Statistical Phenomena in Astronomy

Lajos G. Balázs

August 11, 2004

Contents

Part I

Mathematical introduction

Chapter 1

Nature of astronomical information

The information we receive from Cosmos is predominantly in the form of electromagnetic radiation. An incoming plain wave can be characterized by the following quantities:

$$\mathbf{n}\,(direction),\;\; \lambda\,(wavelength),\;\; polarization.$$

These physical quantities determine basically the possible observational programs:

1. Position $\Longrightarrow$ Astrometry
2. Distr. of photons with $\lambda \Longrightarrow$ Spectroscopy
3. Number of photons $\Longrightarrow$ Photometry
4. Polarization $\Longrightarrow$ Polarimetry
5. Time of observation $\Longrightarrow$ variability
6. Distr. of photons in Data space $\Longrightarrow$ Statistical studies

In the reality, however, not all these quantities can be measured simultaneously. Restriction is imposed by the existing instrumentation.

By observing and storing the photons of the incoming radiation typically we get a data cube defined by $(\alpha, \delta, \lambda)$. The measuring instrument has some finite resolution in respect to the parameters of the incoming radiation. Consequently, the data cube can be divided into cells of size of the resolution. The astronomical objects can be characterized by isolated domains on the

α, δ plane. A real object can be more extended than one pixel in this plane. An object is called point source if it occupies an isolated pixel.

Each pixel in the α, δ plane can have a set of non-empty cells according to the different λ values. A list of non-empty pixels can be ordered into a matrix form having columns of properties (α, δ, and the set of λs) and rows referring to the serial number of objects. This structure is the 'Data Matrix' which is the input of many multivariate statistical procedures.

Table 1.1: Structure of the Data Matrix: m means the number of properties and n runs over the cases.

α_1	δ_1	λ_{11}	λ_{12}	$\cdots$	λ_{1m}
α_2	δ_2	λ_{21}	λ_{22}	$\cdots$	λ_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
α_n	δ_n	λ_{11}	λ_{12}	$\ldots$	λ_{nm}

1.1 Brief summary of multivariate methods

1.1.1 Factor Analysis

A common problem in the multivariate statistics whether the stochastic variables described by different properties are statistically independent or can be described by a less number of physically important quantities behind the data observed. The solution of this problem is the subject of the factor analysis.

Factor analysis assumes a linear relationship between the observed and the background variables. The value (factor scores) and number of background variables, along with the coefficients of the relationship (factor loadings) are outputs of the analysis. The basic model of factor analysis can be written in the following form:

$$X_j = \sum_{k=1}^{m} a_{jk} F_k + u_j \qquad , \qquad (j = 1, \cdots, p). \tag{1.1}$$

In the formula above X_j mean the observed variables, p the no. of properties, m the no. of hidden factors (normally $m < p$), a_{jk} the factor loadings, F_k the factor scores, and u_j called individual factors. The individual factors represent that part of the observed variables which are not explained by the common factors.

A common way to solve the factor problem uses the Principal Components Analysis (PCA). PCA has many similarity with the factor analysis, however, its basic idea is different. Factor analysis assumes that behind the observed ones there are hidden variables, less in number, responsible for the correlation between the observed ones. The PCA looks for uncorrelated background variables from which one obtains the observed variables by linear combination. The number of PCs equal to those of the the observed variables. In order to compute the PCs one have to solve the following eigenvalue problem:

$$\mathbf{R}\,\mathbf{a} = \lambda\,\mathbf{a} \tag{1.2}$$

where $\mathbf{R}$, $\mathbf{a}$ and λ mean the correlation matrix of the observed variables, its eigenvector and eigenvalue, respectively. The components of the $\mathbf{a}$ eigenvectors gives the coefficients of the linear relationship between the PCs and the observed variables. The PC belonging to the biggest eigenvalue of $\mathbf{R}$ gives the most significant contribution to the observed variables. The PCs can be ordered according the size of the eigenvalues. In most cases the default solution of the factor problem is the PCA in the statistical software packages (BMDP, SPSS, ...). Normally, if the observed variables can be described by a less number of background variables (the starting assumption of the factor model) there is a small number of PCs having large eigenvalue and their linear combination reproduce fairly well the observed quantities. The number of large eigenvalues gives an idea on the number of the hidden factors. Keeping only those PCs having large eigenvalues offers a solution for the factor model. This technique has a very wide application in the different branches of observational sciences. For the astronomical context see Murtagh & Heck (1987).

The factor model can be used successfully for separating cosmic structures physically not related to each other but projected by chance on the same area of the sky. We will return to the details later on when dealing with case studies.

1.1.2 Cluster Analysis

Factor analysis is dealing with relationships between properties when describing the mutual correlations of observed quantities by hidden background variables. One may ask, however, for the relationship between cases. In order to study the relationship between cases one have to introduce some measure of similarity. Two cases are similar if their properties, the value of their observed quantities, are close to each other.

"Similarity", or alternatively "distance" between l and k cases, is a function of two X_j^l, X_j^k set of observed quantities (j is running over the properties describing a given case). Conventionally, if $l = k$, i. e. the two cases are identical, the similarity $a(X_j^l, X_j^k) = 1$ and the distance $d(X_j^l, X_j^k) = 0$. The mutual similarities or distances of cases form a similarity or distance matrix.

Forming groups from cases having similar properties according to the measures of similarities and the distances is the task of cluster analysis. There are several methods for searching clusters in multivariate data. There is no room here to enter into the details. For the astronomical context see again (Murtagh & Heck , 1987). Typical application of this procedure is the recognition of celestial areas with similar properties, based on multicolor observations. The procedure of clustering in this case is a searching for pixels on the images taken in different wavelengths but having similar intensities in the given colors.

In the following we try to demonstrate how these procedures are working in real cases.

Chapter 2

The basic equation of stellar statistics

2.1 Formulation of the problem

It is a basic problem in astronomy that we can not measure quantities directly, e.g. linear extension, space velocity or absolute brightness. Instead, we can get only their distance-dependent apparent values (angular diameter, proper motion, apparent brightness). In general, there exists a relationship in the form of

$$e = e(E, r) \quad , \tag{2.1}$$

where e, E and r are the measured value, the true value, and the distance, respectively. Following the law of full probabilities we may write:

$$\phi(e) = \int \phi(e|r) g(r) dr \quad . \tag{2.2}$$

In this expression $\phi(e)$, $\phi(e|r)$ and $g(r)$ are the probability density of e, the conditional probability of e, given r, and the probability density of r, respectively. Since the measured distance-dependent values can be written in the form $e = E\beta(r)$ in the cases we consider, where $\beta(r) = r^{-n}$ if we neglected interstellar absorption and cosmological effects, one can reduce $e = e(r, E)$ to $z = x + y$ with a suitable logarithmic transformation. ($n = 1$ for angular diameters and proper motions while $n = 2$ for apparent brightness).

Supposing the statistical independence of y and z we get the convolution equation between the probability densities of these variables:

$$h(z) = \int f(z - y) g(y) dz \quad . \tag{2.3}$$

This equation is also called the basic equation of stellar statistics.

2.2 Solutions of the basic equation

The task is to find $g(y)$ if $h(z)$ and $f(x)$ are given. The convolution equation is a Fredholm type integral equation. Its formal solution can be obtained by Fourier transformation. This solution has little practical use because we generally know only a statistical sample drawn from $h(z)$. There are several methods for obtaining the moments of $g(y)$, its approximation in form of orthogonal series, or its numerical values at discrete points. Particular attention will be devoted to computing the maximum likelihood estimate of the discrete values of g(y) via the EM algorithm.

2.2.1 Eddington's solution

Let us assume that $f(x)$ and $g(y)$ are the probability densities of the statistically independent random variables x and y.

By definition $z = x + y$, and the probability density $h(z)$ of the observed quantity z is connected with $f(x)$ and $g(y)$ by the convolution equation.

Suppose that $h(z)$ has derivatives $h^{(k)}(z)$ of all orders and $f(x)$ has central moments, μ_k. We look for the solution of the convolution equation in the form of a series, $g(z) = \sum\limits_0^\infty \gamma_k h^{(k)}(z)$. (Eddington 1913; uniqueness of the solution and convergence of the series are discussed by Kurth 1952).

Writing

$$h(z) = \int\limits_{-\infty}^{+\infty} f(x) g(z-x) dx \tag{2.4}$$

and expanding g in a Taylor series around z we obtain

$$h(z) = \int\limits_{-\infty}^{+\infty} f(x) \sum_{k=0}^{\infty} \frac{(-x)^k}{k!} g^{(k)}(z) dx. \tag{2.5}$$

Integrating term by term and supposing $Ex = 0$ (E here meaning expectation value) we get

$$h(z) = \sum_{k=0}^{\infty} g^{(k)}(z) \frac{(-1)^k}{k!} \mu_n. \tag{2.6}$$

Now we substitute the formal series

$$h(z) = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \gamma_m h^{(k+m)}(z) \frac{(-1)^k}{k!} \mu_k. \tag{2.7}$$

Introducing $n = k + m$, the double sum can be written in the form

$$h(z) = \sum_{n=0}^{\infty} h^{(n)} \sum_{k=0}^{n} \gamma_{n-k} \frac{(-1)^k}{k!} \mu_k. \tag{2.8}$$

The equation is satisfied if on the right side the coefficients of $h^{(n)}$ except $n = 0$ equal zero. This means

$$\begin{aligned} \gamma_0 &= \quad 1 \\ \sum_{k=0}^{n} \gamma_{n-k} \frac{(-1)^k}{k!} \mu_k &= \quad 0, \qquad n \neq 0. \end{aligned} \tag{2.9}$$

Based on these equalities the coefficients γ_k of the formal series can be expressed by the central moments μ_k of f

$$\begin{aligned} \gamma_0 &= 1 \qquad\qquad , \\ \gamma_1 &= 0, \\ \gamma_2 &= -\tfrac{1}{2}\mu_2 = -\tfrac{1}{2}\sigma^2, \\ \gamma_3 &= \tfrac{1}{6}\mu_3, \\ \gamma_4 &= -\tfrac{1}{24}\mu_4 + \tfrac{1}{4}\mu_2^2, \qquad etc. \end{aligned} \tag{2.10}$$

For small values of $\mu_2 = \sigma^2$ the above equalities suggest approximating the solution $g(y)$ by the function

$$h(y) - \frac{1}{2}\sigma^2 h^{''}(y). \tag{2.11}$$

In real cases, however, it is rather inconvenient to compute the second derivative $h^{''}(y)$ numerically. One may therefore replace it by the ratio

$$\alpha^{-2}[h(y+\alpha) + h(y-\alpha) - 2h(y)], \tag{2.12}$$

where α is a small positive number. As shown by the Taylor theorem, the difference between $h^{''}$ and the above formula is of the fourth order in α. Using this expression for the second derivative one obtain the following approximation

$$g_2(y) = h(y) - \frac{1}{2}(\frac{\sigma}{\alpha})^2[h(y+\alpha) + h(y-\alpha) - 2h(y)]. \tag{2.13}$$

In an analogous way one can derive a "five-point formula", a "seven point formula" etc. in order to get the higher order derivatives. In applications, however, derivatives of order higher than 4 are not used because of the increasing uncertainty of the estimation of moments with increasing order.

2.2.2 Solution by Fourier transformation

A formally simple method of analytic solution of the basic equation is the Fourier transform one. Taking the Fourier transform of both sides of the equation yields

$$\hat{h}(\nu) = \sqrt{2\pi}\ \hat{f}(\nu)\ \hat{g}(\nu) \tag{2.14}$$

where ν is the frequency. It is easy to get from this equation an analytic solution

$$g(y) = \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \frac{\hat{h}(\nu)}{\hat{f}(\nu)} e^{i\nu y} d\nu. \tag{2.15}$$

In real cases, however, $h(z)$ is known only in terms of a sample drawn from it and one may approximate the original probability density function with an arbitrarily small error as the sample size n goes to infinity. A given sample size sets a bound within which the different solutions of the basic equation are statistically indistinguishable.

Suppose we have a solution obtained by applying the above formula. One may add to $\hat{g}(\nu)$ a $\delta(\nu)$ function in the frequency domain which has a constant value in a $\{\nu_0, \nu_0 + \Delta\nu\}$ region and 0 otherwise. Since the existence of the Fourier transformation requires $\int \hat{f}(\nu)^2 d\nu < \infty$, $\sqrt{2\pi}\ \hat{f}(\nu)\ [\hat{g}(\nu) + \delta(\nu)]$ approaches arbitrarily closely to $\hat{h}(\nu)$, and after Fourier transformation $h(z)$, as ν_0 goes to infinity.

The Fourier counterpart of $\delta(\nu)$ is a high frequency oscillating function which one may add to the $g(y)$ solution without violating the statistical bound around $h(z)$, i.e. $\int [\hat{g}(\nu) + \delta(\nu)] exp(i\nu x) dx$ may also be accepted as a solution. Increasing the sample size n decreases the statistical bound as $\sqrt{n}$. Halász (1984) showed, however, that the bound around the $g(y)$ solution decreases only as $log(n)$ at the same time.

An obvious way out of this difficulty appears to be to apply some cut-off frequency ν_0, i.e to integrate in the $\{-\nu_0; \nu_0\}$ domain instead of $\{-\infty; +\infty\}$. Proceeding in this way, however, one has some problem in selecting the 'best value' for ν_0 which makes this procedure somewhat arbitrary.

Another, at first glance appealing, approach appears to be developing the $h(z)$ probability density into an orthogonal series:

$$h(z) = \sum_{k=0}^{\infty} a_k H_k(z)\alpha(z), \quad where \;\; a_k = \frac{1}{k!}\int_{-\infty}^{+\infty} H_k(z)h(z)dz. \tag{2.16}$$

Here the $\alpha(z) = exp(-z^2/2)/\sqrt{2\pi}$ and $H_k(z)$ are Hermite polynomials. Applying the Fourier transformation to the Hermite functions $H_k(z)\alpha(z)$ we obtain

$$\hat{H}_k(\nu) = \sqrt{2\pi}(i\nu)^k\alpha(\nu). \tag{2.17}$$

(For further details see Kendall & Stuart, 1973, Vol1, pp. 156-157). The solution of the basic equation by the Fourier method reduces then to solving the equations

$$\alpha(z)H_k(z) = \int_{-\infty}^{+\infty} f(z-y)g_k(y)dx, \quad g_k(y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \frac{(i\nu)^k\alpha(\nu)}{\hat{f}(\nu)}e^{i\nu y}d\nu. \tag{2.18}$$

If these integrals exist, we get the solution in form of $g(y) = \sum_{k=0}^{\infty} a_k g_k(y)$. This approach seems to overcome the difficulty of high frequency noise occurring in the solution. Computing the higher order a_k coefficients, however, one get increasing uncertainty due to the sampling error of $h(z)$.

2.2.3 Malmquist's solution

In deriving the basic equation we used the law of full probability. Formally, based on this theorem, we may write the $g(y)$ solution as

$$g(y) = \int g(y|z)h(z)dz \tag{2.19}$$

where $g(y|z)$ is the conditional probability assuming z, the measured quantity, is given. If one succeeds in developing $g(y|z)$, the solution is given by computing the integral given above.

In some important cases, as Malmquist (1924, 1936) recognized, the momenta can be developed without knowing $g(y|z)$ itself. Let z be a unit random variable and $h(z)$ its probability density. It may be represented as

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}[1 + \gamma_3 H_3(z) + \gamma_4 H_4(z) + ...], \tag{2.20}$$

where

$$\begin{aligned} \gamma_3 = & \quad \tfrac{1}{\sqrt{3!}}\mu_3, \\ \gamma_4 = & \quad \tfrac{1}{\sqrt{4!}}(\mu_4 - 3) \end{aligned} \tag{2.21}$$

and μ_3, μ_4 are the central moments of z of the third and fourth order. $H_k(z) - s$ are the orthogonal functions of the "Hermite's polynomials" and the series obtained is called the "Gram-Charlier series" (cf. Kendall & Stuart, 1973, Vol 1, pp 156-157). If z is not a unit variate the corresponding series can be obtained using the unit variate $\zeta = (z - \mu)/\sigma$ by the appropriate transformation. Computing the corresponding moments of the conditional probability we may develop its Gram-Charlier series and, consequently, the $g(y)$ solution.

According to Bayes theorem

$$g(y|z) = \frac{\chi(y,z)}{\int \chi(y,z)dy} = \frac{\chi(y,z)}{h(z)}. \tag{2.22}$$

Following the notations of the first paragraph we may write

$$\chi(y,z) = f(z-y)g(y) \quad and \quad g(y|z) = \frac{f(z-y)g(y)}{h(z)}. \tag{2.23}$$

In a number of practically important cases $f(x)$ is a gaussian function, i.e. $f(x) = exp(-(x-\mu)^2/2\sigma_0^2)/\sqrt{2\pi}\sigma_0$. In this case one need not know $g(y)$ when computing the momenta of the conditional probability $g(y|z)$. By definition

$$\mu(z) = \int y g(y|z)dz = \frac{\int y e^{\frac{-(z-y-\mu)^2}{2\sigma_0^2}} g(y)dy}{\int e^{\frac{-(z-y-\mu)^2}{2\sigma_0^2}} g(y)dy}. \tag{2.24}$$

Differentiation of $\log(h(z))$ yields

$$\frac{d}{dz}\log(h(z)) = \frac{\int \frac{-(z-y-\mu)}{\sigma_0^2} e^{\frac{-(z-y-\mu)^2}{2\sigma_0^2}} g(y)dy}{\int e^{\frac{-(z-y-\mu)^2}{2\sigma_0^2}} g(y)dy} \tag{2.25}$$

and we get

$$\mu(z) = z - \mu + \sigma_0^2 \frac{d}{dz}\log(h(z)). \tag{2.26}$$

Similar computation gives an expression for the $\sigma(z)$ standard deviation of $g(y|z)$

$$\sigma(z)^2 = \sigma_0^2[1 + \frac{\sigma_0^2}{2}\frac{d^2}{dz^2}\log(h(z))]. \tag{2.27}$$

Higher order moments can be obtained in an analogous way. Substituting these moments into the expression of the Gram-Charlier series we get finally $g(y|z)$ and thus $g(y)$, after performing the integration.

The expressions for the moments of the $g(y|z)$ conditional probability have important consequences, often overlooked in practical applications. If z is given by the observations, then the expected value of y is not simply $z - \mu$, where μ is the expected value of x, but has to be corrected by the third term in the expression for $\mu(z)$. This term is called the *Malmquist correction.* If one determined spectroscopically the absolute magnitude M_0 of a star of apparent magnitude m the expected value of the distance modulus is not simply $m - M_0$ but a value corrected with the Malmquist's term.(For more details see Mihalas and Binney (1981)).

2.2.4 Lucy's algorithm

Lucy (1974) proposed an algorithm for solving the integral equation connecting $g(y), g(y|z)$ and $h(z)$, in an iterative way. Let us assume that at some stage of the iteration we get $g^{(r)}(y), g^{(r)}(y|z)$. The estimate $(r+1)$ is

$$g^{(r+1)}(y) = \int g^{(r)}(y|z)h(z)dz, . \tag{2.28}$$

where, applying Bayes theorem of conditional probabilities,

$$g^{(r)}(y|z) = \frac{g^{(r)}(y)h(z|y)}{h^{(r)}(z)}. \tag{2.29}$$

with

$$h^{(r)}(z) = \int h(z|y)g^{(r)}(y)dy. \tag{2.30}$$

In the above equalities $h(z|y)$ denotes the conditional probability of z, assuming y is given. In the case when y and $z - y$ are statistically independent $h(z|y) = f(z-y)$. Since $g(y)$ is a probability density it should fulfil the following relations:

$$\int g(y)dy = 1 \quad \textit{and} \quad g(y) \geq 0. \tag{2.31}$$

One may readily show that this iterative procedure conserves these constraints. Eliminating $h(z|y)$ from the above equations we obtain

$$g^{(r+1)}(y) = g^{(r)}(y) \int \frac{h(z)}{h^{(r)}(z)} h(z|y) dy \tag{2.32}$$

from which one can easily recognize that $g^{(r+1)} \geq 0$ if $g^{(0)} \geq 0$. Proof of the conservation of the normalization constraint is straightforward by integration of $g^{(r+1)}(y)$ and using the normalizations of the $g^{(r)}(y|z)$ and $h(z)$ probabilities.

Actually, we know $h(z)$ only in terms of a sample drawn from it and we might be concerned at the loss of information involved in forming a histogram from the data in order to get an $\tilde{h}(z)$ estimate for $h(z)$. The sample itself can be represented by the expression

$$\tilde{h}(z) = \frac{1}{N} \sum_{n=1}^{N} \delta(z - z_n), \tag{2.33}$$

where N, z_n and $\delta(z - z_n)$ are the size of the sample, the individual measures and Dirac's delta function, respectively. Substituting this expression into the iterative scheme yields

$$g^{(r+1)}(y) = \frac{1}{N} \sum_{n=1}^{N} g^{(r)}(y|z). \tag{2.34}$$

In practice the computation of $g(y)$ is performed on a set of values y_m and the iterative scheme can be interpreted as a procedure for estimating the unknown parameters $g_m = g(y_m)$. In this view the integral expression for $h^{(r)}$ will be approximated by the following sum

$$h^{(r)}(z) = \sum_{j=1}^{m} g_j^{(r)} h(z|y_j) \Delta y. \tag{2.35}$$

The expression obtained in this way allows us to make a comparison with the maximum likelihood (ML) estimation of unknown parameters. If we have a sample of measures z_n then the likelihood function can be written in the form

$$L(g_j) = \sum_{i=1}^{n} \log(h(z_i)). \tag{2.36}$$

ML means to maximize $L(g_j)$ over the given sample subject to the constraint $\sum g_j = 1$. Applying a λ Lagrange multiplier one has to maximize the following function

$$G(g_j; \lambda) = L(g_j) + \lambda(\sum_{j=1}^{m} g_j - 1). \tag{2.37}$$

One finds that the multiplier $\lambda = -n$. Performing the differentiation one gets equations for g_j-s maximizing $L(g_j)$

$$0 = \sum_{i=1}^{n} \frac{h(z_i|y_j)}{\sum\limits_{j=1}^{m} g_j h(z_i|y_j)} - n \qquad j = 1, ..., m \tag{2.38}$$

and multiplication with g_j yields

$$g_j = \frac{1}{n} \sum_{i=1}^{n} \frac{g_j h(z_i|y_j)}{\sum\limits_{j=1}^{m} g_j h(z_i|y_j)} \tag{2.39}$$

which is identical with the last expression of the iteration scheme when it converges ($g_j = g_j^{(r+1)} = g_j^{(r)}$). Lucy showed that his iteration algorithm monotonically increases the likelihood and converges to the ML solution. Numerical experiments showed, however, that the ML solution did not give the best fit to $g(y)$ while approximating $\tilde{h}(z)$. This result also presents some hint for the inherent uncertainties of the solution procedures of the basic equation.

2.2.5 The EM algorithm

We have mentioned already in previous sections that in a number of practical cases $h(z)$ is not given analytically but only in a form of a sample drawn from it. To get the solution, therefore, is rather a statistical estimation problem than an analytical procedure for solving integral equations. The true nature of the problem is nonparametric. When one approximates the integral $h(z) = \int h(z|y)g(y)dy$ by the sum $\sum\limits_{j=1}^{m} h(z|y_j)g(y_j)\Delta y$, containing $g(y_j)$ (and maybe y_j) as unknown parameters, (Δy can be absorbed into g_j) the solution of the basic equation will be reduced to a parameter estimation problem. The number of parameters, however, is fully arbitrary. Nevertheless, there is a minimum number of parameters still giving a solution within the statistical bound of the sample of $h(z)$.

Replacing the integral with a sum means a discretization of the y statistical variable into y_j, $j = 1, ..., m$, values. The y_j variable takes these values with $g(y_j) = g_j$ probabilities. We assume furthermore there exists an unobserved vector $\mathbf{u} = (u_1, ..., u_m)$ whose components are all zero except for one (the j-th) equal to unity indicating the unobserved y_j state associated with the actually observed z. In the case when both the $\mathbf{u}$ and z variables were observed the data would be complete. Their joint probability density would be

$$p(z,\mathbf{u}) = e^{\sum\limits_{j=1}^{m} u_j[\log(g_j)+\log(h(z|y_j)]}. \tag{2.40}$$

Since the elements of the vector are zero except a certain u_j, which is unity, $p(z,u)$ reduces to $g_j h(z|y_j)$ if j is given. To perform a ML estimation we define the likelihood function in the usual form

$$\begin{aligned} L(g_1,...,j_m) = & \sum_{i=1}^{n} \log(p(z_i,\mathbf{u}^{(i)})) = \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} u_j^{(i)}[\log(g_j)+\log(h(z_i|y_j))]. \end{aligned} \tag{2.41}$$

Maximizing the likelihood with respect to the g_j-s one should take into account the constraint $\sum g_j = 1$ with a Lagrange multiplier λ. Differentiation with respect to the parameters g_j yields

$$0 = \sum_{i=1}^{n} \frac{u_j^{(i)}}{g_j} + \lambda = \frac{n_j}{g_j} + \lambda, \qquad j = 1,...,m \tag{2.42}$$

where n_j is the frequency of occurrence of the j-th state in the sample. Summing the equations gives $\lambda = -n$ and substituting it into the equations we get $g_j = n_j/n$ as a ML estimation.

In the case of incomplete data one does not have observations of the states j. Dempster et al. (1977) showed, however, that a ML estimation is still possible if instead of $\mathbf{u}$ we use its conditional expectation assuming that z is given. Proceeding in this way one may estimate g_j-s and the solution is also the ML solution of the case with complete data specification. By definition the conditional probability of $\mathbf{u}$ is given by

$$v(\mathbf{u}|z) = \frac{p(z,\mathbf{u})}{h(z)} \qquad \textit{and} \qquad E(u_j|z) = \int u_j \frac{p(z,\mathbf{u})}{h(z)} du_j. \tag{2.43}$$

Since u_j has only two values, 0 or 1, the integral is simplified to the form

$$E(u_j|z) = \frac{p(z,\mathbf{u})}{h(z)} = \frac{g_j h(z|y_j)}{\sum\limits_{j=1}^{m} g_j h(z|y_j)}. \tag{2.44}$$

Considering these conditional expected values of the u_j-s as complete data we get the ML estimation in the form

$$g_j = \frac{1}{n}\sum_{i=1}^{n} E(u_j|z) = \frac{1}{n}\sum_{i=1}^{n} \frac{g_j h(z|y_j)}{\sum\limits_{j=1}^{m} g_j h(z|y_j)}. \tag{2.45}$$

Suppose we have an estimate $g_j^{(r)}$ of g_j; the next approximation $g_j^{(r+1)}$ can be obtained if we substitute $g_j^{(r)}$ into the right side and $g_j^{(r+1)}$ into the left side. In this way we recovered Lucy's algorithm, which is a special case of the EM algorithm, a more general framework for performing ML estimation in cases of incomplete data specification. The convergence and general properties of this algorithm are given by Dempster et al. (1977).

The EM algorithm consists of two steps:

E step: One replaces the unobserved part of the data with the expected values based on their conditional probabilities assuming that the observed part is given.

M step: Using the complete data specification obtained in this way one gets the corresponding ML estimation.

Dempster et al. (1977) showed that performing the E and M steps sequentially the procedure monotonically converges to the ML solution.

Up to this point the y_j parameters have not been included in the parameters to be estimated. The general framework of the EM algorithm, however, allows this generalization without any difficulty. We show this only in a special case when $h(z|y_j) = exp(-(z-y_j)^2/2\sigma_j^2)/\sqrt{2\pi}\sigma_j$. In the case of a gaussian distribution the ML estimation of the y_j and σ_j parameters is the arithmetic mean of z and $(z-y_j)^2$. When we are dealing with incomplete data these relations correspond to

$$y_j = \frac{\sum\limits_{i=1}^{n} z_i E(u_j|z)}{\sum\limits_{i=1}^{n} E(u_j|z)} \tag{2.46}$$

and

$$\sigma_j^2 = \frac{\sum\limits_{i=1}^{n} (z_i - y_j)^2 E(u_j|z)}{\sum\limits_{i=1}^{n} E(u_j|z)}. \tag{2.47}$$

If σ is independent of j then we may add a third equation to the equations for y_j and σ_j

$$\sigma^2 = \sum_{j=1}^{m} g_j \sigma_j^2. \tag{2.48}$$

These equations could be added to the M step. The estimation of the parameter σ_j is much more uncertain than of y_j in particular at small sample sizes.

2.2.6 Dolan's matrix method

When the integral in the basic equation is approximated by a sum an obvious idea is to solve the resulting system of n linear equations with respect to the unknowns $g(y_j)$, setting $n = m$

$$\sum_{j=1}^{m} f(z_i - y_j)g(y_j)\Delta y = \tilde{h}(z_i), \qquad i = 1, ..., n. \tag{2.49}$$

$\tilde{h}(z_i)$ is obtained from the observed sample of $h(z)$.

This method is rather popular among astronomers due to its simplicity and flexibility since we do not have any particular basis for assumptions about the functions f, g, h. (For further details see Bok (1937)). There are, however, serious problems in practical applications. Unless the sample size is large, the solution of this system of linear equations often yields negative $g(y_j)$-s violating the $g(y_j) \geq 0$ property of probability densities (cf. Trumpler and Weaver 1953, p. 112-114).

The equations can be written in the more compact matrix form $\mathbf{FG} = \mathbf{H}$ where $\mathbf{F}, \mathbf{G}, \mathbf{H}$ are $n \times n$, $n \times 1$ and $n \times 1$ matrices, respectively. The solution is simply given by $\mathbf{G} = \mathbf{F^{-1}H}$ requiring the inversion of $\mathbf{F}$. The large errors occurring in the solution $\mathbf{G}$ can be understand if we recognized that application of $\mathbf{F}$ on $\mathbf{G}$ means some kind of "smoothing". The application of the inverse transformation on $\mathbf{H}$ amplifies the stochastic fluctuations inherent in obtaining $\mathbf{H}$.

The solution of the system of linear equations requires the inversion of the $\mathbf{F}$ matrix. Unfortunately, in a number of cases some of the eigenvalues of $\mathbf{F}$ are small making the numerical inversion very unstable. To avoid the problem of matrix inversion Dolan (1974) proposed the technique of apodisation (see e.g., Lloyd, 1969). Apodisation constructs the $\mathbf{G}$ matrix from the observed $\mathbf{H}$ using the known properties of the gaussian "smoothing" function. A discussion of the numerical accuracy of this procedure is given by Dolan (1972). As shown by Lloyd, the matrix $\mathbf{G}$ has elements

$$g_j = h_j - K\{\frac{1}{2}[h_{j+l} + h_{j-l}] - h_j\} \tag{2.50}$$

where l is the "apodisation length" and K is an "instrument factor" whose value depends on the shape of the smoothing function. For a gaussian function $K = (2/\pi)(\sigma/l)^2$. The apodisation length, l, should thus be chosen to the nearest integer number to the dispersion σ . The expression given above strongly resembles that obtained in the first approximation of Eddington's solution.

2.3 Concluding remarks

All the methods discussed in the present chapter represent different approaches to the same problem. As we mentioned in the introduction there is a typical situation in statistical astronomy where the probability densities of the distance-dependent measurable quantities, of its true physical counterparts, and of some distance measures are connected by the convolution equation.

To get the probability density function of the unobserved background variables (e.g. the distance), however, is not a simple analytical procedure of solving a Fredholm-type integral equation of the first kind. Namely, the probability density function $h(z)$ of the observed variable z is not given but has to be estimated somehow from a sample obtained from the observations.

From the analytic properties of the Fredholm integral equations of the first kind it follows that a small deviation from $h(z)$ is amplified by the solution process in obtaining $g(y)$. This uncertainty is inherent in the equation and can not be overcome by a suitable choice of procedure. In this view the result of Halász (1984) is very significant, i.e. the improvement of the statistical accuracy in getting $h(z)$ by increase of the sample size n means an improvement in $g(y)$ proportional only to $\log(n)$.

Solution procedures may differ in some practical aspects: how they approximate the integral in the equation or the numerical derivatives of $h(z)$ (e.g. in the case of the Eddington or Malmquist methods). These practical aspects could cause particular uncertainties not inherent in other procedures.

The approximation of the integral with a finite sum reduces the solution to the problem of the estimation of unknown parameters behind an observed sample. However, the parameters to be estimated are not in the true sense the values of the unknown $g(y)$ function but its integral over Δy. Consequently, the accuracy of the estimated parameters follows the general $\sqrt{n}$ rule of parameter estimation, when $n \to +\infty$.

The EM algorithm and its particular case the Lucy algorithm have the nice property of retaining $g(y) \geq 0$ for the probability density function of y. Analytic solutions often suffer from violating this basic property, in particular when the sample size is not large.

The computational efforts are very different among the procedures discussed in this chapter. Obviously, the first approximation of the Eddington solution and the formula given by Dolan are the simplest also from that point of view. The EM algorithm on the other hand requires much more computing time and memory, in particular when the sample size is large, but it has the clear benefit of a sound theoretical basis in mathematical statistics.

This discussion does not attempt to give strict advice on performing solutions in particular cases. Our aim was rather to call attention to important aspects when trying to find the solution. Our feeling is, however, that the true essence of the problem is more stochastic than analytic and those procedures are superior which retain this basic characteristic.

Part II

Statistical study of extended sources

Chapter 3

Separation of Components

3.1 Separation of the Zodiacal and Galactic Light

The IRAS space mission covered the whole sky in four (12, 25, 60, 100 μm) wavelengths. In particular, the 12 and 25 μm images were dominated by the thermal emission of the Zodiacal Light (ZL) having a characteristic temperature around $250\,K$. The contamination of the Galactic Dust thermic radiation by the ZL is quite serious close to the Ecliptic. Assuming that both radiation are coming from optically thin media the observed infrared intensities are sums of those coming from these two components. We may assume furthermore the distribution of the intensity of thermal radiation on the sky coming from the Galactic component has some similarities when observed at the given wavelengths and the same holds also for the ZL. Identifying the radiation coming from these two physically distinct components with the hidden variables in Eq. (1.1) and the incoming intensity with the observed ones the separation of the ZL and the Galactic radiation can be translated into the general framework of factor analysis.

In the case of the IRAS images the **R** correlation matrix has a size of 4×4 by cross correlating the four (12, 25, 60 and 100 μm) images. We selected a field of $15° \times 15°$ (corresponding to 500×500 pixels) in the Perseus close to the ecliptic and containing the California Nebula, IC 348 and the Pleiades.

Solving Eq. (1.2) in this case we got the results summarized in Tab. 3.1. One can infer from this table that there are two large eigen values indicating the presence of two important factors. The last two columns of the table give the a_{jk} factor coefficients for Eq. (1.1). The first factor dominates the radiation at 12 and 25 μm while the second one does it at 60 and 100 μm. Computing the factor values from the observed data (the measured 12, 25,

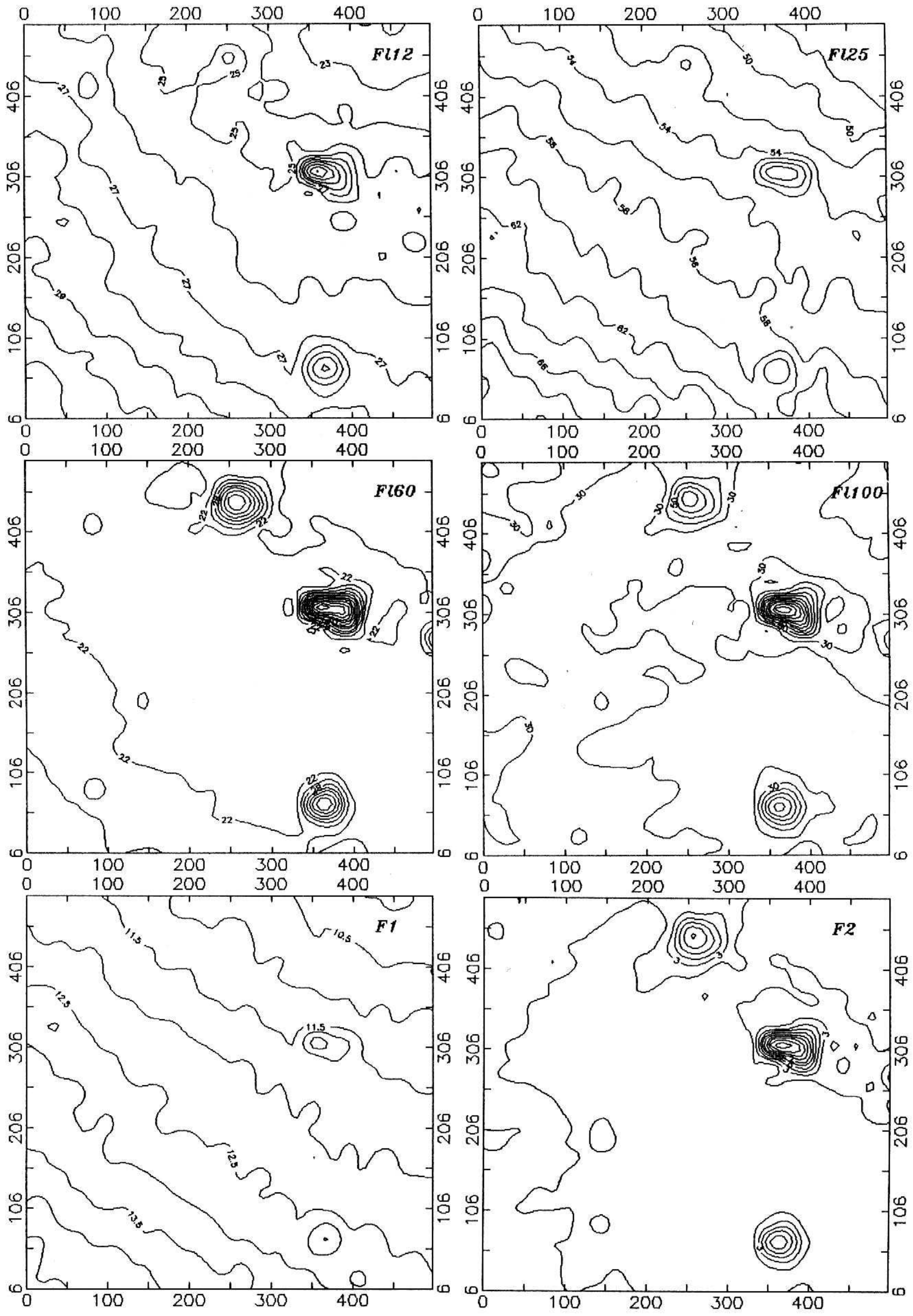

Figure 3.1: Input IRAS (12, 25, 60, and $100\mu m$) images of the factor analysis and the resulted two factor pictures. The coordinates are measured in pixels. The objects are the California Nebula, IC 348 and the Pleiades, in descending order. Note the strong trend in F_1 representing the ZL while F_2 displays the Galactic component (Balázs & Tóth , 1991).

Table 3.1: Results of factor analysis. There are two large eigen values indicating the presence of two important factors. The last two columns of the table give the a_{jk} factor coefficients for Eq. (1.1). (Balázs et al. , 1990)

eigen value	cum. percent.	Variable	1. factor	2. factor
2.4818	62.0	F_{12}	0.9637	0.2089
1.3910	96.8	F_{25}	0.9917	0.0458
0.1003	99.3	F_{60}	0.3625	0.9044
0.0268	100.0	F_{100}	0.0409	0.9819

60 and 100 μm intensities) one gets the two images as shown in Fig. 3.1, along with the originals (Balázs & Tóth , 1991).

In order to define regions of similar physical properties we performed cluster analysis in the $\{F_1; F_2\}$ factor plane. These two factors define a two-dimensional subspace in the four-dimensional color space. The 1-st factor almost fully explains the 25 μm flux, which is heavily dominated by the Zodiacal Light and therefore represents its influence in different colors. The second factor, in contrast, describes the effect of the radiation coming from the galactic dust which produces most of the 100 μm emission. Performing cluster analysis altogether 10 regions were defined, however this figure was arbitrary. The result is given in Fig. 3.2. The basic features of this plot are the two 'fingers' pointing upwards and nearly horizontally. These 'fingers' may be identified with the Zodiacal Light (dominating F_l) and the galactic radiation (dominating F_2).

The dust emission is basically thermal. We computed the total infrared emission by adding the fluxes in the four bands:

$$F = F_{12} + F_{25} + F_{60} + F_{100} \tag{3.1}$$

Assuming a dust emission law in the form of $B(T)/\lambda^{\alpha}$ where $B(T)$ is the black body (BB) radiation at T temperature, λ the wavelength and α depend on the physical properties of the emitting dust, we put $\alpha = 1$. However, recent studies of the far infrared radiation of the ZL with the ISO satellite indicate nearly BB radiation (Leinert et al. , 2002), i.e $\alpha = 0$. The specification of α influences the numerical results obtained, of course, but our goal is only to demonstrate the link between the statistical procedure and the physical quantities.

The F_i/F ratios (i is 12, 25, 60 or 100) depend only on T if a region determined by one characteristic temperature. Supposing the validity of the dust emission law given above we computed the loci of such regions in Fig. 3.2,

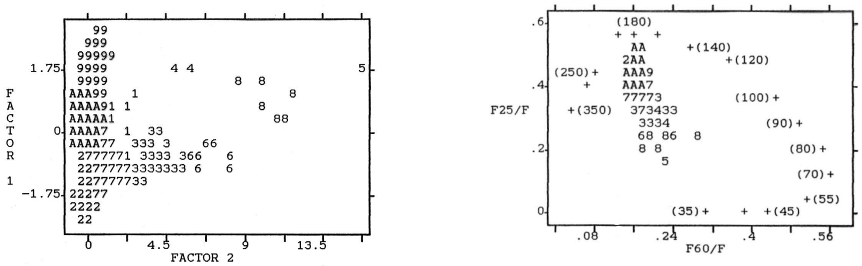

Figure 3.2: Character plot of regions (clusters) of similar properties in the $\{F_1; F_2\}$ factor plane (left panel). The identical symbols mean physically similar regions. The basic features of this plot are the two 'fingers' pointing upwards and nearly horizontally. These 'fingers' may be identified with the Zodiacal Light (dominating $Factor\,l$) and the galactic radiation (dominating $Factor\,2$) (Balázs et al. , 1990). Distribution of duster members in the $\{F25/F; F6O/F\}$ plane (right panel). The coding of dusters is the same as in the left panel. The loci of dust low $\alpha = 1$ radiations of different temperatures are marked with crosses. The numbers in parentheses are the respective temperatures. Note that the wedge-shaped distribution of symbols representing real measurements points towards about 40 K and 200 K dust temperatures (Balázs et al. , 1990).

marked with crosses the sources of different temperatures in the line of sight. As a consequence, the real points in Fig. 3.2 are not on the theoretically computed line but deviate from it according to the relative intensity of superimposed sources of different temperatures. One gets a wedge-shaped distribution of symbols representing real measurements pointing towards about 40 K and 200 K dust temperatures. This distribution can be obtained from the superimposed ZL and Galactic sources with these characteristic temperatures.

3.2 Sturucture and Dynamics of the Cepheus Bubble

3.2.1 Basic characteristics

Analysing the IRAS $60\mu m$ and $100\mu m$ maps of Cepheus, Kun, Balázs and Tóth (1987, hereafter KBT) reported that the area $l = 90^o$ to 106^o, $b = +3^o$ to $+10^o$ is faint at far-infrared wavelengths, and is encircled by a ring-shaped region of enhanced infrared emission, the 'Cepheus Bubble'. They compared the IRAS maps with Dubout-Crillon's (1976) $H\alpha$ photographs, and demonstrated that the brightest parts of the infrared ring correspond to the HII regions IC 1396, Sh2-129, Sh2-133, Sh2-134, and Sh2-140 as well as to fainter or smaller HII regions not listed in Sharpless' (1959) catalogue. Since most of the listed bright HII regions were placed at about 900 pc from the Sun in the literature, KBT accepted this value for the Cepheus Bubble as well.

The infrared ring is probably related to the 6^o loop structure discovered at $l = 103^o, b = +4^o$ by Brand and Zealey (1975) as semi-circular filamentary dust lanes and faint emission nebulosity on the POSS prints. They associate this structure with Cep OB2 by noting that to the west the dust seems to interact with IC1396, although in their Table 1 erroneously Cepheus OB1 is given as related OB association. In the area encircled by the infrared ring a deficiency in HI column density was observed by Simonson and van Someren Greve (1976, hereafter SVSG) in the $[-20, -6]$ kms^{-1} radial velocity range. The infrared ring coincides positionally with the OB association Cepheus OB2, which consists of two subgroups: the older and more dispersed one, Cep OB2a ($t \geq 8 \times 10^6 yrs$, de Zeeuw and Brand (1985)) occupies the interior of the ring, while the younger and smaller subgroup, Cep OB2b ($t = 4 \times 10^6 yrs$, SVSG) is situated at the edge of the ring. Since the distance of Cepheus OB2 is also about 900 pc, KBT proposed a physical link between the association and the bubble. A similar link was assumed earlier by SVSG, interpreting the low HI content in the region as a sign of the full ionization of the interstellar gas by the Cep OB2 association.

KBT proposed that the Cepheus Bubble had been created by the strong stellar wind/UV radiation and the subsequent supernova explosion of the most massive star in the older subgroup Cep OB2a. The exploded star was perhaps the former companion of the runaway star λ Cep whose proper motion points backwards approximately to the centre of the bubble. From tracing back this motion, KBT estimated an age of $3 \times 10^6 yrs$ for the bubble. The proper motion measured by the Hipparcos satellite does not significantly change this age estimate. The formation of the younger subgroup, Cep OB2b could have been triggered by stars of Cep OB2a via stellar wind and propagating ionization fronts as proposed by KBT and Patel et al. (1995). This trigger could also be responsible for the birth of OB stars and cold embedded IRAS sources along the periphery of the Cepheus Bubble (KBT, Balázs & Kun, 1989).

De Zeeuw et al. (1999) determined the distances of nearby OB associations using proper motions and parallaxes measured by Hipparcos. Their work on 49 members of Cep OB2 resulted in a distance of $559 \pm 30pc$, significantly lower than the $900pc$ assumed before. However, the recalibration of the upper main sequence, suggested by Hipparcos, would also lessen the distance of those objects along the periphery of the bubble (HII regions and reflection nebulae) which were used as distance indicators for the Cepheus Bubble by KBT. Thus the arguments for the physical connection of the bubble with Cep OB2 are probably not affected by the new distance values. Therefore further investigations are needed before accepting the shorter Hipparcos scale.

A better understanding of the history of this Cepheus region requires more detailed mapping of the interstellar matter (including better distance estimates) as well as information on the large scale motions. Velocity information can also help to separate distinct interstellar features projected on the IRAS images. Recently, Patel et al. (1998) conducted a $CO\,(1-0)$ spectral line survey in Cepheus and discussed the origin and evolution of the Cepheus Bubble on the basis of the overall distribution of molecular gas. They also performed an HI survey and concluded that the bulk of the interstellar gas associated with the bubble is in atomic form. In this section we investigate the large scale morphology and kinematics of the Cepheus Bubble by a detailed analysis of the distribution of atomic hydrogen.

In the following we analyse HI 21 cm measurements taken from the Leiden/Dwingeloo survey, in order to identify the atomic gas component of the Cepheus Bubble. In addition to the more traditional methods, the analysis of the HI maps is also performed by using multivariate statistical methods. After connecting these new pieces of information into a coherent picture, we speculate about the possible origin of the Cepheus Bubble.

3.2.2 Distribution of neutral Hydrogen

The HI data were taken from the Leiden/Dwingeloo HI survey (Hartmann & Burton, 1997). The angular and velocity resolution of the spectra are 36' and $1.03\,kms^{-1}$, respectively, covering the velocity range $[-450, +450]\,kms^{-1}$. The observed positions are distributed on a regular grid with steps of 0.5^o both in l and b. This grid provides a spatial resolution of up to 3 times higher than obtained by SVSG, although the Leiden/Dwingeloo sampling is somewhat coarser than that of the HI data set of Patel et al. (1998).

HI distribution in the channel maps In the region $l = 95^o$ to 110^o, $b = 0^o$ to $+15^o$ HI emission is dominated by a narrow galactic plane layer of $b < +6^o$ at any radial velocity between $-110\,kms^{-1}$ and $20\,kms^{-1}$. Although HI emission is detectable over a large range in both b and V_{LSR}, at $b > +6^o$ the most prominent emission features appear in two discrete velocity intervals at $[-155, -135]\,kms^{-1}$ and $[-45, +20]\,kms^{-1}$. The HI structure at $V_{LSR} \simeq -145\,kms^{-1}$ belongs to the extended Outer Arm high velocity cloud (Wakker and van Woerden, 1991), whose study is beyond the scope of this work.

We focus on the $[-45, +20]\,kms^{-1}$ velocity range, and search for hydrogen structures possibly associated with the infrared ring found in the IRAS maps. We display in Fig. 3.3 and 3.4 a series of HI maps by integrating the spectra over $4\,kms^{-1}$ velocity intervals between $-38\,kms^{-1}$ and $+10\,kms^{-1}$. The maps show that the bulk of HI emission arises from the $[-14, +2]\,kms^{-1}$ velocity range. Since the infrared ring is expected to correlate with the projection of the most prominent HI features on the sky, we plotted in Fig. 3.5 the hydrogen emission integrated between $-14\,kms^{-1}$ and $+2\,kms^{-1}$. The figure shows a well-defined closed ring around the low emission region $l = 101^o$ to $l = 105^o$, $b = +2^o$ to $b = +2^o$, a result also published by SVSG and by Patel et al. (1998).

An inspection of the HI maps of Fig. 3.3, 3.4 reveals loop structures in several velocity regimes. The most prominent ring structure, with sharp inner edge in the direction of the Cepheus Bubble, appears in the $V_{lsr} = [-14, -10]\,kms^{-1}$ range. A similar ring-like pattern is clearly recognizable at more negative velocities as well. Between $-26\,kms^{-1}$ and $-14\,kms^{-1}$, Fig. 3.3 shows a low emission area ('hole'), bounded by stronger HI emission regions, and the whole structure extends over about parallel to the galactic plane. Although the center of this higher negative velocity hole ($l \simeq 103^o, b \simeq +8^o$) is at slightly higher galactic latitude than that of the ring in Fig. 3.5, the transition between these two loop structures is continuous in the velocity space (Fig. 3.3, 3.4), providing a strong evidence for their physical link. At even higher negative velocities ($V_{LSR} \simeq -26\,kms^{-1}$) the upper boundary of the hole is fragmented, and the loop structure is no longer visible. The fragments are, however, still recognizable at more negative velocities, roughly following the trend that fragments of higher negative radial velocities appear closer to the center of the former ring. At $V_{LSR} \simeq -30\,kms^{-1}$ even these fragments disappear. The interpretation of these results in terms of an expanding shell is given in Sect. 3.2.3.

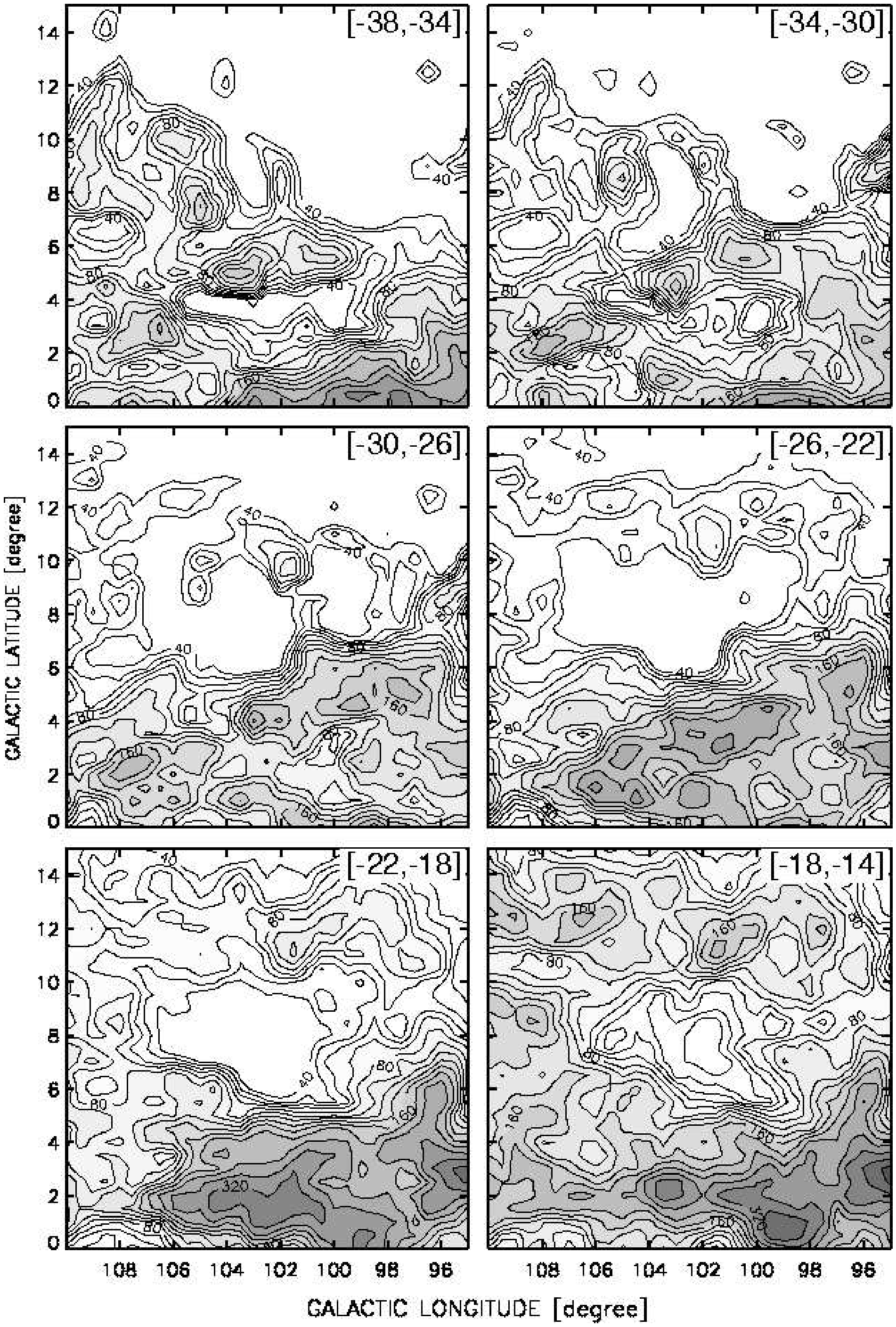

Figure 3.3: HI 21 cm maps toward the Cepheus Bubble in the $[-38,-14]\,kms^{-1}$ velocity range, integrated over $4\,kms^{-1}$ intervals. The lowest contour corresponds to 40 $Kkms^{-1}$.

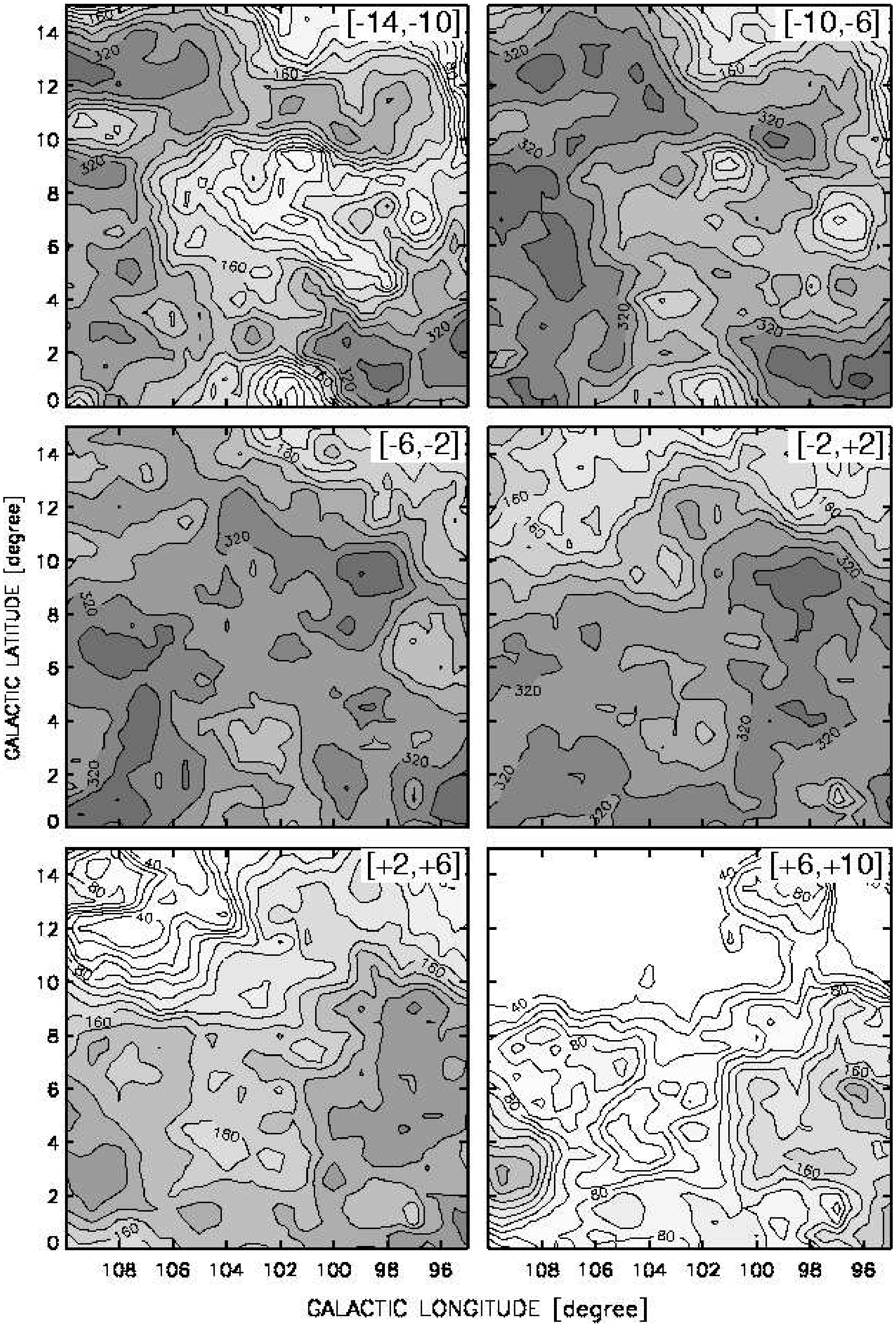

Figure 3.4: Continuation of the previous Figure in the $[-14, +10]\,kms^{-1}$ velocity interval.

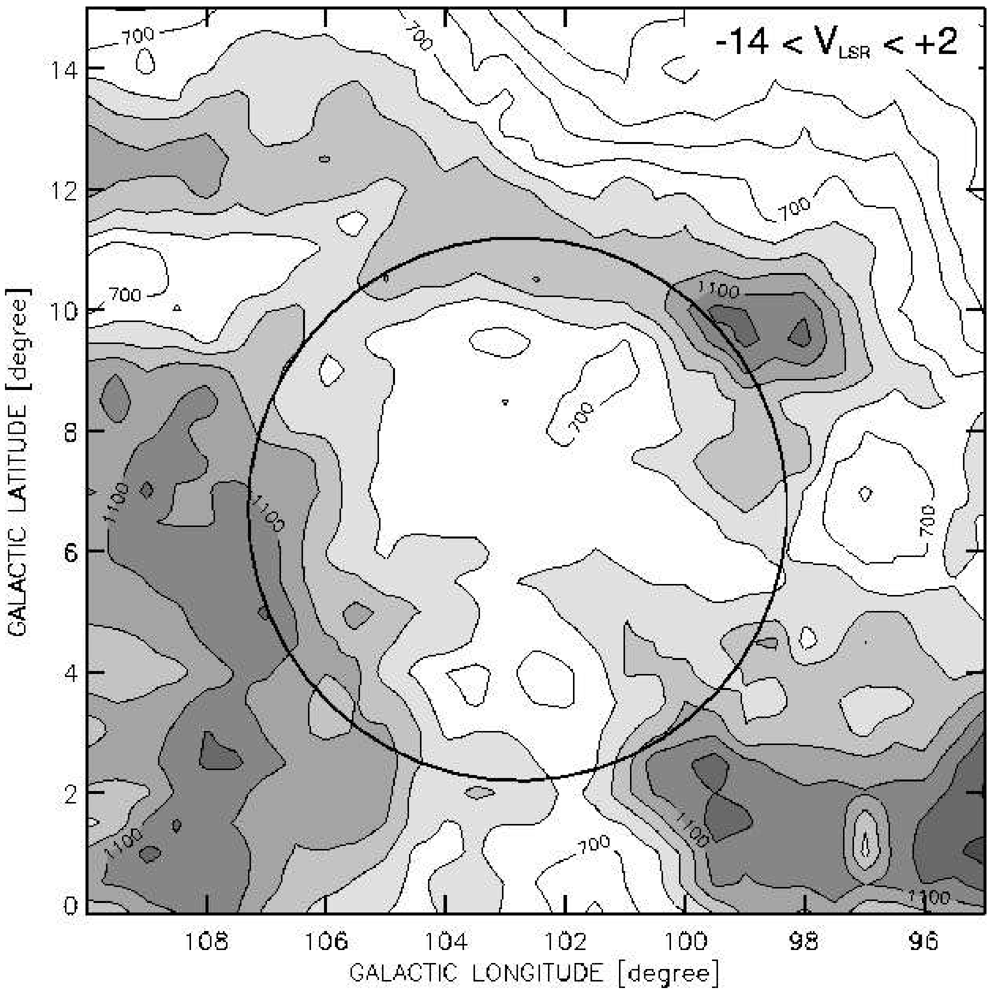

Figure 3.5: *HI* 21 *cm* intensity integrated between -14 and +2 kms^{-1}. The lowest contour is 300 $Kkms^{-1}$, the contour interval is 100 $Kkms^{-1}$. The circle around the low emission region marks the approximate outer boundary of the infrared ring.

So far we identified the significant cloud complexes related to the Cepheus Bubble by visual inspection of the HI maps. This method, however, is not automatic, can be somewhat subjective, and works less efficiently in regions where the resolution of the kinematic distances, resulting from the differential rotation of the Galaxy, is poor (like in the Cepheus region which is close to the $l = 90^o$ tangent point). Visual inspection may also fail to identify structures which extend over very large radial velocity ranges due to internal and/or peculiar motions. In the next subsection we use a multivariate statistical method for identifying the main structures in the data cube representing the Cepheus Bubble, free from subjective bias.

Multivariate analysis of the HI channel maps The positional and velocity data of the neutral hydrogen form a data cube $\{l, b, v\}$. We assume that the HI emission is optically thin, and the observed channel maps are weighted superpositions of k components which represent the main hydrogen cloud complexes, i.e.:

$$I_i = \sum_{j=1}^{k} a_{ij} F_j \qquad\qquad i = 1, \dots, n \tag{3.2}$$

where I_i, a_{ij} and F_j are the measured intensities in the channel maps, the weighting coefficients, and the contributions of the components, respectively, and n is the number of the channel maps. Normally, we may assume that $k < n$. Description of the observed variables by linear combination of hidden variables (factors) is a standard procedure of multivariate statistics.

We make the assumption that the correlation between the F_j components is negligible. This assertion enables us to apply the principal components analysis (PCA) for finding the number of significant components (factors) and their numerical values, using standard techniques implemented in statistical software packages. The PCA represents the observed variables ($T_b \Delta v$ values in our case) as linear combinations of non-correlated background variables (principal components). We note that although PCA is often used for finding the factors, there are many other techniques for obtaining a factor model. PCA and factor analysis represent two different procedures, strongly related but not identical.

PCA obtains the factors by solving the eigenvalue equation of a matrix built up from the correlations of the observed quantities. The components of the obtained eigenvectors serve as coefficients of the factors (the significant principal components in this technique) in the equation given above. The eigenvalues $\lambda_1, \dots \lambda_n$ give some hint for the 'importance' of the corresponding components. The $\lambda_i / \sum\limits_{j=1}^{n}$ and $\sum\limits_{j=1}^{n} \lambda_i / \sum\limits_{j=1}^{n}$ ratios indicate what percentage of the variance of the observed variables can be explained by the i-th principal component (explained percentage) and by the linear combination of the first i principal components (cumulative percentage). For further details of this technique see Murtagh and Heck (1987). PCA is a standard procedure of many statistical software packages. Balázs et al. (1990) used this technique to separate the galactic background from the zodiacal light.

We analyzed a matrix built up from the mutual correlations between the HI channel values. We used altogether 43 channels in the [-38,+6] kms^{-1} region, corresponding to a sampling frequency of about 1 kms^{-1}. Table 3.2 shows the eigenvalues and the explained percentages of the principal components, as well as their cumulative percentages. We found that the 6 major principal components having eigenvalues larger than 1.0 can describe 95.4% of the variance of the observed HI channel maps. We kept these principal components for getting the factors describing the observed HI distribution. The results demonstrate that the 43 channel maps can be represented by only 6 maps created by the PCA, while the remaining 37 maps carry mainly redundant information, and may be dropped from the further data analysis. However, the PCA does not guarantee that a factor map contains only physically related objects: if two independent HI clouds appear exactly in the same velocity range, they would be included in the same factor.

Table 3.2: Results of the factor analysis on the HI data of the Cepheus Bubble. There are 6 eigenvalues > 1 reproducing 95.4 % of the total variance of the data.

PC	Eigenvalue.	Pct. of Var. [%]	Cum.Pct. [%]
1	20.41	47.5	45.5
2	7.80	18.3	65.8
3	5.87	13.7	79.5
4	3.41	7.9	87.4
5	1.87	4.3	91.8
6	1.56	3.6	95.4
7	0.66	1.5	96.7
8	0.52	1.2	98.1
⋮	⋮	⋮	⋮
43	0.00	0.00	100.0

Fig. 3.6 presents maps of the 6 factor values, and Fig. 3.7 shows the weighting coefficients for these 6 factors as a function of radial velocity. Fig. 3.7 reveals that each factor has a well-defined radial velocity interval

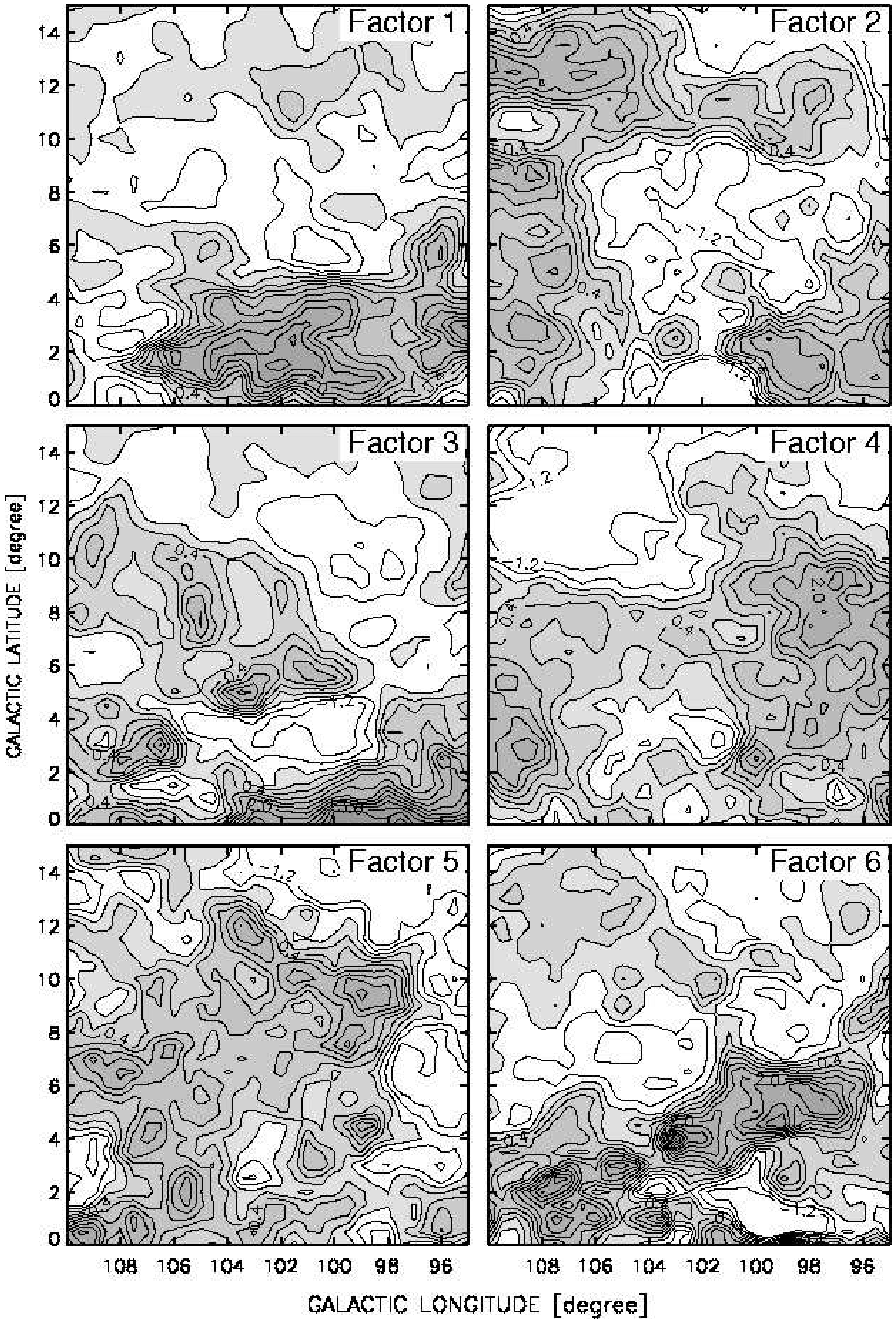

Figure 3.6: Maps of the factor values. The lowest contour is -1.2, the contour interval is 0.4. The bubble itself is described predominantly by Factor 2.

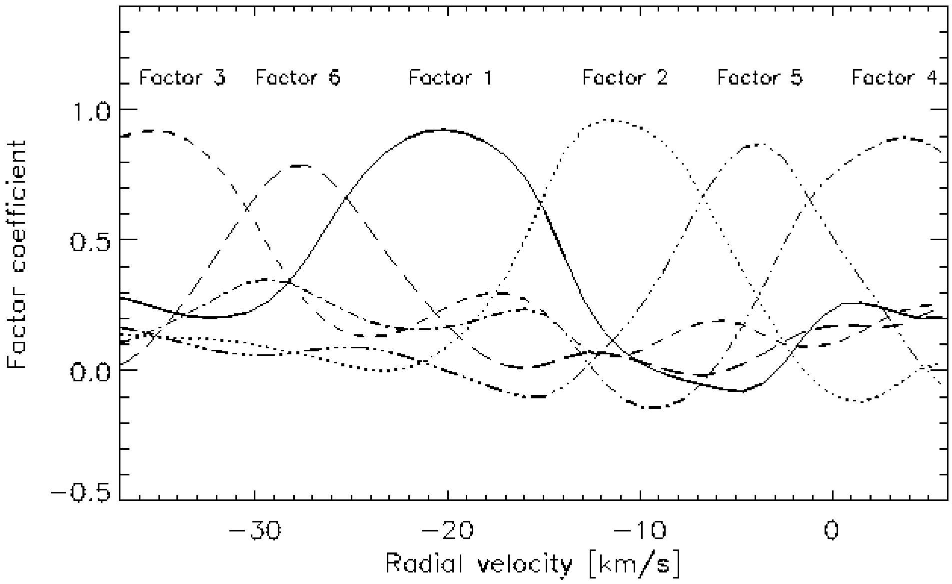

Figure 3.7: Factor weighting coefficients vs. radial velocity of the channel maps

where it is dominant and where the contributions of the others are almost negligible.

In the following we compare the results of the PCA with those derived in the previous subsection. The well defined loop structure in the velocity range $[-14, -10]\, kms^{-1}$ (Fig. 3.4) can easily be identified with Factor 2, by both their patterns and their velocity ranges. Factor 1, although dominated by a very strong feature at $b \leq +5^o$, contains also the loop visible in the maps of Fig. 3.3 between $-26\, kms^{-1}$ and $-14\, kms^{-1}$. Fragments of this loop towards the centre can be associated with Factor 3. On the more positive velocity side, Factor 5 is dominated by a concentration towards the interior of the ring, although at this velocity significant foreground contamination due to local HI can be expected. Factor 4, which is important only at more positive velocities, and the weak Factor 6 apparently do not carry substantial information on the bubble.

We found that all prominent emission structures, recognized in the HI maps of Fig. 3.3, 3.4, were identified by the PCA as well, and the results of the multivariate analysis could be converted into useful physical information. This approach offers an objective way to get an unbiased estimate of the characteristic radial velocities of the most significant structures, which is not given by the visual inspection. The method also shows how to reduce the size of our data cube without losing too much information, and therefore it could be used for automatic analysis of larger data sets, too.

***HI* distribution in the position-velocity space** The existence of an extended depression in the hydrogen emission at high negative velocities is also evident from Fig. 3.8, a position-velocity diagram taken perpendicularly to the galactic plane at $l = 102^o$. The figure shows a large hole between $b = 3^o$ and 11^o, $V_{LSR} = -37$ and $-4\,kms^{-1}$, and suggests that most HI features in this velocity range (including the fragments at $V_{LSR} < -26\,kms^{-1}$) belong to a large interstellar structure, forming a closed loop around the hole. Similar cuts at $l = 100^o$ and $l = 104^o$ reveal similar structures but with a somewhat smaller diameter, justifying our choice of $l = 102^o$ as the main cross section of the bubble.

We propose to interpret the observed spatial-velocity distribution as radial expansion of a 3-dimensional shell. According to this interpretation, the regular ring patterns in the $[-26, -6]\,kms^{-1}$ velocity range (Fig. 3.3, 3.4) correspond to different cross sections of the shell, while the blue-shifted fragments at $V_{LSR} \simeq -37\,kms^{-1}$ represent its approaching part. The shift of the fragments toward the hole's center at larger negative radial velocities is consistent with the expansion model, which predicts that at blue-shifted velocities an expanding shell appears in the form of concentric rings of apparent radii decreasing with radial velocity. The receding wall of the expanding shell, however, is not easy to identify. Fig. 3.8 suggests that the receding side is seen at $V_{LSR} \simeq -4\,kms^{-1}$, but this emission could be seriously contaminated by HI emission from the solar neighborhood expected at $V_{LSR} \simeq 0\,kms^{-1}$. The map of Factor 5 (Fig. 3.6), however, which contains emission having central velocity of $\simeq -4\,kms^{-1}$, reveals a mass concentration towards the interior of the bubble instead of the picture of a homogeneous foreground emission. This result may indicate that the shell is closed at $V_{LSR} \simeq -4\,kms^{-1}$.

The apparent center of the shell lies at $V_{LSR} \simeq -20\,kms^{-1}$ and $b \simeq +7^o$ (Fig. 3.8). This velocity is significantly more negative than the $V_{LSR} \simeq -12\,kms^{-1}$ derived in the previous subsection as the characteristic velocity of the best defined cross section of the bubble. Fig. 3.8, however, shows that the emission of the shell along the velocity axis is asymmetric, concentrating towards more positive radial velocities (the measured brightness temperatures are approximately 15 K and 60 K in the directions of the approaching and receding sides of the shell, respectively). This asymmetric mass distribution may indicate that the Bubble was formed at the near side of a large cloud complex. Assuming optically thin emission, the 1:4 brightness temperature ratio between the approaching and receding sides is transformed into the same ratio for the corresponding column densities. In order to determine

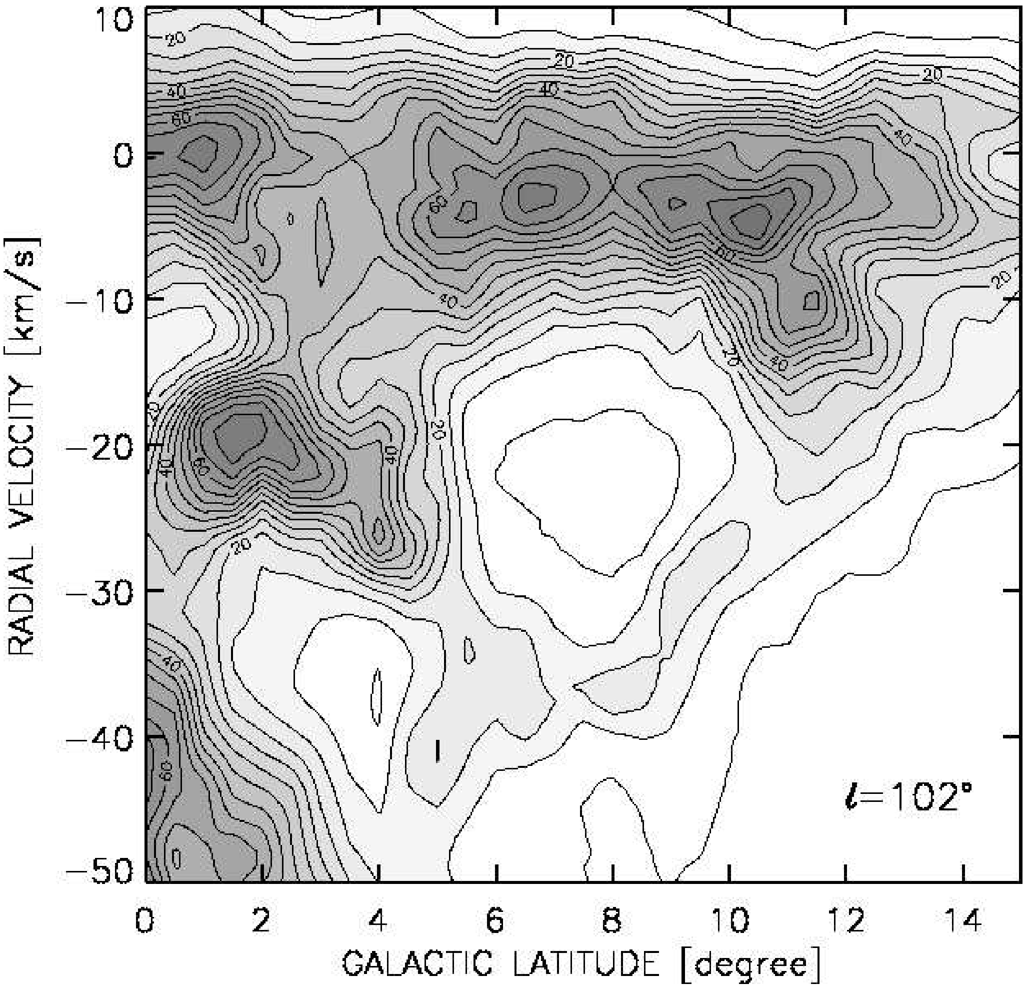

Figure 3.8: Position-velocity map taken perpendicularly to the galactic plane at $l = 102^o$.

the true central velocity of the expansion, we weighted the radial velocities of the approaching ($\simeq -37\,kms^{-1}$) and receding ($\simeq -4\,kms^{-1}$)side with the densities. The result is $V_{LSR} \simeq -10kms^{-1}$, close to the $V_{LSR} \simeq -12\,kms^{-1}$ yielded by the PCA in the previous subsection, but in clear contradiction with the value of $-2kms^{-1}$ given by Patel et al. (1998), who assumed that the systemic radial velocity of the Bubble is identical to that of the ionized gas of $IC1396$. This difference in the systemic velocities may explain also the discrepancy between the kinetic energies obtained by them and those calculated below. Following the same procedure, i.e. weighting the velocities with the corresponding HI densities, we also calculated an 'effective expansion velocity' of $\simeq 10.2\,kms^{-1}$. These velocities will be used to model the possible physical origin of the bubble.

3.2.3 Physical nature of the bubble

Possible scenarios for the origin and evolution of the Cepheus Bubble have been proposed by KBT and Patel et al. (1998). In the following we discuss how the 3-dimensional expanding HI shell, described in the present work, fits into these scenarios.

Mass and energetics We determined the mass of the HI shell by summing up the column densities of the HI gas (Hartmann & Burton 1997) over the apparent area of the shell. In order to separate the emission related to the shell from foreground and background emissions we multiplied the measured $T_b\Delta v$ channel values by the corresponding factor weighting coefficients from factors 2, 3 and 5, obtained in Sect. 3.2.2. In this way we obtained a mass of $M \simeq 2.7 \times 10^5 M_\odot$, including a factor of 1.4 to account for the total mass to HI mass ratio (Brown et al. 1995). This value is close to the value of $3 \times 10^5 M_\odot$ obtained by Patel et al. (1998). Including the mass of molecular gas ($1 \times 10^5 M_\odot$, Patel et al. 1998) would not change significantly the total mass.

From the estimated mass and observed size of the shell we obtain an initial ambient density of $n \sim 5.3\,cm^{-3}$ for the interstellar medium in this region. The mass of the bubble, together with the effective expansion velocity derived in Sect. 3.2.2, provides an estimate of the total kinetic energy of $E_{kin} \sim 2.7 \times 10^{50} erg$ for the expanding shell. Following Weaver et al. (1977), we assume that about 20% of the total energy deposited into the ISM is converted into kinetic energy. With this assumption the total energy of the bubble is $E_{kin} = 1.4 \times 10^{51} erg$.

The derived total energy is close to the canonical value of a supernova explosion of $10^{51} erg$, thus such an event is a possible origin of the expansion.

In order to check if other sources were also able to provide the required energy of the expansion, we estimated the total contribution of stellar winds from some 20 O-type and early B-type members of Cep OB2a. The stars selected are located within a circle of 3^o (the approximate angular diameter of the internal cavity) around HD 207198, and we assume that all are members of the association. The mechanical luminosity was calculated by estimating the mass loss from the spectral type and the luminosity vs. mass loss relation of Garmany et al. (1981) and from the analytical expression given by Kudritzki (1998). This gives $3 \times 10^{35} ergs^{-1}$ for the mechanical luminosity of the OB stars. Assuming that this luminosity was constant during the lifetime of Cep OB2a ($\simeq 8 \times 10^6 yrs$, Sect. 3.2.1), the total power exerted amounts to $7.5 \times 10^{50} erg$. This figure agrees within a factor of 2 with the total energy required for the expansion of the bubble. Considering the uncertainties in our computations, we conclude that the integrated stellar wind from the existing early-type stars in the interior of the bubble during a period of $\simeq 8 \times 10^6 yrs$ could also power the observed expansion.

Kinematics and age Massive O-type stars affect their environments via UV radiation, stellar wind, and supernova explosion. UV photons ionize the interstellar gas and develop an HII region, as well as homogenize the surrounding medium by photo-evaporating the nearby clouds and/or removing them via the 'rocket effect' (McKee et al. 1984). This homogenization process enables us to describe the temporal evolution of the expansion by the analytical formulae of Weaver et al. (1977) and Chevalier (1974) (see also Tenorio-Tagle and Bodenheimer (1988)). In both the stellar wind bubble (SWB) and supernova remnant (SNR) scenarios the evolution of the size is given by a power law of $R(t) \sim t^p$, where $p = 0.31$ in case of SNR in post-Sedov phase and $p = 0.6$ in the case of SWB. The expansion velocity of the radius of the ring pattern is given by the time derivative of $R(t)$, i.e. $V_{exp} \sim p\, t^{p-1}$. Division of $R(t)$ by $V_{exp}(t)$ gives a simple equation for obtaining t, the age of the Bubble. We emphasize that none of the mechanical luminosity of the stellar wind, the explosion energy of the SN or the density of the ambient interstellar matter enters into the final expression of t, depending only on p, the observed size and expansion velocity of the shell. The ages obtained from this simple formula give therefore facing values of $t = 1.7 \times 10^6 yrs$ (SNR) and $t = 3.1 \times 10^6 yrs$ (SWB) for the expansion age of the Bubble.

The age derived for a SWB, however, is only 40% of the lifetime of Cep OB2a, and during this period the energy injected into the ISM via stellar wind is only $3 \times 10^{50} erg$. This energy is significantly lower than the

$E = 1.4 \times 10^{51} erg$ required for the expansion of the bubble. It would also be difficult to explain why the creation of the SWB does not coincide with the birth of most of the early-type stars. From these arguments we think that a supernova explosion occurring about $t = 1.7^6 yrs$ ago is a more straightforward explanation for the origin of the expansion. The derived expansion age of a SNR is also consistent with the kinematic age of the runaway star $\lambda\, Cep$ ($2.8 \pm 1 \times 10^6 yrs$, Stone (1979)), which was proposed by KBT to be the former companion of the exploded star.

Evolution of the Cepheus Bubble As was discussed in the previous paragraph, the observed large scale expansion of the Cepheus Bubble is probably due to a relatively recent supernova explosion in the Cep OB2a association. The true age of the bubble, however, may be significantly larger, because the strong stellar wind and UV radiation from the progenitor of the supernova, along with other OB stars in Cep OB2a, is expected to create a large cavity already long before the supernova event. Both KBT and Patel et al. (1998) propose that the younger subgroup of the association Cep OB2b has been triggered by the older subgroup Cep OB2a. Since the age of the most massive star of the younger subgroup is about $4 \times 10^6 yrs$ (Patel et al. 1998), an extended cavity/shell structure of approximately the present size was already present $4 \times 10^6 yrs$ ago.

In this picture the supernova has exploded in an already existing low density cavity. It is likely that by the time of the supernova event the shell around the cavity was already fragmented and its expansion practically stopped (Patel et al. 1998). The expanding shock front of the supernova, however, revived the old shell by forcing its fragments into motion again. The shock front also interacted with those regions which existed before the explosion, such as IC 1396, and possibly Sh2-140, NGC 7129. This interaction can explain the relatively sharp and well-defined inner edge of the Cepheus Bubble. The supernova shock might have also influenced the structure of the star forming regions all along the Cepheus Bubble and triggered the recent wave of star formation indicated by the IRAS measurements (Balázs & Kun , 1989; Patel et al. , 1998).

Chapter 4

Star count study of the extinction

4.1 Extinction map of L1251

The **Cepheus Flare** is an extended complex of molecular clouds, where luminous stars are clearly absent, and it harbors dark clouds forming low mass stars (for the details see the comprehensive study of Kun (1998) on this region). **L1251** ($\alpha_{2000} = 22^h36^m\!.1; \delta_{2000} = +75^o16'$) is a dark cloud in this region at a distance of $300 \pm 50pc$ (Kun & Prusti , 1993) and apparently belongs to this complex. Already at the advent of molecular radio astronomical studies in the late sixties the cloud was detected among the ten brightest OH emission sources on the sky by Cudaback & Heiles (1969). The cloud was also listed among sources of strong formaldehyde emission (Dieter , 1973; Sume et al. , 1975).

After these early successes of molecular radio observations L1251 apparently escaped the attention of radio observers. Although Kun (1982) discovered a number of $H\alpha$ emission stars associated with L1251, it appeared again in molecular radio studies in the late eighties (Sato & Fukui , 1989; Zhou et al. , 1989; Benson & Myers, 1989). Zhou et al. (1989) recognized an ammonia core in the dense part of the cloud and comparing the large number of $H\alpha$ objects associated with the cloud Sato & Fukui (1989) concluded that the star formation efficiency is anomalously high in L1251. From all of these facts it was obvious that the star formation processes in this region needed further detailed investigations. This motivated Kun & Prusti (1993) to study the properties and distribution of faint IRAS point sources along with the $H\alpha$ emission objects. Their study indicated that L1251 has been forming low-mass stars with an efficiency higher than usually encountered in dark clouds. The eastern, head region of the cloud has been found to contain more evolved YSOs than the western (tail) side.

A census of dense cores carried out by Tóth & Walmsley (1996) based on the NH_3 1.3 cm line using the 100 m dish at Effelsberg resulted in detection of eight ammonia cores with typical size of $FWHM \approx 2'$ ($0.2pc$ at $350pc$ distance). Five of the cores were found to be gravitationally bound. L1251 was recently surveyed in several mm lines, the structure and kinematics of the cores were studied e.g. by Caselli et al. (2002), Lee et al. (1999), Lee et al. (2001) and Nikolić et al. (2003). Both infall and outward motions were detected in the gravitationally unstable cores.

Tóth et al. (1995) tried to explain the shape of the cloud, as seen in the CO observations (Goodman et al. , 1993; Sato et al. , 1994), by hydrodynamical modelling assuming an encounter of a dense molecular cloud with an external shock, probably originated from a nearby supernova explosion. Reality of this assumption gets some support from the detection of a soft X-ray excess region eastwards from the cloud by Grenier et al. (1989). Further confirmation of an SN explosion comes from the space motion of the runaway star HD203854 (Kun et al. , 2000).

Although the dense gas component of the cloud is well studied, much less is known about the distribution and properties of interstellar dust in and around L1251. Recently Kandori et al. (2003) studied the extinction of the dust component using $B,\ V,\ R,\ I$ star counts and found high R_V values indicating grain growth in the head of the cloud.

In the present work we study the spatial distribution of dust, the mass, and the basic physical properties in the cloud by means of optical extinction maps in $B,\ V,\ R,\ I$ colors. For obtaining direct information on the spatial distribution of dusty material the study of star counts is still one of the most reliable approach. The basic aim of the recent chapter is to carry out such kind of analysis.

4.2 Input data

In order to study the surface distribution of the optical extinction we obtained star counts in $B,\ V,\ R,\ I$ colors based on photographic observations with the 60/90/180 cm Schmidt telescope of Konkoly Observatory. The plates were digitized with a pixel size of 20 μm in a 1.5×1.5 degree field around L1251 using the PDS microdensitometer of Vienna Observatory in 1991. We scanned 4, 3, 3, and 2 plates in B, V, R_J, I_J color, respectively. The scans were processed with the ROMAPHOT photometric programme integrated into the MIDAS data analysis package. The plates were calibrated via CCD observations performed with the 1.23 m Ritchey-Chretien telescope of the German-Spanish Observatory, Calar Alto (Balázs et al., 1992). The limiting magnitude of the photographic survey was 19.0, 18.5, 17.5, and 16.5 in B, V, R_J, and I_J, respectively. Although these figures are less than could be obtained from the Digitized Sky Survey maps (see e.g. the extinction maps of Cambrésy (1999)), the well defined color system, the much better photometric calibration and four colors, however, make a sense for using our data. The estimated completeness of star counts was about 1.5 magnitude above the detection limit of the plates. The final star count maps of the region were obtained by counting the stars in each color on pixels of $6' \times 6'$ size and a $2'$ mesh of the star count maps was selected in both directions. (This resolution approximately corresponds to those of the 100 μm IRAS maps). Kun (1982) and Kun & Prusti (1993) lists 12 $H\alpha$ objects (candidate pre-main sequence stars) apparently associated with the cloud. Their effect on the R and I star counts might have significance. Consequently, we have omitted them from the further analysis.

4.3 Extinction maps from star counts

Several studies indicated empirically that the surface distribution of star counts is an excellent tracer of optical extinction. (e.g. Dickman (1978); Cernicharo et al. (1985)). This means that there is a simple linear relationship between the logarithmic star counts and the a_{cl} extinction of a dust cloud:

$$a_{cl} = a \times \log(N(m)) + b(m) \tag{4.1}$$

where $N(m)$ is the cumulative star count up to a given m limiting magnitude, the a constant and $b(m)$ depend on the Galactic longitude and latitude; $-b/a = \log(N_0(m))$ measures the logarithmic star count in an extinction-free region (Dickman , 1978).

In the following we calibrate this expression using a multivariate statistical method, the k-means clustering, and a maximum likelihood procedure. Based on this calibrated relationship we assign extinction values to each star count pixel yielding an extinction map of the cloud. The procedure gives as a byproduct the distance of L1251 which enables us to calculate the mass of the cloud. Since we derive the extinction of the cloud in different colors we also discuss the ratio of the selective to total extinction.

4.3.1 Definition of areas of similar extinction

The *B, V, R, I* star count data on the pixels defined above represent a distribution of points in a four dimensional *B, V, R, I* parameter space. Following the relationships between star counts and extinction we assumed that the regions of equal extinction have equal star counts on the maps. In other words looking for areas of equal extinction means searching for points of similar coordinates, i.e. lying close to each other in the four dimensional parameter space, made up from the pixel values of the star count maps.

The linear relationship between the logarithmic star counts and interstellar extinction predicts a one dimensional manifold in the four dimensional parameter space, stretched by the extinction, and there is a Poisson noise superimposed on it by the star counts. Therefore, to convert star counts into optical extinction we had to divide the whole star count map into regions of equal counts. According to Equation (4.1) the logarithms of the star counts are scaling linearly with the extinction; therefore we used the logarithmic *B, V, R, I* star counts in this procedure.

In order to divide the points into groups of similar extinction in the parameter space we invoked the technique of k-means clustering (see e.g. Murtagh & Heck (1987)). K-means clustering orders the points in the parameter space into k groups. The k number of the groups should be specified before running the clustering procedure. Assigning the points to any of the groups proceeds on the basis of some distance measure between the points. We used squared Euclidean distance.

There are no definite criteria for fixing the value of k. By trial we selected $k = 5$ enabling clear separation of the high extinction regions from those of low extinction and ensuring enough stars in each cluster for reliable analysis. Figure 4.1 shows the result of the clustering, reprojected onto the celestial sphere. Areas of same grey level in the map represent pixels belonging to the same cluster and consequently, having similar extinction. Table 4.1 summarizes the number of stars in each subregion (cluster) in each color. One can infer from the data of this table that the 5th cluster is scarcely populated, therefore we excluded it from the further analysis.

The central, densest part of L1251 is clearly separated from the surrounding lower extinction region. Already at the first glance of this figure one gets the impression of a flying bullet of supersonic speed in respect to an ambient medium. The less dense area surrounding the bullet-like main body of the cloud has a form of a bow shock. Accepting this view one can estimate the relative speed of the cloud to the ambient medium. The interaction of the cloud with its surroundings is probably the key issue in understanding the history of star formation. We will return to this problem later on in Sect. 4.4.

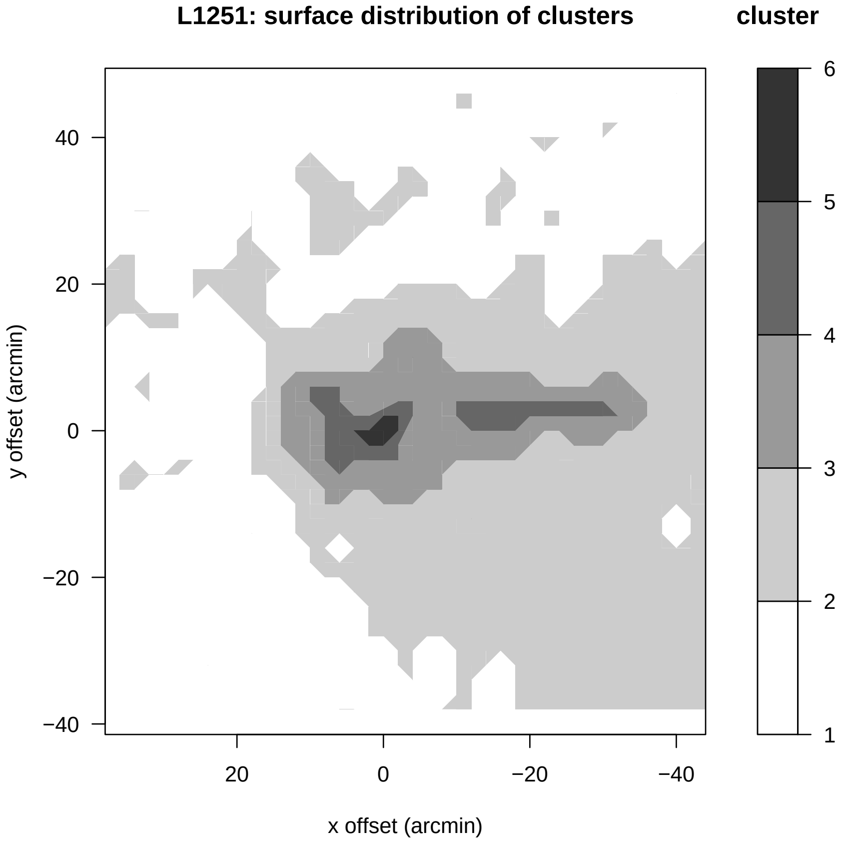

Figure 4.1: Result of the clustering, reprojected onto the celestial sphere. The areas of same grey level represent pixels in the map belonging to the same cluster and having similar extinction. We displayed the cloud in rectangular coordinates in a plane tangential to the celestial sphere. North is at the top and East is to the left. The (0,0) position corresponds to $\alpha_{2000} = 22^h36^m\!.1$; $\delta_{2000} = +75^o16'$.

Table 4.1: Summary of the star count analysis of L1251, number of stars in each cluster and in each color

Field	*Area* (*SqD*)	N_B	N_V	N_R	N_I
1.	0.9880	3982	3929	3939	2723
2.	0.8770	1883	2045	2186	1530
3.	0.1515	101	143	146	138
4.	0.0605	7	13	14	15
5.	0.0090	1	2	4	4

4.3.2 Modelling the effect of extinction on the star counts

Modelling differential star counts

For calibrating star counts in terms of optical extinction we attempted to model the observed apparent magnitude distribution of the stars. The starting point of our approach was the Galactic model of Wainscoat et al. (1992). Following the basic ideas of this model we assumed that the spatial distribution of the stars in the region investigated can be satisfactorily described by superposition of exponential disks, corresponding to different types of stars, and a spheroidal component. The exponential disks were defined by their scale heights and local stellar densities near the Sun. We used the data of Wainscoat et al. (1992) for the characteristic values of the exponential disks in the model.

In order to model the effect of the absorbing cloud on the star counts we assumed that besides the obscuring matter associated with L1251 there is no other significant dust cloud in the line of sight. This assumption is quite reasonable due to the high ($+15^o$) galactic latitude of L1251. As in the Wainscoat et al. (1992) model we assumed that the diffuse component of interstellar extinction has a form of

$$a_{diff}(r) = a_0 \sec(b)(1 - \exp(-r\sin(b)/h)) \tag{4.2}$$

where a_0 is a constant depending on the color selected, b is the Galactic latitude of the cloud and h is the scale height of the obscuring material. We added to the extinction described above a further component in the form of a step function.

$$a(r) = a_{diff}(r) + a_{cl}(r); \quad a_{cl}(r) = \begin{cases} 0, & \text{if } r < r_{cl} \\ a_{cl}, & \text{if } r > r_{cl} \end{cases} \tag{4.3}$$

Table 4.2: Summary of the star count analysis of L1251, extinction

Field	*Area (SqD)*	a_B	σ_{a_B}	a_V	σ_{a_V}	a_R	σ_{a_R}	a_I	σ_{a_I}
1.	0.9980	0.45	0.05	0.30	0.05	0.12	0.05	0.10	0.03
2.	0.8770	1.55	0.10	1.25	0.07	0.85	0.07	0.63	0.07
3.	0.1515	3.30	0.15	2.65	0.15	2.20	0.15	1.75	0.2
4.	0.0605	6.5	0.80	5.2	0.80	> 5.3	-	4.5	0.4

The distance of the cloud, r_{cl}, and the cloud extinction, a_{cl}, are constants to be estimated within a cluster obtained by the k-means clustering in the star count parameter space.

Maximum likelihood estimation of the distance and extinction

The model described in the previous section allows us to get the distance and extinction of the cloud using the maximum likelihood (ML) estimation. According to our model assumption the probability density of the apparent magnitude of stars in our observed sample is given by

$$A(m|a_{cl}, r_{cl}) = \omega \sum_{sp} \int_0^\infty D_{sp}(r)\Phi(m - 5\log(r) - 5 - a(r|a_{cl}, r_{cl}))r^2 dr \quad (4.4)$$

In the above formula ω is a suitably chosen normalizing constant, $D_{sp}(r)$ means the spatial density, Φ is the luminosity function assumed to have a Gaussian form and sp runs over the spectral types represented in the Wainscoat et al. (1992) model. We used Monte Carlo simulation in order to get a distribution of apparent magnitudes corresponding to the probability density of $A(m|a_{cl}, r_{cl})$. We used the MC simulated data to calculate the numerical values of A for the ML procedure.

The likelihood function in our case can be written as

$$L(a_{cl}, r_{cl}) = \sum_{i=1}^{n} \log(A(m_i|a_{cl}, r_{cl})) \quad (4.5)$$

where m_i-s are the observed apparent magnitudes in one of the colors in our sample. Maximization of $L(a_{cl}, r_{cl})$ in respect to a_{cl} and r_{cl} yields the ML estimation of the extinction and the distance of the cloud.

Performing the ML estimation within all groups given by the k-means clustering and in all colors, separately, we obtained the results summarized in Tables 4.2 and 4.3. The results summarized in Table 4.2 enable us to

Table 4.3: Summary of the star count analysis of L1251, distance moduli; the weighted mean of the data gives 7.58 ± 0.2

Field	*Area (SqD)*	r_B	σ_{r_B}	r_V	σ_{r_V}	r_R	σ_{r_R}	r_I	σ_{r_I}
1.	0.9880	7.45	0.60	>8.50	-	-	-	-	-
2.	0.8770	7.50	0.35	7.75	0.55	<7.6	-	<7.4	-
3.	0.1515	7.75	0.30	7.65	0.45	7.65	0.45	<7.4	-
4.	0.0605	7.05	0.7	<7.4	-	<7.4	-	<7.3	-

convert the star counts in different colors into extinction. We return to this calibration in the following subsection.

One may use the calculated distance moduli in Table 4.3 to get an estimate for the distance of L1251. We computed a weighted mean of the data in the table using weights inversely proportional with σ. It resulted in a distance modulus of 7.58 ± 0.2 corresponding to 330 ± 30 pc.

Confidence interval for the parameters estimated

The ML estimation allows a straightforward way to obtain the confidence interval for the estimated parameters, the extinction and distance. Denoting the value of the parameters maximizing the likelihood function with a_{cl}^{max}, r_{cl}^{max} and with a_{cl}^{true} , r_{cl}^{true} their true values we have asymptotically if the sample size goes to infinity

$$2[L(a_{cl}^{max}, r_{cl}^{max}) - L(a_{cl}^{true}, r_{cl}^{true})] = \chi_k^2; \quad k = 2 \tag{4.6}$$

In general, k equals the number of parameters estimated and χ_k^2 is a χ-square variable with k degrees of freedom (for the proof of this theorem see Kendall & Stuart (1973)). The probability that the true values of the parameters are within a certain region in the k dimensional parameter space is given by

$$P(\chi_k^2 \leq \chi_{0k}^2) = 1 - \delta \tag{4.7}$$

The projection of this k dimensional domain which is given by the $\chi_k^2 \leq \chi_{0k}^2$ inequality yields the confidence interval of the individual parameter values estimated by the ML procedure. The χ_{0k}^2, δ pairs are tabulated and one may find them in the text books (see e.g Kendall & Stuart (1973)). The confidence intervals corresponding to the 1σ levels are given in the σ columns of the tables. In the case of *Field* 4 due to the small numbers of stars the confidence interval is not closed towards higher extinctions and lower distances.

Table 4.4: Linear least squares fitting of the $a_{cl} = a \times \log(N(m)) + b(m)$ extinction - star count relation

color	a	σ_a	$b(m)$	σ_b
B	-4.08	0.23	5.78	0.21
V	-4.08	0.16	5.67	0.15
R	-4.39	0.30	5.86	0.28
I	-4.52	0.18	5.31	0.16

4.3.3 Star count - extinction conversion

Verifying the linear $\log N(m)$ - extinction relationship

The results yielded by the ML analysis enabled us to verify the star count - extinction relationship in each color in our study. Assuming the functional form of the formula given by Eq. (4.1) one can get its constants by a linear least squares fitting of the star counts versus extinction given in Table 4.2. We summarized the results of the least squares fitting in Table 4.4.

Table 4.4 clearly demonstrates that the linear relationship fits nicely to the data obtained from the ML analysis. It means that the postulated linearity was convincingly recovered from the ML analysis performed.

In the B and V colors the slope of the relationship is the same while in R and in I, in particular, significantly differ. In the literature the inverse value of a is usually given. Using the value of a obtained for B would give a bias of about 0.44 $\log[N_0(m)/N(m)]$ *mag* in the estimation of the optical extinction, in the I color.

4.3.4 Surface distribution of the obscuring material

Based on the calibration procedure one may assign extinction to each pixel in the star count maps, in all of the four colors studied. Fig. 4.2 shows the contour maps of the extinction obtained in this way. The main body of the cloud jumps out with a contrast of several magnitudes from the less obscured region behind the bow shock.

The overall distribution of the obscuring material, obtained from our study and that of Kandori et al. (2003), has a reasonable good correlation. There are, however, remarkable differences between them, in particular in the densest part of the cloud. The probable reason for these discrepancies lies in the different ways used in obtaining extinction maps from the surface distribution of the stars in the region of L1251. Both studies had a mesh of 2' resolution but applied different kind of smoothing. We used a boxcar of 6'x 6'

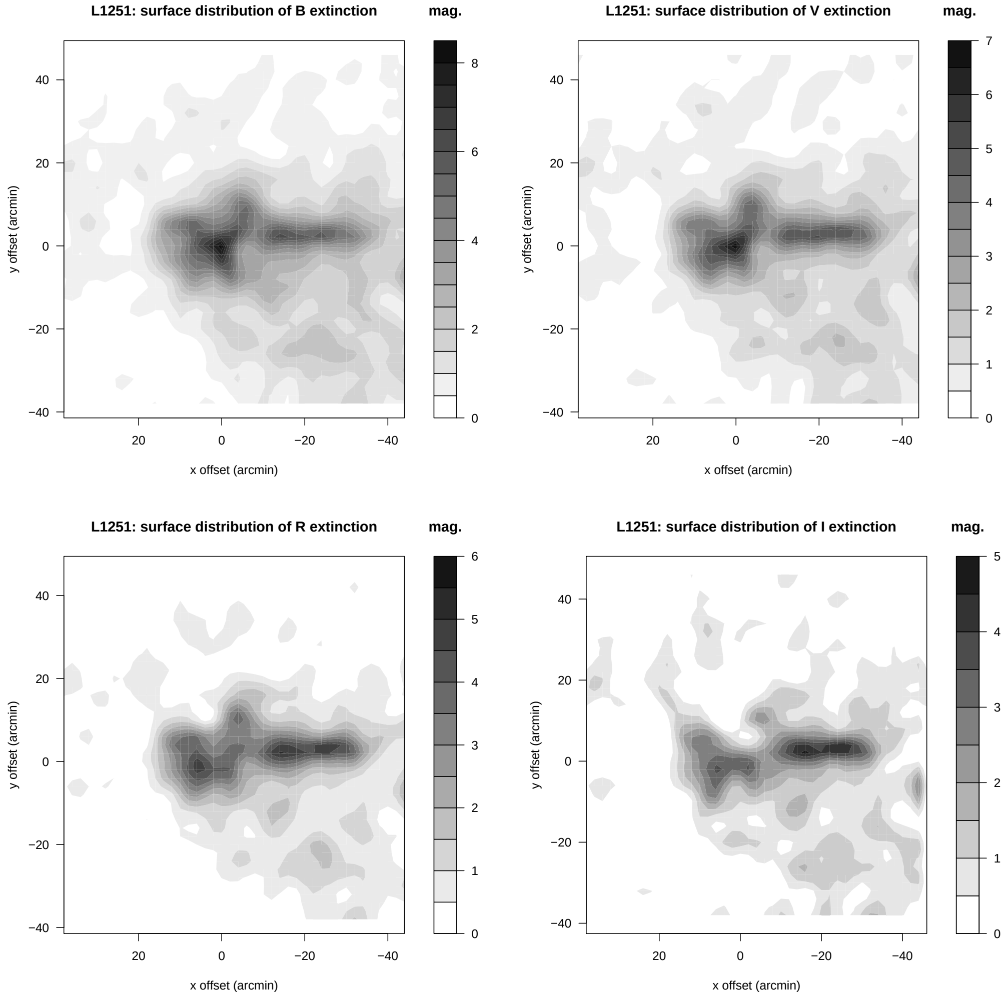

Figure 4.2: Contour maps of the extinction in the B, V, R and I color. The 'flying bullet' form of the main body and the bow shock displayed in Fig. 4.1 is clearly visible. The coordinates of the (0,0) position are same as in Fig. 4.1

size while the other study a Gaussian filtering of 6' resolution. It is probably important to note that the tails of the Gaussian filter give contributions from much wider areas, in particular in the densest part where the star counts have low values.

A further reason for the discrepancy might originate from the method of converting the star count maps into extinction. Our method accounts for the foreground stars whose contribution increases the counts and decreases correspondingly the estimated extinction values. We made simulations of the Wainscoat et al. (1992) model assuming a limiting magnitude of about 19 mag and 6 mag of extinction at the distance of L1251. The conventional way of the star count extinction conversion resulted only in a 5 mag. value, i.e. one magnitude less than the true (6 mag.) extinction.

4.4 Discussion

4.4.1 Shape of the cloud

We have already indicated above that the extinction map derived from the star counts makes the impression of a body flying with a hypersonic speed across an ambient medium.

The shape of L1251 was accounted by Tóth et al. (1995) for a shock wave passing the cloud and produced by a nearby supernova. They showed that the cooling by the H_2 molecules plays an important role in the formation of the observable shape of the cloud.

The presence of a bow shock as indicated by the inspection of the extinction maps suggests another type of the cloud-environment interactions. The observed form of the bow shock in our case bears a remarkable resemblance to that of a blunt body flying with hypersonic speed. The blunt body solutions of the hypersonic flows are standard topics of the textbooks (see e.g. Hayes & Probstein (1959). For the astrophysical context of the problem see the works of Rózyczka & Tenorio-Tagle (1985) and Canto & Raga (1998)). One can identify the tail of L1251 with a wake of the head of the cloud, typical of the blunt body hypersonic streaming patterns.

In the following we do not try to fit the form of the bow shock using some solution of the blunt-body problem. This solution would make a better understanding of the role of different significant physical parameters defining a particular fit, however this is beyond the scope of the present work. An obvious significant parameter would be the Mach number of the flow which is easy to calculate from the angle between the two asymptotes of the bow shock (Hayes & Probstein , 1959). The Mach number is given by $M = \sin(\alpha/2)^{-1}$

where α is the angle between the asymptotes. It yielded $M \approx 2$ in our case.

It is clear already at the first glance that the bow shock is not symmetric to the main axis of the cloud. The symmetry axis of the bow shock has a tilt of about 10 deg to the main axis of L1251. Probably it is conclusive to note that the symmetry axis of the bow shock points toward the center of the bubble discovered by Grenier et al. (1989). Kun et al. (2000) found a runaway star which also might have been from the SN explosion probably responsible for the bubble and derived an age of 10^6 yrs in this way. According to this picture the bow shock resulted from an encounter of the cloud with the wind coming from the interior of the bubble.

4.4.2 Properties of the obscuring material

The estimation of the total extinction in different colors makes it possible to get the value of the total to selective extinction $R_V = a_V/E_{B-V}$. The scatterplot between a_V and a_B is displayed in Fig. 4.3. The relationship between a_V and a_B can be written in the form of $a_B = (1 + 1/R_V) \times a_V$. We marked with lines (labelled with the corresponding R_V values) in Fig. 4.3 the relationships between a_V and a_B assuming that $R_V = 3$ (close to the canonical value for the general interstellar matter) and $R_V = 6$. The line assuming that the interstellar absorption does not depend on the color (grey approximation) is also marked with $R_V = inf$.

Figure 4.3 clearly demonstrates that above $a_V = 3$ (i.e. within the main central body of L1251) the points depart from the $R_V = 3$ line and approach the lines of higher R_V as one moves to higher extinction. Departure from the canonical $R_V = 3.1$ towards higher values is a common behavior of dense clouds (see the review paper of Mathis (1990)).

Postulating a relationship in the form of $a_B = \alpha \times a_V$, where $\alpha = (1 + 1/R_V)$, and substituting R_V values between 3-6, we can infer that α depends only weakly on a_V and may have approximately a form of $\alpha(a_V) = \alpha_0 + \alpha_1 \times a_V$. Assuming that $R_V = 1 + 1/\alpha_0 = 3.1$ we made a least squares fitting to get α_1. In this way we obtained $R_V(a_V) = 1/(0.32 - 0.019 \times a_V)$ and $a_B = \alpha(a_V) \times a_V$ gave a reasonable good fit to the points in Fig. 4.3.

Based on this result we may assign a value of R_V to every pixel in the a_V extinction map of L1251. By assigning R_V to every pixel we obtained a map of the total to selective extinction as given in Fig. 4.3. One may infer from this map that the high values of R_V are concentrated only in the densest parts of L1251.

The canonical dependence of the interstellar extinction on the wavelength is realized by the standard relation of Savage & Mathis (1979) which can be well represented by a linear function of the extinction on $1/\lambda$ in the range of

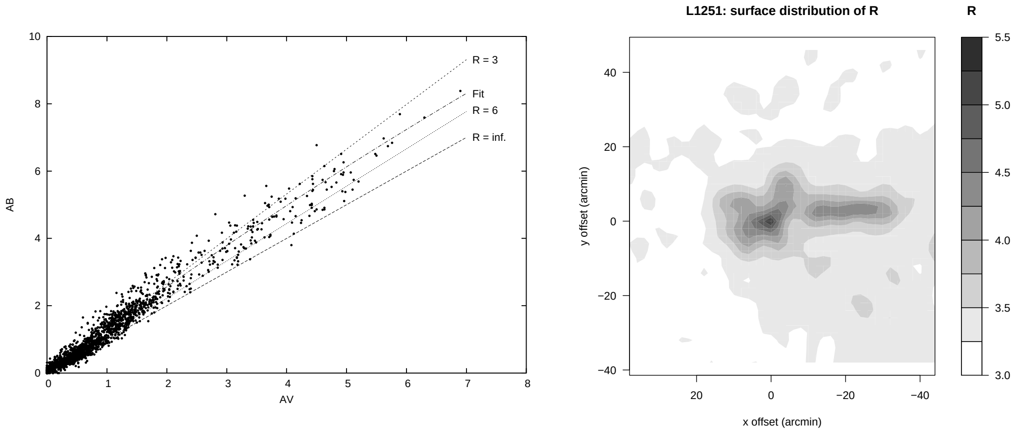

Figure 4.3: Scatterplot between a_V and a_B (left panel). The relationship between a_V and a_B can be written in the form of $a_B = (1 + 1/R) \times a_V$. We marked with lines (labelled with the corresponding R values) the relationships between a_V and a_B assuming that $R = 3$ (close to the canonical value for the general interstellar matter) and $R = 6$. The line assuming that the interstellar extinction does not depend on the color (grey approximation) is also marked with $R = inf$. Assigning R_V to every pixel we obtained a map of the total to selective extinction (right panel). One may infer from this map that the high values of R_V are concentrated only in the densest parts of L1251. The coordinates of the (0,0) position are same as in Fig. 4.1

the *B, V, R, I* colors used in our study.

Departure from the standard $R_V = 3.1$ value above $a_V = 3$ in dense interstellar clouds, as we found in the case of L1251, was also obtained by Whittet et al. (2001) in the Taurus region. They claimed that for extinctions $a_V > 3$ real changes in grain properties may occur, characterized by observed R_V values in the range of 3.5-4.0. A simple model for the development of R_V with a_V suggested that R_V may approach values of 4.5 or more in the densest regions of the cloud. According to Whittet et al. (2001) the transition between "normal" and "dense cloud" extinction occurs at $a_V \approx 3.2$, a value coincident with the threshold extinction above which H_2O-ice is detected on grains within the cloud.

The R_V values derived in our analysis correspond to those obtained by Kandori et al. (2003) within the limits of the statistical errors. There are differences, however, in the surface maps between the two studies. For obtaining R_V values we used the relationship between the a_B and a_V values so our surface map reflects basically that of the extinction in the V color. In the contrary, Kandori et al. (2003) applied an adaptive averaging of $B - V$ and $V - I$ color indices by varying the size of the smoothing window, keeping constant the number of stars in it but on the cost of the spatial resolution, in particular in the densest part of the cloud.

4.4.3 Mass of the cloud

Knowing the distance of L1251 we converted the a_V extinction values into the mass using the empirical formula given by Dickman (1978). According to this formula

$$M = (\alpha d)^2 \mu \frac{N_H}{a_V} \sum_i a_V(i) \tag{4.8}$$

where $M, \alpha, d, \mu, a_V(i)$ are, the mass of the cloud, the angular size of a pixel, the distance, the mean molecular mass and the extinction of a pixel, respectively; $N_H/a_V = 1.87 \times 10^{21} cm^{-2} mag^{-1}$ and $N_H = N_{HI} + 2N_{H_2}$. Based on the Dickman's formula we computed the mass of the different subregions of the cloud as given in Table 4.5.

Fields 3, 4 and 5 represent the main body of L1251. Based on Table 4.5 the total mass of this part of the cloud amounts to 371 $M_\odot$. This value can be compared with the figure obtained by Sato et al. (1994) based on $C^{18}O$ measurements, taking into account that the field occupied by our main body lies completely inside the region covered by the $C^{18}O$ study and contains only 85 % of the mass calculated from it. Re-scaling this fraction of mass onto the distance we obtained in this work one gets $422 M_\odot$, surprisingly close to

Table 4.5: Mass of the cloud (using the formula of Dickman (1978))

$Field$	$Area\,(SqD)$	$Mass\,(M_\odot)$	$No.\,of\,pix.$
1.	0.9880	153.34	887
2.	0.8770	490.48	784
3.	0.1515	212.60	136
4.	0.0601	132.05	54
5.	0.0089	26.07	8

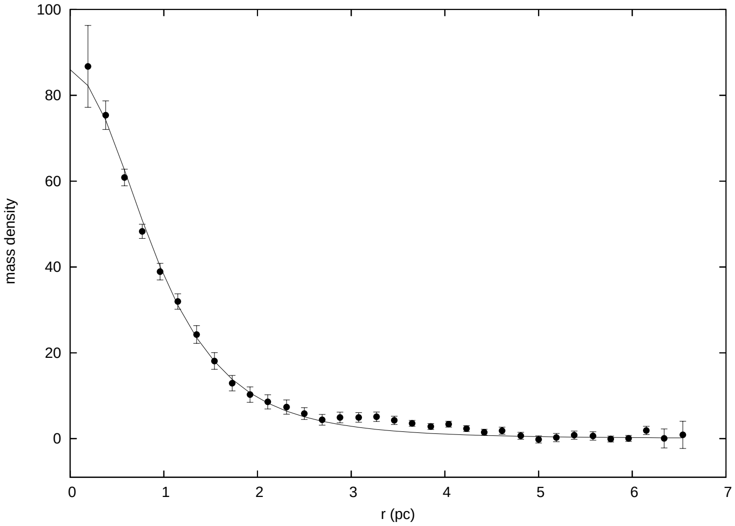

Figure 4.4: Fitting the radial distribution of the surface mass density by a Schuster sphere. Up to 2.5 pc distance from the center the fit is very good. Beyond this distance, however, there is an excess of the observed mass density due to the tail of the cloud.

our estimated value. Studying a somewhat larger area than ours Lee (1994) obtained 610 $M_\odot$ from his ^{12}CO and ^{13}CO measurements.

4.4.4 Mass model of the head of L1251

We mentioned already in Sect. 4.4.1 that the shape of the cloud can be represented as a hypersonic flow around a blunt body (the head of the cloud) with a wake forming the tail. In the following we try to model the head assuming spherical symmetry.

Recently, using the data obtained with VLT of the ESO Alves et al. (2001) modelled the spatial structure of the Bok globule Barnard 68. They supposed the globule to be an isothermal Bonnor-Ebert sphere (Bonnor , 1956; Ebert , 1955) in pressure equilibrium with the external much hotter and less dense interstellar medium.

We have not required isothermality in our analysis. The isothermal solution is a specific case of the class of polytropic spheres studied extensively by Emden (1907). The isothermal solution represents the case when the polytropic index $n = \infty$. Actually, the mass of an isothermal sphere is infinite and the limit between the finite and infinite mass solutions is $n = 5$. Although the mass is finite in this case the size of the sphere is infinite. One obtains finite size solutions only in the $n < 5$ case. Recently, Medvedev & Rybicki (2001) studied the properties of polytropic spheres near $n = 5$ and found them very suitable to characterize the structure of molecular cloud cores.

We fixed the mass center of the sphere at the maximum value of a_V. This choice is quite reasonable in view of Equation (4.8). After fixing the center of the sphere we averaged the a_V values on concentric annuli. This procedure gave a radial profile of the extinction of the head of L1251. This profile can be converted into the surface mass density by Equation (4.8).

Evaluating the radial profile obtained from the observed data we projected the $n = 5$ polytropic sphere (Schuster, 1883) for getting its surface mass density distribution. The Schuster sphere has two free parameters to be adjusted: the central mass density and a scale parameter. After adjusting these parameters we get the fit displayed in Fig. 4.4. Up to 2.5pc distance from the center the fit is very good. Beyond this distance, however, there is an excess of the observed mass density due to the tail of the cloud which significantly distorts the spherical symmetry of the head.

The density parameter of the fitted Schuster sphere, amounting $45M_{\odot}pc^{-3}$ ($3.06 \times 10^{-21}gcm^{-3}$), gives an estimate for the central mass density of the head of L1251. Comparing the finite size polytropic solutions of $n < 5$ with those of the Schuster sphere (i.e. $n = 5$) one can infer that they concentrate more mass at finite distances and give a much worse fit to the points observed.

Assuming a polytropic gas sphere one can compute the radial velocity dispersion profile from the fitted density profile, based on the polytropic equation of state. The p/ρ ratio of the pressure and mass density gives the T temperature, i.e. the velocity dispersion. The velocity dispersion projected onto the celestial sphere can be directly compared with that observed. The projected $FWHM$ of the V-profile of a Schuster sphere is displayed in Fig. 4.5, along with those of the NH_3 molecule measured by Tóth & Walmsley (1996).

There is a considerable scatter of the measured line widths in the head of the cloud around the mean which is well matched by the Schuster solution. The Schuster curve, however gives an unexpectedly good fit in the tail region.

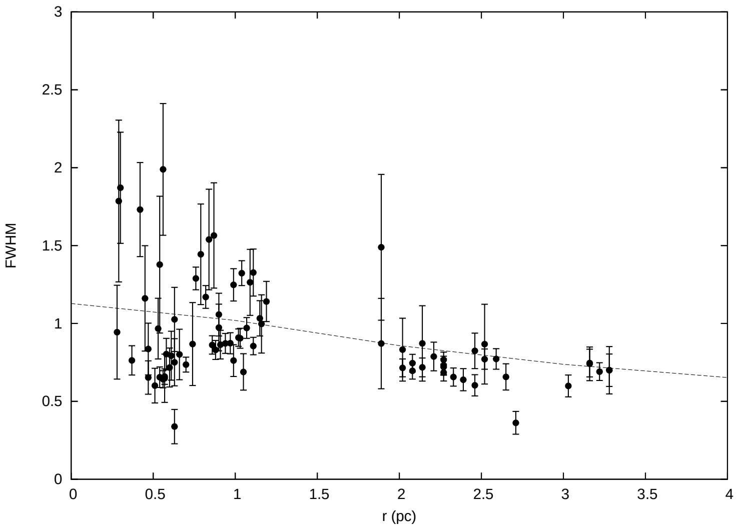

Figure 4.5: $FWHM$ of the projected velocity profile of a Schuster sphere (dashed line) and the NH_3 molecule measured by Tóth & Walmsley (1996). There is a considerable scatter of the line widths in the head of the cloud around the mean which is well matched by the Schuster solution. The Schuster curve gives an unexpectedly good fit in the tail region.

This good fit is not expected far from the head since due to the tail the mass distribution drastically differs from the spherical symmetry. However, we may assume that our predicted $FWHM$s of the NH_3 line refers to the initial and those measured by Tóth & Walmsley (1996) to the final stage of the formation of density enhancements in the tail region. We conclude therefore that the formation of structures in the tail left the linewidth practically unchanged. This indicates an isothermal contraction. However, the question remains open: what kind of instability played a significant role in forming the density enhancements in the tail. We note that according to Tóth & Walmsley (1996) the thermal energy is dominating over the turbulent energy in the NH_3 cores of the tail region and thus we can not exclude the scenario that thermal instability played a significant role in the cloud fragmentation.

Part III

Statistics of point sources

Chapter 5

Classification of stellar spectra

5.1 Formulation of the problem

5.1.1 The quantitative measure of similarity

In order to make some quantitative comparison one has to introduce some measure of similarity between the different spectra. This quantitative value is the greatest (usually taken as unity) when comparing the spectra with themselves and it is less otherwise. In the case of two spectra S_1, S_2 the quantitative measure of similarity δ fulfil the following inequalities:

$$\delta(S_1, S_2) \leq \delta(S_1, S_1) = \delta(S_2, S_2) = 1 \quad . \tag{5.1}$$

It is easy to make a transformation mapping the similarity into the ρ distance:

$$\rho = \delta(S_1, S_1) - 2\delta(S_1, S_2) + \delta(S_2, S_2) \tag{5.2}$$

If we have three spectra S_1, S_2 and S_3 then the ρ distance should fulfil the usual triangle criteria:

$$\begin{aligned} \rho(S_1, S_2) &= \rho(S_2, S_1) \\ \rho(S_1, S_1) &= 0 \\ \rho(S_1, S_3) &\leq \rho(S_1, S_2) + \rho(S_2, S_3) \end{aligned} \tag{5.3}$$

An obvious definition for ρ is the quadratic Euclidean distance, i.e.

$$\rho(S_1, S_2) = \int (S_1 - S_2)^2 d\lambda. \tag{5.4}$$

This distance definition requires identical spectra for $\rho = 0$. For the purpose of classification, however, one should consider two spectra identical if they differ only in a multiplicative and/or an additive constant. For quantitative classification, therefore the spectra have to be normalized:

$$Z = \frac{S - \overline{S}}{\sqrt{\int (S - \overline{S})^2 d\lambda}} \tag{5.5}$$

Where $\overline{S} = \int S d\lambda$. (This normalization differs from those normally used in spectroscopy where the intensity at a certain wavelength, e.g. at 5550 $\mathring{A}$, is taken equal with unity). With this normalization

$$\overline{Z} = \int Z d\lambda = 0 \quad and \quad \int Z^2 d\lambda = 1. \tag{5.6}$$

Let us suppose that we have two spectra $S_1(\lambda)$ and $S_2(\lambda)$ then after normalization

$$\rho(Z_1, Z_2) = \int (Z_1 - Z_2)^2 d\lambda = 2(1 - R_{12}). \tag{5.7}$$

Where $R_{12} = \int Z_1 Z_2 d\lambda$ is called the correlation coefficient between Z_1 and Z_2. The correlation coefficient varies between 1 and -1. It reaches 1 when $Z_1 = Z_2$ and -1 when $Z_1 = -Z_2$.

If we have a set of spectra $S_1, ..., S_n$ the pair-wise correlations between the elements of the set define a matrix R_{ij} and correspondingly a matrix of distances. Analyzing the structure of the distance matrix, i.e. looking for possible groupings according to the distances, is the task of different clustering algorithms. The algorithm results in the characteristic spectra within the groups and can be used as templates for further classification. One can find more details on this topic in textbooks dealing with different clustering algorithms (see e.g. Murtagh and Heck, 1987).

5.1.2 Factorization of the spectra, number of significant physical quantities

The Euclidean distance we introduced gives equal weights for all wavelengths involved, regardless of the role the different physical quantities play in the actual form of the spectra to be classified. (Nevertheless, In practical cases the spectral features attributed to certain physical quantities are restricted to particular wavelengths and their influence might be hidden in the noise when computing pure Euclidean distances. In these cases one have to use weighting functions in order to enhance the effect of the important quantities on the distances. If one does not have apriori knowledge on the spectra the use of weighting functions could be the second approach after using pure Euclidean distances.)

Normally, the output of the recent spectrographs is an array of intensities where the number of elements of the array is much higher than the number of significant physical quantities behind the observed shape of the spectra. We may represent each spectrum by a point in a multidimensional parameter space where the dimension of the space equals to the number of the elements of the observed array and, correspondingly, the intensities give the coordinates of the point. The points representing the spectra will populate a subspace of a dimension corresponding to the number of physical quantities having influence on the shape of the spectrum.

If we have a set of arrays representing the observed spectra $\{\mathbf{S}_1, ..., \mathbf{S}_n\}$ they always can be expressed as a linear combination of orthogonal and normalized arrays.

In order to study the number and nature of these significant physical quantities (i.e. effective temperature, chem. abundances) we develop each array, representing the spectra, into an ortogonal series; i.e. we assume that

$$\mathbf{S}_j = a_{1j}\mathbf{e}_1 + ... + a_{nj}\mathbf{e}_n \quad ; \quad (\mathbf{e}_k, \mathbf{e}_l) = \begin{cases} 0, & \text{if} \quad k \neq l \ , \\ 1, & \text{if} \quad k = l \ . \end{cases} \tag{5.8}$$

Where $\mathbf{e}_1, ..., \mathbf{e}_n$ are orthogonal normalized arrays. Any of the n arrays can always be represented by means of the linear combination of n orthogonal arrays but in the case of a small number of significant physical quantities behind the spectra we may suppose that the number of orthogonal arrays k, necessary to reproduce the basic properties of the observed ones, is smaller than n. Namely, as we mentioned, due to the physical quantities playing a significant role in the shape of the spectra the points representing each spectrum in the parameter space populate a subspace of lower dimension than the original parameter space representing the observed arrays. This subspace, consequently, is stretched by smaller number of linearly independent arrays than the whole parameter space.

It can be shown that the distance $\rho(\mathbf{S}_i, \mathbf{S}_i^{(m)})$ between the $i-th$ member of the observed n arrays, $\mathbf{S}_i$, and a linear combination of a set of normalized orthogonal arrays $\{\mathbf{e}_j, \quad j = 1, ..., m < n\}$, $\mathbf{S}_i^{(m)} = a_{i1}\mathbf{e}_1 + ... + a_{im}\mathbf{e}_m$ reaches a minimum at a fixed $m < n$, with a suitably selected set of the vectors $\{\mathbf{e}_j\}$, if the $a_{i1}, ..., a_{im}$ coefficients are given by the elements of the first $m < n$ eigenvectors in the solution of the

$$\mathbf{R}\phi_k = \lambda_k \phi_k \tag{5.9}$$

eigenvalue equation. $\mathbf{R}$ is an $n \times n$ matrix consisting of the pair-wise correlations between the arrays representing the n observed spectra; ϕ_k means the $k-th$ eigenvector and λ_k is the corresponding eigenvalue.

If $m = n$ than $\mathbf{S}_i = \mathbf{S}_i^{(m)}$ and the series $\mathbf{S}_i^{(m)}$ is called the principal component representation of $\mathbf{S}_i$. The procedure attempting to reproduce the basic properties of $\mathbf{S}_i$ by $m < n$ components is called factor analysis (Sec. 1.1.1). The factor analysis based on principal components is the default of many statistical software packages (e.g. SPSS). One can find further details on factor analysis in the textbooks (see e.g. Murtagh and Heck, 1987).

Performing factor analysis we succeeded in representing the original n observed arrays by the linear combination of arrays of a less number. The a_{ij} coefficients depend on the physical quantities (e.g. effective temperature, luminosity, chemical abundance) responsible for the basic characteristics of the observed spectra. If the physical quantities have only a small change across the given set of observed spectra one may assume that their relation to the a_{ij} coefficients is linear and the number of significant principal components, the factors, is the same as those of the physical quantities. In that way, therefore, factor analysis helps to derive the number of significant physical quantities in the observed spectra.

5.2 Classification of Hα objects in IC1396

$H\alpha$ emission objects are important optical tracers of places of active star formation. As observations testify they are usually overabundant in regions (molecular clouds, stellar associations), known as sites of significant star formation activity. This relationship is well established by many surveys on wide field telescopes equipped with objective prisms of low spectral resolution. In a number of cases, however, due to the low spectral resolution, this technique yields only objects with some suspect of $H\alpha$ emission and useful for further follow up investigations.

IC1396 is an extended HII region hosting the younger part of the Cepheus OB2 association and the open cluster Tr 37. The HII region is excited by the luminous OB stars of the association and in particular by HD 206267, an O6Vf type star in a Trapezium-like system. The spectroscopic study of luminous stars (Simonson, 1968) indicated that this younger part of Cepheus OB2 has an age of about 3 million years. There are some evidence that the interstellar matter of IC1396 was processed by a nearby supernova explosion (Clayton and Fitzpatrick, 1987). The supernova might have been a member of the older subgroup of Cep OB2 which is surrounded by a giant infrared ring (Kun et al., 1987) connecting several well known star forming regions

(Sh2-140, NGC 7129), among them IC1396 as well (see also Sect. 3.2). The supernova explosion could have a significant influence on the star forming activity in all of these areas. A detailed study of the IRAS point sources along the giant ring (Balázs and Kun, 1989) revealed that the brightest sources are located at IC1396. There are several other indications that the formation of massive stars is still going on in this region (Baars and Wendker, 1976; Sugitani et al., 1990). All of these facts make IC1396 an important laboratory of the different physical phenomena playing significant role in the formation of high mass stars.

The lower mass stellar population in IC1396 is not so well studied as in other sites of massive star formation like NGC2244, NGC2264, or NGC6530. Marschall and van Altena (1987) estimated kinematic membership for Tr37 and Marschall et al. (1990) measured photoelectric UBV colors for 120 probable members. They concluded that Tr37 contains several pre-main sequence stars. Dolidze and Vyazovov (1959) published 125 $H\alpha$ emission stars in the field of IC1396. Wackerling (1970) and Dolidze (1975) listed 45 objects while Kun (1986) and Kun and Pásztor (1990) did so far the most comprehensive study in this field identifying 220 $H\alpha$ emission stars. There are only two spectroscopically identified pre-main sequence objects, LkHα349 and LkHα349$/c$ in the Herbig- Bell Catalogue (1988).

Although the distribution of the interstellar matter is very inhomogeneous the average foreground absorption is about 1.5 mag in the direction of IC1396. The emission line objects in the typical magnitude interval of 13-16 mag of these surveys correspond, therefore, to F-G type stars of 1-2 solar masses, at the 800-1000 pc distance of the region.

Comparing these different surveys one gets only a very limited correspondence between the objects identified by different authors. The reason for that is partly lying in the temporal variability of the emission nature of these stars and partly in the uncertainty of recognizing the appropriate spectral features at low resolution near the detection limit.

The present work was a first attempt to get medium resolution Cassegrain spectra on a limited sample of objects listed as $H\alpha$ stars in IC1396 and try to study their main spectral characteristics.

5.3 Observational data

We selected 35 stars of 11.2-15.2 mag from the list of Kun & Pásztor (1990). The spectra were obtained with an UAGS Cassegrain spectrograph equipped with an YMK91B image intensifier attached to the 2.6 m telescope of the Byurakan Observatory and were recorded on Kodak 103aO films. We utilized a diffraction grating of 651 *lines/mm* giving a dispersion of $100\AA/mm$. We scanned the spectra with the PDS 1010 microdensitometer of the Byurakan Observatory using a $50 \times 50\,\mu m$ slit and $25\,\mu m$ step length. We scanned the background and the calibration spectra (He-Ne-Ar) on both side of the stellar spectra.

The spectra were analyzed with the help of the AIDA image processing system developed in Byurakan. The wavelength scales were linearized using the built in procedure of the AIDA system and all of the spectra were normalized to $5556\AA$.

Together with the program stars we observed comparison stars of known energy distribution on each night in order to control the changes of the wavelength dependency of the sensitivity of the system. In most cases we used BD $+28^o$ 4211 and in some cases BD $+40^o$ 4032. The sensitivity curve of the system was slightly varying from night to night but it was nearly constant between $4500 - 6400\AA$.

5.4 Statistical analysis of the spectra

Due to the moderate resolution it is not possible to make detailed quantitative analysis of the spectra; one may only classify the individual stars into different categories, based on some characteristic spectral features. In the case of supervised classification one compare the given spectra with a set of apriori given templates and assigns to the objects those templates which reveal the highest degree of similarity. On the contrary, the unsupervised classification does not require a set of templates but it is a byproduct of the appropriate algorithm.

There are several procedures to perform unsupervised classification. The procedures require a definition of some measure of similarity (or distance) between the individual objects which are spectra in our case. Before running any of these algorithms we extract some basic properties of the observed spectra by applying factor analysis.

In accordance with Sec. 5.1.2 to perform factor analysis we assume that the observed spectra are arrays where the wavelength runs in integer steps along the elements of the array. Let us have a set of observed spectra in the form of arrays $\{\mathbf{S}_1, ..., \mathbf{S}_n\}$. It can always be represented as a sum of n orthogonal arrays, i.e.:

$$\mathbf{S}_i = \sum_{i=1}^{n} a_{ij}\mathbf{e}_j \tag{5.10}$$

where the $\mathbf{e}_j$ vectors are pair-wise orthogonal and called principal com-

Table 5.1: Resulted eigenvalues of the principal components analysis

No.	Eigen val.	Pct.(%)	Cum.pct.(%)
1	27.991	80.0	80.0
2	2.251	6.4	86.4
3	.727	2.1	88.5
4	.647	1.8	90.3
5	.597	1.7	92.0
6	.528	1.5	93.5
7	.446	1.3	94.8
8	.279	.8	95.6
9	.217	.6	96.2
10	.165	.5	96.7
⋮	⋮	⋮	⋮
35	.007	.0	100.0

ponents if the a_{ij} coefficients consist of the elements of the eigenvectors of $\mathbf{R}$, the matrix of the pair-wise correlations between the $\mathbf{S}_i$ arrays. Factor analysis attempts to represent the original arrays by the sum of $m < n$ components. If the number of significant physical quantities responsible for the main characteristics of the spectra is smaller than n we may suppose that it is also reflected in a smaller number of significant components in the sum given above.

5.4.1 Results of the factor analysis

Using principal components analysis on the observed spectra one can get two large eigenvalues of $\mathbf{R}$ describing 86.4 % of the variances of the observed arrays (see Tab. 5.1). This figure can be obtained by computing η_m according to the following expression

$$\eta_m = 1 - \frac{\sum\limits_{j=1}^{m} \lambda_j}{\sum\limits_{j=1}^{n} \lambda_j} \tag{5.11}$$

Where $\lambda_1 \geq, ..., \geq \lambda_n$ are the eigenvalues of the correlations matrix and $m < n$. η_m is the relative error in estimating the measured arrays by means of the linear combination of m eigenvectors of $\mathbf{R}$. In our case $\eta_m = 0.864$.

Besides the a_{ij} coefficients the factor analysis algorithm yielded also the factor values, the elements of the $\mathbf{e}_j$ arrays. Since the linear combination of

$\mathbf{e}_j$-s gives the measured spectra the arrays themselves have some similarity to real spectra. However, they do not necessarily represent real spectra. They represent real spectra only in that case if at some $\mathbf{S_i}$ only one of the a_{ij} coefficients differs significantly from zero. We displayed in Fig. 5.1 the $\mathbf{e}_1$ and $\mathbf{e}_2$ factor arrays belonging to the first two eigenvectors of the $\mathbf{R}$ correlation matrix. The factor arrays are normalized to unit standard deviation and zero mean value and this explains the range and scale on the vertical axis in the Figure.

After normalizing to unit variance the original observed arrays the equality $\sum_{i=1}^{n} a_{ij}^2 = 1$ is always holding. Keeping only the significant members of this sum we get that fraction of the original variances which is explained by the significant principal components, the factors. This fraction is called communality. In our case we have two factors. Since the communalities are always less than unity $a_{i1}^2 + a_{i2}^2 < 1, (i = 1, ..., n)$, the inequality holds for all the points representing real spectra in the $\{a_1; a_2\}$ plane. In Fig. 5.2 we displayed the distribution of a_1, a_2 values obtained from the factor analysis. Fig. 5.2 clearly shows that the vast majority of the points approaches closely the unit circle confirming the validity of the two-factor representation of the original spectra. Some points, however, are well inside this circle. Inspection them individually reveals that they are underexposed noisy spectra.

The fact that the majority of the points in the $\{a_1; a_2\}$ plane approaches the unit circle indicates that there is only one significant physical variable responsible for the main properties of the observed spectra.

5.4.2 Clustering spectra according to the factor coefficients

First inspection of the stars in the $\{a_1; a_2\}$ plane shows three major groups. The biggest one, however, may be splitted into two smaller ones. We ordered our stars, therefore, into four groups. In order to find the members of the groups we carried out k-means clustering (see e.g. Murtagh and Heck, 1987). After finding the group members we computed the average spectra within the groups.

Since similar spectra mean similar factor coefficients the average spectra within the groups may serve as templates for further classification. Fig. 5.3a-d show the template spectra obtained in this way.

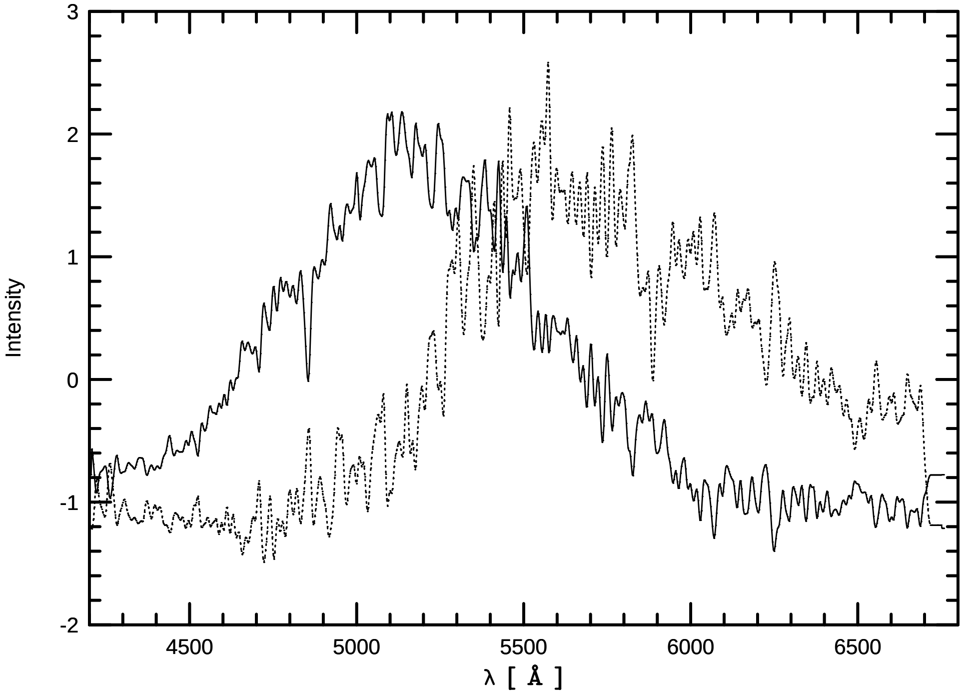

Figure 5.1: The first (solid line) and second (dotted line) most significant factors resulted from the principal components analysis. The two factors are orthogonal i.e. $\int e_1(\lambda)e_2(\lambda)d\lambda = 0$. The observed spectra can be represented as linear combinations of the factors: $S(\lambda) = a_1e_1(\lambda) + a_2e_2(\lambda)$. The factors are normalized to unit standard deviation and zero mean value and this explains the range and scale on the vertical axis in the Figure.

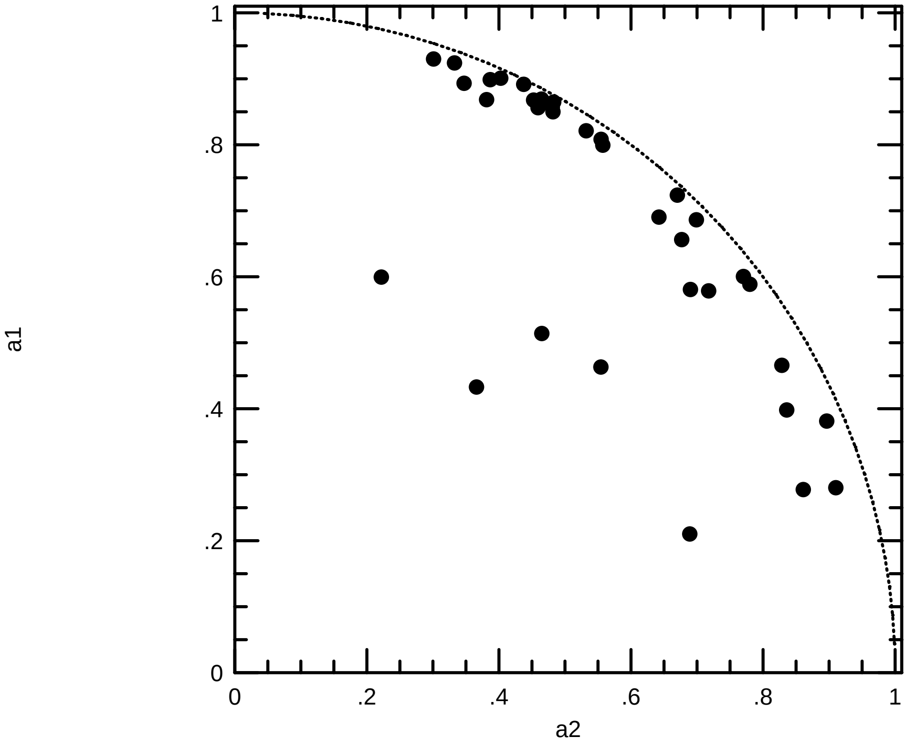

Figure 5.2: Distribution of the factor coefficients on the $\{a_1; a_2\}$ plane. If the spectra are well represented by the first two factors $a_1^2 + a_2^2$ is close to the unity. The Figure clearly shows that the vast majority of the points approaches closely the unit circle confirming the validity of the two-factor representation of the original spectra. Some points, however, are well inside this circle. Inspection them individually reveals that they are underexposed noisy spectra.

5.4.3 Classification of the mean spectra

To identify different features on our mean spectra obtained in the previous paragraph we used templates given in the Bonner Spectral Atlas (Seitter, 1975). In order to get accurate spectral types for the template spectra we compared them with the spectra listed in the spectral library of Jacoby and Hunter (1984). Spectra in the library having similar strength of the characteristic lines to those of our program stars have intensity maxima shifted systematically towards shorter wavelengths. The shift of the intensity maxima of the template spectra in comparison with the library spectra of similar line strength can be accounted for the effect of interstellar reddening. Multiplying the library spectra with the Whitford's reddening law (Whitford , 1958) and varying the value of the visual absorption A_V gives a much better fit for our template spectra.

Spectrum of group 1: A conspicuous $H\beta$ line is visible in absorption. Nevertheless, one cannot recognize $H\alpha$ because it would be near to the red edge of the spectrum which is somewhat underexposed. Otherwise the spectrum is dominated by neutral metallic lines (FeI, MgI,). There is some suspect for the $NaI/5890-96\AA$ doublet. The best matching library spectrum is *A8*.

Spectrum of group 2: The $H\beta$ line is nearly as conspicuous as at *group 1.*. In the contrary, the FeI and MgI lines in the $5160-90\AA$ range are definitely stronger than in the previous case. The $NaI/5890-96\AA$ doublet is clearly visible. Following the same procedure as before we obtained *F6* spectral type.

Spectrum of group 3: The strength of the neutral metallic lines increases in comparison with the previous groups. There is some hint for the $H\beta$ line in absorption. The $NaI/5890-96\AA$ doublet is very pronounced. The resulted spectral type is *G9*.

Spectrum of group 4: The appearance of $H\beta$ is similar to *group 3* but the characteristic metallic lines are more enhanced. The shift of the intensity maximum towards the red corresponds to the enhancement of the neutral metallic lines so it might be accounted for the differences of effective temperatures between the template spectra of *groups 3 and 4*. Applying the same procedure as in the case of previous groups we got *K8* for this template.

Fig. 5.3 shows the mean spectra obtained by averaging within the groups obtained.

The spectral type and effective temperature of the template spectra and the mean values of factor coefficients within the four groups define a relationship which enables us to assign spectral types and effective temperature to all of the program stars, after obtaining the coefficients from factor analysis. Based on the spectral types obtained and the V and BV data of Kun and Pásztor (1990) we calculated the interstellar reddening for our program stars.

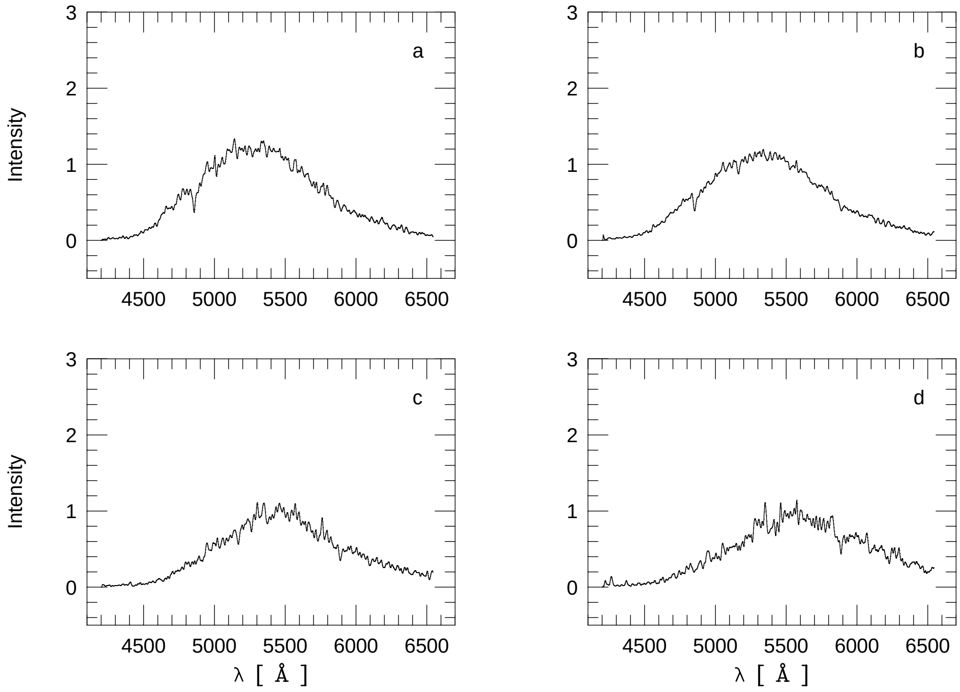

Figure 5.3: a-d: Spectra obtained by averaging within the groups defined by clustering the stars on the $\{a_1; a_2\}$ plane. The physical parameters of the spectra are summarized in Table 5.2.

Table 5.2: Physical parameters of the groups obtained

Group	$T_{eff}(K^o)$	*Sp*	$B-V$	E_{B-V}	a_1	a_2
1	7500	A8	0.70	.47	.90	.36
2	6000	F6	1.03	.53	.85	.49
3	5300	G9	1.78	.98	.64	.71
4	3920	K8	2.04	.90	.36	.87

Averaging the interstellar absorption within the spectral groups defined we got the mean reddening for the template spectra. Table 5.2 summarizes the spectral type, effective temperature and interstellar absorption of the template spectra obtained and Table 5.3 gives these data for the program stars, along with α, δ *and* V listed in the paper of Kun and Pásztor (1990).

Inspecting the data of Table 5.2 reveals that the reddening is lower in *groups 1* and *2* and higher in *groups 3* and *4*. This means that the basic physical parameter responsible for the differences between the template spectra is a combined effect of the reddening and the effective temperature. It is worth mentioning that the reddening obtained from fitting the templates with the library spectra is systematically lower by about 0.2 mag than those obtained from the V and BV photometric data.

Although all of our program stars were picked up in previous surveys by some suspected $H\alpha$ emission on small scale spectra the template spectra show no definitive evidence for $H\alpha$ emission. However, it is quite usual at repeated small scale spectral surveys on the same field that stars which were detected at the first coverage do not show any remarkable evidence of emission on the next occasion. Repeated small scale spectral observations revealed that the correlation length in time of the suspected $H\alpha$ emission is in the order of few days (Balázs et al., 1987). Among our 35 program stars only Kun 193 shows very strong Hydrogen emission (Fig. 5.4).

5.4.4 Location of the program stars in the HR diagram

In order to put our program stars onto the HR diagram we corrected their V colors, given by Kun and Pásztor (1990), for the interstellar extinction obtained in the previous paragraph. Marschall and van Altena (1987) listed proper motions for 10 of our stars and 9 seemed to be members of Tr37. We assume therefore that the vast majority of our objects belong to IC1396 and, consequently, have a distance modulus of 9.5 mag (Simonson, 1968). The bolometric correction was taken from the table published by Johnson (1966).

Plotting our stars on the $\{\log_{10}(L); \log_{10}(T_{eff})\}$ plane (Fig. 5.5), along

Table 5.3: List of the program stars. A semicolon after the spectral type marks uncertain classification.

No.	α_{1950}	δ_{1950}	V	BV	E_{BV}	T_e	*Sp.*	a_1	a_2
8	21 27 1.8	56 15 31	14.02	2.21	.88	3940	K8	.398	.836
15	21 28 44	56 45 35	14.80	2.41	1.04	3860	M3	.280	.910
17	21 28 56	56 45 27	14.60	1.48	1.04	6380	F4	.865	.465
25	21 30 11	56 24 40	14.49	2.11	.69	3740	M0	.278	.861
41	21 33 2.4	56 19 29	13.18	1.07	.52	6000	F8	.821	.532
66	21 35 1.5	57 41 45	14.51	2.19	1.85	6820	F1:	.433	.366
68	21 35 17	57 22 01	13.33	.44	.18	7220	F0	.868	.381
315	21 35 19	57 17 41	12.68	1.56	.69	5100	K1:	.210	.689
73	21 36 26	57 25 18	14.68	1.39	.11	4080	K7	.466	.828
189	21 36 39	57 07 41	13.06	1.76	1.07	5580	G6	.724	.670
190	21 36 43	57 08 56	13.23	1.23	.72	6140	F6:	.514	.465
191	21 36 52	57 09 38	14.30	.77	.34	8980	A2:	.599	.222
83	21 37 46	57 21 17	14.24	1.18	.76	6480	F3	.868	.453
84	21 37 48	57 10 25	13.84	1.06	.58	6260	F5	.850	.482
85	21 37 51	57 10 16	14.45	1.34	.58	5400	G8	.686	.699
193	21 37 56	56 45 33	14.29	1.37	.65	5500	G7	.656	.677
87	21 37 56	57 20 44	14.07	.88	.61	7200	F0	.899	.387
89	21 38 13	57 12 33	13.80	.59	.12	6260	F5	.865	.483
198	21 38 13	57 13 10	14.54	.62	.42	7980	A6	.924	.333
91	21 38 34	57 04 37	14.55	1.88	.84	4680	K3	.589	.780
330	21 38 39	57 28 08	11.21	.27	-.20	6260	F5	.857	.481
197	21 38 42	57 23 37	12.91	.40	.02	6640	F2	.892	.437
196	21 38 42	57 27 12	13.20	.65	.44	7720	A7	.893	.347
93	21 38 47	57 17 44	13.48	.45	.15	7000	F1	.901	.403
94	21 38 54	57 18 06	15.03	2.06	.63	3720	M1	.381	.896
204	21 39 22	58 05 23	13.34	2.19	1.53	5660	G5	.690	.642
97	21 39 26	57 16 60	14.16	.91	.33	5920	G0	.799	.557
98	21 39 30	57 45 50	15.13	1.34	.74	5840	G2:	.463	.554
205	21 39 49	57 49 59	13.77	1.09	.85	8460	A5	.930	.301
103	21 40 7.1	57 59 38	14.52	2.12	1.36	5380	G9	.581	.690
104	21 40 7.4	57 58 07	13.86	1.48	.91	5940	G0	.808	.555
110	21 40 43	57 47 28	15.04	1.78	.95	5200	K0	.579	.718
112	21 41 3.9	57 07 27	14.05	1.20	.76	6380	F4	.869	.464
114	21 41 25	56 13 56	15.17	.08	-.35	6420	F4	.856	.459
115	21 41 28	56 15 06	14.30	1.76	.76	4780	K3	.600	.770

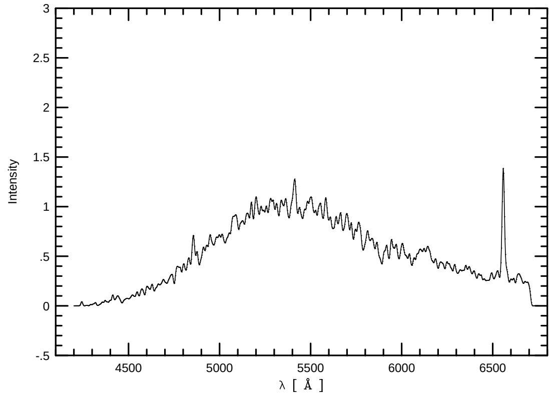

Figure 5.4: The spectrum of Kun 193. It is the only star in the sample showing very conspicuous $H\alpha$ emission.

with the evolutionary tracks and isochrones of Forestini (1994), one may infer that they are pre-main sequence objects with masses of $0.5M_\odot < M < 3M_\odot$ and $10^5 < t < 10^7$ years age.

Simonson (1968) derived an age of 3×10^6 years for this part of the CepOB2 association.

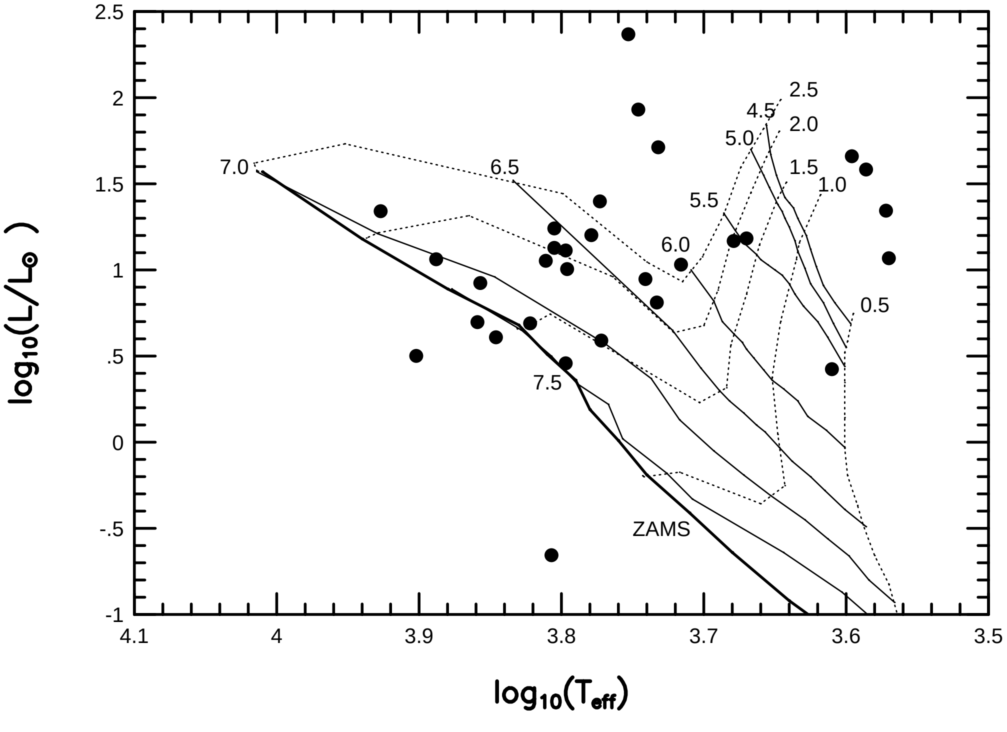

Figure 5.5: Distribution of our program stars on the $\{\log_{10}(L); \log_{10}(T_{eff})\}$ plane. Solid lines mark the isochrones and dotted ones the evolutionary tracks with constant masses. Numbers at the top of the lines show the logarithm of the corresponding stellar ages and masses, respectively. Stars with uncertain classification are omitted from this plot.

Chapter 6

Angular distribution of GRBs

6.1 True nature of GRBs

The true physical nature of gamma-ray bursts (GRBs) is one of the tantalizing enigmas of the recent astrophysics. Although since their first detection (Klebesadel et al., 1973) there were several suggestions trying to give a clear explanation for their origin, no definite answer has been given yet (cf. Paczyński, 1995). Recently, the successful identifications made by the Beppo-SAX satellite, followed by the detection of optical counterparts (van Paradijs et al., 1997), seem to give firm support for the models putting them definitely into the cosmological distances. The alternative Galactic origin seems to be ruled out (a survey of the question of distances may be found, e.g., in Paczyński (1995)). However, the small number of optically identified events is far from being enough to characterize the properties of the whole burst population. On the other hand, the existence of cosmological distances of GRBs seems to be definite.

In addition, even before this identification, indirect observational evidences were known for the cosmological origin. These evidences were based mainly on the modifications of $< V/Vmax >$ test (cf. Norris et al., 1994; Mészáros P. & Mészáros A., 1995; Norris et al., 1995; Nemiroff 1995; Horváth et al., 1996), and on the study of the time dilatation (cf. Norris et al., 1995; Mészáros A. & Mészáros P., 1996; Stern 1996; Mészáros A. et al., 1996; Che et al., 1997). A further important indirect support of cosmological origin is based on the observed isotropy on the sky (Briggs, 1993; Syer & Saha, 1994; Briggs, 1995; Tegmark et al., 1996a,b; Briggs et al., 1996). All these papers suggest that the angular distribution is isotropic, because there are no statistically significant departures from the isotropy. Also the separations of GRBs into the different subclasses either due to the duration (Kouveliotou et al., 1993; Belli, 1995; Dezalay et al., 1996) or due to the fluence on channel above the 300 keV (Pendleton et al., 1997) do not change the situation; the proposed subclasses alone also seem to be distributed isotropically. Hence, it can well occur that the total number of observed GRBs is a mixture of a wide variety of physically different objects, but all GRBs should be at cosmological distances due to their isotropic angular distribution on sky, and due to other direct and indirect supports.

In the following we find a clear anisotropy of all GRBs, and then separately of the short GRBs, too. On the other hand, the long GRBs seem to be still distributed isotropically. To do this we will use both the standard analysis based on the spherical harmonics and also the so called binomial test.

6.2 Mathematical skeleton of the problem

Testing the isotropy on the celestial sphere one may use several methods. Nevertheless, strictly from the mathematical point of view, the necessary condition for the isotropy is the stochastic independency of the sky distribution of the bursts on their observed physical properties. It means that, if $f(l,b,x_1,\dots,x_n)dFdx_1\dots dx_n$ is the probability of finding an object in the $dF = \cos b\, dl\, db$ infinitesimal solid angle and in the $(x_1, x_1+dx_1,\dots,x_n,x_n+dx_n)$ interval, one must have

$$f(l,b,x_1,\dots,x_n) = \omega(l,b)g(x_1,\dots,x_n) \tag{6.1}$$

Here $0 \le l \le 360^o, -90^o \le b \le 90^o$ give the celestial positions in Galactic coordinates, $x_n\,(n \ge 1)$ measure the physical properties (peak fluxes, fluences, durations, etc...) of GRBs and g is their probability density. It may well be assumed that in the case of isotropy the distribution of GRBs fulfils this equation (cf. Briggs et al., 1996; Tegmark et al., 1996a,b).

However, statement (6.1) is only a necessary but not a sufficient condition for isotropy. Isotropy means that also $\omega(l,b) = \frac{1}{(4\pi)}$. Hence, in this case, for N observed GRBs the events $dN = N\omega dF$, i.e. the expected number of GRBs in an infinitesimal solid angle, is not depending on $[l,b]$. In other words, the isotropy means that the probability of observing a burst in a solid angle $0 < \Omega \le 4\pi$ (Ω is in steradians) is given by $\frac{\Omega}{4\pi}$ and is independent on its location on the celestial sphere. This follows immediately from (6.1), if one does integration over l and b to obtain, first, the solid angle Ω, and, second, the whole sky. Then the ratio of two results gives $\frac{\Omega}{4\pi}$, and the concrete form of g is unimportant.

The most frequently used procedure to test the isotropy of GRBs is based on the spherical harmonics (Briggs 1993, 1995; Briggs et al., 1996; Tegmark et al., 1996a,b). The key idea is the following. In general case one may decompose the function $\omega(l,b)$ into the well-known spherical harmonics. One has:

$$\omega(l,b) = (4\pi)^{-1/2}\omega_0$$

$$-(\frac{3}{(4\pi)})^{1/2}(\omega_{1,-1}\cos b\sin l - \omega_{1,1}\cos b\cos l + \omega_{1,0}\sin b)$$

$$+(\frac{15}{(16\pi)})^{1/2}(\omega_{2,-2}\cos^2 b\sin 2l + \omega_{2,2}\cos^2 b\cos 2l$$

$$-\omega_{2,-1}\sin 2b\sin l - \omega_{2,1}\sin 2b\cos l)$$

$$+(\frac{5}{(16\pi)})^{1/2}\omega_{2,0}(3\sin^2 b - 1) + \textit{higher order harm.} \tag{6.2}$$

The first term on the right-hand side is the monopole term, the following three ones are the dipole terms, the following five ones are the quadrupole terms (cf. Press et al. 1992; Chapt. 6.8). Nevertheless, ω is constant for isotropic distribution, and hence on the right-hand side any terms, except for ω_0, should be identically zeros. To test this hypothesis one may proceed, e.g., as follows. Let there are observed N GRBs with their measured positions $[l_j, b_j]$ $(j = 1, 2, \ldots, N)$. In this case ω is given as a set of points on the celestial sphere. Because the spherical harmonics are orthogonal functions, to calculate the $\omega_{\{\}}$ coefficients one has to compute the functional scalar products. For example, $\omega_{2,-1}$ is given by

$$\omega_{2,-1} =$$

$$-(\frac{15}{(16\pi)})^{1/2}\int\limits_{-\pi/2}^{\pi/2}\cos b\,db\int\limits_{0}^{2\pi}\omega(l,b)\sin\,2b\,\sin\,l dl$$

$$= -(\frac{15}{(16\pi)})^{1/2}N^{-1}\sum_{j=1}^{N}\sin\,2b_j\sin\,l_j. \tag{6.3}$$

Because ω is given only in discrete points, the integral is transformed into an ordinary summation (cf. Kendall & Stuart (1973)). In the case of isotropy one has $\omega_{2,-1} = 0$, and hence $N^{-1}\sum\limits_{j=1}^{N}\sin\,2b_j\sin\,l_j = 0$. therefore, the expected mean of sin $2b_j$ sin l_j values is zero. One has to proceed similarly to any other $\omega_{\{\}}$ coefficient.

In order to test the zero value of, e.g., $\omega_{2,-1}$ one has to calculate, first, $\sin 2b_j \sin l_j$ for any $j = 1, 2, \ldots, N$ and, second, the mean, standard deviation and Student's t variable (cf. Press et al. 1992, Chapt. 14). Finally, third, one has to ensure the validity of zero mean from Student test. As far as it is known, no statistically significant anisotropy of GRBs were detected yet by this procedure (cf. Briggs et al., 1996; Tegmark et al., 1996a,b).

Nevertheless, there are also other ways to test the isotropy. An extremely simple method uses the binomial distribution. In the remaining part of this section we explain this test (see also Mészáros A., 1997).

In order to test the anisotropy by this method one may proceed as follows. Let us take an area on the sky defined by a solid angle $0 < \Omega < 4\pi$ (Ω is in steradians). In the case of isotropy the p probability to observe a burst within this area is $p = \frac{\Omega}{(4\pi)}$. Then, obviously, $q = 1 - p$ is the probability to have it outside. Observing $N > 0$ bursts on the whole sky the probability to have k bursts (it may be $k = 0, 1, 2, \ldots, N$) within Ω is given by the binomial (Bernoulli) distribution taking the form

$$P_p(N, k) = \frac{N}{k!(N-k)!} p^k q^{N-k} \qquad (6.4)$$

This distribution is one of the standard probability distributions discussed widely in statistical text-books (e.g. Trumpler & Weaver, 1953; Kendall and Stuart, 1973; about its use in astronomy see, e.g., Mészáros, 1997). The expected mean is Np and the expected variance is Npq. One may also calculate the integral (full) probability, too, by a simple summation.

In our case we will consider N GRBs, and we will test the hypothesis whether they are distributed isotropically on the sky. Assume that k_{obs} is the observed number of GRBs at the solid angle Ω. If the apriori assumption is the isotropy, i.e $p = \frac{\Omega}{(4\pi)}$, then one may test whether the observed number k_{obs} is compatible with this apriori assumption. Of course, any $0 \leq k_{obs} \leq N$ can occur with a certain probability given by the binomial distribution. But, if this probability is too small, one hesitates seriously to accept the apriori assumption.

Consider the value $|k_{obs} - Np| = k_0$. The value k_0 characterizes "the departure" of k_{obs} from the mean Np. Then one may introduce the probability

$$P(N, k_{obs}) = 1 - P_{p,tot}(N, (Np + k_0)) + P_{p,tot}(N, (Np - k_0)) \qquad (6.5)$$

$P(N, k_{obs})$ is the probability that the departure of k_{obs} from the Np mean is still given by a chance.

Table 6.1: Student test of the dipole and quadrupole terms of 2025 GRBs. In the first column the coefficients defined in Eq. (6.2) are given. In the second column the Student t is provided. The third column shows the probability that the considered terms are still zeros.

	t	%
$\omega_{1,-1}$	1.51	13.4
$\omega_{1,1}$	1.77	7.7
$\omega_{1,0}$	0.71	47.7
$\omega_{2,-2}$	2.76	0.6
$\omega_{2,2}$	1.54	12.1
$\omega_{2,-1}$	3.26	0.1
$\omega_{2,1}$	0.98	33.3
$\omega_{2,0}$	0.36	71.9

In order to test the isotropy of the GRBs celestial distribution we will divide the sky into two equal areas, i.e we will choose $p = 0.5$. It is essential to note here that neither of these two parts must be simply connected compact regions.

6.3 Anisotropy of all GRBs

In order to test the isotropy of GRBs we will test the three dipole and five quadrupole terms in accordance with the method described in the previous section. We will consider all GRBs that have well-defined angular positions. Up to the end of year 1997 there were 2025 such objects at Current BATSE catalog (Meegan et al. 1997; Paciesas et al. 1998). The results are summarized in Table 6.1.

We see that, except for the terms defined by $\omega_{2,-1}$ and $\omega_{2,-2}$, the remaining six terms may still be taken to be zero. This means that there is a clear anisotropy defined by term $\sim \sin 2b \sin l$. The probability that this term is zero is than 0.1%. It is practically sure that the second quadrupole term being proportional to $\sim \cos^2 b \sin 2l$ is non-zero, too.

The anisotropy defined by $\omega_{2,-1}$ may be defined in another form, too. This quadrupole term has a positive sign, when both $\sin 2b$ and $\sin l$ have the same signs, and has a negative sign, when $\sin l$ and $\sin 2b$ have opposite signs. Therefore, let us define two parts of the sky having the same sizes (2π steradians). The first one is defined by Galactical coordinates $b > 0, l > 180^o$ and $b < 0, 0 < l < 180^o$. This means that this first part is in fact composed from two separated "sky-quarters". The second part is then given by $b > 0, 0 < l < 180^o$ and $b < 0, l > 180^o$. This means that this second part is also given by two separated "sky-quarters". Then the detected quadrupole anisotropy suggests that there should be an essential difference, e.g., in the number of GRBs in these two parts.

In order to test again this expectation we will do the test based on Bernoulli distribution. We divide the whole sky into these two parts, and hence we expect a Bernoulli distribution with $p = 0.5$ for $N = 2025$. As it is noted in Sect. 6.2, it is certainly allowable that these parts are composed from several subregions.

A straightforward counting of GRBs in these regions shows that 930 GRBs are in the first one and 1095 are in the second. Assuming $p = 0.5$ the binomial (Bernoulli) test gives a 0.03% probability that this distribution is caused only by a chance. Hence, the relatively smaller number in the first region compared with the second one is not a chance, and the distribution of all GRBs is anisotropic with a high probability.

Clearly, concerning the consequences of the intrinsic anisotropy of GRBs, one must be still careful. Instrumental effects of the BATSE experiment may also play a role, and in principle it can also occur that the detected anisotropy is caused exclusively by instrumental effects. To be as correct as possible, one may claim that, in essence, there can be three different causes of this observed anisotropy: a. The anisotropy is purely caused by the BATSE's nonuniform sky exposure (in other words, the intrinsic angular distribution of GRBs is still isotropic, and the observed anisotropy is a pure instrumental effect); b. The anisotropy is purely given by the intrinsic anisotropy of GRBs, and the instrumental effects are unimportant; c. The anisotropy is given both by instrumental effects and also by the intrinsic anisotropy. To be sure that there is also an intrinsic anisotropy of GRBs, one must be sure that the possibility a. does not occur. In what follows, when we will speak about "the possibility a.", we will consider this one.

It is well-known that the sky exposure of the BATSE instrument is nonuniform. This question is described and discussed in several papers (cf. Briggs, 1993; Fishman et al., 1994; Tegmark et al., 1996a,b; Briggs et al., 1996). The BATSE sky coverage depends on the declination only in the equatorial coordinate system (Tegmark et al., 1996b) in such a manner that the probability of detection is about 10% higher near the pole than near the equator. This behavior in Galactic coordinates predicts excess numbers of GRBs just in sky-quarters given by $b > 0, 0 < l < 180^o$ and $b < 0, l > 180^o$. Hence, in principle, it is well possible that the observed anisotropy is caused by a pure instrumental effect. The purpose of Sect. 6.4 is to show that this is not the case, and the possibility a. should be excluded.

6.4 Different distribution of short and long GRBs

If there is an intrinsic isotropy indeed, then Eq. (6.1) will continue to fulfill, and $\omega(l, b)$ itself will reflects the non-uniformity of sky-coverage. Then at the first part the number of observed GRBs should be smaller, because the integrations of $\omega(l, b)dF$ giving the first and second part, respectively, do not give the same values. Their ratio should be $\simeq 930/1095 = 0.85$. This ratio should be obtained for any subset of GRBs, when the choice of this subset is based on some physical properties of the bursts, because the function g does not enter into the calculation for X. In other words, if the level of anisotropy depends on the physical parameters of GRBs, then "the possibility a." should be excluded.

To be extra cautious it is also necessary to remember that some bias may arise for the dimmest GRBs, because for them it is not necessarily true that the total exposure time is exactly proportional to the observed number of sources due to the varying threshold limit of BATSE (see Tegmark et al., 1996a for the discussion of this question). To avoid this bias the simplest procedure is not to take into account the dimmest GRBs.

All this allows to exclude the possibility a. quite simply. To do this it is necessary to take some subsets of GRBs, and to verify for any of them that the ratio X is roughly 0.85. For security, it is also necessary to omit the dimmest GRBs.

For the sake of maximal correctness we will not use any ad hoc criterions to define such subsets, but we will exclusively use criterions which were introduced earlier by others. First, we exclude any GRBs having the peak fluxes on 256 ms trigger smaller than 0.65 $photons/(cm^2 s)$. The truncation with this threshold is proposed and used in Pendleton et al. (1997). Second, from the remaining GRBs we exclude GRBs which have no defined duration T_{90} (for the definition of this duration see Kouveliotou et al. (1993)). Third, we also exclude the bursts which have no f_3 value, which is the fluence on the energy channel [100, 300] keV. These truncations are also necessary, because we will consider the subsets, for which the criterions use T_{90} and f_3.

932 GRBs remain, and from them 430 are at the first part, and 502 are at the second one. Hence, here X = 430/502 = 0.86. There is no doubt that this "truncated" sample of GRBs is distributed similarly to that of whole sample with 2025 GRBs. The probability that this distribution is given by a chance is here 2%. (Note that, of course, here the same X does not give the same probability, because there is a smaller number in the sample.)

932 GRBs will be separated, first, into the "short" and "long" subclasses

Table 6.2: Results of the binomial test of subsamples of GRBs with different durations. N is the number of GRBs at the given subsample, k_{obs} is the observed number GRBs at the first part in this subsample, and % is the probability in percentages that the assumption of isotropy is still valid.

sample	N	k_{obs}	$(N - k_{obs})$	%
all GRBs	932	430	502	2.0
$T_{90} < 1s$	206	82	124	0.43
$T_{90} < 2s$	251	103	148	0.55
$T_{90} < 10s$	372	154	218	0.11
$T_{90} > 2s$	681	327	354	32
$T_{90} > 10s$	560	276	284	77
$T_{90} > 15s$	507	247	260	59

(cf. Kouveliotou et al., 1993), and, second, into the No-High-Energy (NHE) and High-Energy (HE) bursts (Pendleton et al., 1997).

The boundary between the short and long bursts is usually taken for $T_{90} = 2s$ (Kouveliotou et al., 1993; Belli, 1995; Dezalay et al., 1996). Nevertheless, this boundary at $T_{90} = 2s$ is not so precise (e.g. in Katz & Canel (1996) $T_{90} = 10s$ is used). In addition, this boundary gives no definite strict separation, because at class $T_{90} < 2s$ long bursts, and at class $T_{90} > 2s$ short bursts are also possible, respectively (cf. Belli, 1995). Therefore, in order to test more safely the distribution of two subclasses, we will consider also the case when the boundary is at $T_{90} = 10s$. We will consider also the subsamples of GRBs having $T_{90} < 1s$, and $T_{90} = 15s$, respectively, because then they contain surely only short and long bursts, respectively. The results are shown in Table 6.2.

Table 6.2 shows, e.g., that there are 251 GRBs with $T_{90} < 2s$, and 681 GRBs with $T_{90} > 2s$. Then, from the short GRBs 103 are at the first part of sky and 148 at the second one. This gives X = 103/148 = 0.70. It seems that at the first part there is even a smaller portion of shorter GRBs than that of the all GRBs. The probability that this is a chance is given by 0.55%. On the other hand, from the long GRBs 327 are at the first part and 354 at the second one. This gives X = 327/354 = 0.92. Hence, it seems immediately that for the long GRBs the isotropy is still an acceptable assumption. The binomial test quantifies: there is a 32% probability that this distribution is given by a chance. Doubtlessly, the long GRBs are distributed more isotropically than the short ones; there is no statistically significant departure from isotropy for long GRBs. The subsamples $T_{90} < 1s$ and $T_{90} < 15s$ confirm this expectation; the boundary at $T_{90} = 10s$ also does not change the conclusion.

Table 6.3: Results of the binomial test of NHE-HE subsamples of GRBs. The subsamples NHE1, HE1, NHE2, HE2 are explained in the text.

sample	N	k_{obs}	$(N - K_{obs})$	%
all GRBs	932	430	502	2.0
NHE1	131	52	69	14.9
HE1	693	327	366	14.9
NHE2	168	69	99	2.5
HE2	764	361	403	13.8

One may claim that the anisotropy of short GRBs is statistically significant, but for long GRBs it is not.

Doubtlessly, the short and long subclasses are distributed differently. This also excludes the possibility a.; the observed anisotropy of all GRBs cannot be caused exclusively by instrumental effects. It is difficult to imagine an instrumental effect which leads to isotropy of long GRBs and to anisotropy of short GRBs.

Pendleton et al. (1997) introduces the subclasses of HE and NHE bursts. The criterion depends on the ratio f_4/f_3, where f_3 is the fluence on energy channel [100,300] keV and f_4 is the fluence on the energy channel [> 300] keV. (From this it is also clear, why we needed non-zero f_3 . On the other hand, f_4 can be vanishing; these GRBs are simply NHE bursts.) Application of this criterion is not so simple, because for a great portion of GRBs there are large uncertainties of the values of f_4 due to their errors. Concretely, for 693 GRBs the value of f_4 is bigger than the corresponding error of this; for 131 GRBs there is no f_4; for the remaining 108 GRBs there are some values of f_4 at the current BATSE catalog (Meegan et al., 1997), but they are smaller than their errors. Hence, 693 GRBs can be taken as HE bursts (HE1 subsample). 131 GRBs having no f_4 may be taken as NHE bursts (NHE1 subsample). We did binomial tests for these two subclasses. Separation of the remaining 108 GRBs into the HE and NHE subclasses is not so clear. We consider artificially the boundary as follows: If the value of f_4 is bigger than the half of error, then we have HE; otherwise NHE. Applying this criterion we will have 168 NHE (NHE2 subsample) and 764 HE bursts (HE2 subsample). For them the binomial tests were also done. The results are collected in Table 6.3.

Table 6.3 gives an ambiguous result. Due to the smaller number in subclasses no anisotropy is confirmed yet on a satisfactorily high level of significance. In addition, contrary to the short-long separation, there is no obvious difference between HE and NHE classes.

6.5 Discussion

The quadrupole anisotropy reported in the previous section is an unexpected and new result. As far as it is known no anisotropy terms were detected yet (cf. Tegmark et al., 1996a,b). Probably this situation was given by the fact that the majority of these isotropy studies concentrated the effort into dipole and $\omega_{2,0}$ quadrupole terms, which are expected to differ from zero, if the GRBs are arisen in the Galaxy.

The essentially different angular distribution of short and long GRBs suggests that their separation into these two subclasses has a deeper cause. It is well-known that in average the short bursts have higher hardnesses (hardness $=f_3/f_2$, where f_2 is the fluence on energy channel [50, 100] keV). Katz & Canel (1996) have also shown that the $< V/V_{max} >$ values are different; the smaller value for longer GRBs suggests that they are on average at bigger cosmological distances. Keeping all this in mind it seems to be definite that these two types are physically different objects at different cosmological scales.

Contrary to this, we did not find any significant difference in the angular distribution of HE and NHE subclasses. This is an unclear result, because in Pendleton et al. (1997) it is clearly stated that the $< V/V_{max} >$ values are different for the HE and NHE subclasses, and hence they should also be at different distances. The isotropy tests do not confirm this expectation. This also means that the separation based on the most energetic channel remains unclear. In fact, the question of fourth channel is highly topical recently, because Bagoly et al. (1998) shows - independently on Pendleton et al. (1997) - that f_4 alone is an important quantity. The question of most energetic channel trivially needs further study, and is planned to be done.

At the end of Sect. 6.3 we pointed out that the dependence of the BATSE detection probability on the declination may mimic some sort of anisotropy. Therefore, without a detailed correction in accordance with the sky exposure function one may state only the presence of an intrinsic anisotropy from the different behavior of short and long bursts. The requirement of the study of this correction for both subclasses is trivial, and is planned to be done in the near future.

One may also speculate that the short GRBs can arise in the Galaxy and the long ones at cosmological distances. (About the cosmological origin of long GRBs there seems to exist no doubt; see, cf., Mészáros A. et al., 1996). The existence of non-zero $\omega_{2,-1}$ and probably also of $\omega_{2,-2}$ quadrupole terms with the simultaneous zeros for other dipole and quadrupole terms is a strange behavior for any objects arising in the Galaxy. E.g., it is highly complicated to have an $\omega_{2,-1}$ term, and simultaneously not to have the $\omega_{2,0}$

term, if the sources have arisen in the Galaxy. Simply, any objects in the Galaxy should have fully different anisotropy terms (for further details and for the survey of earlier studies of isotropy see, e.g., Briggs (1993), (1995); Briggs et al. (1996); Tegmark et al. (1996a,b), Meegan et al. (1996)).

Remark also the following. The so called transition scale to homogeneity (cf. Mészáros A., 1997) is minimally of size $\simeq 300h^{-1}\,Mpc$ (h is the Hubble constant in units $100km/(s\,Mpc)$). This means that up to this distance an inhomogeneous and anisotropic spatial distribution is not only possible but it is even expected. In addition, at the last time several observational implications, both from the distribution of galaxies and from the anisotropy of cosmic microwave background radiation, highly query the fulfilment of homogeneity and isotropy even up to the Hubble scale (Lauer & Postman, 1994; Slechta & Mészáros A., 1997; Mészáros A. & Vansek, 1997; Coles, 1998; Sylos-Labini et al., 1998). Hence, as far as anisotropy concerns we think that the different distribution of short and long GRBs and their cosmological origin are not in contradiction, and we further mean that all GRBs are at cosmological distances.

Chapter 7

Classification of GRBs

7.1 Formulation of the problem

It is a great challenge to classify the gamma-ray bursts (GRBs) into subclasses. Mazets et al. (1981) and Norris et al. (1984) suggested that the duration alone might be a good classification parameter in order to separate the objects into subclasses. Using the First BATSE Catalog Kouveliotou et al. (1993) found a bimodality in the distribution of the logarithms of the durations. This bimodality is highly pronounced, if one uses the parameter T_{90} (the time during which 90% of the fluence is accumulated (Kouveliotou et al. , 1993)) for characterizing the durations of GRBs (McBreen et al. , 1994; Koshut et al. , 1996; Belli , 1997; Pendleton et al. , 1997). Today it is widely accepted that the physics of these two subgroups (also "subclasses" or simply "classes") are really different, and these two kind of GRBs are really different phenomena (Norris et al. , 2001; Balázs et al. , 2003). Note that the high redshifts and the cosmological distances are directly confirmed for the long bursts only; while for the short ones there are only indirect evidences for the cosmological origin (Mészáros , 2001, 2003).

Using the Third BATSE Catalog (Meegan et al., 1996) Horváth (1998) has shown that the distribution of the logarithms of the durations of GRBs ($\log T_{90}$) could be well fitted by a sum of three Gaussian distributions. He finds statistically unlikely (with a probability $\sim 10^{-4}$) that there are only two subgroups. Somewhat later several authors (Mukherjee et al., 1998; Hakkila et al., 2000c; Balastegui et al. , 2001; Rajaniemi & Mähönen , 2002; Borgonovo , 2004) included more parameters into the burst analysis (e.g. peak-fluxes, fluences, hardness ratios, etc.). A cluster analysis in this multidimensional parameter space suggests the existence of the third ("intermediate") subgroup as well (Mukherjee et al., 1998; Hakkila et al., 2000c).

On the other hand, the physical existence of the third group is, however, still not convincingly proven. For example, Hakkila et al. (2000) believes that the third subgroup is only a deviation caused by a complicated instrumental effect, which can reduce the durations of some faint long bursts. However, the sky distribution of the third subgroup is anisotropic (Mészáros et al. , 2000a,b; Litvin et al. , 2001); i.e. different to that of the long GRBs alone (Mészáros & Štoček , 2003). The logN-logS distribution may also differ from those of the other groups (Horváth, 1998). All this means that the existence of the third intermediate subgroup is acceptable, but its physical meaning, importance and origin is less clear than those of the other subgroups. Hence, its further study is highly required.

Using the Principal Component Analysis (PCA) Bagoly et al. (1998) has shown that there are only two major quantities necessary (the so called Principal Components; PCs) to characterize all the properties of the bursts in the BATSE Catalog. The first PC accounts for the duration and the fluence; the second one for the peak-fluxes. Consequently, the problem of the choice of the relevant parameters describing GRBs *is basically a two dimensional problem.* For the statistical analysis the choice of two independent parameters is enough; they may, but not necessarily, be the two principal components. Recently Balázs et al. (2003) has used a two dimensional analysis and has shown that the two subgroups (short and long ones) are really different phenomena. The most remarkable difference is seen if the hardness (Meegan et al., 1996) and the duration are used as the two different independent parameters. The short and long GRBs are different in both these independent variables. All this means that only two - relevantly and correctly chosen - parameters should be enough for the classification and determination of the subgroups, too. Therefore, a detailed two-dimensional analysis - using the duration and hardness - in the classifications is well supported and highly needed. This will be done here concentrating our effort with respect to the third intermediate subgroup.

Several questions should be answered concerning the intermediate subgroup. First, will the statistical analysis, using only these two different parameters, reconfirm the existence of the intermediate subgroup? Second, if this question is positively answered, then one has to show that either further subgroups exists, or not. This is a fully open question. Mukherjee et al. (1998) claims that only three subgroups are necessary; on the other hand, Cline et al. (1999) proclaims the existence of a fourth subgroup having very short durations. Third, one also has to define the quantities, in which this third subgroup is different. Hardness and the duration seem to be the best parameter; but other choices may also be possible. This should be clarified. Fourth, the method - allowing to assign a given concrete GRB to a given subgroup - should be developed, too. The simple separation with respect to T_{90} is surely rough, of course. A better one is clearly needed. Fifth, the fraction of this third intermediate subgroup in the whole BATSE Catalog should also be determined more exactly. Sixth, as it was noted, the intermediate subclass shows a remarkable anisotropy in the angular sky distribution. Confirmation or the rejection of this behaviour is also required. Seventh, does the intermediate group really represent a third type of bursts different from both the short and long ones by its astrophysical origin?

Searching for answers to these questions is the aim of this chapter. The observational data from the BATSE Current GRB Catalog (Meegan et al., 2001) will be used, in which there are 2702 GRBs; from them 1956 GRBs have both measured hardnesses and durations. These 1956 GRBs define the sample studied in this chapter.

7.2 Mathematics of the two-dimensional fit

We will study the distribution of GRBs in the $[\log T_{90}, \log H_{32}]$ plane. Previously Belli (1997) used this plane to separate the bursts. She suggested that the curve $H_{32} = 2T_{90}^{0.5}$ gave a better division than the cut $T_{90} = 2$ s between the short and long GRBs.

We may assume that the observed probability distribution of the GRBs in this plane is a superposition of the distributions characterizing the different types of the bursts being present in the sample. Introducing the notations $x = \log T_{90}$ and $y = \log H_{32}$ and using the law of full probabilities (Rényi , 1962) we may write

$$p(x, y) = \sum_{l=1}^{k} p(x, y|l) p_l. \tag{7.1}$$

In this equation $p(x, y|l)$ is the conditional probability density assuming that a burst belongs to the l-th class. p_l is the probability for this class in the observed sample ($\sum\limits_{l=1}^{k} p_l = 1$), where k is the number of classes. In order to decompose the observed probability distribution $p(x, y)$ into the superposition of different classes we need the functional form of $p(x, y|l)$. The probability distribution of the logarithm of durations can be well fitted by Gaussian distributions, if we restrict ourselves to the short and long GRBs, respectively (Horváth (1998)). We assume the same for the y coordinate as well. With this assumption we obtain for a certain l-th class of GRBs a two dimensional Gaussian in the form of

$$p(x,y|l) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \times \exp\left[-\frac{1}{2(1-r^2)}\left(\frac{(x-a_x)^2}{\sigma_x^2} + \frac{(y-a_y)^2}{\sigma_y^2} - \frac{C}{\sigma_x\sigma_y}\right)\right], \tag{7.2}$$

where $C = 2r(x-a_x)(y-a_y)$; a_x, a_y are the means, σ_x, σ_y are the dispersions, and r is the correlation coefficient (Trumpler & Weaver (1953); Chapt. 1.25). Hence, a certain class is defined by 5 independent parameters, a_x, a_y, σ_x, σ_y, r, which are different for different l. If we have k classes, then we have $(6k-1)$ independent parameters (constants), because any class is given by the five parameters of Eq.(7.2) and the weight p_l of the class; one weight is not independent, because it holds $\sum_{l=1}^{k} p_l = 1$. Sum of k functions defined by Eq.(7.2) gives the theoretical function of the fit. In Balázs et al. (2003) this fit for $k=2$ was used, and procedure for $k=2$ was described in more details. However, that paper used fluence instead of hardness. We will make here similar calculations for $k=3$ and $k=4$, as well.

7.3 Confirmation of the intermediate group

We mentioned in the previous section that by decomposing $p(x,y)$ into the superposition of $p(x,y|l)$ conditional probabilities one divides the original population of GRBs into k subgroups; at least from the mathematical point of view. Dissolving the left-hand-side of Eq.(7.1) into the sum of the right-hand-side, one needs the functional form of $p(x,y|l)$ distributions, and k has to be fixed, too. Because we assume that the functional form is a bivariate Gaussian distribution (see Eq.(7.2)), our task is reduced to the estimation of its parameters, to the values of k and p_l-s.

In order to get the unknown constants in Eq.(7.2) we use the Maximum Likelihood (ML) procedure of parameter estimation (Balázs et al. , 2003). Assuming a set of N observed $[x_i, y_i]$, $(i=1,\ldots,N)$, values (N is the number of GRBs in the sample for our case, which here is 1956) we may define the Likelihood Function in the usual way, after fixing the value of k, in the form

$$L = \sum_{i=1}^{N} \log p(x_i, y_i)\ , \tag{7.3}$$

where $p(x_i, y_i)$ has the form given by Eq.(7.1). Similarly, as it was done by Balázs et al. (2003), the EM (Expectation and Maximization) algorithm is used to obtain the $a_x, a_y, \sigma_x, \sigma_y, r$ and p_l parameters at which L reaches its

Table 7.1: Results of the EM algorithm. $k = 2 \quad L_{max} = 1193$

l	p_l	a_x	a_y	σ_x	σ_y	r
1	.280	-.233	.740	.541	.259	.049
2	.720	1.488	.396	.471	.237	.128

Table 7.2: Results of the EM algorithm. $k = 3 \quad L_{max} = 1237$

l	p_l	a_x	a_y	σ_x	σ_y	r
1	.245	-.301	.763	.525	.251	.163
2	.109	.637	.269	.474	.344	-.513
3	.646	1.565	.427	.416	.210	-.034

maximum value. We made the calculations at different values of k in order to see the improvement of L as we increase the number of the parameters to be estimated.

Tables 7.1-7.3 summarize the results of the fits for $k = 2, 3, 4$.

Similarly to Sec. 4.3.2 the confidence interval of the parameters estimated can be given on the basis of the following theorem. Denoting with L_{max} and L_0 the values of the Likelihood Function at the maximum and at the true value of the parameters, respectively, one may write asymptotically as the sample size $N \to \infty$ (Kendall & Stuart, 1973)

$$2(L_{max} - L_0) \approx \chi^2_m, \tag{7.4}$$

where m is the number of the parameters estimated ($m = 6k - 1$ in our case), and χ^2_m is the usual m-dimensional χ^2 function (Trumpler & Weaver, 1953). Moving from $k = 2$ to $k = 3$ the number of parameters m is increasing by 6 (from 11 to 17), and L_{max} is growing from 1193 to 1237. Since $\chi^2_{17} = \chi^2_{11} + \chi^2_6$ the increase in L_{max} with a value of 44 corresponds to a value of 88 for

Table 7.3: Results of the EM algorithm. $k = 4 \quad L_{max} = 1243$

l	p_l	a_x	a_y	σ_x	σ_y	r
1	.234	-.307	.752	.524	.246	.215
2	.060	.441	.426	.637	.440	-.871
3	.060	.623	.262	.325	.325	-.095
4	.646	1.569	.426	.410	.211	-.034

a χ^2_6 distribution. The probability for $\chi^2_6 \geq 88$ is extremely low ($< 10^{-10}$), so we may conclude that the inclusion of a third class into the fitting procedure is well justified by an extremely high level of significance.

Moving from $k = 3$ to $k = 4$, however, the improvement in L_{max} is only 6 (from 1137 to 1143) corresponding to $\chi^2_6 \geq 12$, which may happen by a chance with a probability of 6.2 %. Hence, the inclusion of the fourth class is *not* justified. We may conclude from this analysis that the superposition of three Gaussian bivariate distributions satisfactorily describes the observed distribution.

This means that the 17 constants for $k = 3$ in Table 7.2 define the parameters of the three subgroups. We see that the mean hardness of the intermediate subclass is very low - the third subclass is the softest one. Because $p_2 = 0.109$, 11% of all GRBs belongs to this subgroup. This value is very close to the previous ones (Mukherjee et al., 1998; Hakkila et al., 2000c; Horváth , 2002).

7.4 Mathematical classification of GRBs

7.4.1 The method

Based on the calculations in the previous Section we resolved the $p(x, y)$ probability density of the observed quantities into a superposition of three Gaussian distributions. Using this decomposition we may classify *any* observed GRB into the classes represented by these subgroups. In other words, we develop a method allowing us to obtain, for any given GRB, its three membership probabilities defining how likely this GRB belongs to the short, intermediate and long subgroups, respectively. The sum of these three probabilities is unity, of course. For this purpose we define the following $I_l(x, y)$ indicator function, which assigns a membership probability for each observed burst in a given l class as follows:

$$I_l(x, y) = \frac{p_l p(x, y|l)}{\sum\limits_{l=1}^{k} p_l p(x, y|l)}. \tag{7.5}$$

According to Eq.(7.5) each burst may belong to any of the classes with a certain probability. In this sense one cannot assign a given burst to a given class with a certainty, but with a given probability. This type of classification is called "fuzzy" classification (McLachlan & Basford , 1988). Although, any burst with a given $[x, y]$ could be assigned to all classes with a certain probability, one may select that l at which the $I_l(x, y)$ indicator function

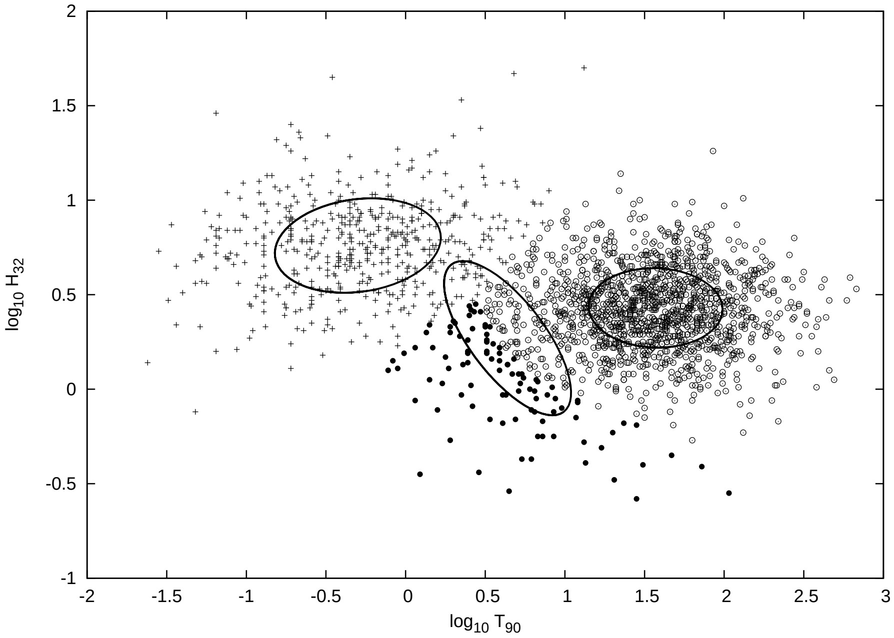

Figure 7.1: Distribution of $N = 1956$ GRBs in the $\{\log_{10} T_{90}; \log_{10} H_{32}\}$ plane. The 1σ ellipses of the three Gaussian distributions are also shown, which were obtained in the ML procedure. The different symbols (crosses, filled circles and open circles) mark bursts belonging to the short, intermediate and long classes, respectively.

reaches its maximum value. Figure 7.1 shows the distribution of GRBs in the $\{\log_{10} T_{90}; \log_{10} H_{32}\}$ plane, in which the classes obtained in this way are marked with different symbols. The 1σ ellipses of the three Gaussian distributions are also shown.

7.4.2 Application of the fuzzy classification

Inspecting Figure 7.1 one may recognize immediately that the domain within the ellipse of the intermediate group is only partly populated by GRBs belonging to this class according to the classification procedure described above. The remaining part is dominated by GRBs classified into the short and, in particular, the long subgroup. In other words, the ellipse of the third subgroup contains an essential amount of GRBs, which should belong either to the long subgroup or to the short subgroup. Due to the 'fuzzy' classification some probability was also assigned to the other classes. Based on the analytical expressions of the components one may easily calculate the contribution of any other groups within the ellipse of a given class by summing up the

$I_l(x, y)$ values of different l-s for the bursts lying in this particular region.

The reliability of the classification can be characterized by counting the different classes of the GRBs lying within the 1σ ellipse of a given Gaussian component. If the classification was correct only those GRBs are within the ellipse of a given l which have classes corresponding to this component. Denoting with n_l the number of GRBs within the ellipse belonging to subclass l one gets $n_1 = 298$, $n_2 = 291$, $n_3 = 827$. The rows of Table 7.4 give the number of GRBs of all classes within the 1σ ellipses of the short, intermediate and long Gaussian components. The first row shows that in the ellipse, which defines the short subgroup, there are 298 GRBs; in accordance with the fuzzy classification all have the biggest probability assigning them to the short subgroup. Similarly, the third row shows that in the ellipse, which defines the long subgroup, there are 827 GRBs; all in accordance with the fuzzy classification have the biggest probability assigning them to the long subgroup. But in the second row, which defines the number 291 of GRBs being in the ellipse defining the intermediate subgroup, only 65 bursts have the biggest probability assigning them to the intermediate subgroup; further 67 (159) GRBs should belong to the short (long) subclass.

Table 7.4: Number of GRBs classified by the procedure described in the text, within the 1σ ellipses of $l = 1, 2, 3$ Gaussian components.

l	short	interm.	long	Total
1.	298	-	-	298
2.	67	65	159	291
3.	-	-	827	827
Total	365	65	986	1416

Table 7.4 demonstrates convincingly that the classification of the short and long GRBs are very reliable, since they do not have excursors from the other two classes. This means that GRBs within the ellipse of the first and third class (first and third row in Table 7.4) were classified well as short and long, respectively. In contrast, the ellipse of the intermediate component (second row) contains a significant number of the two remaining classes, in particular from the long ones. This is caused predominantly by the closeness of the most numerous long class to the intermediate one.

There are $N - (n_1 + n_2 + n_3) = 440$ GRBs sprinkling over a much larger area outside the ellipses. In this region the Gaussian components give low probabilities. The indicator function can have still a large value, however, because there are small numbers in both the nominator and denominator of the right-hand-side of Eq.(7.5). Although the classification of these bursts is formally correct, it is less reliable than those within the ellipses.

7.5 Physical differences between classes

The mathematical deconvolution of the $p(x, y)$ joint probability density of the observed quantities into Gaussian components does not necessarily mean that the physics behind the classes obtained mathematically is really different. It could well be possible that the true functional form of the distributions is not exactly Gaussian and the algorithm of deconvolution inserts only formally a third one in order to get a satisfactory fit.

Recently Balázs et al. (2003) found compelling evidences that there is a significant difference between the short and long GRBs. This might indicate that different types of engines are at work. The relationship of long GRBs to the massive collapsing objects is now also observationally well established (Mészáros , 2003), and the relation between the comoving and observed time scales is well understood (Ryde & Petrosian , 2002). The short bursts may be identified with neutron star (or black hole) mergers (Mészáros , 2001). So the mathematical classification of GRBs into the short and long classes - obtained also here (see Table 7.1 for $k = 2$) and in Balázs et al. (2003) - is also physically justified.

An important question has to be answered in this context: Does the intermediate group of GRBs, obtained in the previous paragraph from the mathematical classification, really represent a third type of bursts being physically different from both the short and long ones?

The classification into the short, intermediate and long classes proceeds mainly according to the durations. One may infer from Table 7.2 that these three classes differ also in the hardness. The difference in the hardness between the short and long group is well known (Kouveliotou et al. , 1993). According to these data the intermediate GRBs are the softest ones among the three classes. This different small mean hardness and also the different average duration suggest that the intermediate subgroup should also be a different phenomenon; simply, both in hardness and in the duration also the third subgroup is differing from the remaining two ones. On the other hand, no correlation exists between the hardness and the duration within the short and the long classes. More precisely, no correlation exists for the long subgroup; very weak for the short subgroup (see Table 7.2). In other words, these two quantities may be taken as two independent variables, and the short and long subgroups are different in *both* these variables.

In contrast, there is a strong anticorrelation between the hardness and

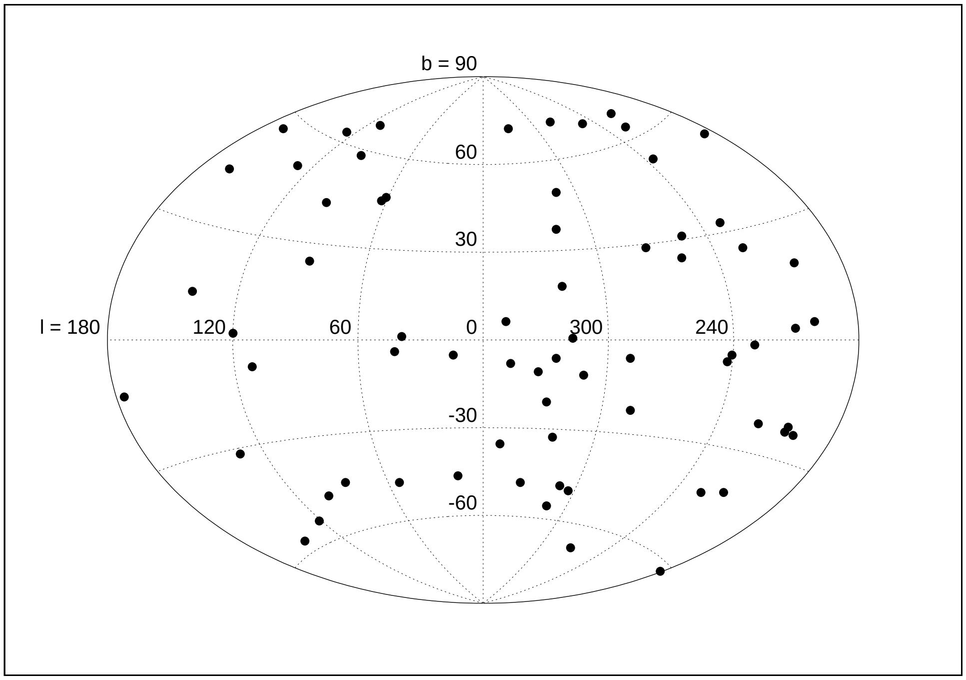

Figure 7.2: Distribution of 68 GRBs of the intermediate subclass in Galactic coordinates; the faintest bursts with $P_{256} < 0.65\, photons/(cm^2\, s)$ are omitted.

the duration within the intermediate class. This is a surprising new result, and because the hardness and the duration are *not* independent in the third subgroup, one simply may say that only one significant physical quantity is responsible for accounting the hardness and the duration within the intermediate subgroup. Consequently, the situation is quite different here, because one needs two independent variables for describing the remaining two other subgroups. This is a strong constraint for modelling the third subgroup. Hence, the question of the true nature of the physics in the intermediate subgroups remains open, and - trivially - it needs further detailed studies.

In previous papers (Mészáros et al. , 2000a,b) evidences were found that the angular sky distribution of the GRBs with duration $2\,\mathrm{s} < T_{90} < 10\,\mathrm{s}$ is not isotropic. In order to avoid the incompleteness problems, we clipped here the bursts with the peak-flux on the 256 ms trigger $P_{256} < 0.65\, photons/(cm^2\,\mathrm{s})$ (Pendleton et al. , 1997) from our new definition of the intermediate subclass, and plotted the rest of the intermediate class into the celestial sphere as displayed in Figure 7.2. Inspecting it, one may infer that from the 68 GRBs 24 are in the $0 < l < 180$ degree range, while 44 are in the $180 < l < 360$ degree range. Assuming that this difference occurs by chance, i.e. both ranges have equal probability, we get from the binomial test that on a 2 % significance level the assumption of isotropy should be rejected. This anisotropic behaviour is in accordance with the earlier results obtained by by Mészáros et al. (2000a, 2000b) and by Litvin et al. (2001).

Chapter 8

Physical difference between GRBs

8.1 Basic characteristics of GRBs

The simplest grouping of gamma-ray bursts (GRBs), which is still lacking a clear physical interpretation, is given by their well-known bimodal duration distribution. This divides bursts into long ($T \gtrsim 2$ s) and short ($T \lesssim 2$ s) duration groups (Kouveliotou et al. , 1993), defined through some specific duration definition such as T_{90}, T_{50} or similar. The bursts measured with the BATSE instrument on the Compton Gamma-Ray Observatory are usually characterized by 9 observational quantities, i.e. 2 durations, 4 fluences and 3 peak fluxes (Meegan et al., 1996; Paciesas et al. , 1999; Meegan et al., 2001). In a previous paper (Bagoly et al. , 1998) we used the principal components analysis (PCA) technique to show that these 9 quantities can be reduced to only two significant independent variables, or principal components (PCs). These PCs can be interpreted as principal vectors, which are made up of some subset of the original observational quantities. The most important PC is made up essentially by the durations and the fluences, while the second, weaker PC is largely made up of the peak fluxes. This simple observational fact, that the dominant principal component consists mainly of the durations and the fluences, may be of consequence for the physical modelling of the burst mechanism. In this chapter we investigate in greater depth the nature of this principal component decomposition, and, in particular, we analyze quantitatively the relationship between the fluences and durations implied by the first PC. In our previous PCA treatment of the BATSE Catalog (Paciesas et al. , 1999) we used logarithmic variables, since these are useful for dealing with the wide dynamic ranges involved. Since the logarithms of the durations and the fluences can be explained by only one quantity (the first PC), one might suspect the existence of only one physical variable responsible for both of these observed quantities. The PCA assumes a linear relationship between the observed quantities and the PC variables. The fact that the logarithmic durations and fluences can be adequately described by only one PC implies a proportionality between them and, consequently, a power law relation between the observed durations and fluences.

We analyze the distribution of the observed fluences and durations of the long and the short bursts, and we present arguments indicating that the intrinsic durations and fluences are well represented by log-normal distributions. The implied bivariate log-normal distribution represents an ellipsoid in these two variables, whose major axis inclinations are statistically different for the long and the short bursts. An analysis of the possible biases and complications is made, leading to the conclusion that the relationship between the durations and fluences appears to be intrinsic, and may thus be related to the physical properties of the sources themselves. We calculate the exponent in the power-laws for the two types of bursts, and find that for the short bursts the total emitted energy is weakly coupled to the intrinsic duration, while for the long ones the fluences are roughly proportional to the intrinsic durations. The possible implications for GRB models are briefly discussed.

8.2 Analysis of the duration distribution

Our GRB sample is selected from the Current BATSE Gamma-Ray Burst Catalog according to two criteria, namely, that they have both measured T_{90} durations and fluences (for the definition of these quantities see Meegan et al. (2000a), henceforth referred to as the Catalog). The Catalog in its final version lists 2041 bursts for which a value of T_{90} is given. The fluences are given in four different energy channels, F_1, F_2, F_3, F_4, whose energy bands correspond to $[25, 50]\ keV$, $[50, 100]\ keV$, $[100, 300]\ keV$ and $> 300\ keV$. The "total" fluence is defined as $F_{tot} = F_1 + F_2 + F_3 + F_4$. We restrict our sample to include only those GRBs, which have $F_i > 0$ values in the first three channels; i.e. F_1, F_2, F_3 are given.

Concerning the fourth channel, whose energy band is $> 300\ keV$, if we had required $F_4 > 0$ as well, this would have reduced the number of eligible GRBs by $\simeq 20\%$. Hence, we decided to accept also these bursts with $F_4 = 0$, rather than deleting them from the sample. (With this choice we also keep in the sample the no-high-energy (NHE) subgroup defined by Pendleton et al. (1997).) Our choice of $F \equiv F_{tot}$, instead of some other quantity as the main variable, is motivated by two arguments. First, as discussed in Bagoly et al.

(1998), F_{tot} is the main constituent of one of the two PCs which represent the data embodied in the BATSE Catalog, and hence it can be considered as a primary quantity, rather than some other combination or subset of its constituents. Second, Petrosian and collaborators in a series of articles (Efron & Petrosian , 1992; Petrosian & Lee , 1996; Lee & Petrosian, 1996, 1997) have also argued for the use of the fluence as the primary quantity instead of, e.g., the peak flux. Using such defined F_{tot}, from the sample only such GRBs are deleted, which have no measured F_{tot}. Because also the peak fluxes are needed, too, we are left with $N = 1929$ GRBs, all of which have defined T_{90} and F_{tot}, as well as peak fluxes P_{256} on the 256 ms trigger scale. If the peak flux P_{64} on the 64 ms trigger scales is needed, then the sample contains $N = 1972$ GRBs. These are the samples studied in this chapter.

The distribution of the logarithm of the observed T_{90} displays two prominent peaks[1], which is interpreted as reflecting the existence of two groups of GRBs (Kouveliotou et al. , 1993; Norris et al. , 2000). This bimodal distribution can be well fitted by the sum of two Gaussian distributions (Horváth, 1998) indicating that both the long and the short bursts are individually well fitted by pure Gaussian distributions in the logarithmic durations. The fact that the distribution of the BATSE T_{90} quantities within a group is log-normal is of interest, since we can show that this property may be extended to the *intrinsic* durations as well. Let us denote the observed duration of a GRB with T_{90} (which may be subject to cosmological time dilatation), and denote with t_{90} the duration which would be measured by a comoving observer, i.e. the intrinsic duration. One has

$$T_{90} = t_{90} f(z), \tag{8.1}$$

where z is the redshift, and $f(z)$ measures the time dilatation. For the concrete form of $f(z)$ one can take $f(z) = (1 + z)^k$, where $k = 1$ or $k = 0.6$, depending on whether energy stretching is included or not (Fenimore & Bloom , 1995; Mészáros & Mészáros , 1996). If energy stretching is included, for different photon frequencies ν the t_{90} depends on these frequencies as $t_{90}(\nu) = t_{90}(\nu_o)(\nu/\nu_o)^{-0.4} \propto \nu^{-0.4}$, where ν_o is an arbitrary frequency in the measured range (i.e. for higher frequencies the intrinsic duration is shorter). The observed duration at ν is simply $(1+z)$ times the intrinsic duration at $\nu \times (1+z)$. Thus, $T_{90}(\nu) = t_{90}(\nu(1+z))(1+z) = t_{90}(\nu_o)(\nu(1+z)/\nu_o)^{-0.4}(1+z) = t_{90}(\nu)(1+z)^{0.6}$. Hence, when stretching is included, $f(z) = (1+z)^{0.6}$ is used.

---

[1] As we claimed in the previous chapter there are also evidences for the existence of a third intermediate subgroup (Horváth, 1998; Mukherjee et al., 1998; Hakkila et al. , 2000a; Horváth , 2002), which shows a distinct sky angular distribution (Mészáros et al. 2000a,b). We do not deal with this third group here.

Taking the logarithms of both sides of Eq. 8.1 one obtains the logarithmic duration as a sum of two independent stochastic variables. According to a mathematical theorem of Cramér (Cramér , 1937; Rényi , 1962), if a variable ζ - which has a Gaussian distribution - is given by the sum of two independent variables, e.g. $\zeta = \xi + \eta$, then both ξ and η have Gaussian distributions. (In practical cases, however, this holds, of course, only if the variances of ξ and η are comparable. If the variance of, say, ξ is much smaller than the variance of η, then both the variables ζ and η may have a normal distribution - but nothing can be said about the distribution of ξ. It can, but also need not be Gaussian.) Therefore, the Gaussian distribution of $\log T_{90}$ - confirmed for the long and short groups separately (Horváth, 1998) - implies that the same type of distribution exists for the variables $\log t_{90}$ and $\log f(z)$. However, unless the space-time geometry has a very particular structure, the distribution of $\log f(z)$ cannot be Gaussian. This means that the Gaussian nature of the distribution of $\log T_{90}$ must be dominated by the distribution of $\log t_{90}$ *alone*, and therefore the latter must then necessarily have a Gaussian distribution. In other words, the variance of $f(z)$ must be much smaller than the variance of $\log t_{90}$. This must hold for both duration groups separately. This also implies that the cosmological time dilatation should not affect significantly the observed distribution of T_{90}, which therefore is not expected to differ statistically from that of t_{90}. We note that several other authors (Wijers & Paczyński, 1994; Norris et al., 1994; Norris et al. , 1995) have already suggested that the distribution of T_{90} reflects predominantly the distribution of t_{90}. Nevertheless, our argumentation based on the mathematical Cramér theorem is new.

One can check the above statement quantitatively by calculating the standard deviation of $f(z)$, using the available observed redshifts of GRB optical afterglows. The number of the latter is, however, relatively modest, and, in addition, so far they have been obtained only for long bursts. There are currently upwards of 21 GRBs with well-known redshifts (Greiner, 2002). The calculated standard deviation is $\sigma_{\log f(z)} = 0.17$, if we assumed $\log f(z) = \log(1+z)$. Comparing the variance $\sigma^2_{\log f(z)}$ with that of the group of long burst durations which gives $\sigma_{\log T_{90}} = 0.5$, one infers that the variance of $\log f(z)$, or of $\log(1+z)$, can explain maximally only about $(0.17/0.50)^2 \simeq 12\%$ of the total variance of the logarithmic durations. (If $f(z) = (1+z)^{0.6}$, then the variance of $\log f(z)$ can only explain an even smaller amount, because $\sigma_{\log f(z)} = 0.6 \times 0.17$.) This comparison supports the conclusion obtained by applying Cramér's theorem to the long duration group. For the short duration group, since this does not so far have measured redshifts, one can rely only on the theorem itself.

8.2.1 Comparison of T_{90} and T_{50} statistical properties

In order to check, whether there is some influence of the time dilatation on the distribution of T_{90} or T_{50}, we compare here the basic properties of these two quantities in our sample for the long and the short bursts, separately. We grouped the data, using the 256 ms peak flux values, into 0.2 bins in $\log P_{256}$, and summarized in Tables 8.1 and 8.2 the mean values and the corresponding standard deviations of the logarithmic durations of GRBs in each peak flux bin. We stress that this does not include any equalization of the noise level in the various bins, and is not intended as a test of the time dilatation hypothesis, but rather as a test of whether dilatation would have any effect on our results.

Inspecting the durations of long ($T_{90} > 2s$) GRBs summarized in Table 8.1 one sees that, except from the brightest and faintest bins, there is no significant difference in $\log T_{90}$. The decrease of the duration in the faintest bin is probably due to the biasing of the determination, namely, the fainter parts of the bursts cannot be discriminated against the background, and therefore the duration obtained is systematically shorter. There is a remarkable homogeneity and no trend in the standard deviations of the $\log T_{90}$.

In the case of the long burst T_{50} durations, this quantity shows an increasing trend towards the bursts of fainter peak flux. The shortening in the faintest bin is probably also due to selection effects. Similarly to the $\log T_{90}$ values, the same homogeneity can be observed in the standard deviations also in case of $\log T_{50}$. The standard deviations are almost the same in both $\log T_{90}$ and $\log T_{50}$. One can test whether, within our analysis methodology and with our sample, there is a significant difference among the binned T_{90} values, and whether the slight trend in the T_{50} significantly differs from zero. To evaluate the significance of these data we performed a one way analysis of variance with the ANOVA program from a standard SPSS package. The ANOVA compares the variances within sub-samples of the data (in our case within bins), with the variances between the sub-samples (bins).

In the case of $\log T_{90}$ the probability that the difference is accidental is 66%. In the case of the T_{50} durations the same quantities (variances within and between bins) gives a probability of 98.5% for being a real difference between bins, or a probability of 1.5% that there is no difference between the bins. This figure gives some significance for the reality of a trend in the data; however, this value of 0.2 explains less than 1/6 of the variance of T_{50} within one bin. We may conclude that even in this case the variance is mainly intrinsic.

Inspecting the same data in the case of the short duration bursts (Table 8.2) we come to a similar conclusion, i.e. there is no sign of trends in the

Table 8.1: GRBs of long duration ($T_{90} > 2s$).

$\log P_{256}$	$\log T_{90}$	$\log T_{50}$	$\sigma_{\log T_{90}}$	$\sigma_{\log T_{50}}$	No. of GRBs
-.50	1.24	.85	.48	.47	49
-.30	1.42	1.00	.47	.50	230
-.10	1.48	1.08	.49	.53	309
.10	1.46	1.02	.51	.57	272
.30	1.51	1.01	.52	.61	194
.50	1.43	.94	.51	.59	161
.70	1.45	.96	.48	.56	104
.90	1.42	.83	.54	.62	56
1.10	1.41	.83	.50	.49	44
1.30	1.44	.88	.50	.53	34
>1.40	1.21	.68	.41	.50	29

durations of the different bins. Dropping the two faintest bins, which are definitely affected by biases, and dropping the poorly populated brightest bins, we arrive by the analysis of variances with ANOVA to probabilities of 53 % and 92.1 % for the difference being purely accidental between bins in T_{90} and T_{50}, respectively.

8.3 Analysis of the fluence distribution

The observed total fluence F_{tot} can be expressed as

$$F_{tot} = \frac{(1+z)E_{tot}}{4\pi d_l^2(z)} = c(z)E_{tot}. \tag{8.2}$$

Here E_{tot} is the total emitted energy of the GRB at the source in ergs, the total fluence has dimension of erg/cm^2, and $d_l(z)$ is the luminosity distance corresponding to z for which analytical expressions exist in any given Friedmann model (Weinberg , 1972; Peebles , 1993). (We note that the considerations in this chapter are valid for any Friedmann model. Note also that the usual relation between the luminosity and flux is given by a similar equation without the extra $(1 + z)$ term in the numerator. Here this extra term is needed because both the left-hand-side is integrated over the observer-frame time and the right-hand-side is integrated over the time at the source (Mészáros & Mészáros , 1995).)

Assuming as the null hypothesis that the $\log F_{tot}$ of the short bursts has a Gaussian distribution, for the sample of 447 bursts with $T_{90} < 2\,s$, a χ^2 test

Table 8.2: GRBs of short duration ($T_{90} < 2s$).

$\log P_{256}$	$\log T_{90}$	$\log T_{50}$	$\sigma_{\log T_{90}}$	$\sigma_{\log T_{50}}$	No. of GRBs
-.50	-.57	-.87	.55	.60	7
-.30	-.65	-1.01	.53	.57	43
-.10	-.40	-.77	.49	.51	103
.10	-.35	-.74	.35	.32	105
.30	-.33	-.75	.39	.41	75
.50	-.27	-.69	.35	.36	54
.70	-.29	-.72	.36	.34	25
.90	-.35	-.76	.39	.36	22
1.10	-.18	-.72	.44	.39	7
1.30	-.74	-1.21	.31	.43	5
>1.40	-.72	-.90	.00	.00	1

with 26 degrees of freedom gives an excellent fit with $\chi^2 = 20.17$. Accepting the hypothesis of a Gaussian distribution within this group, one can apply again Cramer's theorem similarly to what was done for the logarithm of durations. This leads to the conclusion that either *both* the distribution of $\log c(z)$ and the distribution of $\log E_{tot}$ are Gaussian ones, or else the variance of one of these quantities is negligible compared to the other, which then must be mainly responsible for the Gaussian behavior. Because $\log c(z)$ hardly can have a log-normal distribution, the second possibility seems to be the situation. In any case, one may conclude that the intrinsic fluence (i.e. the total emitted energy) should be distributed log-normally.

In the case of the long bursts, a fit to a Gaussian distribution of logarithmic fluences does not give a significance level, which is as convincing as for the short duration group. For the 1482 GRBs with $T_{90} > 2\,s$ a χ^2 test on $\log F_{tot}$ with 22 degrees of freedom gives a fit with $\chi^2 = 35.12$. Therefore, in this case the χ^2 test casts some doubt on normality but only with a relatively high error probability of 3.5% for rejecting a Gaussian distribution (Trumpler & Weaver, 1953; Kendall & Stuart, 1973; Press et al. , 1992). This circumstance prevent us from applying Cramér's theorem directly in the same way as we did with the short duration group. Calculating the variance of $\log c(z)$ for the GRBs with known 21 redshifts (Greiner, 2002) one obtains $\sigma_{\log c(z)} = 0.43$. For the GRBs of long duration, however, one obtains $\sigma_{\log F_{tot}} = 0.66$. The ratio of these variances equals $(0.43/0.66)^2 \simeq 42\%$, i.e. more than half of the variance of F_{tot} is not explained by the variance of $c(z)$. (If one takes into account the energy stretching even a larger fraction remains unexplained). In other words, a significant fraction of the total variance of F_{tot} has to be intrinsic. It is worth mentioning that the unexplained part of the variance of F_{tot} corresponds nicely to the value obtained in Section 8.4 making use the EM algorythm.

Despite these difficulties, there is a substantial reason to argue that the intrinsic distribution of total emitted energies is distributed log-normally for the long subgroup, too. There are also other authors arriving to the same conclusion (McBreen et al., 2002, Quilligan et al, 2002). This point of view may be supported by the following argument.

The Gaussian behaviour of $\log c(z)$ can almost certainly be excluded. One can do this on the basis of the current observed distribution of redshifts (Greiner, 2002), or on the basis of fits of the number vs. peak flux distributions (Fenimore & Bloom , 1995; Ulmer & Wijers , 1995; Horváth et al., 1996; Reichart & Mészáros , 1997). In such fits, using a number density $n(z) \propto (1+z)^D$ with $D \simeq (3-5)$, one finds no evidence for the stopping of this increase with increasing z (up to $z \simeq (5-20)$). Hence, it would be contrived to deduce from this result that the distribution of $\log c(z)$ is normal. In order to do this, one would need several ad hoc assumptions. First, the increasing of number density would need to stop around some unknown high z. This was studied (Mészáros & Mészáros , 1995; Horváth et al., 1996; Mészáros & Mészáros , 1996), and no such effect was found. (For the sake of preciseness it must be added here that these fits were done for the whole sample of GRBs. But, because GRBs are dominated by the long ones, conclusions from these fits should hold for the long subgroup, too.) Second, even if this were the case, above this z the decrease of $n(z)$ should mimic the behavior of a log-normal distribution for $c(z)$, without any obvious justification. Third, below this z one must again have a log-normal behavior for $c(z)$, in contradiction with the various number vs. peak flux fits. Fourth, this behavior should occur for any subclass separately. Hence, the assumption of log-normal distribution of $c(z)$ appears highly improbable.

Having a highly improbable log-normal distribution of $\log c(z)$, which variance is surely not negligible, a 3.5% error probability (i.e the probability that we reject the hypothesis of normality but it is still true) from the goodness-of-fit is still remarkable. One may argue that, if the distribution of the total emitted energy were not distributed log-normally, then the two non-normal distributions together would give a fully wrong χ^2 fit for $\log F_{tot}$; under this condition even the 3.5% probability would not be reachable. Of course, this argumentation is more or less heuristic, and - as the conclusion - one cannot say that the log-normal distribution of E_{tot} is confirmed similarly unambiguously in both subgroups. In the case of long subgroup questions still remain, and they will still be discussed.

In addition, even in the case of short GRBs the situation is not so clear yet. The argument based on the Cramér theorem for the short GRBs should also be taken with some caution. As shown in Bagoly et al. (1998), the stochastic variable corresponding to the duration is independent from that of the peak flux. This means that a fixed level of detection, given by the peak fluxes, does not have significant influence on the shape of the detected distribution of the durations (Efron & Petrosian , 1992; Wijers & Paczyński, 1994; Norris et al., 1994; Norris et al. , 1995; Petrosian & Lee , 1996; Lee & Petrosian, 1996, 1997). In the case of the fluences, however, a detection threshold in the peak fluxes induces a bias on the true distribution, since fluences and durations are stochastically not independent. Therefore, the log-normal distribution recognized from the data does not necessarily imply the same behaviour for the true distribution of fluences occurring at the detector. In other words, observational biases may have important roles; in addition, for both subgroups. A discussion of these problems can be found in a series of papers published by Petrosian and collaborators (Efron & Petrosian , 1992; Petrosian & Lee , 1996; Lee & Petrosian, 1996, 1997; Lloyd & Petrosian , 1999). In what follows, we also will study the biases together with the fitting procedures.

8.4 Correlation between the fluence and duration

In the previous Sections we presented firm evidences that the observed distribution of the durations is basically intrinsic. We argued furthermore that a significant fraction of the variance of the fluences is also intrinsic. We proceed a step further in this Section and try to demonstrate that there is a relationship between the duration and the fluence which is also intrinsic. There are two basic difficulties in searching the concrete form of this relationship (if there is any at all): first, we observe only those bursts which fulfill some triggering criteria and, second, the observed quantities are suffering from some type of bias depending on the process of detection. Several papers discuss these biases (Efron & Petrosian , 1992; Lamb et al. , 1993; Lee & Petrosian, 1996; Petrosian & Lee , 1996; Lee & Petrosian, 1997; Stern et al. , 1999; Paciesas et al. , 1999; Hakkila et al. , 2000b; Meegan et al, 2000b). In the following we will address these issues in a new way.

The detection proceeds on three time scales: the input signal is analyzed on 64, 256 and 1024 *ms* resolution. The counts in these bins of these scales are compared with the corresponding 17 second long averaged value. There are eight detectors around the BATSE instrument. If at least one of the three peak intensities in the second brightest detector exceeds 5.5 sigma of the threshold computed from the averaged signal the burst will be detected. In case of the bursts of long duration (at least several seconds) the differences in the time scales of detection do not play an important role since the vast majority of the events were triggered on the 1024 ms scale and the detection proceeded if the peak exceeded the threshold on this time scale. In contrast, at the bursts of short duration - when T_{90} could be much shorter then the time scale of the detection - the situation could be drastically changed. Looking at the data of the BATSE the bursts of duration of $T_{90} < 2s$ are mixtures of those triggered on different time scales.

Among bursts triggered on the same time scale the detection proceeds when the corresponding peak flux exceeds the threshold. In the case of bursts, which are shorter than their triggering time scale, the corresponding peak fluxes are given by the fluence itself. This has the consequence that the threshold in the peak flux means the same for the fluence, i.e. it results a horizontal cut on the fluence - duration plane and a bias in the relationship between these quantities. In order to minimize this effect we will use the peak flux on the 64 ms time scale in our further analysis. The BATSE had a spectral response on the detected γ radiation. It had the consequence that different measured values were assigned to bursts having the same total energy at the entrance of the detector if the incoming photons had different spectral distributions

The duration of a GRB is only a lower limit for its intrinsic value since a certain fraction of the burst can be buried in the background noise. Therefore any relationship recognized among the observed fluence and duration is not necessarily representative for those between the corresponding intrinsic quantities. In the next paragraphs we address these issues in more details.

8.4.1 Effect of the detection threshold on the joint probability distribution of the fluence and duration

In the following we will study the effect of the detection threshold on the joint probability distribution of the observed fluence and duration. In order to put this effect into a quantitative basis we use the law of full probabilities (see e.g. Rényi (1962)). Let $P(F_{tot}, T_{90})$ be the joint probability density of the fluence and duration. Using this theorem any of the probability densities on the right side can be written in the form of

$$P(F_{tot}, T_{90}) = \int\limits_0^\infty P(F_{tot}, T_{90}|p)G(p)dp, \tag{8.3}$$

where p is the peak flux at any of the 64 ms, 256 ms and 1024 ms time scales, $P(F_{tot}, T_{90}|p)$ is the joint (bivariate) probability density of the fluence and duration (assuming that p is given), and $G(p)$ is the probability density of p. This means that, if there are N bursts in the sample, then $NP(F_{tot}, T_{90}) d\log F_{tot} d\log T_{90}$ is the expected number of observed GRBs in the infinitesimal intervals $[\log F_{tot}, (\log F_{tot} + d\log F_{tot})]$ and $[\log T_{90}, (\log T_{90} + d\log T_{90})]$, respectively. Among the bursts triggered on a given time scale $G(p)$ represents an unbiased function above p_{th}, the peak flux corresponding to the detection threshold. Below this limit, however, $G(p)$ is biased by the process of detection. It inserts also a bias on the joint probability density of the observed fluence and duration. Nevertheless, the kernel $P(F_{tot}, T_{90}|p)$ represents some intrinsic relationship between these two quantities, and it is free from the bias of $G(p)$. Following our discussion given above, we use in the following the peak fluxes of the 64 ms time scale.

8.4.2 Intrinsic relationship between the fluence and the duration

We demonstrated in an earlier paper (Bagoly et al. 1998) that the logarithms of the peak flux and the duration represent two independent stochastic variables and the logarithmic fluence can be well approximated as the linear combination of these variables:

$$\log F_{tot} = a_1 \log T_{90} + a_2 \log p + \epsilon, \tag{8.4}$$

where a_1, a_2 are constants, and ϵ is a noise term (later on we will see that a_1 may depend on the duration, i.e. it is different for the bursts of short and long duration). One may confirm this statement by inspecting the Tables 8.1, 8.2. They demonstrates convincingly that, independently of the choice of the peak flux, the standard deviation and the mean value of the duration is not changed significantly. This expression reveals that - fixing the peak intensity - the distribution of the fluences reflects basically the distribution of the durations. Since the probability density of the durations is a superposition of two Gaussian distributions, the same should hold also for the fluences. Consequently, we may assume that the joint conditional probability distribution of the fluence and duration consists of a superposition of two two-dimensional Gaussian distributions. One such distribution takes the form

$$
\begin{aligned}
f(x,y) = {} & \tfrac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \\
& \times \exp\left[-\tfrac{1}{2(1-r^2)}\left(\tfrac{(x-a_x)^2}{\sigma_x^2}+\tfrac{(y-a_y)^2}{\sigma_y^2}-\tfrac{2r(x-a_x)(y-a_y)}{\sigma_x\sigma_y}\right)\right] , \qquad (8.5)
\end{aligned}
$$

where $x = \log T_{90}$, $y = \log F_{tot}$ a_x, a_y are the means, σ_x, σ_y are the dispersions, and r is the correlation coefficient (Trumpler & Weaver (1953); Chapt. 1.25). In our case one needs a weighted sum of two such bivariate distributions. This means that 11 free parameters should be determined (two times 5 parameters for the both distributions; the 11th independent parameter is the weight of the, say, first subgroup). This also means that two r correlation coefficients should be obtained, which may be different for the two subgroups.

The parameters a_x, σ_x, characterizing the distribution of the duration do not depend on the peak flux, because T_{90} and p are independent stochastic variables. In the case, when the r-correlation coefficient differs from zero, the semi-major axis of the dispersion ellipse represents a linear relationship between $\log T_{90}$ and $\log F_{tot}$, with a slope of $m = \tan\alpha$, where

$$\tan 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2} \ . \qquad (8.6)$$

This linear relationship between the logarithmic variables implies a power-law relation of form $F_{tot} = (T_{90})^m$ between the fluence and the duration, where m may be different for the two groups. Replacing the $G(p)$ probability density by the empirical distribution of the measured peak fluxes, one may write the joint probability density of the fluence and duration in the form of

$$
\begin{aligned}
P(F_{tot},T_{90}) = {} & \int\limits_0^\infty P(F_{tot},T_{90}|p)G(p)dp \\
& \simeq \sum_{i=1}^{N} P(F_{tot},T_{90}|p_i) \simeq \sum_{l=1}^{k} b_l P(F_{tot},T_{90}|p_l), \qquad (8.7)
\end{aligned}
$$

i.e. the integral is approximated by a sum of k separate terms (bins), in which b_l is the number of GRBs at the given bin.

The k is the number of bins at the right-hand-side, and is somewhat arbitrary. Trivially, bigger k leads to a better approximation of the integral. On the other hand, bigger k leads to the situation, when in one single bin the number of GRBs b_l is smaller. Hence, k should be small in order to get enough number of GRBs in each bin for making statistics, but not too small in order to have good approximation of the integral.

8.4.3 Maximum Likelihood estimation of the parameters via EM algorithm

One finds in the Tables 8.1, 8.2 the computed mean values and standard deviations of the logarithmic durations for the short and long bursts, respectively. These Tables clearly suggest that, except for the faintest bins where we expect serious biases in the duration and fluence due to the detection close to the background, the standard deviations do not differ significantly between the bins. Dividing the sample into short and long bursts by the cut of $T_{90} < 2s$ and $T_{90} > 2s$, we may assume that these subsamples are dominated by only one Gaussian distribution and we may compute its parameters in a simple way as given below.

If the $P(F_{tot}, T_{90}|p)$ conditional probability density is a pure Gaussian one, then the Maximum Likelihood estimation of its parameters would be very simple, because they can be obtained by computing the mean values, standard deviations and the correlation between the fluence and duration. In the reality, however, this probability density is a superposition of two Gaussians, and the simple cut at $T_{90} = 2s$ is hardly satisfactory. The proper way to estimate the free parameters is not so simple. For this reason, similarly to Sec. 7.3 we will use a procedure, called EM algorithm (**E**xpectation and **M**aximalization), which terminates at the Maximum Likelihood solution (Dempster et al. 1977).

If we knew, which of the bursts belong to the short and long duration groups, we may add a $\{i_1, i_2\}$ two dimensional indicator variable to each GRB having the value of $\{1, 0\}$ in the case if a burst was short, and similarly $\{0, 1\}$ if it was long. The sample means of T_{90} weighted with i_1 would give the ML estimation of a_x of the first Gaussian distribution (i.e. $a_x = \sum_{j=1}^{N} i_{1j} x_j / \sum_{j=1}^{n} i_{1j}$). The same hold for the other parameters. Weighting with i_2 would give the parameters of the second Gaussian distribution. Hence 10 parameters of the two distributions would be well calculable. The 11th parameter would also be trivially calculable, because the fraction of first subgroup should simply be $\sum_{i=1}^{N} i_1/N$. Hence, if the values of the $\{i_1, i_2\}$ indicator variable were known, the ML parameters would be well calculable.

If the parameters of the two Gaussians were given, one could compute the $\{p_1, p_2\}$ membership probabilities of a burst to each of the two groups. Replacing the indicator variable by these probabilities one may calculate new parameters in the same way as was done assuming $\{i_1, i_2\}$ were given. Then one may again calculate new $\{i_1, i_2\}$, and again the new parameters. This iteration is exactly the procedure, what EM algorithm is doing. One gives an initial estimate for the parameters of the two Gaussian distributions. Then one estimates the membership probabilities (E step). Weighting with the

Table 8.3: Number of GRBs within the 0.2 wide strata of the logarithmic 64 *ms* peak fluxes.

Serial No.	$\log P_{64}$	total No. of GRBs	No. of GRBs with $T_{90} < 2s$	No. of GRBs with $T_{90} < 0.064s$
1.	-0.6 - -0.4	5	1	0
2.	-0.4 - -0.2	113	5	0
3.	-0.2 - 0.0	385	44	1
4.	0.0 - 0.2	434	104	4
5.	0.2 - 0.4	365	126	8
6.	0.4 - 0.6	254	79	3
7.	0.6 - 0.8	166	47	2
8.	0.8 - 1.0	95	34	0
9.	1.0 - 1.2	74	22	0
10.	1.2 - 1.4	39	6	0
11.	1.4 - 1.6	19	5	0
12.	1.6 - 1.8	15	2	0
13.	1.8 - 2.0	6	1	0
14.	2.0 <	2	0	0

membership probabilities one obtains the new ML estimation of the parameters (M step). Repeating these steps successively one proceeds to the ML solution of the parameter estimation (Dempster et al. 1977).

In order to fit the $[\log T_{90}, \log F_{tot}]$ data pairs with the superposition of two two-dimensional Gaussian bivariate distributions we splitted the Catalog into subsamples with respect to 64 *ms* peak fluxes. The strata were obtained by taking 0.2 wide strips in the logarithmic peak fluxes. Table 8.3 summarizes the number of GRBs within the strata. In addition, also the number of GRBs with $T_{90} < 2s$ and with $T_{90} < 0.064s$ are given there. The first one shows that, roughly, which fraction of GRBs belong to the short subgroup in the given strata, and the second one shows which fraction is maximally biased.

In the fitting procedure we omitted bins No. 1.–3., being affected by selection bias, and also No. 9.–14., being scarcely populated. We performed the ML fitting in the bins No. 4.–8., making use the EM algorithm. Table 8.4 summarizes the results of the ML fitting for the short GRBs, and Table 8.5 for the long GRBs, respectively. On Figures 8.1-8.5 the results of fitting for bins No. 4.–8. are shown. The ellipses define the 1-sigma and 2-sigma regions, respectively.

The slopes of short GRBs obtained for the bins No.7. and No.8. differ remarkably from those in bins No. 4.–6. They are based, however, on a small

Table 8.4: Results of the ML fitting for the short GRBs using the EM algorithm. Weighted mean for m is $m = 0.81 \pm 0.06$.

Strip	freq.	a_x	a_y	σ_x	σ_y	r	No. GRBs	$m = tan\alpha$
4.	.293	-.199	-6.587	.549	.502	.593	434	0.86
5.	.418	-.275	-6.488	.575	.503	.591	365	0.80
6.	.321	-.365	-6.244	.486	.497	.515	254	1.04
7.	.332	-.188	-5.921	.510	.420	.342	166	0.58
8.	.358	-.325	-5.910	.440	.347	.279	95	0.46

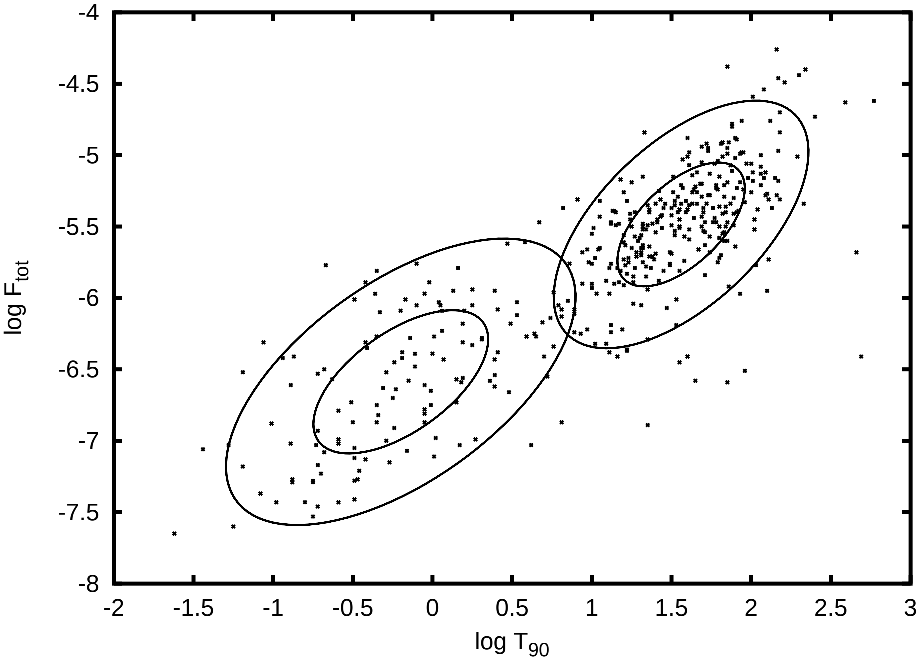

Figure 8.1: The best ML fits of the two log-Gaussian distributions for the faintest sample No.4 with $N = 434$.

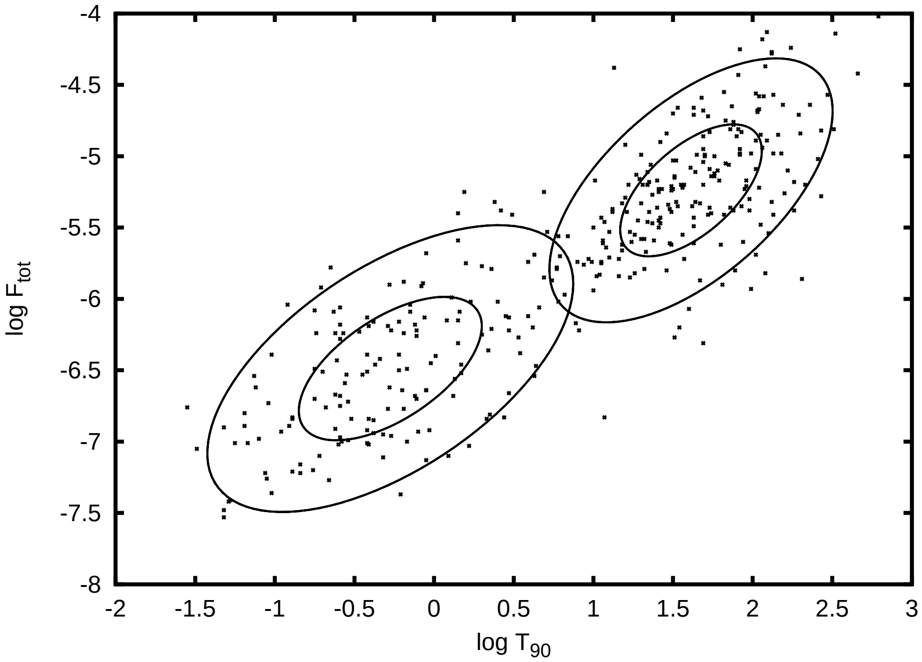

Figure 8.2: The best ML fits of the two log-Gaussian distributions for the sample No.5 with $N = 365$.

number of bursts; hence, the r parameter is highly uncertain. Since we used for weighting the number of GRBs within the given bin their, contribution to the final result is marginal. We noted above that the duration and peak flux are independent stochastic variables. Since the sample was splitted into subsamples by the peak flux, this means that the parameters of Gaussian distributions referring to T_{90} either in the Table 8.4 or in Table 8.5 should be identical within the statistical uncertainty of estimation.

Inspecting a_x in these Tables - which summarize the results of the EM algorithm - clearly demonstrates that their difference is much less than σ_x. It is also possible to compare the mean slopes obtained by weighting the results for the short and long GRBs, respectively, in order to test the significance of the difference between these groups. One may compute a $\chi^2 = (m_1 - m)^2/\sigma_1^2 + (m_2 - m)^2/\sigma_2^2$ variable based on the assumption that the m_1, m_2 slopes of the short and long GRBs differs from the m weighted mean only by chance. Making this assumption one obtains $\chi^2 = 22.2$ indicating that the null hypothesis, i.e. $m_1 = m_2$, should be rejected on a 4.7σ significance level. *The two slopes are different.*

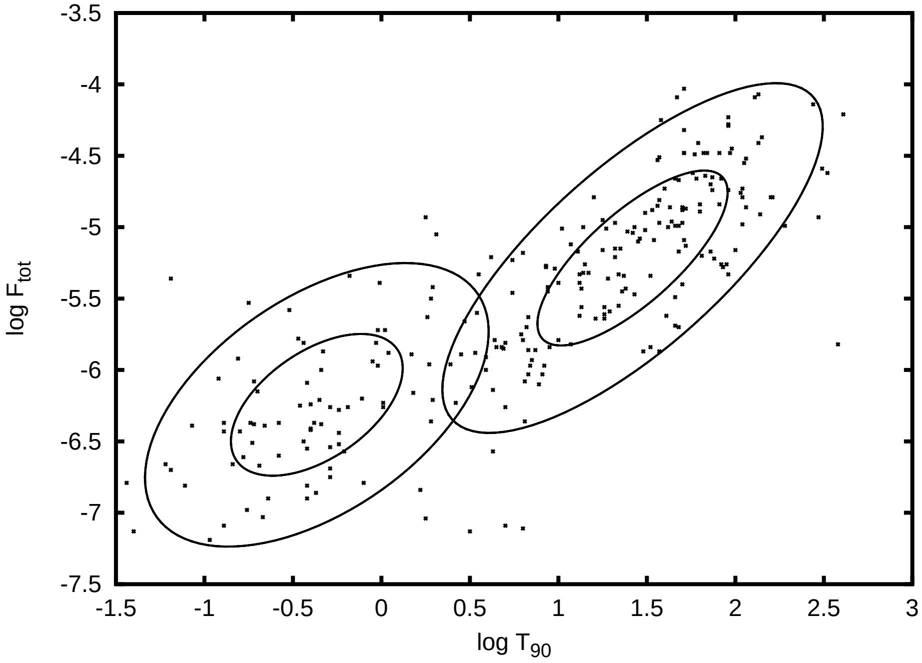

Figure 8.3: The best ML fits of the two log-Gaussian distributions for the sample No.6 with $N = 254$.

Table 8.5: Results of the ML fitting for the long GRBs using the EM algorithm. Weighted mean for m is $m = 1.11 \pm 0.03$.

Strip	freq.	a_x	a_y	σ_x	σ_y	r	No. GRBs	$m = tan\alpha$
4.	.707	1.560	-5.485	.400	.434	.586	434	1.15
5.	.582	1.613	-5.239	.445	.463	.599	365	1.07
6.	.679	1.419	-5.216	.538	.613	.753	254	1.19
7.	.668	1.468	-4.894	.448	.459	.610	166	1.04
8.	.642	1.391	-4.779	.541	.531	.656	95	0.97

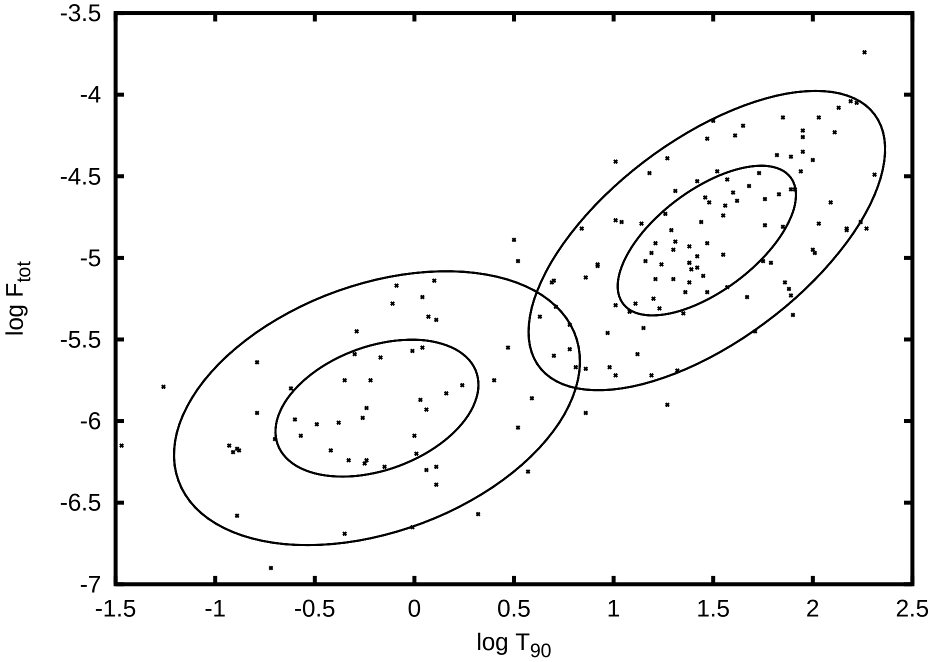

Figure 8.4: The best ML fits of the two log-Gaussian distributions for the sample No.7 with $N = 166$.

8.4.4 Possible sources of the biases

The relationships derived in the previous Subsection refer to the observed values of GRBs. There is a dilemma, however, how representative they are for the true quantities of GRBs not affected by the process of detection. We mentioned already several major source of bias. Here we summarize them again:

- Some GRBs below the threshold remain undetected. Therefore, the stochastic properties of the observed part of the true joint distribution of $\{\log T_{90},\ \log F_{tot}\}$ are not necessarily relevant for the whole population.

- Observed duration refers to the detected part of GRBs. The real duration might be much longer.

- There is a similar bias also for the fluence.

- Additionally, due to the limited spectral response of BATSE, a significant fraction of the high energy part of the fluence may remain unobserved.

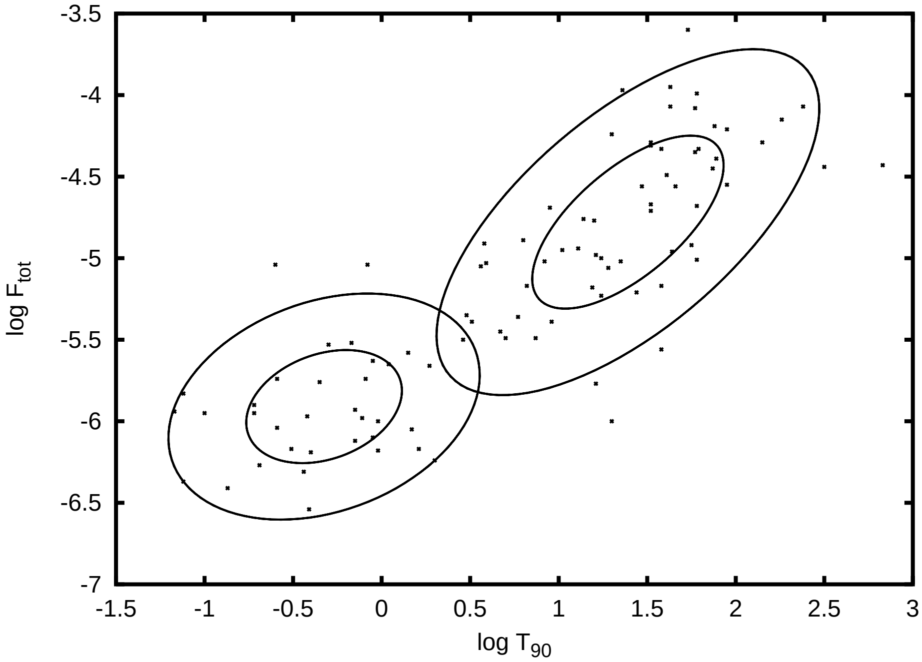

Figure 8.5: The best ML fits of the two log-Gaussian distributions for the brightest sample No.8 with $N = 95$.

- There is a special bias at short GRBs. At GRB, where the duration is shorter than the time resolution of detection, there is a one-to-one correspondence between the peak flux and the fluence.

Effect of the threshold

Using the law of full probability we decomposed the observed joint probability distribution of $\{\log T_{90},\ \log F_{tot}\}$ into the distribution of the peak flux and a conditional probability, assuming p is given. Since the detection proceeds on three different time scales, one does not expect a sharp cut on $G(p)$ but the distortion is more complicated in the reality. Although the observational threshold may seriously affect the detected form of $G(p)$, it need not necessarily modify $P(F_{tot}, T_{90}|p)$. The detection threshold, however, may distort also the fluence and duration themselves, and in this way also the form of $P(F_{tot}, T_{90}|p)$.

True vs. observed duration

Depending on the light curve of the GRBs a significant fraction of the outbursts may remain unobserved. So the duration derived from the observed part is only a lower limit for the true one. Approaching the detection threshold this effect should become more and more serious. Assuming a Gaussian form for $P(F_{tot}, T_{90}|p)$ one expect a systematic change in the parameters as one is approaching the threshold. Inspecting the mean values and standard deviations of the duration in the Tables 8.1, 8.2, one may really recognize this effect in the three faintest bins. In the remaining part of the sample, however, there is a remarkable homogeneity in the mean value and standard deviation of duration. It is also worth mentioning that the same is true in Table 8.4 and 8.5 summarizing the result of the ML fitting. So one may conclude that this bias does not play a significant influence in the 4.–8. bins used for our calculations.

True vs. observed fluence

Similarly to the duration also the observed fluence might be a lower bound depending on the light curve of the burst. Although fixing p resulted in a similar functional (Gaussian) form of the fluence as of the duration, its mean value a_y differs from bin to bin due to the dependence of F_{tot} on the peak flux. Its standard deviation σ_y, however, shows a noticeable homogeneity within the limits of statistical uncertainty. Again, this implies a constancy in the functional form of $P(F_{tot}, T_{90}|p)$ in the bins studied. The only exception is perhaps the bin No. 8 for the short GRBs, where the standard deviation and the r correlations coefficient seems to depart considerably from the others in Table 8.4 One may test the significance of the excursion of σ_y in bin No. 8 by performing a F test (see e.g. Kendall & Stuart (1973)). Computing the $F = \sigma_8^2/\sigma_5^2$ value, where the indexes refer to the serial number of bins, one obtains $F = 2.11$ indicating significant difference on the 99.9 % level. Except for this significant excursion in the 8th bin, the σ_y values are statistically identical implying that the functional relationship between F_{tot} and T_{90} is not significantly influenced by the process of detection in the bins studied.

Bias from the spectral response

BATSE were observing in four energy channels. Even the highest energy channel was not able to detect the hardest parts of the bursts. A significant fraction of the incoming energy might remain unobserved. In principle, there is a possibility for estimating the amount of unobserved part of radiation by supposing a spectral model for the GRB. Fitting this model to the values measured in the four energy channels one may get an estimate for the unobserved part. Supposing, the energy distribution of the bursts can be described by two power laws separated by an E_p energy Lloyd & Petrosian

(1999) did a four parameter fit (two powers, E_p and an amplitude) for GRBs detected by BATSE. A basic trouble at this approach appears in the fact that numbers of points and parameters to be fitted are identical and, consequently, any uncertainty in the measured values has a very sensitive impact on the parameters estimated. Moreover, a significant fraction of GRBs does not have a reliable fluence in the high energy channel which exceeds at least the 3σ level of the background. In particular, it is true for the No. 4.–8. bins.

It is well-known that the short bursts are harder in the average than the long ones. Consequently, the fraction of the unobserved part of the energy spectrum may have a negative correlation with the duration in the case of this subgroup. The detected part of the fluence experiences therefore a positive correlation, assuming there is no intrinsic relationship between the duration and the true total fluence. In the case of a real intrinsic relationship between these quantities, the apparent correlation from the spectral bias may have a contribution to the real one. One may expect that the spectral bias is more serious at bursts, where the whole high energy fluence is buried into the background noise. So one expect a gradual change in the slope of the relationship between F_{tot}, and T_{90} as one proceeds from the faint bursts to the brighter ones. Table 8.6 and Table 8.7 summarize the frequency of bursts having different S/N ("signal-to-noise") ratios within the studied peak flux bins, separately.

It is clear from Table 8.7 that the long faint bins are dominated by bursts with no significant high energy fluence. The contrary is true for the brighter ones. Proceeding from the faint burst to the bright ones one does not see a gradual change in the slope of the $\{\log F_{tot}, \log T_{90}\}$ relationship. Hence, we may conclude that the spectral bias makes only a marginal contribution, and the correlation observed is close to the real one. For the short bursts (Table 8.6), in the contrary, a significant change is observed, which might be interpreted as a clear sign of spectral bias. It implies, furthermore, that the real slope, if any, is smaller than the observed one. This fact strengthens the conclusion on the difference between the short and long GRBs with respect of the $\{\log F_{tot}, \log T_{90}\}$ relationship. .

Bias from the finite time resolution

We mentioned above the detection proceeded on three ($64\ ms, 256\ ms$ and $1024\ ms$) time scales. The incoming photons were binned in these time scales and the bin having the maximum count rate were used for triggering the detection. The bursts having $T_{90} < 64\ ms$, however, consist of only one bin, consequently, the fluence and the peak flux are based on the same

Table 8.6: Frequency of S/N ratios of fluences in the high energy channel for GRBs of $T_{90} < 2s$ (the integer numbers given in the header are the truncated S/N values).

Bin	S/N				Row
	.00	1.00	2.00	> 3.00	Total
1.	0	1	0	0	1
2.	3	0	1	1	5
3.	19	10	7	8	44
4.	57	15	14	18	104
5.	53	28	14	31	126
6.	17	16	17	29	79
7.	4	3	4	36	47
8.	2	4	4	24	34
9.	0	0	3	19	22
10.	1	0	1	4	6
11.	0	0	0	5	5
12.	0	0	0	2	2
13.	0	0	0	1	1
Column	156	77	65	178	476
Total	32.8	16.2	13.7	37.4	100.0%

Table 8.7: Frequency of S/N ratios of fluences in the high energy channel for GRBs of $T_{90} > 2s$ (the integer numbers given in the header have the same meaning as in Table 8.6)

Bin	S/N				Row
	.00	1.00	2.00	> 3.00	Total
1.	2	1	0	1	4
2.	71	19	7	11	108
3.	193	58	37	53	341
4.	135	65	42	88	330
5.	72	33	40	94	239
6.	33	23	16	103	175
7.	12	9	10	88	119
8.	1	2	7	51	61
9.	1	1	1	49	52
10.	0	0	0	33	33
11.	0	0	0	14	14
12.	0	0	0	13	13
13.	0	0	0	5	5
14.	0	0	0	2	2
Column	520	211	160	605	1496
Total	34.8	14.1	10.7	40.4	100.0 %

incoming photons on this time scale. If the incoming photons of a burst had the same energy fixing p would mean fixing F_{tot} as well and equation 4 is no longer valid since T_{90} does not have any impact on the fluence observed. By fixing p this effect degenerate the distribution of F_{tot} into one point and it does no longer reflects the distribution of T_{90} we supposed. In the reality, however, the energies of the incoming photons have a wide range and this effect is not so pronounced.

As the duration covers an increasing number of bins of 64 ms the particular bin representing the peak flux has a decreasing impact on the value of the fluence. In Table 8.8 we gave some stochastic parameters (mean, standard deviation, correlation) of the joint distribution of $\log F_{tot}$ and $\log p_{64}$ within the first 10 bins of T_{90} of 64 ms, in order to see the possible quantitative differences. Except the mean value of $\log F_{tot}$ in the first bin, which deviate from the sample value at about 1σ level there is no striking differences between the parameters. For testing the possible differences between the bins in Table 8.8 we did a multivariate analysis of variance (MANOVA) which compares the variances and covariances of variables within the bins and between them. The analysis resulted in a difference on the 99.5 % significance level. The MANOVA module of the SPSS software package was used for these calculations [2]. Repeating the calculation but abandoning the first bin, the suspected outlier, the significance dropped back to 50.4 % inferring that the distributions in bins 2.–10. were identical within the limits of statistical uncertainty. Even if we treated the excursion of the bin No. 1 as a real effect there is only a small number of GRBs in it (see Table 8.3) which do not affect the final results in Table 8.4 and 8.5.

Summing up the discussions we performed in this subsection on the different bias we may conclude that either they do not have a significant impact on the final result (i.e. there is a significant difference in the $\{\log F_{tot}, \log T_{90}\}$ correlation between the short and long GRBs) or the observed difference in the relationship is even enhanced in the reality if we considered the bias properly.

8.5 Discussion

We have presented evidence indicating that there is a power-law relationship between the logarithmic fluences and the logarithmic T_{90} durations of the GRBs in the Current BATSE Catalog, based on the EM maximum likelihood estimation of the parameters of the bivariate distribution of these measured quantities. This relationship holds for both subclasses of GRBs

---

[2] SPSS is a registered trademark. See SPSS home page ( ) in references

Table 8.8: Mean values and standard deviations of the total fluences and the 64 *ms* peak fluxes within the first ten 64 *ms* bin of the T_{90} duration. Except the fluence in the first bin all the values do not differ from those of the entire sample, within the limits of statistical uncertainties.

	$\log F_{tot}$		$\log p_{64}$			
Bin	mean	st. dev.	mean	st. dev.	corr. coeff.	no. of GRBs
1.	-7.0243	.5043	.3535	.1912	.7625	17
2.	-6.6280	.5109	.3815	.2447	.6333	33
3.	-6.6756	.5122	.3690	.2379	.6335	37
4.	-6.4863	.5408	.4090	.2629	.5793	36
5.	-6.5480	.4428	.3691	.2460	.7125	26
6.	-6.5804	.5637	.3482	.2370	.6117	16
7.	-6.4492	.3823	.4191	.2278	.2241	31
8.	-6.3312	.4756	.4278	.2802	.6868	20
9.	-6.4532	.4517	.3348	.2382	.6780	16
10.	-6.4292	.3826	.3682	.2258	.5620	16
entire sample	6.5599	.5015	.3828	.2395	.5822	248

separately. As shown in Tab. 8.1 and 8.2, the dispersions of the T_{90} do not differ significantly from those of the T_{50} distributions, and therefore the same correlations and the same power-law relations would be expected if one used the T_{50} instead of the T_{90}. We have also evaluated the possible impact of instrumental biases, with the results that the conclusions do not change significantly when these effects are taken into account.

An intriguing corollary of these results is that the exponents in the power-law dependence between the fluence and the duration differs significantly for the two groups of short ($T_{90} < 2$ s) and long ($T_{90} > 2$ s) bursts at a 4.7σ level. As shown in Sec. 8.4.4, this also means that the same power law relations hold between the total energy emitted (E_{tot}) and the intrinsic durations (t_{90}) of the two groups.

While an understanding of such power-law relations in terms of physical models of GRB would require more elaborate considerations, we note that there is substantial evidence indicating the two classes of bursts are physically different. First, there is the fact that short burst are harder (Kouveliotou et al., 1993); this is confirmed also by the analysis of Mukerjee et al. (1998). Then, there is evidence that the spectral break energies of short bursts are larger than for long bursts (Paciesas et al., 2001). The short bursts have a different spectral lag vs. luminosity ratios than log bursts (Norris, Scargle

& Bonnell, 2001). Finally, the number of sub-pulses, and the soft-to-hard evolution is different depending on the duration (Gupta et al. , 2002).

The results obtained here are compatible with a simple interpretation where the bursts involve a wind outflow leading to internal shocks responsible for the gamma-rays (Rees & Mészáros , 1994; Piran , 1999), in which the luminosity is approximately constant over the duration t of the outflow, so that both the total energy E_{tot} and the fluence F_{tot} are $\propto t$. If an external shock were involved, e.g. Mészáros & Rees (1993); Piran (1999), for a sufficiently short intrinsic duration (impulsive approximation) there would be a simple relationship between the observed duration and the total energy, $t \propto E^{1/3}$, resulting from the self-similar behavior of the explosion and the time delay of the pulse arrival from over the width of the blast wave from across the light cone. This relationship is steeper than the one we deduced for long bursts.

The fluence – duration relation of GRBs which we have discussed here appears to be physical, and it is significantly different for the short and the long bursts. For the short ones, the total energy released is proportional to the $m = 0.81$ power of duration of the gamma ray emission, while for the long ones it is proportional roughly to the of $m = 1.11$ power of the duration. This may indicate that two different types of central engines are at work, or perhaps two different types of progenitor systems are involved. It is often argued that those bursts for which X-ray, optical and radio afterglows have been found, all of which belong to the long-duration group, may be due to the collapse of a massive stellar progenitor (Paczyński , 1998; Fryer et al., 1999). The short bursts, none of which have as of August 2002 yielded afterglows, may be hypothetically associated with neutron star mergers (Fryer et al., 1999) or perhaps other systems. While the nature of the progenitors remains so far indeterminate, our results provide new evidence suggesting an intrinsic difference between the long and short bursts, which probably reflects a difference in the physical character of the energy release process. This result is completely model-independent, and if confirmed, it would provide a potentially useful constraint on the types of models used to describe the two groups of bursts.

Part IV

Summary and theses

Summary

1. Mathematical introduction

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 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.

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

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, reveal 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.

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

Classification of stellar spectra I made medium resolution ($100\AA/mm$) 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.

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.

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 hardnesses 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.

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.

Theses

1. I performed principal components analysis in the IRAS field PL51 ($\alpha = 4h, \delta = +30^o$). The analysis resulted in two significant components corresponding to the Zodiacal Light and Galactic radiation. I estimated two characteristic temperatures of $200\,K$ and $40\,K$, respectively.

2. I investigated the spatial and velocity distribution of atomic hydrogen associated with the Cepheus Bubble, a giant dust ring around the older part of the Cep OB2 association. Using HI 21 cm data, taken from the Leiden/Dwingeloo survey, I have reached the following results:

2/1 I identified the main HI structures associated with the dust ring in the velocity range [-14,+2] kms^{-1}.

2/2 At least three quarters of the ring seemed to be located at 900 pc, and I suggested the existence of a physically contiguous, almost complete ring around the older subgroup of the Cep OB2 association.

2/3 The HI data revealed an expanding shell with a radial velocity difference of $V \simeq 33kms^{-1}$ between the approaching and receding sides. In order to compare its kinematics with the analytical theoretical results of expanding shells I introduced an 'effective expansion velocity' of $\simeq 10kms^{-1}$ taking into account the 1:4 density ratio between the approaching and receding sides. The size and kinematics of this shell can best be modelled by a SN explosion occured at about 1.7 Myrs ago.

2/4 I pointed out that the time scale of the SNR expansion was considerably lower than the age of several parts of the Cepheus Bubble, suggesting that the supernova was probably exploded in an already existing cavity. The expanding shock front then interacted with the already existing star forming regions Sh2-140, IC 1396, and NGC 7129 at the inner edge of the Bubble, leading to a new wave of star formation.

3. I studied the B, V, R, I star count maps of L1251 in order to derive the spatial structure of the obscuring matter in the cloud. I assumed that the pixel values in the star count maps form a one dimensional manifold in the $BVRI$ four dimensional parameter space and defined areas of similar extinction by means of multivariate $k - means$ clustering. After defining areas of equal extinction I derived the amount of obscuration using a maximum likelihood procedure based on a Monte Carlo simulation of the Wainscoat et al. (1992) model. As a byproduct of the model fitting I obtained the $330 \pm 30\,pc$ distance of the cloud.

3/1 The extinction maps clearly showed the main body of the cloud and a less dense region having a form of a bow shock of a blunt body. The form of the bow shock allowed me to calculate the approximate Mach number ($M \approx 2$) of the streaming around the head of L1251.

3/2 Comparing the obscuration of the dust in different colors I calculated the dependence of the total to selective extinction (R_V value) on the visual extinction. The total to selective extinction exceeded $R_V = 5$ in the densest part of the cloud.

3/3 Using Equation (4.8) for converting the optical extinction distribution into surface mass density I obtained the mass of 371 $M_\odot$ for the cloud, in a reasonably good agreement with the 410 $M_\odot$ figure obtained by Sato et al. (1994), based on $C^{18}O$ measurements.

3/4 Assuming a spherical symmetry for the head of L1251 I computed the radial distribution of the mass and compared it with a polytropic model of $n = 5$ (Schuster sphere). Up to $r = 2.5\,pc$ from the center of the mass the fit is excellent but beyond this distance the observed points start to depart remarkably from the fit due to the drastic distortion of the spherical symmetry by the tail of L1251.

3/5 In the head of L1251 the Schuster model fit matches well the mean $FWHM$ of the 1.3 cm NH_3 line measured by Tóth & Walmsley (1996). It gives an unexpectedly good fit in the tail region indicating that isothermal contraction played a significant role in forming the density enhancements.

4. I made a medium resolution ($100\mathring{A}/mm$) spectroscopy for 35 stars from the list of Kun and Pásztor (1990). I developed a statistical method based on factor analysis and k-means clustering to get templates for classifying the spectra. The main points obtained in my study might be summarized as follows:

4/1 The spectra of our program stars are dominated by neutral metallic lines (FeI, MgI, NaI). Comparing our templates, obtained by factor analysis and k-means clustering, with those published in the library of standard spectra (Jacoby and Hunter l984) I obtained spectral types of A2-M3 for our program stars.

4/2 Except Kun 193 the stars do not show conspicuous Hydrogen emission. Since all of the objects studied was picked up by some suspect of $H\alpha$ emission on small scale spectra the weakness of this feature in my case might be explained by the transient behavior of emissivity.

4/3 Using the proper motion data of Marschall and van Altena (1987) I inferred that the vast majority of our stars belong to Tr37.

4/4 Comparing the loci of my program stars on the $\{Log(L); Log(T_{eff})\}$ plane with the evolutionary tracks of Forrestini (1994) I concluded that they are in the pre-main sequence evolutionary stage and have a mass of $0.5M_\odot < M < 3M_\odot$ and an age of $10^5 < t < 10^7$ years.

5. Studying the angular distribution of GRBs on the celestial sphere I obtained the following results:

5/1 The sky distribution of 2025 GRBs is anisotropic.

5/2 This anisotropy is not caused exclusively by instrumental effects.

5/3 Separating GRBs into the short and long subclasses it is shown that the short ones are distributed anisotropically, but the long ones can still be distributed isotropically.

5/4 I conjecture that the anisotropic distribution of short GRBs does not query their cosmological origin.

6. Using bivariate duration-hardness fittings I obtained the following results:

6/1 Increasing of k from 2 to 3 shows that the introduction of the third subgroup is real. This means that three subgroups of GRBs should exist. This confirms the earlier results of several authors.

6/2 Increasing of k from 3 to 4 shows that the introduction of the fourth subgroup is not needed. This means that *only* three subgroups should exist. This result is in accordance with Mukherjee et al. (1998).

6/3 From the fitting procedures it follows that the duration and the hardness are good quantities for the classification of GRBs. Remarkably, the intermediate subclass is in average even softer than the long subgroup.

6/4 I developed a method allowing to define for any GRB the probabilities determining how is it belonging to a given subclass. 11% of GRBs in the Current BATSE Catalog should belong to the intermediate subclass.

6/5 The anisotropy of the intermediate subclass is again detected on the 2 % significance level.

7. I have presented quantitative arguments in supporting two new results:

7/1 There is a power law relation between the fluence and duration of GRBs. For the short subgroup one obtains $m \simeq (0.46 - 1.04)$ with the most probable value around $m \simeq 0.81$. (In the reality, however, this value could be much smaller due to a possible strong spectral bias). For the long subgroup one obtains $m \simeq (0.97 - 1.19)$ with the most probable value around $m \simeq 1.11$. The difference is significant on the 4.7σ level.

7/2 For the short ones, the total energy released is weakly depending on the duration of the gamma ray emission, while for the long ones it is proportional roughly to the duration. This may indicate that two different types of central engines are at work, or perhaps two different types of progenitor systems are involved. It is often argued that those bursts for which X-ray, optical and radio afterglows have been found, all of which belong to the long-duration group, may be due to the collapse of a massive stellar progenitor (Paczyński , 1998; Fryer et al., 1999). The short bursts, none of which have as of yielded afterglows, may be hypothetically associated with neutron star mergers (Fryer et al., 1999) or perhaps other systems.

References to the Theses

1. Balázs, L.G., Kun, M., & Tóth, V., 1990, in 'The Galactic and Extragalactic Background Radiation', IAU Symposia No. 139, ed. S. Bowyer and C. Leinert, Kluwer Academic Publishers, Dordrecht, Holland, 214

   Balázs, L.G., & Tóth, L.V., 1991, in the 'Physics and Composition of Interstellar Matter', ed. J. Krelowski and J. Papaj, Institute of Astronomy Nicolaus Copernicus University, Torun , 135

2. Ábrahám P., Kun M., Balázs L.G., Holl A., & Frontó A., 1993, A&A, 268, 230

   Ábrahám, P., Balázs, L.G., & Kun, M., 2000, A&A, 354, 645

   Balázs, L. G. and Ábrahám, P.: 1996, IAU Symposia, No.169, p.623

   Balázs, L.G., & Kun, M., 1989, AN, 310, 385

   Kun, M., Balázs, L.G., & Tóth, I., 1987, Ap&SS, 134, 211

3. Balázs, L.G., Eisloeffel, J., Holl, A., Kelemen, J., & Kun, M., 1992, A&A, 255, 281

   Balázs, L.G., Ábrahám P., Kun, M., Kelemen, J., & Tóth, L.V., 2004, A&A, in press

4. Balázs, L.G., Melikyan, N.D., Melnikov, S.Yu., Shevchenko, V.S., 1987, IBVS No. 3099

   Balázs, L.G., Garibjanjan, A.T., Mirzoyan, L.V., Hambarjan, V.V., Kun, M., Frontó, A., Kelemen, J., 1996, A&A, 311, 145

5. Balázs, L.G., Mészáros, A., & Horváth, I., 1998, A&A, 39, 1

   Balázs, L.G., Mészáros, A., Horváth, I., & Vavrek, R., 1999, A&A Suppl., 138, 417

6. Bagoly, Z., Mészáros, A., Horváth, I., Balázs, L.G., & Mészáros, P., 1998, ApJ, 498, 42

   Balázs, L.G., Bagoly, Z., Horváth, I., Mészáros, A., & Mészáros, P., 2003, A&A, 381, 417

   Mészáros, A., Bagoly, Z., Horváth, I., Balázs, L.G., & Vavrek, R., 2000b, ApJ, 539, 98

7. Bagoly, Z., Mészáros, A., Horváth, I., Balázs, L.G., & Mészáros, P., 1998, ApJ, 498, 42

   Balázs, L.G., Bagoly, Z., Horváth, I., Mészáros, A., & Mészáros, P., 2003, A&A, 381, 417

Acknowledgements

Without the collaboration of several friends and colleagues this work never could come into existence. In the first line I am indebted to the late Laszló Detre who taught me to honor the observational facts in astronomy. I would like to express my thanks to Béla Balázs for introducing me into the statistical astronomy. Particular thanks belong to Béla Szeidl due to his many years support and encouragement in my work. My humble progress in the field of mathematical statistics would not be possible without the numerous discussions with Gábor Tusnády.

All the results presented in this dissertation were born in close collaboration with a high number of colleagues. It is not possible to rank them by significance, I list them therefore in alphabetic order: Péter Ábrahám, Zsolt

Bagoly, András Frontó, András Holl, István Horváth, János Kelemen, Mária Kun, Margit Paparó, László Pásztor, Viktor Tóth, Roland Vavrek and many others from abroad. I apologize for not mentioning all of them by name.

I am indebted for several colleagues whom contribution did not appear directly in the scientific results but without their work no results would be possible. In the first line I would like to mention Géza Virághalmy for his tireless work to develop and maintain the instrumentations. I am grateful to Pál Decsy, Ilona Gaál, Ilona Kálmán, Zsuzsa Litvay for assisting me in my work. The library plays a unique role in the scientific research. My distinguished thanks is devoted therefore to Magda Vargha for her many decades struggle for the up to day availability of all the relevant information in astronomy. I am also indebted to József Márton for continuing this line. I apologize for not mentioning all of the past and recent workers of the institute by name whom activity has been keeping the institute in working.

This work could never be completed without the warm atmosphere of my family. My greatest thank is devoted to my wife, Márta, to my children Borbála and Gergely for their continuous encouragement and their believe in completing this dissertation.

I dedicate this work to the memory of my parents.

Bibliography

Alves, J. ., Lada, C. J., & Lada, E. A., 2001, Nature, 409, 159

Ábrahám, P., Kun, M., Balázs, L.G., Holl, A., & Frontó, A., 1993, A&A, 268, 230

Ábrahám, P., Balázs, L.G., & Kun, M., 2000, A&A, 354, 645

Baars, J.W.M., & Wendker, H.J., 1976, A&A, 49, 473.

Bagoly, Z., Mészáros, A., Horváth, I., Balázs, L.G., & Mészáros, P., 1998, ApJ, 498, 42

Balázs, L.G., Bagoly, Z., Horváth, I., Mészáros, A., & Mészáros, P., 2003, A&A, 381, 417

Balázs, L.G., Eisloeffel, J., Holl, A., Kelemen, J., & Kun, M., 1992, A&A, 255, 281

Balázs, L.G., Garibjanjan, A.T., Mirzoyan, L.V., Hambarjan, V.V., Kun, M., Frontó, A., Kelemen, J., 1996, A&A, 311, 145

Balázs, L.G., & Kun, M., 1989, AN, 310, 385

Balázs, L.G., Kun, M., & Tóth, V., 1990, in 'The Galactic and Extragalactic Background Radiation', IAU Symposia No. 139, ed. S. Bowyer and C. Leinert, Kluwer Academic Publishers, Dordrecht, Holland, 214

Balázs, L.G., Melikyan, N.D., Melnikov, S.Yu., Shevchenko, V.S., 1987, IBVS No. 3099

Balázs, L.G., Mészáros, A., & Horváth, I., 1998, A&A, 39, 1

Balázs, L.G., Mészáros, A., Horváth, I., & Vavrek, R., 1999, A&A Suppl., 138, 417

Balázs, L.G., & Tóth, L.V., 1991, in the 'Physics and Composition of Interstellar Matter', ed. J. Krelowski and J. Papaj, Institute of Astronomy Nicolaus Copernicus University, Torun , 135

Balastegui, A., Ruiz-Lapuente, & Canal, R., 2001, MNRAS, 328, 283

Belli, B.M., 1995, Ap&SS 231, 43

Belli, B.M., 1997, ApJ, 479, L31

Benson, P. J., & Myers, P. C., 1989, ApJS, 71, 89

Blitz, L., Fich M., & Stark A.A., 1982, ApJS 49, 183

Bok, B.J., 1937, *'The Distribution of Stars in Space'*, U. Chicago Press

Bonnor, W. B., 1956, MNRAS, 116, 351

Borgonovo, L., 2004, A&A, in press, astro-ph/0402107

Brand, P.W.J.L., & Zealey, W.J., 1975, A&A 38, 363

Briggs, M.S., 1993, ApJ, 407, 126

Briggs, M.S., 1995, Ap&SS, 231, 3

Briggs, M.S., Paciesas, W.S., Pendleton, G.F., et al., 1996, ApJ, 459, 40

Brown, A.G.A., Hartmann, D., & Burton, W.B., 1995, A&A, 300, 903

Burton, W. B., & Hartmann, D., 1994, Ap&SS, 217, 189

Cambrésy, L., 1999, A&A, 345, 965

Canto, J., & Raga, A., 1998, MNRAS, 297, 383

Caselli, P., Benson, P. J., Myers, P. C., & Tafalla, M., 2002, ApJ, 572, 238

Cernicharo, J., Bachiller, R., & Duvert, G., 1985, A&A, 149, 273

Che, H., Yang, Y., Wu, M., & Li, T.P., 1997, ApJ 477, L69

Chevalier, R.A., 1974, ApJ 188, 501

Clayton, G.C., & Fitzpatrick, E.L., 1987, AJ, 92, 157.

Cline, D.B., Matthey, C., & Otwinowski, S., 1999, ApJ, 527, 827

Coles, P., 1998, Nature 391, 120

Cramér, H., 1937, '*Random variables and probability distributions'*, Cambridge Tracts in Mathematics and Mathematical Physics, No.36 (Cambridge University Press, Cambridge)

Crampton, D., & Fisher, W.A., 1974, Publ. Dom. Astrophys. Obs., 14, 283

Cudaback, D. D., & Heiles, C., 1969, ApJ, 155, L21

Dame, T.M., 1999, personal communication

Dame, T.M., Ungerechts H., Cohen R.S., et al., 1987, ApJ 322, 706

de Zeeuw, P.T., & Brand, J., 1985, In: "Birth and Evolution of Massive Stars and Stellar Groups", Boland W., van Woerden H. (eds.) p. 95

de Zeeuw, P.T., Brown, A.G.A., de Bruijne, J.H.J., et al., 1999, AJ 117, 354

Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977, J. Roy. Stat. Soc., B, 39, 1

Dezalay, J.P., Lestrade, J.P., Barat, C., et al., 1996, ApJ, 471, L27

Dickman, R. L., 1978, AJ, 83, 363

Dieter, N. H., 1973, ApJ, 183, 449

Dolan, J.F., 1972, Ap&SS, 17, 472

Dolan, J.F., 1974, A&A, 35, 105

Dolidze, M.V., 1975, Bull. Abast. Obs., 47, 3.

Dolidze, M.V., & Vyazovov, V.V., 1959, Bull. Abast. Obs., 24, 3.

Dubout-Crillon, R., 1976, A&AS, 25, 25

Ebert, R., 1955, Zeitschrift Astrophysics, 37, 217

Eddington, A.S., 1913, MNRAS, 73, 359

Efron, B., & Petrosian, V. 1992, ApJ, 339, 345

Emden, V. R., 1907, *'Gaskugeln'* (Leipzig)

ESA 1997, The Hipparcos and Tycho Catalogues, ESA SP-1200

Fenimore, E.E., & Bloom, J.S., 1995, ApJ, 453, 16

Fich, M., Treffers, R.R., & Dahl, G.P., 1990, AJ 99, 622

Fishman, G.J., Meegan, C.A., Wilson, R.B., et al., 1994, ApJS, 92, 229

Forestini, M., 1994, A&A, 285, 473

Fryer, C.L., Woosley, S.E., & Hartmann, D. H. 1999, ApJ, 526, 152

Garmany C.D., Olson G.L., Conti P.S., & van Steenberg M.E., 1981, ApJ 250, 660

Goodman, A. A., Benson, P. J., Fuller, G. A., & Myers, P. C., 1993, ApJ, 406, 528

Greiner, J., 2002, http://www.aip.de/ jcg/grbgen.html

Grenier, I. A., Lebrun, F., Arnaud, M., Dame, T. M., & Thaddeus, P., 1989, ApJ, 347, 231

Gupta, V., Das Gupta, P. & Bhat, P.N., 2002, astro-ph/0206402

Hakkila, J., Haglin, D.J., Roiger, R.J., Mallozzi, R.S., Pendleton, G.F., & Meegan, C.A., 2000a, in 'Gamma-Ray Bursts'; 5th Huntsville Symp., eds. R.M. Kippen, R.S. Mallozzi, G.J. Fishman, AIP, Melville, p.33

Hakkila, J., Meegan, C.A., Pendleton, G.N., Mallozzi, R.S., Haglin, D.J., & Roiger, R.J., 2000b, in 'Gamma-Ray Bursts'; 5th Huntsville Symp., eds. R.M. Kippen, R.S. Mallozzi, G.J. Fishman, AIP, Melville, p.48

Hakkila, J., Haglin, D.J., Pendleton, G.N., Mallozzi, R.S., Meegan, C.A., Roiger, R.J., 2000c, ApJ, 538, 165

Halász, G., 1984, private communication

Hartmann, D., & Burton, W.B., 1997, Atlas of Galactic HI, Cambridge Univ. Press

Hayes, W. D., & Probstein, R. F., 1959, *'Hypersonic Flow Theory'* (Academic Press, New York and London)

Herbig, G.H., & Bell, K.R., 1988, Lick. Obs. Bull., No. 1111

Hildebrand, R.H., 1983, QJRAS 24, 267

Horváth, I., 1998, ApJ, 508, 757

Horváth, I., 2002, A&A, 392, 791

Horváth, I., Mészáros, P., & Mészáros, A., 1996, ApJ 470, 56

Jacoby, G.H., & Hunter, D.A., 1984, ApJS, 56, 257.

Jarrett, T.H., Dickman, R.L., & Herbst, W., 1989, ApJ, 345, 881

Johnson, H.L., 1966, ARAA, 4, 193.

Kandori, R., Dobashi, K., Uehara, H., Sato, F., & Yanagisawa, K., 2003, AJ, 126, 1888

Katz, J.I., & Canel, L.M., 1996, ApJ 471, 915

Kendall, M.G., and Stuart, A., 1973, '*The AdvancedTheory of Statistics*', Charles Griffin & Co. Ltd., London &High Wycombe

Klebesadel, R.W., Strong, I.B., & Olson, R.A., 1973, ApJ, 182, L85

Koshut, T.M., et al., 1996, ApJ, 463, 570

Kouveliotou, C., Meegan, C.A., Fishman, G.J., et al., 1993, ApJ, 413, L101

Kouveliotou, C., et al., 1993, ApJ, 413, L101

Kouveliotou, C., et al., 1993, ApJ, 413, L101

Kudritzki, R.P., 1998, In: Proc. 8th Canary Winter School, 1996, Aparicio A., Herrero A., Sanchez F. (eds.) Cambridge Univ. Press.

Kun, M., 1982, Astrophysics, 18, 37

Kun, M., 1986, Ap&SS, 125, 13.

Kun, M., 1995, Ap&SS, 224, 73

Kun, M., 1998, ApJS, 115, 59

Kun, M., Balázs, L.G., & Tóth, I., 1987, Ap&SS, 134, 211 (KBT)

Kun, M., & Pásztor, L., 1990, Ap&SS, 174, 13

Kun, M., & Prusti, T., 1993, A&A, 272, 235

Kun, M., Vinkó, J., & Szabados, L., 2000, MNRAS, 319, 777

Kurth, R., 1952, Z. Astroph., 31, 115

Lamb, D.Q., Graziani, C., & Smith, I.A., 1993, ApJ, 413, L11

Lauer, T.R., & Postman, M., 1994, ApJ, 425, 418

Lee, C. W., Myers, P. C., & Tafalla, M., 1999, ApJ, 526, 788

Lee, C. W., Myers, P. C., & Tafalla, M., 2001, ApJS, 136, 703

Lee, Y., 1994, Journal of Korean Astronomical Society, 27, 159

Lee, T., & Petrosian, V., 1996, ApJ, 470, 479

Lee, T., & Petrosian, V., 1997, ApJ, 474, 37L

Leinert, C., Ábrahám, P., Acosta-Pulido, J., Lemke, D., & Siebenmorgen, R., 2002, A&A, 393, 1073

Litvin, V.F., Matveev, S.A., Mamedov, S.V., & Orlov, V.V., 2001, Pis'ma v Astronomicheskiy Zhurnal, 27, 495

Lloyd, K.H., 1969, Am. J. Phys., 37, 329

Lloyd, N.L., & Petrosian, V., 1999, ApJ, 511, 550

Lucy, L.B., 1974, AJ, 79, 745

Malmquist, K.G., 1924, Medd. Lund Astron. Obs. Ser. II., No 32, 64

Malmquist, K.G., 1936, Stockholm Obs. Medd, No 26

Marschall, L.A., Comins, N.F., & Karshner, G.B., 1990, AJ, 99, 1536

Marschall, L.A., & Van Altena, W.F., 1987, AJ, 94, 1

Mathis, J. S., 1990, ARAA, 28, 37

Mazets, E.P., Golenetskii, S.V., Il'Inskii, V.N., Panov, V.N., Aptekar, R. L., Gur'yan, Y. A., et al., 1981, Ap&SS, 80, 3

McBreen, B., Hurley, K.J., Long, R., & Metcalfe, L., 1994, MNRAS, 271, 662

McKee, Ch.F., Van Buren,D., & Lazareff, B., 1984, ApJ, 278, L115

McLachlan, G.J., & Basford, K.E., 1988, *'Mixture models'*, (New York: Marcel Dekker)

Medvedev, M. V., & Rybicki, G., 2001, ApJ, 555, 863

Meegan, C.A., et al., 1996, ApJS, 106, 65 (3B BATSE Catalog)

Meegan, C.A., Pendleton, G.N., Briggs, M.S., et al., 1996, ApJS 106, 65 (The Third BATSE Gamma-Ray Burst Catalog)

Meegan, C.A., Pendleton, G.N., Briggs, M.S., et al., 1997, Current BATSE Gamma-Ray Burst Catalog, http://www.batse.msfc.nasa.gov/data

Meegan, C.A., Hakkila, J., Johnson, A., Pendleton, G., & Mallozzi, R.S., 2000b, in Gamma-Ray Bursts; 5th Huntsville Symp., eds. R.M. Kippen, R.S. Mallozzi, G.J. Fishman, AIP, Melville, p.43

Meegan, C.A., Pendleton, G.N., Briggs, M.S., et al., 2001, The BATSE Current Gamma-Ray Burst Catalog, http://gammaray.msfc.nasa.gov/batse/grb/catalog/currernt

Mészáros, A., 1997, A&A, 328, 1

Mészáros, A., Bagoly, Z., Horváth, I., & Mészáros, P., 1996, J. Korean Astron. Soc. 29, S43

Mészáros, A., Bagoly, Z., & Vavrek, R., 2000a, A&A, 354, 1

Mészáros, A., Bagoly, Z., Horváth, I., Balázs, L.G., & Vavrek, R., 2000b, ApJ, 539, 98

Mészáros, A., & Mészáros, P., 1996, ApJ, 466, 29

Mészáros, A., & Mészáros, P., 1996, ApJ, 466, 29

Mészáros, A., & Štoček, J., 2003, A&A, 403, 443

Mészáros, A., & Vanysek, V., 1997, A&A, 319, 371

Mészáros, P., 2001, Science, 291, 79

Mészáros, P., 2003, Nature, 423, 809

Mészáros, P., & Mészáros, A., 1995, ApJ, 449, 9

Mészáros, P., & Rees, M. J., 1993, ApJ, 405, 278

Mészáros, P., & Rees, M.J., 1997, ApJ, 482, L29

Mihalas, D., & Binney, J., 1981, *'Galactic Astronomy'*, W.H. Freeman and Co., San Francisco

Mukherjee, S., Feigelson, E.D., Jogesh Babu, G., Murtagh, F., Fraley, Ch., & Raftery, A., 1998, ApJ, 508, 314

Murtagh, F. & Heck, A., 1987, *'Multivariate data analysis'* (Astrophysics and Space Science Library, Dordrecht: Reidel, 1987)

Nemiroff, R.J., 1995, PASP, 10, 1131

Nikolić, S., Johansson, L. E. B., & Harju, J., 2003, A&A, 409, 941

Norris, J.P., Bonnell, J.T., Nemiroff, R.J., et al., 1995, ApJ, 439, 542

Norris, J.P., Cline, T.L., Desai, U.D., Teegarden, B.J., 1984, Nature, 308, 434

Norris, J.P., Nemiroff, R.J., Scargle, J.D., et al., 1994, ApJ, 424, 540

Norris, J.P., Scargle, J.D., & Bonnell, J.T., 2000, in 'Gamma-Ray Bursts in the Afterglow Era', Proc. of the Intern. Workshop, Rome, Italy, 17-20 Oct. 2000, eds. E. Costa et al., Springer, p.40

Norris, J.P., Scargle, J.D., & Bonnell, J.T., 2001, in 'Gamma-Ray Bursts in the Afterglow Era', Proc. Int. Workshop held in Rome, Italy, eds. E. Costa et al., ESO Astrophysics Symp. (Berlin: Springer), p. 40

Paciesas, W. S., Meegan C.A., Pendleton G.N., et al., 1999, ApJS, 122, 465 (4B BATSE Catalog)

Paczyński, B., 1995, PASP, 107, 1167

Paczyński, B., 1998, ApJ, 497, L45

Patel, N.A., Goldsmith, P.F., Snell, R.L., Hezel, T., & Xie, T., 1995, ApJ. 447, 721

Patel, N.A., Goldsmith, P.F., Heyer, M.H., Snell, R., & Pratap, P., 1998, ApJ, 507, 241

Pedlar, A., 1980, MNRAS 192, 179

Peebles, P.J.E. 1993, *'Principles of Physical Cosmology'* (Princeton University Press, Princeton)

Pendleton, C.N., Paciesas, W.S., Briggs, M.S., et al., 1997, ApJ, 489, 175

Petrosian, V., & Lee, T., 1996, ApJ, 467, L29

Piran, T., 1999, Phys.Rep., 314, 575

Press, W. H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T., 1992, *Numerical Recipes* (Cambridge University Press, Cambridge)

Racine, R., 1968, AJ, 73, 233

Rajaniemi, H.J., & Mähönen, P., 2002, ApJ, 566, 202

Rees, M.J., Mészáros, P., 1994, ApJ, 430, L93

Rees, M.J., & Mészáros, P., 1994, ApJ, 430, L93

Reichart, D.E., & Mészáros, P., 1997, ApJ, 483, 597

Rényi, A., 1962, *'Wahrscheinlichtkeitsrechnung'* (Berlin: VEB Deutscher Verlag der Wissenschaften)

Rózyczka, M., & Tenorio-Tagle, G., 1985, A&A, 147, 220

Ryde, F., & Petrosian, V., 2002, ApJ, 578, 290

Sato, F., & Fukui, Y., 1989, ApJ, 343, 773

Sato, F., Mizuno, A., Nagahama, T., et al., 1994, ApJ, 435, 279

Savage, B.D., & Mathis, J.S., 1979, ARAA, 17, 73

Schmidt-Kaler, T.H., 1982, In: Landolt-Brnstein New Series, Group 6, Vol. 2b, Stars and Star Clusters, Schaifers K., Voigt H.H. (eds.) (Berlin: Springer)

Schuster, A., 1883, British Assoc. Report, 427

Seitter, W.C., 1975, *'Atlas for Objective Prism Spectra'*, erd. Dümmler Verlag, Bonn

Sharpless, S., 1959, ApJS, 4, 257

Simonson, S.C., 1968, ApJ, 154, 923

Simonson, S.C., & van Someren Greve, H.W., 1976, A&A 49, 343 (SVSG)

Slechta, M., & Mészáros, A., 1997, Ap&SS 249, 1

Snowden, S.L., Freyberg, M.J., Plucinsky, P.P., et al., 1995, ApJ, 454, 643

SPSS inc: http://www.SPSS.com

Stern, B.E., 1996, ApJ, 464, L111

Stern, B., Poutanen, J., & Svensson, R., 1999, ApJ, 510, 312

Stone, R.C., 1979, ApJ, 232, 520

Sugitani, K., Fukui, Y., Mizuno, A., & Ohashi, N., 1989, ApJ, L87

Sume, A., Downes, D., & Wilson, T. L., 1975, A&A, 39, 435

Syer, D., & Saha, P., 1994, ApJ, 427, 714

Sylos-Labini, F., Montuori, M., Pietronero, L., 1998, Phys. Rep., 293, 61

Tegmark, M., Hartmann, D.H., Briggs, M.S., Hakkila, J., & Meegan, C.A., 1996a, ApJ, 466, 757

Tegmark, M., Hartmann, D.H., Briggs, M.S., & Meegan, C.A., 1996b, ApJ, 468, 214

Tenorio-Tagle G., & Bodenheimer P., 1988, ARAA, 26, 145

Tóth, L.V., Horváth, A., & Haikala, L.A., 1995, Ap&SS, 233, 175

Tóth, L.V., Mattila, K., Haikala, L., & Balázs, L.G., 1993, in ASP Conf. Ser. 52: 'Astronomical Data Analysis Software and Systems II', 462

Tóth, L.V., & Walmsley, C.M., 1996, A&A, 311, 981

Trumpler, R. J., & Weaver, H. F., 1953, *Statistical Astronomy* (Berkeley: University of California Press)

Ulmer, A., & Wijers, R.A.M.J., 1995, ApJ, 439, 303

Usov, V., & Chibisov, G., 1975, SvA, 19, 115

van Paradijs, J., Groot, P.J., Galama, T., et al., 1997, Nature, 386, 686

Wackerling, L.R., 1970, Mem. Roy. Astron. Soc., 73, 153

Wainscoat, R.J., Cohen, M., Volk, K., Walker, H.J., & Schwartz, D.E., 1992, ApJS, 83, 111

Wakker, B.P., & van Woerden, H., 1991, A&A, 250, 509

Weaver, R., Castor, J., McCray, R., Shapiro, P., & Moore, R., 1977, ApJ, 218, 377

Weikard, H., Wouterloot, J.G.A., Castets, A., Winnewisser, G., & Sugitani, K., 1996, A&A, 309, 581

Weinberg, S., 1972, *'Gravitation and Cosmology'* (Wiley, New York)

Wheelock, S., Gautier, N., Chillemi, J., et al., 1994, IRAS Sky Survey Atlas, Explanatory Supplement, Infrared Processing and Analysing Center, Jet Propulsion Laboratory

Whittet, D.C.B., Gerakines, P.A., Hough, J.H., & Shenoy, S.S., 2001, ApJ, 547, 872

Wijers, R.A.M.J., & Paczyński, B., 1994, ApJ, 437, L107

Wilson, R.E., 1963, *'General Catalogue of Stellar Radial Velocities'*, Carnegie Institute of Washington, 601, Washington D.C.

Whitford, A.E., 1958, AJ, 63, 201

Yonekura, Y., Dobashi, K., Mizuno, A., Ogawa, H., & Fukui, Y., 1997, ApJS, 110, 21

Zhou, S., Wu, Y., Evans, N. J., Fuller, G. A., & Myers, P. C., 1989, ApJ, 346, 168