$\log T_{90}$	−0.316	−0.498	−0.154	−0.212	0.640	−0.349	0.213
$\log$ *Fluence*	−0.400	−0.378	0.120	−0.055	−0.442	−0.181	−0.555
Eigenvalue	4.314	1.761	1.224	0.464	0.189	0.038	0.009

from the BATSE Catalog, there is only one further important PC, and this one is a complicated linear combination of all variables. In other words: Although the correlations among the variables are generally weak, these correlations are enough to ensure a coupling among the variables.

The $\log E_{peak} + \log T_{90} + \log$ *Fluence* correlation is well known. This correlation in essence ensures that $\log E_{peak}$ alone is not independent. The correlation $\log(1+z)$ and α also ensures that the redshift alone is also *not* independent—a quite remarkable behavior. (As is seen on fig. 1, the positive correlation between $\log(1+z)$ and α is obvious, because the trend—increasing of z with increasing α—is well seen. Contrary to this, no such trend is seen between z and β.) Similarly, the remaining correlations are enough to lower the number of important variables.

Note that, interestingly, β stands practically alone in the fourth PC (Comp4). Hence, the only variable, which can define *alone* a PC, is the quantity β. But, in accordance with [1], this PC has already an eigenvalue smaller than 1.0.

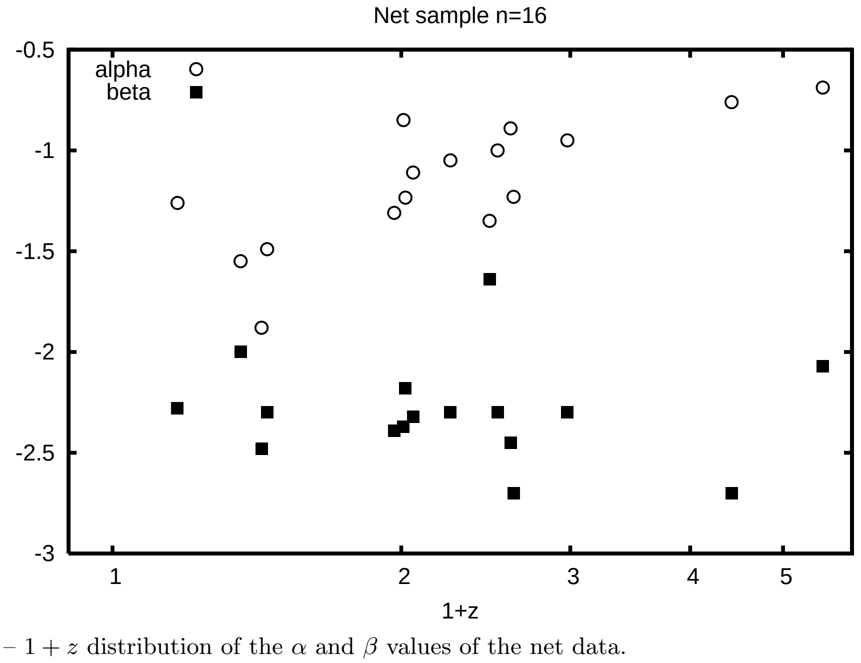

Fig. 1. – $1+z$ distribution of the α and β values of the net data.

3. – Discussion

The importance of the PCA is given also by the fact that it uses exclusively the observational data, and any model of GRBs must respect the reduction of important variables. Here we briefly discuss the consequences of PCA on the models of GRBs.

The prompt emission in GRBs can in most cases be described by various models of optically thin synchrotron emission. Theoretical considerations suggest that, apart from this non-thermal emission, a thermal emission component, from the photosphere of the outflow, could be equally strong in the gamma-ray band [5, 6]. In [7] it was found that the observed spectra can indeed be described by a thermal black-body superimposed on a non-thermal spectrum. The relative strengths of these two components vary. A large α is measured if the thermal component is dominant, while a lower value (more like a synchrotron value) is found if the non-thermal component dominates. The variations in α therefore could suggest a variation in importance of the photosphere.

Note that α values are measured on the *integrated* spectra. These are necessarily softer than the instantaneous spectra [8]. It is the instantaneous spectrum that reflects the physical radiation process shaping the spectrum. We therefore have that $\alpha_{rad.proc.} > \alpha_{int.}$. The strong correlation between α and $1 + z$ therefore suggests that the thermal component becomes more dominant at larger redshifts. In [5] it was showed that a combination of the variation time scale at the central engine and the dimensionless entropy of the outflow determines this relation. Low variability at the central engine combined with large Lorentz factors should therefore have been dominant at large redshifts. [9] also found this α-*vs.*-$1 + z$ correlation. They explained it in terms of the $E_{peak} \propto E_{rad}^{0.5}$ correlation, where E_{rad} is the total emitted energy of GRBs.

4. – Conclusions

The PCA method indicates that three PCs are enough to describe the variability of the data. Hence, a reduction of the eight variables to three is strongly suggested. The strong correlation between α and $1 + z$ suggests that the thermal emission component becomes more dominant at larger redshifts.

* * *

Thanks are due to the valuable discussions with C.-I. Björnsson, I. Csabai, C. Fransson, J. P. Uldall Fynbo, P. Mészáros and G. Tusnády. This research was supported through OTKA grants T034549 and T48870, and by a grant from the Wenner-Gren Foundations (A.M.).

REFERENCES

[1] Jolliffe I.T., *Principal Component Analysis* (Springer) 1986.
[2] Bagoly Z. *et al.*, *ApJ*, **498** (1998) 342.
[3] Meegan C.A. *et al.*, `http://www.gammaray.msfc.nasa.gov/batse`.
[4] Greiner J., `http://www.mpe.mpg.de/~jcg/grbgen.html`.
[5] Mészáros P. and Rees M., *ApJ*, **541** (2000) L5.
[6] Daigne F. and Moshkovich R., *MNRAS*, **342** (2000) 587.
[7] Ryde F., *ApJ*, **614** (2004) 827.
[8] Ryde F. and Svensson R., *ApJ*, **512** (1999) 639.
[9] Amati L. *et al.*, *A&A*, **390** (2002) 81.