

A PRINCIPAL COMPONENT ANALYSIS OF THE 3B GAMMA-RAY BURST DATA

Z. BAGOLY,¹ A. MÉSZÁROS,^{2,3} I. HORVÁTH,⁴ L. G. BALÁZS,³ AND P. MÉSZÁROS⁴

Received 1997 July 11; accepted 1997 December 11

ABSTRACT

We have carried out a principal component analysis for 625 gamma-ray bursts in the BATSE 3B catalog for which nonzero values exist for the nine measured variables. This shows that only two out of the three basic quantities of duration, peak flux, and fluence are independent, even if this relation is strongly affected by instrumental effects, and these two account for 91.6% of the total information content. The next most important variable is the fluence in the fourth energy channel (at energies above 320 keV). This has a larger variance and is less correlated with the fluences in the remaining three channels than the latter correlate among themselves. Thus a separate consideration of the fourth channel and an increased attention paid to the related hardness ratio $H43$ appear useful for future studies. The analysis gives the weights for the individual measurements needed to define a single duration, peak flux, and fluence. It also shows that, in logarithmic variables, the hardness ratio $H32$ is significantly correlated with peak flux, while $H43$ is significantly anticorrelated with peak flux. The principal component analysis provides a potentially useful tool for estimating the improvement in information content to be achieved by considering alternative variables or for performing various corrections on available measurements.

Subject heading: gamma rays: bursts — methods: data analysis — methods: statistical

1. INTRODUCTION

Extensive databases on gamma-ray burst (GRB) properties such as the BATSE 3B catalog (Meegan et al. 1996) contain a wealth of statistical information. However, to translate that into useful knowledge about the physics of GRBs requires a significant amount of interpretational effort. In its simplest form, the 3B archive, and foreseeably also its later updates, contain nine entries per event, including several different definitions of three basic physical measurements, duration, fluence, and peak flux. One of the principal questions that must be asked is how many of these entries are truly important. Another question is, which subset or combination of these nine quantities contains the maximum amount of nonredundant information. In other words, what is the number of significant physical quantities responsible for the observed variables. Numerous analyses of the 3B catalog and its predecessors have been made (e.g., Kouveliotou et al. 1993; Lamb, Graziani, & Smith 1993; Mitrofanov et al. 1994; Norris et al. 1994; Briggs et al. 1995; Norris et al. 1995; Loredo & Wasserman 1995; Emslie & Horack 1995), using different techniques. These papers have looked at various statistical properties of the bursts within a broader context. However, none of them appears to have investigated the above questions with a method specifically designed to answer them. In this paper we address these questions specifically, within the framework of the principal component analysis (PCA), which is particularly suited for this task. The PCA is a well-known statistical method (e.g., Morrison 1967; Jolliffe 1986; Murtagh & Heck 1987) that has wide applications in engineering, artificial intelligence, geophysics, biophysics, and also in some areas of astronomy (e.g., Connolly et al. 1995; Balázs et al. 1996). However, to

our knowledge, this method has not so far been used to any significant extent in the field of GRBs (a partial exception is Mukherjee, Feigelson, & Babu 1997). We show that the PCA can provide useful insights into the statistical properties of the variables in the 3B catalog that could simplify the investigation of the physical nature of GRBs.

Broadly speaking, there are three main benefits to be gained from an analysis using the principal components of a data set. First, without any essential loss of information, it allows one to reduce the total number n of observed variables to a lower value $m < n$, which can be considered to be statistically uncorrelated with each other (i.e., one can obtain a smaller number m of “highly significant” variables). Second, it identifies among the remaining $(n - m)$ variables a smaller number that may contain some degree of further information. Finally, it allows one to obtain, in some sense, information about the character of parameters of the sources themselves (e.g., Connolly et al. 1995; Balázs et al. 1996).

The nine entries of the 3B database for each GRB consist of two durations, T_{50} and T_{90} , which contain 50% and 90% of the burst counts, respectively (Kouveliotou et al. 1993; Koshut et al. 1996); four fluences (time-integrated energy fluxes) $F1$, $F2$, $F3$, $F4$, defined over different energy channels; and three measures of the peak flux (each summed over the four energy channels), measured over three different resolution timescales (64, 256, and 1024 ms). Thus the initial number of variables for the PCA is $n = 9$. There is, of course, some incompleteness in the catalog in that not all nine quantities are available for all the GRBs in the catalog. There are several possible ways to deal with this problem in statistical studies (e.g., Jolliffe 1986, p. 219). However, here we will not address the incompleteness issue, choosing instead to use a subset of GRBs for which all nine entries are nonzero. There are 625 such GRBs in the 3B catalog, and the PCA is done here on these. In § 2 a PCA is done for the subspace of the four fluences only, to address the question of their independence and to probe the information contained in related quantities such as hardness ratios. In § 3 we perform a PCA of the full $n = 9$ variables in the

¹ Eötvös University, Laboratory for Information Technology, Múzeum krt. 6-8, Budapest, H-1088, Hungary.

² Department of Astronomy, Charles University, 150 00 Prague 5, Švédská 8, Czech Republic.

³ Konkoly Observatory, Budapest, Box 67, H-1525, Hungary.

⁴ Department of Astronomy and Astrophysics, Pennsylvania State University, 525 Davey Lab, University Park, PA 16802.

catalog. In § 4 we discuss and summarize the results. In Appendix A a similar statistical method called factor analysis is also briefly presented, completing the results of § 3.

2. PRINCIPAL COMPONENT ANALYSIS OF THE FOUR FLUENCES

A PCA for the fluences in four channels is straightforward, using the method of the correlation matrix described in Jolliffe (1986). A PCA by means of the correlation matrix is the default option of a number of statistical packages. The motivation for using correlations instead of covariances is based on the fact that the observed variables might have very different scales. The correlation uses variables normalized with standard deviations and mean values, thus overcoming the scale disparity problem. This is important for GRBs, which have durations ranging over 6 orders of magnitude. Also, throughout this paper we will deal with the logarithms of the quantities involved. This leads to a linear functional dependence between two quantities in the case when one quantity is proportional to an arbitrary power of another quantity. This is a reasonable assumption in our case, since a number of models of GRBs predict a power-law relationship between the observable quantities (see Nemiroff 1994; Paczyński 1995). For this reason, the use of logarithms seems to be a justified procedure. (We note, however, that we have also performed the entire analysis directly on the quantities themselves, as opposed to on their logarithms, and the basic conclusions remain unchanged.) The correlation matrix for the logarithms of the four fluences is immediately calculable from the BATSE 3B catalog and is presented in Table 1.

The correlation matrix is symmetric and is written, as usual, in a “triangle form” (see, e.g., Jolliffe 1986, p. 34). The values in Table 1 are straightforward: for example, 0.97 in the first row and second column is the correlation coefficient between the logarithms of the fluences in the first and second channels, etc. The correlation coefficients are calculated with the classical Pearson’s formula (Press et al. 1992), the diagonal matrix elements being always unity (Jolliffe 1986). The values in Table 1 indicate that there are extremely strong correlations among the fluences of the first three channels. On the other hand, the fourth channel is obviously less correlated with the remaining ones, although the degree of correlation is still significant. In Table 1 and in the other tables with correlation matrices in this paper, the symbol “†” indicates a $\geq 99.9\%$ probability that the two quantities considered are correlated, while the symbol “+” indicates a correlation probability of between 99% and

99.9%. (The probability of the existence of correlation is calculated using equation [14.5.2] from Press et al. 1992.)

The four variables $\log Fi$ constitute a four-dimensional vector space, and the principal components are, in essence, the eigenvectors of the correlation matrix in this space, i.e., they determine orthogonal directions. Usually, the principal components are also unit vectors (i.e., the sum of the squares of the four coefficients is 1), but one can multiply them with an arbitrary nonzero constant without any loss of generality. The relative weights of the four variables are proportional to the coefficients, e.g., if the coefficients are equal (in four dimensions they are 0.5), then the four variables are equally important (they have the same weight). To calculate the principal components (hereafter PCs) one may use, e.g., the singular-value decomposition algorithm (a numerical routine is available in Press et al. 1992) or the iteration method described in Morrison (1967). The results are shown in Table 2.

From Table 2 (first row), the first PC is given by the following linear combination of the four basic variables used here: $0.51 \log F1 + 0.52 \log F2 + 0.52 \log F3 + 0.43 \log F4$. This first PC accounts for $\sim 86\%$ of the total statistical information (which is 100 % when the four PCs are taken into account). One can see that the first PC is a unit vector in which the weights of the four $\log Fi$ are almost identical. However, the remaining three principal components (the next three rows of Table 2) are more complicated combinations of the four $\log Fi$. For instance, the second PC accounts for $\sim 12\%$ of the total statistical information content in this space, and it is given approximately by $-\log F1 - \log F2 + 2 \log F4$. This second PC is dominated by $\log F4$, owing to the much larger weight given to it. We also see that, in essence, the relative importance of the second PC comes from the fact that $F4$ does not correlate as strongly with the remaining three fluences as do these among themselves. Another way to look at the second PC is as a quantity involving hardness ratios related to $F4$, i.e., $-\log F1 - \log F2 + 2 \log F4 = \log F4^2 / (\log F1F2) = \log H41H42$, where $Hij = Fi/Fj$ ($i, j = 1, 2, 3, 4$), which is a hardness ratio. The hardness ratio $H32$ is more generally used than the other simple ratios such as $H21$, etc., in discussions of the GRB data (e.g., Kouveliotou et al. 1993). For completeness it is necessary to consider six different hardnesses, of which only three are independent (e.g., $H43, H32, H21$). The remaining three hardnesses are obtainable from them ($H42 = H43H32$; $H41 = H42H21$; $H31 = H32H21$). We see that the product of two hardnesses is a PC, i.e., it does not correlate with the remaining three PCs.

We note that the BATSE 3B fluences have associated errors, which are listed in the catalog (Meegan et al. 1996). They are calculated by the BATSE group, taking into account both systematical and statistical effects. The sizes of these errors are sometimes large, and there are large varia-

TABLE 1
CORRELATION MATRIX BETWEEN THE LOGARITHMS OF THE FLUENCES

| Quantity | log F1 | log F2 | log F3 | log F4 |
|-------------|--------|--------|--------|--------|
| log F1..... | 1 | 0.97† | 0.90† | 0.62† |
| log F2..... | | 1 | 0.94† | 0.65† |
| log F3..... | | | 1 | 0.75† |
| log F4..... | | | | 1 |

NOTE.—Correlation matrix between the logarithms of the fluences in the four channels for 625 GRBs in the 3B catalog. The † symbol (usually written as ** in the statistical literature) indicates that the quantities in that row and column are correlated with a probability $\geq 99.9\%$.

TABLE 2
PRINCIPAL COMPONENTS OF THE LOGARITHMS OF THE FOUR FLUENCES

| Percentage ^a | log F1 | log F2 | log F3 | log F4 |
|-------------------------|--------|--------|--------|--------|
| 85.75 | 0.51 | 0.52 | 0.52 | 0.43 |
| 11.75 | -0.37 | -0.32 | -0.04 | 0.87 |
| 2.00 | 0.56 | 0.06 | -0.78 | 0.23 |
| 0.50 | 0.53 | -0.79 | 0.31 | -0.05 |

^a Percentage of total variation along each particular PC.

TABLE 3
CORRELATION MATRIX OF THE LOGARITHMS OF THE TWO DURATIONS, FOUR FLUENCES, AND THREE PEAK FLUXES

| Quantity | log T_{50} | log T_{90} | log $F1$ | log $F2$ | log $F3$ | log $F4$ | log P_{64} | log P_{256} | log P_{1024} |
|----------------------|--------------|--------------|----------|----------|----------|----------|--------------|---------------|----------------|
| log T_{50} | 1 | 0.97‡ | 0.80‡ | 0.78‡ | 0.71‡ | 0.44‡ | -0.16‡ | -0.01 | 0.29‡ |
| log T_{90} | | 1 | 0.83‡ | 0.81‡ | 0.74‡ | 0.48‡ | -0.11‡ | 0.05 | 0.35‡ |
| log $F1$ | | | 1 | 0.97‡ | 0.90‡ | 0.62‡ | 0.30‡ | 0.44‡ | 0.69‡ |
| log $F2$ | | | | 1 | 0.94‡ | 0.65‡ | 0.35‡ | 0.49‡ | 0.73‡ |
| log $F3$ | | | | | 1 | 0.75‡ | 0.47‡ | 0.60‡ | 0.80‡ |
| log $F4$ | | | | | | 1 | 0.46‡ | 0.53‡ | 0.63‡ |
| log P_{64} | | | | | | | 1 | 0.97‡ | 0.84‡ |
| log P_{256} | | | | | | | | 1 | 0.93‡ |
| log P_{1024} | | | | | | | | | 1 |

NOTES.—For 625 GRBs in the 3B catalog. A ‡ means that the probability of the existence of correlation (or for a negative sign, the existence of anticorrelation) is greater than 99.9%, while † means that this probability is between 99% and 99.9%.

tions among them. In order to determine the impact of these errors on our analysis we used Monte Carlo simulations. For every burst we obtained new $F1$, $F2$, $F3$, $F4$ fluences, chosen randomly out of a distribution around the original 3B values, the size of these distributions being determined by the specific 3B errors listed for the burst considered. We generated 100 such new data sets and then repeated the whole data analysis procedure for each sample. These simulations show that neither the correlations in Table 1 nor the PCs in Table 2 change by more than 0.3% for different realizations of the errors (see also Appendix B). The reason for this is that the spread of the fluences around the mean (the standard deviation of the F_i) is much larger than either the mean error or the standard deviation of the errors in the F_i . (For some individual bursts, but by no means for all, the errors can be of the order of the fluence, but this is not the case for the averages.) This means that the impact of the listed 3B errors on our PCA calculations is small. (We also note here that Monte Carlo simulations of the data errors were done for all of the other correlation matrices and PCAs presented in this article. In all cases the changes implied by such simulations were not in excess of 1%. Hence, the impact of the errors seems to be unimportant throughout, at least for the present purposes.)

In the present case we have $n = 4$ variables in our vector space, and we ask what the highly significant number $m \leq n$ is. There are different criteria for the definition of m (see Jolliffe 1986). For example, in accordance with Jolliffe’s rule (Jolliffe 1972), m is given by the number of principal components that explain more than $(70/n)\%$ of the variations. This cutoff level is 17.7% in our case, which is clearly not fulfilled for the second PC. Hence $m = 1$, according to

this criterion. Nonetheless, while the third and fourth PCs clearly can safely be assumed to be “fully unimportant,” the second PC, which accounts for $\approx 12\%$ of the variation, cannot be considered “negligible.”

Summarizing this section’s results, one can say that as a rough approximation, most of the information in the four logarithmic fluences is contained in the first PC, which is the sum of the logarithms of the fluences in the four channels. In a second more precise approximation, the information can be represented by two important quantities or PCs, which are (a) the sum of the logarithms of the four fluences, and (b) the logarithm of the fluence in the fourth channel.

We note that the same is true (with similar levels of probability of correlation) when the PCA is done for the 3B fluences themselves, rather than than for their logarithms. In this case the first PC is in essence the total fluence $F = F1 + F2 + F3 + F4$, while the second meaningful PC is $F4$.

3. PRINCIPAL COMPONENT ANALYSIS WITH NINE VARIABLES

In this section we perform a PCA on the 625 bursts in the 3B catalog for which all nine quantities (T_{50} , T_{90} , four fluences, and the peak fluxes on three triggers) are nonzero. Again, we use as our basic vector space the logarithms of the quantities, rather than the quantities themselves. The correlation matrix is given in Table 3. The PCs of this nine-dimensional space and the percent variation involved in each of them are shown in Table 4.

From Tables 3 and 4 we see that the first PC is again roughly given by the sum of all nine logarithmic quantities (durations, fluences, and peak fluxes), with some extra

TABLE 4
PRINCIPAL COMPONENTS OF THE LOGARITHMS OF THE TWO DURATIONS, FOUR FLUENCES, AND THREE PEAK FLUXES

| Percentage ^a | log T_{50} | log T_{90} | log $F1$ | log $F2$ | log $F3$ | log $F4$ | log P_{64} | log P_{256} | log P_{1024} |
|-------------------------|--------------|--------------|----------|----------|----------|----------|--------------|---------------|----------------|
| 64.8 | 0.29 | 0.31 | 0.39 | 0.39 | 0.40 | 0.32 | 0.22 | 0.28 | 0.35 |
| 26.8 | -0.44 | -0.41 | -0.16 | -0.13 | -0.04 | 0.07 | 0.53 | 0.47 | 0.30 |
| 5.1 | -0.07 | -0.07 | -0.19 | -0.18 | 0.03 | 0.93 | -0.09 | -0.14 | -0.19 |
| 1.5 | -0.48 | -0.44 | 0.53 | 0.42 | 0.11 | 0.04 | -0.20 | -0.24 | -0.11 |
| 0.8 | -0.08 | -0.13 | -0.48 | 0.05 | 0.82 | -0.16 | -0.16 | -0.10 | 0.03 |
| 0.5 | -0.12 | 0.01 | -0.04 | -0.13 | -0.15 | 0.06 | -0.62 | 0.04 | 0.75 |
| 0.2 | -0.68 | 0.71 | 0.00 | -0.04 | 0.06 | -0.02 | 0.05 | -0.02 | -0.08 |
| 0.2 | -0.03 | 0.05 | -0.53 | 0.77 | -0.34 | 0.06 | 0.03 | -0.06 | 0.05 |
| 0.1 | -0.02 | -0.01 | -0.01 | 0.10 | 0.01 | 0.00 | -0.46 | 0.78 | -0.41 |

^a Percentage of total variation in each PC.

weight placed on the first three fluences. Because of the different dimensions involved, it has only a formal meaning. The second PC, accounting for 26.5% of the variation, is clearly important, the value being far above Jolliffe's $70/9 = 7.8\%$ cutoff level. This PC is roughly given by the formal difference of the logarithmic peak fluxes and durations. (The sum would be along the direction of the fluence, but the difference is orthogonal, as expected for different PCs.) This means that the duration, peak flux, and total fluence are undoubtedly important quantities, but only two of them are independent. The third PC, practically defined by F_4 alone, accounts for 5.1% of the variation and is already below Jolliffe's level. Hence, $m = 2$. However, because the third PC is just below Jolliffe's level, it might still be of some importance. (This is examined in greater detail in Appendix A.) The fourth PC, at 1.5%, is far below Jolliffe's limit, and its importance should be even more questionable.

It is essential in discussing the PCA of this section to consider also the importance of some instrumental effects. As seen from Table 3, it is interesting that the duration is anticorrelated with the peak flux on 64 ms, noncorrelated on 256 ms, and positively correlated on 1024 ms. In fact, there is controversy among different authors in this respect, since Norris et al. (1994) and (1995) see an anticorrelation on 256 ms, but Mitrofanov et al. (1994) do not. These are problematic questions, and we think that instrumental effects should play a significant role in this behavior. For example, there are strong grounds for arguing that the correlation between the peak flux on 1024 ms and the duration should have an instrumental origin (Lee & Petrosian 1996, 1997). The same situation should occur also for the large correlation between fluences and durations (Lee & Petrosian 1997). Fortunately, these ambiguities do not change the conclusion that there are two important PCs, and that the duration itself is an independent quantity.

One can also, of course, use other quantities as the original variables of the vector space. For instance, we have also performed the same analysis with the quantities themselves, rather than with the logarithms, and the results are essentially similar. There appears to be no general rule for preferring logarithms over the quantities themselves. Taking logarithms has some numerical advantages when dealing with quantities that vary by many orders of magnitude.

Another possibility is, instead of using the four fluences (or their logarithms), to take their ratios (hardnesses). This is possible because there is a 1:1 correspondence between

the four fluences and the four new variables defined by the total fluence and three independent hardnesses. [One has then $F = F_1 + F_2 + F_3 + F_4$, $H_{21} = F_2/F_1$, $H_{32} = F_3/F_2$, and $H_{43} = F_4/F_3$, and these four new variables are defined unambiguously by the four original ones. The inverse is $F_1 = F/(H_{21} + H_{21}H_{32} + H_{21}H_{32} + H_{43}H_{32}H_{21})$, the calculation of the remaining three fluences being obvious.] It is interesting to calculate the correlation matrix among these new quantities (again taking logarithms). These correlation coefficients are presented in Table 5.

The correlation matrix of Table 5 shows several things. First, it is clear that all the hardnesses are anticorrelated (to $\geq 99.9\%$ significance) with the durations. This fact is, of course, not new, because, e.g., in Kouveliotou et al. (1993), the same anticorrelation between H_{32} and T_{90} is also presented for 222 GRBs. Second, it seems also that the hardnesses are not correlated with the total fluence. This result is in principle expected from the discussion in § 2. (In the PCA for the subspace of the four fluences by themselves, we obtained that the total fluence F and the product of H_{41} and H_{42} are PCs, and hence they should not correlate. This also suggests that the individual hardnesses themselves should not correlate strongly with total fluence.) Furthermore, the hardness ratio H_{32} is significantly correlated ($\geq 99.9\%$) with the peak fluxes P_{64} and P_{256} , but interestingly, the hardness ratio H_{43} is *anticorrelated* with the peak flux P_{1024} , also with $\geq 99.9\%$ significance.

A computation of the PCs corresponding to Table 5 shows that the first PC (34% variation) is dominated by the peak flux, the second PC (30%) by the duration (both with contributions from the fluence), and the third PC (15%) by H_{43} .

4. DISCUSSION AND CONCLUSIONS

We have carried out a principal component analysis (PCA) with the nine variables describing 625 GRBs in the BATSE 3B catalog. The results of this analysis may be summarized as follows.

1. A PCA for the $n = 9$ variables identifies a subset of $m = 2$ important variables, i.e., two principal components (PCs) are unambiguously important, when Jolliffe's criterion is applied. These are constructed out of the fluence, peak flux, and duration, implying that only two of the three are independent. This means that in the roughest approximation, it is enough to consider, e.g., a total fluence and a duration, and that these two represent 91.6% of the infor-

TABLE 5
CORRELATION MATRIX FOR LOGARITHMS OF THE DURATIONS, TOTAL FLUENCE, HARDNESSES, AND PEAK FLUXES

| Quantity | $\log T_{50}$ | $\log T_{90}$ | $\log F$ | $\log H_{21}$ | $\log H_{32}$ | $\log H_{43}$ | $\log P_{64}$ | $\log P_{256}$ | $\log P_{1024}$ |
|-----------------------|---------------|---------------|----------|---------------|---------------|---------------|---------------|----------------|-----------------|
| $\log T_{50}$ | 1 | 0.97‡ | 0.65‡ | -0.23‡ | -0.38‡ | -0.33‡ | -0.16‡ | -0.01 | 0.29‡ |
| $\log T_{90}$ | | 1 | 0.68‡ | -0.23‡ | -0.38‡ | -0.33‡ | -0.11‡ | 0.05 | 0.35‡ |
| $\log F$ | | | 1 | -0.08 | -0.06 | 0.03 | 0.47‡ | 0.58‡ | 0.76‡ |
| $\log H_{21}$ | | | | 1 | 0.24‡ | -0.02 | 0.16‡ | 0.11‡ | 0.03 |
| $\log H_{32}$ | | | | | 1 | 0.28‡ | 0.23‡ | 0.17‡ | 0.01 |
| $\log H_{43}$ | | | | | | 1 | 0.02 | -0.05 | -0.18‡ |
| $\log P_{64}$ | | | | | | | 1 | 0.97‡ | 0.84‡ |
| $\log P_{256}$ | | | | | | | | 1 | 0.93‡ |
| $\log P_{1024}$ | | | | | | | | | 1 |

NOTE—Correlation matrix for the logarithms of the two durations; the total fluence $F = F_1 + F_2 + F_3 + F_4$; the hardnesses H_{21} , H_{32} , H_{43} ; and the peak fluxes for 625 GRBs in the 3B catalog. A ‡ means that the probability of the existence of correlation (or for a negative sign, the existence of anticorrelation) is greater than 99.9%, while † means that this probability is between 99% and 99.9%.

mation content in the 3B catalog. While in an ideal case the above dependence is obvious from the definition of these quantities, it is not clear that it should continue to hold for sources in a flux-limited sample with complicated light curves, located at increasing distances and subject to complicated detection biases. However our analysis shows that this result is valid for the 3B sample of sources, even if the relation between the fluence and the duration (or peak flux) is heavily influenced by instrumental effects.

2. From the remaining PCs, the third is definitely non-negligible, although strictly it is already below Jolliffe's cutoff level. This third PC is roughly identical to the fluence in the fourth (highest energy) channel $F4$. This means that in a finer approximation one could take, e.g., duration, total fluence, and $F4$ as variables containing significant information. The fact that $F4$ has a different behavior than the remaining three fluences was also confirmed by a PCA of the subset of four fluences alone. This different behavior of the fourth channel is also manifested by the fact that the hardness ratio $H43$ has a much bigger variance than the commonly used $H32$. The first three PCs account for 96.7% of the total information available in the 3B catalog.

3. A fourth PC is defined, in essence, by the ratio of fluence and peak flux. Thus, in an even finer approximation one would, in addition to the above three PCs, also consider the information provided by considering the fluence and the peak flux separately. However, the latter does not add much new information (only 1.5% more; see Table 4). In other words, in the finest approximation one can take, e.g., the total fluence, duration, $F4$, and peak flux: this contains 98.2% of the total information content of the 3B catalog with nine entries per burst.

4. The most meaningful single value of the duration, fluence, and peak flux for each event that can be constructed from the nine entries in the 3B catalog is obtained by using the weights given in the first line of Table 4 for that physical variable (e.g., a duration would be defined by using relative weights of 0.29 and 0.31 for T_{50} and T_{90} , with an appropriate normalization, etc.).

The above statements are purely statistical and deliberately omit any extra information concerning the operational way in which quantities are measured or theoretical models of what the data may mean. The facts that the fluence and durations are the most important quantities, and that a hardness and the peak flux are also useful, are agreed upon at a qualitative level by most people in the field. However, what is new here is the more rigorous quantification of the level of importance that can be assigned by the PCA method to each quantum of information and the corresponding ordering or prioritization of the different quantities that this affords. This quantification also allows one, for instance, to decide whether some alternative definitions of the basic quantities contain more information than others.

From a statistical information viewpoint, there appears to be no significant preference between, e.g., the durations T_{90} and T_{50} ; and on the same basis, a choice between the peak fluxes on the trigger timescales would appear to be approximately inconsequential. Of course, additional instrumental, operational, or physical model considerations would serve to refine such choices, depending on what one is seeking or on the hypothesis that one wants to test. For example, in Che et al. (1997), remarkable conclusions are

drawn from the differences between the peak fluxes on the 64 and 1024 ms triggers.

Some of the results are unexpected. For instance, the analysis indicates that an allowable approximation would be to combine (add) the fluences in the first three channels and consider them in conjunction with the fluence $F4$ in the fourth channel as the basis vectors for the fluence space. This singling out of $F4$ based on its (statistical) information content appears to be new. As is known, the hardness ratio $H32$ is most often used in statistical analyses of the BATSE data (or sometimes $H21$). However, $H32$ appears to have a significantly smaller variance than $H43$. It seems paradoxical, from the information content viewpoint, to concentrate attention on a variable of relatively small variance while generally ignoring other variables that have a much greater variance, namely $H43$. Of course, a careful consideration is required of whether the greater variance of $H43$ results from greater photon noise or other larger errors in determining $F4$, or from physically interesting facts (e.g., the spectral break occurring in the fourth channel or an additional steepening of the spectral index occurring there, etc.).

A concern here is that (as we understand it) the BATSE channel 4 fluence listed in the 3B catalog is obtained from a fit of a lognormal function across all four channels, and in the range above 300 keV, there is only one data point to anchor the extrapolation of this model shape (and the 3B usage of fluences in ergs further accentuates the uncertainty of this extrapolation). One way to address the importance of errors in $F4$ is by noting the fact that although the error of $F4$ is much higher than that of the remaining three channels, the variance of the logarithms of the F_i (discussed in this work) turned out to be roughly the same. The higher variance in $H43$ is therefore probably explained by the significantly lower correlation between $F4$ and the other three channels (see Table 1). We note here again that, as shown by the Monte Carlo simulations, a consideration of the errors listed in the 3B catalog does not change the singling out of $F4$ and its smaller degree of correlation to the other three (see also Appendix B). Since the first three channels correlate pretty well between themselves, they may be explained by the same PCs. On the contrary, because of its lower correlation with the other three, the fourth channel evidences the need for a further PC to account for it fully. Hence the fourth channel indicates some sort of additional information that is not contained in the other three. We strongly suspect that this additional information is of a physical nature (e.g., related to the difference between high-energy [HE] and no-HE pulses and bursts; Pendleton et al. 1997).

Another reason why $H43$ may be of interest is that it shows a significant anticorrelation with the peak flux, whereas $H32$ shows a correlation (Table 5). The PCA analysis presented here, in any case, suggests that more information may be available from a careful analysis of quantities involving the fourth channel than has been previously realized.

Finally, we note that the PCA offers a simple method for estimating the degree of improvement in information content that is potentially achievable by performing different manipulations of the data set beyond what is made in the 3B catalog or in its future incarnations. For instance, in designing new analyses that involve various corrections (instrumental or otherwise) to the data, one can use the PCA to measure quantitatively the increase, if any, in the

amount of information available from new definition of the variables to be studied. The PCA is therefore a tool of significant potential usefulness in planning data analysis strategies.

This research was partly supported by NASA grants NAG5-2362 and NAG5-2857 (P. M., I. H.); by a Domus

Hungarica Scientiarum Artiumque grant (A. M.); by OTKA grants F14324, T14304, F26666 (I. H.), and T024027 (L. G. B.); and by the Széchenyi Foundation (I. H.). A. M. acknowledges the kind hospitality of Konkoly Observatory, and we are grateful to E. D. Feigelson, E. E. Fenimore, G. Pendleton, and a referee for useful comments and discussion.

APPENDIX A

3B FACTOR ANALYSIS WITH NINE VARIABLES

The purpose of this appendix is to confirm the results of § 3 with the use of the factor analysis (FA). The FA is a statistical method that is closely related to, but not fully identical to, the PCA. (The comparison of these two methods is discussed in detail in, e.g., Jolliffe 1986). In essence, the major distinction between the PCA and the FA comes from the fact that the FA immediately assumes that the observed n variables are linear combinations of only $m < n$ variables, which contain the basic information in the data. The fulfillment of this assumption is then tested. A possible way to test this is the following: if m new variables are sufficient, then by using only these, one can reproduce the correlation matrix from the m quantities themselves with a high accuracy. From the level of this accuracy one may deduce the correctness of the assumption (for details see Jolliffe 1986; Kendall & Stuart 1976).

In our case this statistical method may be useful, because it gives a further independent criterion besides that given by Jolliffe. In § 3 one obtained $m = 2$ for $n = 9$ from Jolliffe's criterion. Nevertheless, as noted there, the third PC was just barely below the threshold of significance, and therefore one may ask whether it should be considered or not. In § 3 the correlation matrix (Table 3) and the results of a PCA (Table 4) are presented, and $m = 2$ is deduced. Taking $m = 2$, one may try to reproduce the correlation matrix using only these two PCs. The results of this procedure are shown in Table 6.

The "reproduced correlation matrix" is presented in the lower left triangle. For comparison, in the upper right triangle the differences ("residuals") between the observed correlations and the reproduced correlations are presented. One sees straightforwardly that the largest differences arise when the correlations involve $F4$. On the diagonal (*asterisks*) one should have values close to unity, if the PCs considered explain the corresponding observed variable well. Clearly, the departures from unity are again much larger for $F4$ than for any other quantity. Performing the test using the quantity given by the equation (43.132) of Kendall & Stuart (1976), we came to the conclusion that the assumption of $m = 2$ probably does not fully account for the quantity $F4$, and therefore a third PC is probably required.

APPENDIX B

CORRELATION COEFFICIENTS AND ERRORS

A measurement of two statistical variables x and y subject to measurement errors e_x and e_y results in specific values

$$\bar{x} = x + e_x, \quad \bar{y} = y + e_y,$$

where x and y are the ideal values in the absence of errors. The covariance of the two variables is

$$\langle \bar{x}, \bar{y} \rangle = \langle x, y \rangle + \langle x, e_y \rangle + \langle e_x, y \rangle + \langle e_x, e_y \rangle,$$

and unless the errors are of an unusual type, in the right-hand side the only nonzero term is $\langle x, y \rangle$. This is because one expects the x (y) and the e_x (e_y) to be independent, so their covariances vanish, leading to $\langle \bar{x}, \bar{y} \rangle = \langle x, y \rangle$. Thus, the effect of errors is expected to be negligible in the covariance.

TABLE 6
RESULTS OF THE FACTOR ANALYSIS OF THE NINE QUANTITIES USED IN § 3

| Quantity | log T_{50} | log T_{90} | log $F1$ | log $F2$ | log $F3$ | log $F4$ | log P_{64} | log P_{256} | log P_{1024} |
|----------------------|--------------|--------------|----------|----------|----------|----------|--------------|---------------|----------------|
| log T_{50} | 0.955* | 0.019 | -0.024 | -0.020 | -0.012 | -0.030 | 0.019 | 0.020 | 0.009 |
| log T_{90} | 0.955 | 0.958* | -0.021 | -0.019 | -0.014 | -0.031 | 0.017 | 0.019 | 0.011 |
| log $F1$ | 0.820 | 0.848 | 0.925* | 0.036 | -0.019 | -0.073 | -0.000 | -0.001 | 0.005 |
| log $F2$ | 0.803 | 0.834 | 0.936 | 0.950* | 0.002 | -0.073 | -0.001 | -0.003 | 0.005 |
| log $F3$ | 0.722 | 0.759 | 0.918 | 0.939 | 0.948* | 0.004 | -0.009 | -0.011 | -0.008 |
| log $F4$ | 0.473 | 0.509 | 0.693 | 0.718 | 0.745 | 0.607* | -0.041 | -0.058 | -0.076 |
| log P_{64} | -0.183 | -0.125 | 0.297 | 0.352 | 0.475 | 0.499 | 0.969* | 0.010 | -0.009 |
| log P_{256} | -0.028 | 0.031 | 0.440 | 0.495 | 0.606 | 0.589 | 0.961 | 0.979* | 0.014 |
| log P_{1024} | 0.281 | 0.336 | 0.680 | 0.726 | 0.805 | 0.709 | 0.849 | 0.916 | 0.955* |

NOTES.—The lower left triangle contains the reproduced correlation matrix; the upper right triangle contains the residuals between the observed correlations and the reproduced correlations; asterisks indicate the diagonal elements of the reproduced correlation matrix.

The correlation coefficient in the absence of errors is

$$r_{xy} = \frac{\langle x, y \rangle}{\sigma_x \sigma_y},$$

where $\sigma_x^2 = \langle x, x \rangle$ and $\sigma_y^2 = \langle y, y \rangle$. In the presence of errors, however, we have

$$\bar{r}_{xy} = \frac{\langle x, y \rangle}{\sigma_x \sigma_y},$$

where we use the fact that in the numerator there is no change, and

$$\sigma_x^2 = \langle x, x \rangle + 2\langle x, e_x \rangle + \langle e_x, e_x \rangle.$$

While, as before, one expects that $\langle x, e_x \rangle = 0$, the last term $\langle e_x, e_x \rangle$ is nonzero, and the same is true for the y variable. However, the order of magnitude of this last nonzero term (the “dispersion of the error”) is very different from that of the dispersion of x , and, in fact, $\langle e_x, e_x \rangle \ll \langle x, x \rangle$. It also follows that the measured correlations are generally smaller than the ideal (error-free) correlations ($|\bar{r}_{xy}| \leq |r_{xy}|$).

Concretely, if x is $F3$, and y is $F4$, then the fact that $\langle e_x, e_x \rangle \ll \langle x, x \rangle$ and $\langle e_y, e_y \rangle \ll \langle y, y \rangle$ can be verified directly from the 3B database. The catalog shows that while in some cases $e_y \simeq y$, the domain of the y (i.e., the spread of the $F4$) is much greater than the domains of the corresponding e_y (the spread of the e_{F4}). In particular, for channels 1, 2, 3, and 4, the standard deviation of the fluences around the mean and the mean error are given by the following pairs ($\sigma_F | \bar{e}$): 2.18 | 0.02, 2.22 | 0.02, 8.88 | 0.06, and 29.93 | 0.64 (the corresponding standard deviations of the errors around the mean error are also small, 0.02, 0.02, 0.07, and 0.74). Thus, the fact that the errors are sometimes comparable to the measured value ($|e_y| \simeq y$, in a fraction of the cases) does not mean that the correlation coefficients vary by much. The effect of the errors on the correlations is nonzero, but from the above it is expected to be small, and there should be no relationship between the magnitude of the difference between r_{xy} and \bar{r}_{xy} and the number of GRBs measured.

This is verified by the Monte Carlo simulations, which used the errors listed in the 3B catalog for every burst to determine a distribution out of which new values of $F1$, $F2$, $F3$, and $F4$ were chosen at random. This was done for every burst in the sample and repeated 100 times in order to calculate new correlation matrices and new PCs. The results varied at most by $\sim 0.3\%$. This shows that the fact that there are large errors in some bursts does not affect the conclusions about the correlation or lack of correlation between $F4$ and the other variables. Similar Monte Carlo simulations were done for the nine-variable case and for all correlations and PCs discussed in this paper. The results never varied by more than 1%.

REFERENCES

- Balázs, L. G., et al. 1996, *A&A*, 311, 145
 Briggs, M. S., et al. 1995, *ApJ*, 459, 40
 Che, H., et al. 1997, *ApJ*, 477, L69
 Connolly, A. J., et al. 1995, *AJ*, 110, 2655
 Emslie, A. G., & Horack, J. M. 1995, *Inverse Problems*, 11, 743
 Jolliffe, I. T. 1972, *Appl. Stat.*, 21, 160
 ———. 1986, *Principal Component Analysis* (New York: Springer)
 Kendall, M., & Stuart, A. 1976, *The Advanced Theory of Statistics* (London: Griffin)
 Koshut, T. M., et al. 1996, *ApJ*, 463, 570
 Kouveliotou, C., et al. 1993, *ApJ*, 413, L101
 Lamb, D., Graziani, C., & Smith, I. 1993, *ApJ*, 413, L11
 Lee, T. T., & Petrosian, V. 1996, *ApJ*, 470, 479
 ———. 1997, *ApJ*, 474, 37
 Lored, T., & Wasserman, I. 1995, *ApJS*, 96, 261
 Meegan, C. A., et al. 1996, *ApJS*, 106, 65
 Mitrofanov, I. G., et al. 1994, in *AIP Conf. Proc.* 307, *Gamma Ray Bursts*, ed. G. Fishman et al. (New York: AIP), 187
 Morrison, D. F. 1967, *Multivariate Statistical Methods* (New York: McGraw-Hill)
 Mukherjee, S., Feigelson, E. D., & Babu, G. J. 1998, preprint
 Murtagh, F., & Heck, A. 1987, *Multivariate Data Analysis, Astrophysics and Space Science Library* (Dordrecht: Reidel), 131
 Nemiroff, R. J. 1994, *Comments Astrophys.*, 17, 4
 Norris, J. P., et al. 1994, *ApJ*, 424, 540
 ———. 1995, *ApJ*, 439, 542
 Paczyński, B. 1995, *PASP*, 107, 1167
 Pendleton, G., et al. 1997, *ApJ*, in press
 Press, W. H., et al. 1992, *Numerical Recipes in Fortran*, 2d ed. (Cambridge: Cambridge Univ. Press)