Multivariate Separation of Cosmic Components

L.G. Balázs
Konkoly Observatory of the Hungarian Academy of Sciences
P.O. Box 67, H-1525 Budapest, Hungary
E-mail: balazs@konkoly.hu

Abstract

Observing and storing the photons of the incoming radiation from the Cosmos typically give 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 is 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. We demonstrated in three cases how these technics can be used for separating physically independent cosmic components projected onto the same celestial area by chance.

1. Introduction: 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:

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

These physical quantities determine basically the possible observational programs:

1. Position $\Longrightarrow$ astrometry
2. Distribution of photons with $\lambda \Longrightarrow$ spectroscopy
3. Number of photons $\Longrightarrow$ photometry
4. Polarization $\Longrightarrow$ polarimetry
5. Time of observation $\Longrightarrow$ variability
6. Distribution 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. 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 called the 'Data Matrix' which is the input of many multivariate statistical procedures.

Table 1: Structure of the Data Matrix: m means the number of properties and n runs over the cases.

$$\begin{array}{cccccc} \alpha_1 & \delta_1 & \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \alpha_2 & \delta_2 & \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_n & \delta_n & \lambda_{11} & \lambda_{12} & \ldots & \lambda_{nm} \end{array}$$

2. Brief summary of multivariate methods

2.1. Factor Analysis

A common problem in the multivariate statistics is 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:

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

In the formula above X_j means the observed variables, m is the number of properties, p is the number of hidden factors (normally $p < m$), a_{jk} denotes the factor loadings, F_k the factor scores, and u_j-s are called individual factors. The individual factors represent those parts 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 is equal to those of the the observed variables. In order to compute the PCs one has to solve the following eigenvalue problem:

$$\underline{R}\,\underline{a} = \lambda\,\underline{a} \tag{2}$$

where $\underline{R}$, $\underline{a}$ and λ mean the correlation matrix of the observed variables, its eigenvector and eigenvalue, respectively. The components of the $\underline{a}$ eigenvectors give the coefficients of the linear relationship between the PCs and the observed variables. The PC belonging to the biggest eigenvalue of $\underline{R}$ gives the most significant contribution to the observed variables. The PCs can be ordered according to the size of the eigenvalues. In most cases the default solution of the factor problem is the PCA in the statistical software packages (e.g. 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.

2.2. 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.

3. Case studies

3.1. Separation of the Zodiacal Light and Galactic Dust Emission

The IRAS mission covered the whole sky in four (12, 25, 60, 100 μm) wavelengths. In particular, the 12 and 25 micron images were dominated by the thermal emission of the Zodiacal Light (ZL) having a characteristic temperature around 250 $K\hbar$. 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.

Table 2: 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). (Balázs et al., 1990)

eigenvalue	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

(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 $\underline{R}$ 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 $15^\circ \times 15^\circ$ (corresponding to 512×512 pixels) in Perseus close to the ecliptic, containing the California Nebula, IC 348 and the Pleiades.

Solving Eq. (2) for this case we got the results summarized in Tab. 2. One can infer from this table that there are two large eigenvalues indicating the presence of two important factors. The last two columns of the table give the a_{jk} factor coefficients for Eq. (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, 60 and 100 μm intensities) one gets the two images as shown in Fig. 1, along with the originals (Balázs et al., 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 micron 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 micron emission. Performing cluster analysis altogether 10 regions were defined, however this figure was arbitrary. The result is given in Fig. 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).

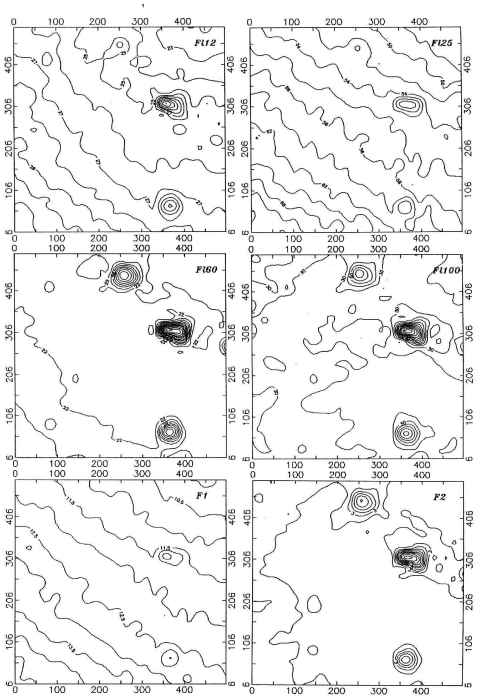

Figure 1: Input IRAS (12, 25, 60, and 100μm) images of the factor analysis and the resulted two factor pictures. The coordinates are measured in pixels. The objects are the California Nebula, IC 348 and the Pleiades, in descending order. Note the strong trend in F_1 representing the ZL while F_2 displays the Galactic component (Balázs et al., 1991).

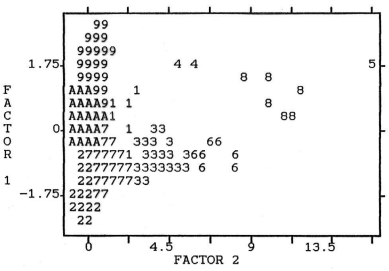

Figure 2. Character plot of regions (clusters) of similar properties in the $\{F_1; F_2\}$ factor plane. The identical symbols mean physically similar regions. The basic features of this plot are the two 'fingers' pointing upwards and nearly horizontally. These 'fingers' may be identified with the Zodiacal Light (dominating $Factor\,1$) and the galactic radiation (dominating $Factor\,2$) (Balázs et al., 1990).

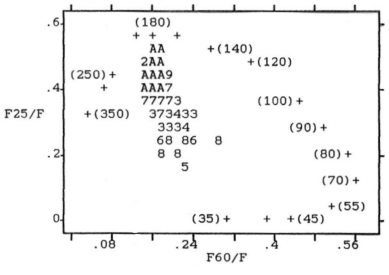

Figure 3. Distribution of duster members in the $\{F25/F; F6O/F\}$ plane. The coding of dusters is the same as in Fig. 2. The loci of dust low $\alpha = 1$ radiations of different temperatures are marked with crosses. The numbers in parentheses are the respective temperatures. Note that the wedge-shaped distribution of symbols representing real measurements points towards dust temeratures of about 40K and 200K (Balázs et al., 1990).

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

$$F = F_{12} + F_{25} + F_{60} + F_{100} \tag{3}$$

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 depends 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, marked with crosses the sources of different temperatures in the line of sight. As a consequence, the real points in Fig. 3 are not on the theoretically computed line but deviate from it according to the relative intensity of superimposed sources of different temperatures. Keeping the same coding of sources as in Fig. 2 one gets a wedge-shaped distribution of symbols representing real measurements pointing towards dust temperatures of about 40 K and 200K. This distribution can be obtained from the superimposed ZL and Galactic sources with these characteristic temperatures.

4. Separation of HI components in the field of L1780

The next case study refers to L1780, a small dust cloud at a high galactic latitude. By analyzing the profile of the HI 21 cm line it was difficult to separate the object from the background since the velocity of the cloud was very close to those of the background.

The cloud was observed with the 100 m telescope at Effelsberg at 209 positions in 82 channels. Formulating the problem of separation with the phraseology of the multivariate statistical analysis we had 209 cases and 82 properties.

Performing PCA yielded 7 eigenvalues > 1 and they were accepted as significant factors. In order to get clear-cut factor pattern we made Varimax rotation. This procedure makes use of the fact that factors are determined only up to an orthogonal transformation. Varimax rotation is an orthogonal transformation

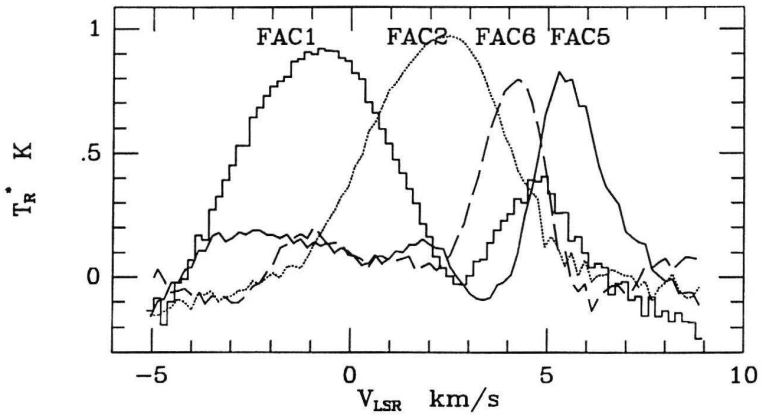

Figure 4: Results of factor analysis in L1780. The factor coefficients are displayed as functions of the channels calibrated to the velocity in the line of sight. Beside the strongest factor (FAC1), the main HI background component, those are displayed which give excess emission in the field of L1780 (FAC2, FAC5, FAC6) (Tóth et al., 1993).

which maximizes the variance of the factor coefficients and usually gives a dominant factor in each observed variables. This dominant factor makes easier to identify the factors with real physical entities.

Inspecting the pictures obtained from the factor scores we found that FAC2, FAC5 and FAC6 indicated excess HI radiation that could be associated with L1780. On the contrary, FAC1, FAC3, FAC4 and FAC7 described the background. The contributions of the different factors to the channel maps are displayed in Fig. 4. Summing up the factors related to the cloud gives the amount of HI associated to L1780 (Tóth et al., 1993).

Using the computed factors associated with the cloud we calculated the HI 21 cm spectra in some characteristic positions of L1780 along with the background as seen in Fig. 5. Note that the difference of the HI spectra across the cloud (i.e. the difference between he position a and c) indicates large scale internal motions.

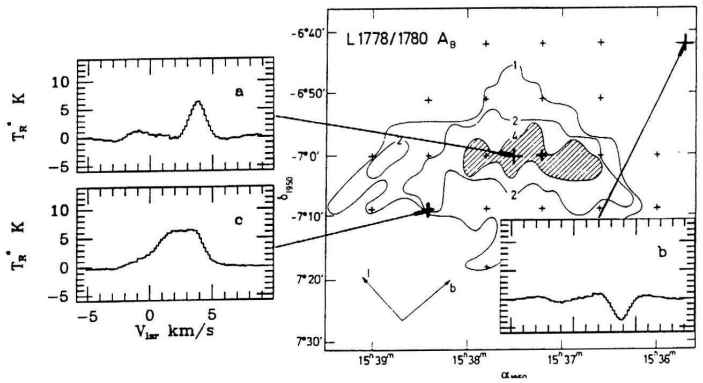

Figure 5: Distribution of the dust in L1780 as obtained from the optical extinction (A_B). The HI spectra in the inserts give velocity profiles at some characteristic parts of the cloud (a, c) and the background (b), respectively. Note that the difference of the HI spectra across the cloud (i.e. the difference between the positions a and c indicates large scale internal motions (Tóth et al., 1993).

5. Multivariate study of the Cepheus Bubble

The Cepheus Bubble was discovered in the IRAS maps (Kun et al., 1987) as a ring about 10 deg. in diameter around Cep OB2 joining several known star forming regions (S140, IC1396, S134, etc). The association of the ring with the star forming regions with known distances (between 800-900 pc) enabled one to estimate the true geometric diameter to be 140-160 pc. The dust responsible for the radiation detected by IRAS, however, is only a tiny fraction of the total mass which is mostly in the form of HI. In order to calculate the mass and internal kinematics of the bubble one of the best choice was to use neutral hydrogen observations. The integrated map of the HI channel intensities clearly showed a ring coinciding with those in the IRAS maps (Fig. 6). We used 43 HI channel maps of the region from the Dwingeloo HI sky survey (Burton & Hartman, 1994) in the [–38 km/s; +10 km/s] range.

Inspection of the channel maps (Fig. 5), starting at –38 km/s and moving towards positive radial velocities, revealed a ring structure starting at –30 km/s, becoming dominant in the [–18 km/s; –10 km/s] range and fading away

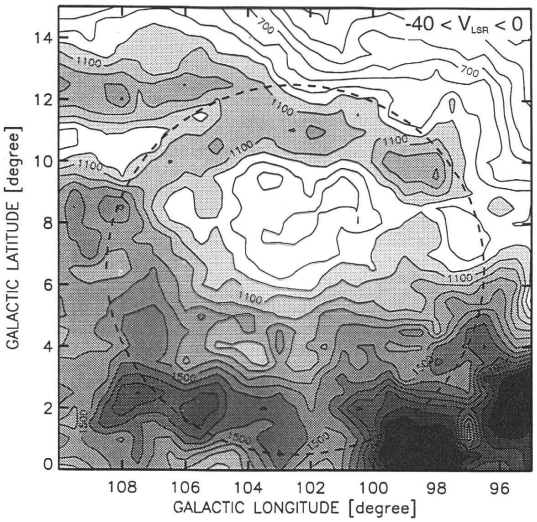

Figure 6: Integrated column density of the HI in the region of the Cepheus Bubble. A dashed circle indicates the outer boundary of the infrared ring (Ábrahám et al., 2000).

afterwards (Fig. 8). In order to separate the HI associated with the Bubble we performed factor analysis based on PCA which yielded 6 main components (see Tab. 3). The factor coefficients, similarly to the case of L1780, could be calibrated for radial velocity and are displayed in Fig. 9. The Figure clearly shows that each factor dominates a certain velocity range. Usage of the images made up from the factor scores (Fig. 10) enabled us to identify the factors in terms of different physical entities of the HI distribution. The main body of the bubble appeared in factor 2 whereas factor 3 and 5 are strong on the area where factor 2 is weak. These factors can be interpreted as different slices of an expanding shell. Identification of the factors with different physical entities of the neutral hydrogen enabled us to separate the HI associated with the Bubble and determine its mass and age (Ábrahám et al., 2000).

6. Conclusions

1. The nature of astronomical information is well suited for multivariate studies.

2. The typical procedures of multivariate methods (factor analysis, cluster analysis) can be applied succesfully for studying different structures in the

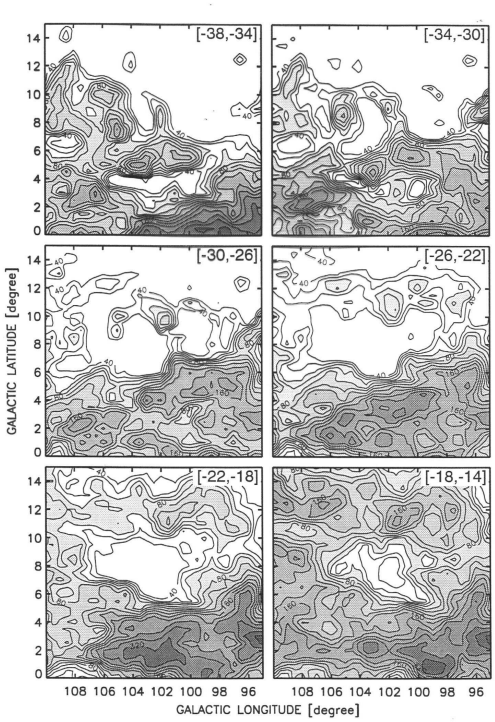

Figure 7: HI channel maps of the Cepheus Bubble in the [-38 km/s; -14 km/s] range. The ring structure appears at -26 km/s and increases in dominance at less negative velocities (Ábrahám et al., 2000).

data cubes.

3. There is no straightforward way to assign physically meaningful objects to the formal statistical results (actually this is one of the basic problems).

4. Special care is needed to separate "ghosts". In some cases physically related structures can be splitted into different mathematical structures.

5. The best results can be expected for problems where the basic mathematical assumptions (e.g. linearity and orthogonality at PCA models) are also physically meaningful.

6. A basic advantage is the existence of professional statistical packages (SPSS, SAS, S-plus, etc.)

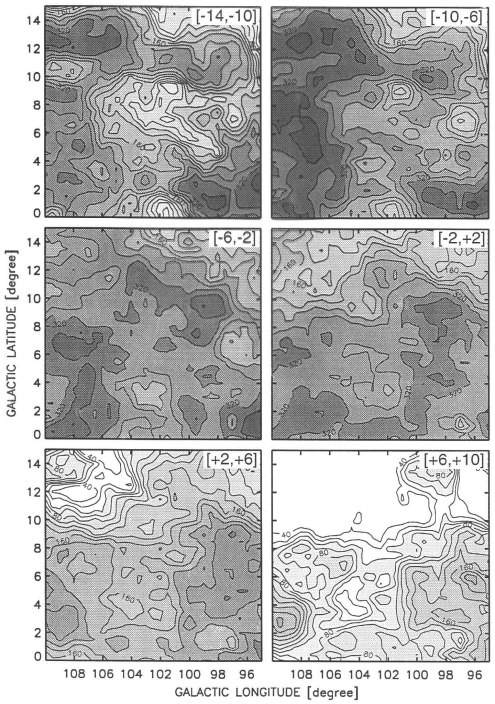

Figure 8: HI channel maps of the Cepheus Bubble in the [-14 km/s; +10 km/s] range. The ring structure fades away towards less negative velocities and completely disappears (Ábrahám et al., 2000).

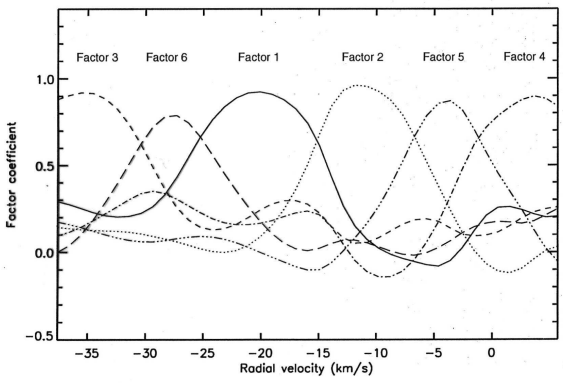

Figure 9: Dependence of the factor coefficients on the radial velocity in the Cepheus Bubble. Each factor dominates a certain range of radial velocities. The 2nd strongest factor can be associated with the main body of the ring (Ábrahám et al., 2000).

Table 3: 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

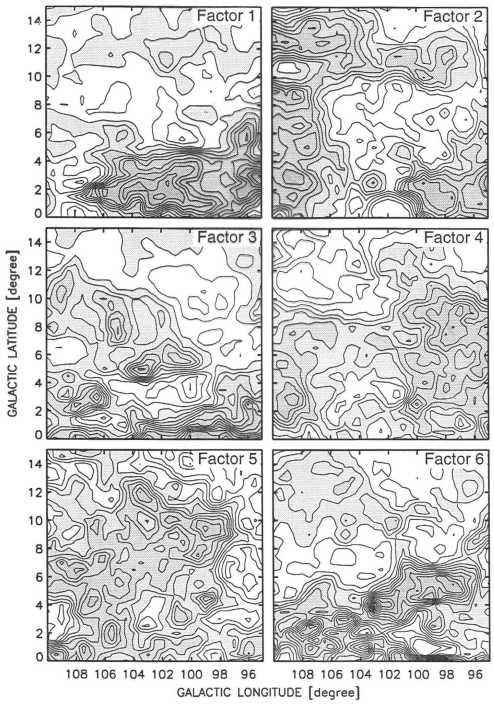

Figure 10: Images made up from the factor scores in the Bubble. The 2nd strongest factor gives the main body of the ring. Images of factors 3 and 5 are strong on the area where factor 2 is weak. These factors can be interpreted as different parts of an expanding shell (Ábrahám et al., 2000).

Acknowledgements

The author is indebted to Dr. Gábor Tusnády (Rényi Institute of Mathematics, Budapest) for the comprehensive discussions in the theory and practice of multivariate statistical methods.

References

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

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

Balázs, L.G., Tóth, L.V., 1990, in the 'Physics and Composition of Interstellar Matter', eds. J. Krelowski & J. Papaj, Institute of Astronomy Nicolaus Copernicus University, Torun , p.135

Burton, W.B., Hartmann, D., 1994, A&ASS, 217, 189

Kun, M., Balázs, L.G., Tóth, I. 1987, A&ASS, 134, 211

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

Murtagh, F., Heck, A., 1987, "Multivariate data analysis", Astrophysics and Space Science Library, Dordrecht: Reidel

Tóth, L.V., Mattila, K., Haikala, L., Balázs, L.G., ASP Conf. Ser. Vol. 52, 'Astronomical Data Analysis Software and Systems II', eds. R.J. Hanisch et al., p.462