PADEU | PADEU **17**, 169 (2006)
ISBN 963 463 557, ISSN 0238-2423

Statistical phenomena in astronomy

Invited Talk

L.G. Balázs
Konkoly Observatory of HAS, H-1525 Budapest, Pf. 67, Hungary

E-mail: balazs@konkoly.hu

Abstract
This paper represents the short version of the author's DSc theses. The whole text is available at the web site of http://www.konkoly.hu/staff/balazs/dissertation.pdf

1 Mathematical introduction

1.1 Nature of astronomical information

Observing and storing the photons of the incoming radiation from the Cosmos typically gives a data cube defined by $(\alpha, \delta, \lambda)$. It is easy to translate this data structure into the formalism of multivariate statistics. A common problem is in the multivariate statistics whether the stochastic variables described by observed properties are statistically independent or can be described by a less number of hidden variables. This is the task of factor analysis. Forming groups from cases having similar properties according to the measures of similarities or the distances is the task of cluster analysis. I demonstrated in several particular cases how these technics can be used for studying structures in the $(\alpha, \delta, \lambda)$ data cube and how to translate the statistical results into true physical quantities.

1.2 The basic equation of stellar statistics

The basic equation of stellar statistics connects the probability density function of a measurable quantity with the probability density of two variables, which can

not be observed directly, by the law of full probabilities. The resulting relation is a Fredholm type integral equation of the first kind. If the two background variables are statistically independent we recover the convolution equation. The analytical solution based on the Fourier transformation is very sensitive to high frequency noise. Eddington's solution attempts to find the unknown function in form of a series $\sum \gamma_j h^{(j)}(z)$. Malmquist's method computes the conditional probability of the unknown variable assuming that the observed variable is given. The statistical aspect of the problem is expressed if one uses the Lucy's algorithm which is a particular form of the more general EM algorithm. Dolan's matrix method solves numerically the matrix equation which approximates the integral equation. Methods are superior which retain the true statistical nature of the problem.

2 Statistical study of extended sources

2.1 Separation of Components

Separation of the Zodiacal and Galactic Light.

Principal components analysis and k-means clustering was utilized to identify different components of cosmic dust. Applying these techniques on the PL51 IRAS maps I recognized two main components with temperatures of about $200\,K$ (Zodiacal Light) and $40\,K$ (Galactic dust).

Structure and Dynamics of the Cepheus Bubble.

The Cepheus Bubble is a giant (10^o in angular diameter) dust ring around the Cep OB2 association. Performing factor analysis on HI 21 cm data, taken from the Leiden/Dwingeloo survey, reveals HI structures in the [-14,+2] kms^{-1} velocity range which can be associated with prominent parts of the dust ring. In the same area the HI maps also show an expanding shell with a well-defined approaching side at VLSR=-37 kms^{-1} and a less well-defined receding side at VLSR=-4 kms^{-1}. The kinematics and size of this shell are best modelled by a supernova explosion, occurring in Cep OB2a at about 1.7 $Myrs$ ago. Since the ages of several parts of the Cepheus Bubble are considerably higher than the age of the expanding shell, the supernova probably exploded in a pre-existing cavity, and its shock front might have interacted with the already existing star forming regions Sh2-140, IC 1396, and NGC 7129, leading to a new wave of star formation there.

2.2 Star count study of the extinction

I studied the ISM distribution in and around the star forming cloud L1251 with optical star counts. A careful calculation with a maximum likelihood based statistical approach resulted in *B, V, R, I* extinction distributions from the star count maps. A distance of 330±30pc was derived. The extinction maps revealed an elongated dense cloud with a bow shock at its eastern side. I estimated a Mach number of $M \approx 2$ for the bow shock. A variation of the apparent dust properties is detected, i.e. the $R_V = A_V/E_{B-V}$ total to selective extinction ratio varies from 3 to 5.5, peaking at the densest part of L1251. The spatial structure of the head of L1251 is well modelled with a Schuster-sphere (i.e. n=5 polytropic sphere). The observed radial distribution of mass fits the model with high accuracy out to $2.5pc$ distance from the assumed center. Unexpectedly, the distribution of NH_3 1.3 cm line widths is also well matched by the Schuster solution even in the tail of the cloud. Since the elongated head-tail structure of L1251 is far from the spherical symmetry the good fit of the linewidths in the tail makes reasonable to assume that the present cloud structure has been formed by isothermal contraction.

3 Statistics of point sources

3.1 Classification of stellar spectra

I made medium resolution ($100\mathring{A}/mm$) spectroscopy of 35 stars, picked up as suspected $H\alpha$ emission objects on small scale spectra, in the IC1396 star-forming region. Statistical studies based on factor analysis and k-means clustering yielded templates for further classification. Using proper motion data published in the literature I suggested that the vast majority of our objects belong to IC1396. Plotting the program stars, along with theoretical evolutionary tracks, onto the $\{Log(L); Log(T_{eff}\}$ plane I concluded that they are pre-main sequence objects of $0.5M_\odot < M < 3M_\odot$ masses and $10^5 < t < 10^7$ years age.

3.2 Angular distribution of GRBs

The isotropy of gamma-ray bursts collected in current BATSE catalog was studied. I showed that the quadrupole term being proportional to $\sin 2b \sin l$ was non-zero with a probability of 99.9%. The occurrence of this anisotropy term was then confirmed by the binomial test even with the probability of 99.97%. Hence, the sky distribution of all known gamma-ray bursts is anisotropic. I

also argued that this anisotropy cannot be caused exclusively by instrumental effects due to the nonuniform sky exposure of BATSE instrument. Separating the GRBs into short and long subclasses, I showed that the short ones are distributed anisotropically, but the long ones seem to be distributed still isotropically. The character of anisotropy suggests that the cosmological origin of short GRBs further holds, and there is no evidence for their Galactic origin.

3.3 Classification of GRBs

The gamma-ray bursts can be divided into three subgroups ("short", "intermediate", "long") with respect to their durations. This classification is somewhat unclear, since the subgroup of the intermediate durations has an admixture of both short and long bursts. A physically more reasonable definition of the intermediate subgroup was presented using also the hardness of the bursts. I showed that the existence of the three subgroups is real, and it was shown that no further subgroups are needed. According to the result the intermediate subgroup is the softest one. From this new definition it follows that 11% of all bursts belong to this subgroup. The intermediate subgroup shows furthermore an anisotropic distribution on the sky. A strong anticorrelation between the hardness and the duration was found - contrary to the short and long subgroups - for this subclass. Despite this difference it is not clear yet whether this subgroup represents a physically different phenomenon.

3.4 Physical difference between GRBs

I argued that the distributions of both the intrinsic fluence and the intrinsic duration of the gamma -ray emission in gamma-ray bursts from the BATSE sample are well represented by log-normal distributions, in which the intrinsic dispersion is much larger than the cosmological time dilatation and redshift effects. I performed separate bivariate log-normal distribution fits to the BATSE short and long burst samples. The bivariate log-normal behavior results in an ellipsoidal distribution, whose major axis determines an overall statistical relation between the fluence and the duration. I showed that this fit provides evidence for a power-law dependence between the fluence and the duration, with a statistically significant different index for the long and short groups. I discuss possible biases, which might affect this result, and argue that the effect is probably real. This may provide a potentially useful constraint for models of long and short bursts.