Astron. Astrophys. Suppl. Ser. **127,** 319-325 (1998)

Covered data structures I

The algorithm

C. Kienel and S. Kimeswenger

Institut für Astronomie der Leopold–Franzens–Universität Innsbruck, Technikerstraße 25, A–6020 Innsbruck, Austria
http://astro.uibk.ac.at

Received September 13, 1996; accepted May 26, 1997

Abstract. Many algorithms separating or detecting groups of similiar objects (for example the extraction of groups lying in a color–color–diagram) are based on two statistical methods: the Kernel Method (Silverman 1986) or the Likelihood Statistic (van der Waerden 1957). These standard methods have one or more restrictions (e.g. known number or differentiability of the groups, ...). We present here a new powerful algorithm and show results worked out with artificial data sets.

The algorithm is based on Recursive Restoration Methods (neither on the Likelihood Statistic (Sutherland & Saunders 1992) nor on the Kernel Method (De Jager et al. 1986)) and allows to detect substructures in a data set, even if they are overlapped or superimposed by any kind of dominating main structure. In comparison to the other methods mentioned above there are no restrictions concerning the form and the dimension of the components lying in the data set.

The algorithm is easy to handle and therefore opens a wide range of applications for many fields of science (see Boller et al. 1992).

Key words: methods: statistical — astronomical data bases: miscellaneous

1. Introduction

Due to new, fast detectors, an increasing amount of information with more and more parameters has to be handled in nearly every branch of science. The extensive information is bound in multidimensional data sets and often consists of mixtures of groupable subsets and errors that might for example be explained by the instruments' or measurement uncertainties.

The **I**nfra**r**ed **A**stronomical **S**atellite (IRAS), started in 1983, detected around 245.000 sources. The **De**ep **N**ear **I**nfrared **S**outhern Sky Survey (DeNIS, Epchtein et al. 1994), started in 1995, will reach a number of objects which is estimated to be around 5 10^8. Such data sets consist of different types of objects, such as galaxies, stars or planetary nebulae. Furthermore the instruments register multiple properties of each source (such as coordinates, different fluxes or magnitudes). One of the main tasks is now the exact separation of different types, by means of known properties, and to get statistical information about objects without such collected properties.

The algorithm presented here is able to separate different structures and to give a probability function which indicates if a source is part of one of the structures.

In order to obtain assumed substructures it is possible to *model* the superimposing main structure with any kind of artificial form (for example multidimensional Gaussian distributions) as well as to use *natural subsets* (for example obtained by special information about a part of the whole data set) as a main structure. The algorithm substracts this main structure from the whole data set, the remaining part (thereafter called *residuals*) opens the view to eventually existing covered structures.

It is possible to improve the model of the main component by using the algorithm in an iterative way.

2. Actual methods used

In order to get the number, form and position of the different structures, as well as an estimation of the error, many solutions exist. Most of these solutions represent a specialization of two main methods:

In many cases the *Maximum Likelihood*–method (*ML*–algorithm) is used to estimate unknown parameters (Sutherland & Saunders 1992). Boller also uses an algorithm based on the Likelihood Satistic and works with four-dimensional (artificial) Gaussian distributions.

The *density–estimation* is another way, often used, to estimate parameters or shapes of distributions building a mixed distribution. One of the algorithms based on the

Send offprint requests to: S. Kimeswenger density estimation is the *Kernel*-method (De Jager et al. 1986).

The specialized solutions requires one or more restrictions which can affect for example:

- the number of substructures,
- special parameterizable models such as Gaussian distributions,
- differentiability or
- the number of dimensions.

In order to categorize objects in a color–color diagram, the most common method is described for example in Walker & Cohen (1988) and Walker et al. (1989). Having a sample of some already identified objects, they try to calculate a distribution–function using different methods. This function often has its base in the normal distribution. Depending on the presentation of the results, the following is often given:

- The parameters of the distribution functions for each type of found group of the diagram.
- A box (drawn in the diagram) giving the maximal limits of the regions where such a type of object should be found.

Walker et al. (1989) give a diagram with boxes specifying the limits mentioned above. Further a table is given containing the parameters μ and σ of Gaussian distributions for many types of objects **and** for each dimension of the color–color diagram ($[12]-[25]$, $[25]-[60]$, $[60]-[100]$).

The disadvantages of such a kind of presentation are the following:

- Referring to the table of Walker et al. (1989) the parameters are presented only for one dimensional Gaussian distributions. Building a multidimensional Gaussian distribution with these parameters leads to errors because of a missing skewness–factor. Further, the relations between the different Gaussian distributions are not given. Therefore it is not possible to calculate contamination ratios or to clearly distinguish between the different groups.
- The graphical presentation with boxes has the same disadvantages as described above. Furthermore the limits never represent natural regions.

The new algorithm can be used with natural regions. It is not necessary to adapt any distribution function to the data set. In order to investigate for example the occupation zones (OZs) of different types of objects in a color–color–diagram, two approaches are conceivable:

1. Use the 4580 so–called *unassociated* IRAS sources defined in Walker & Cohen (1988) as data set I. In order to calculate the underlying substructures, use the same sets s_i of identified objects as given in Walker et al. (1989). Using the algorithm with I and s_i the result will be a distribution function which gives an idea concerning not only the regions where such a group of objects should be found but also the contamination of the different groups.
2. In case of a large dispersion of a group it is possible to use instead of s_i the Gaussian parameters as starting values.

We are currently investigating the capabilities of the algorithm respecting that kind of color–color–diagrams (Kienel & Kimeswenger, in preparation).

3. The algorithm

In this chapter the basic steps of the algorithm will be described. The mathematical details will be shown in the appendix. The following formulae are given in order to have a clear mathematical definition. The whole data set has to have the following form:

$$I(x) = \sum_{i=0}^{n} g_i s_i(x) + \mathrm{err}(x)$$
$$s, \mathrm{err} : R^m \longrightarrow R^m$$
$$n \in N_0, m \in N; \;\; g_i \in R, x \in R^m.$$

This equation can be split into two parts in order to be able to separate one component from the rest of the data set I:

$$\begin{aligned} I(x) = & \quad \textstyle\sum_{i=0}^{n} g_i s_i(x) + \mathrm{err}(x) \\ = & \;\; hf(x) + \textstyle\sum_{i=1}^{n} g_i s_i(x) + \mathrm{err}(x) \end{aligned} \qquad (1)$$

$f(x)$ represents the form and position of the overlapping main structure. $s_i(x), i = 1,..,n$ represents the substructures and h as well as g_i represent a kind of amplitude or multiplier of the structures.

The following inequalities form the restrictions:

$$f(x) \geq 0 \text{ and } s_i(x) \geq 0 \;\; i = 1, n \;\; \forall x \in R^m. \qquad (2)$$

These restrictions normally do not represent a handicap due to two considerations:

1. All the data sets based on histograms (e.g. counting values) have always positive values.
2. It is often possible to shift the data sets to values larger than zero.

The algorithm needs the following main information:

- The whole data set $I(x)$.
- A first estimation $f_0(x)$ of $f(x)$ of the superimposing or overlapping main structure.

The estimation of $f(x)$ concerns only the form and position but not the intensity assumed in the whole data set. *Further* parameters of the algorithm are flags which indicate whether one of the sets has to be smoothed or not.

The algorithm represents an iteration performing the following steps:

3.1. The preparation–phase

Depending on the flags mentioned above, smoothed versions of the sets I and f will be built ($I \rightarrow I_\mathrm{s}$ and $f_0 \rightarrow f_{0\mathrm{s}}$). With these sets, errors (e.g. so–called *outliers*) will be eliminated.

The next step consists in the calculation of an offset Of. If an offset over the whole data set can be found, this offset has to be considered in the following calculations.

$$Of = \max(\min(I_\mathrm{s}(x)), 0). \tag{3}$$

Referring to Eq. (1) the algorithm has to estimate the parameter h. In order to get a starting value h_0 for the iteration, h_0 will be calculated as follows:

$$h_0 = \max(I(x))F_\mathrm{err}, F_\mathrm{err} \geq 1 \in R. \tag{4}$$

The value of the parameter F_err depends on the estimation of the error in the data set. The higher the error (resp. the ratio between the error and the data set) will be assumed, the higher the value of F_err has to be.

After these preparations the iteration starts.

3.2. The iteration

In order to get h_{i+1}, we calculate a function which represents a ratio between the whole data set I_s and the estimated main structure $h_i f_{0\mathrm{s}}$:

$$G(x) = \begin{cases} \frac{I_\mathrm{s}(x)-Of}{h_i f_{0\mathrm{s}}(x)} - 1 & \leftrightarrow f_{0\mathrm{s}}(x) > 0 \\ \text{any value} & \leftrightarrow f_{0\mathrm{s}}(x) = 0. \end{cases} \tag{5}$$

The values of G lie between -1 and any value > 0 (cf. Appendix, points 1 and 2). In these regions where no substructure takes part in the data set, and if the real amplitude h has been found, $G(x) = 0$. The value increases in these regions where the substructures become higher than the estimated main structure. This can lead to extremely high values in the regions where the main structure is much lower than the substructures and the errors (mostly the edges of $f_{0\mathrm{s}}$). In these regions where the estimated main component is higher than the substructures the values of G lie between 0 and 1 ($G \in (0,1)$) (cf. Appendix, point 3).

This function is the basis of the so–called *correcting parameter a*, which represents a kind of a weighted average of special values of G. Two parameters are introduced which select the best values of G in order to build the parameter a:

1. The higher the value of f_0 the more it influences the main structure of the data set I. A parameter H_l takes this fact into account. If there are no substructures, the value of H_l could be set to zero.
2. A parameter G_l is introduced in order to exclude all values of G which exceed a certain value.

Further, a weighting function w_a favours these values of G where f_0 has its highest values (e.g. $\sqrt{f_{0\mathrm{s}}}$).

Based on G, H_l, G_l and the weighting–function w_a, the correcting parameter a has the following form:

$$a = \frac{\sum_{x\in M} G(x)w_a(x)}{\sum_{x\in M} w_a(x)} \tag{6}$$
$$M = \{x : f_{0\mathrm{s}}(x) > H_\mathrm{l} \text{ and } G(x) < G_\mathrm{l}\}\,.$$

With this parameter, a first improvement of the amplitude h_i can be reached:

$$h_{i_1} = h_i(1 + a).$$

After the calculation of h_{i_1} it is possible to build a new set R_res by substracting the estimated main structure from the data set I. There should not remain values less than zero due to condition 2. If they exist, then the main structure has been overestimated. Therefore the set R_res can be used to calculate a second parameter b which corrects such an overestimation:

$$b = \frac{1}{n} \sum_{R_\mathrm{res}<0} \frac{I_\mathrm{s}(x)}{h_{i_1} f_{0\mathrm{s}}(x)} \tag{7}$$

n represents the number of elements of $R_\mathrm{res}(x) < 0$. The new amplitude h_{i+1} is now calculated as follows:

$$h_{i+1} = h_i(1 + a)b \tag{8}$$

3.3. The end of the iteration

The end of the whole iteration is determined by one of the following three factors:

- The difference between h_i and h_{i+1} is smaller than a given limit.
- The variance of all iterated amplitudes is lower than a given limit.
- The variance begins to increase. In this case, the form and/or position of the main component is wrong.

The second and third condition are necessary for the following reasons:

It is possible, that the algorithm does not converge to an exact value. If the algorithm *converges* to two fixed values and pends between these values (near the solution), the variance stops this pending state after a certain time. In case of wrong form or position of the main component, the variance increases after some iterations and does not converge.

3.4. The advantages

- As one of its results the algorithm gives a data set which gives information on the probability of every point of the investigated set (Boller et al. 1992 gives only a probability for the whole set).
- The input-sets are very easy to generate (see Kienel 1996).

- For using the algorithm it is not necessary to be a specialist in mathematics.
- In ordrer to reach the result, the algorithm needs only a few CPU-seconds on out-of-state desktop workstations.

4. Examples

Three examples with generated data sets demonstrate the use of the algorithm. The first two examples deal with one–dimensional Gaussian distributions. The third example handles two–dimensional Gaussian distributions as overlapping main data set and several *blocks* as underlying substructures. All data sets are adapted with a Poisson–distributed error.

In order to be able to present the principle of the algorithm only examples using artificial data sets will be presented. The detection of structures in color–color diagrams by means of the algorithm will be presented in a further paper.

4.1. A small subcomponent

Two Gaussian distributions build the data set I. Table 1 gives the parameters. The last column shows the intensity of each component ($Int = \sum a_i \;\; i = 1, 2$). The intensity of I equals to 4693 ($Int_I = \sum a_1 + \sum a_2 + \mathrm{err}$, the adapted error does not change the sum of the intensities).

Table 1. The parameters of the building components of I

	μ	σ	h	Int
a_1	90	15	120	4512
a_2	99	4	18	181

As an *estimation* of the superimposing main structure, a Gaussian distribution f_0 is generated with the same values (μ, σ) as a_1. The intensity of f_0 is set to 1. Figure 1 shows the sets I, a_1 and a_2. The algorithm has as input values the sets I and f_0. Set I has to be smoothed in contrast to the generated set f_0.

The algorithm runs 4 times and calculates as intensity of the main structure F_4 4332 (4% less than the real intensity of a_1). The intensity of the remaining residuals R_4 reaches 361.

The result concerning the whole data set, $F_4 + R_4$, demonstrates that the algorithm does not change the intensity during the calculation: the algorithm is *intensity–invariant.*

The big difference between the real and the estimated substructure (100%) arises due to the fact that the intensity of the main component is 25 times(!) the intensity of the substructure. If the algorithm reaches the correct intensity of the main component up to **1%** and due to the invariance of the algorithm, the difference has to be exactly **25%** between the original and the calculated substructures.

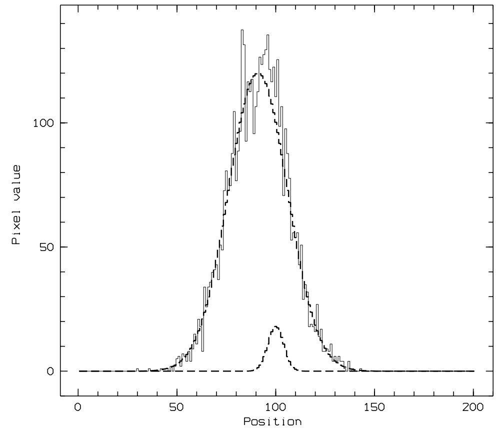

Fig. 1. The main data set I together with the *building* components a_1 and a_2 (broken lines)

After the iteration, a probability function p is calculated in order to be able to distinguish between an error and the real substructure. Actually, this probability function only takes into account a ratio between an estimated maximal error ($c\sqrt{I}$), the real data set I and the found substructure. Figure 2 displays the found substructure R_4 and the function p (which is multiplied with the factor 10 in order to be able to compare it with R_4). p has its highest values in the regions of the real substructure a_i and at the edges of the figure.

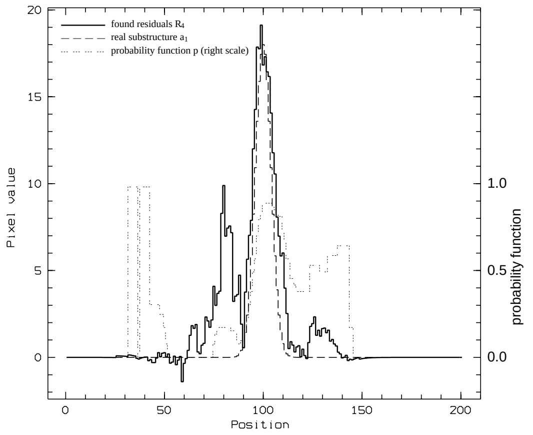

Fig. 2. The residual R_4 together with the probability function p

4.2. Two clusters of similar size

The second example deals with the problem often found in astronomy: Two clustars with about the same size and a non empty intersection. Again two gaussian distributions were used.

Table 2. The parameters of the building components of I

	μ	σ	h	Int
a_1	80	20	40	2005
a_2	120	20	40	2005

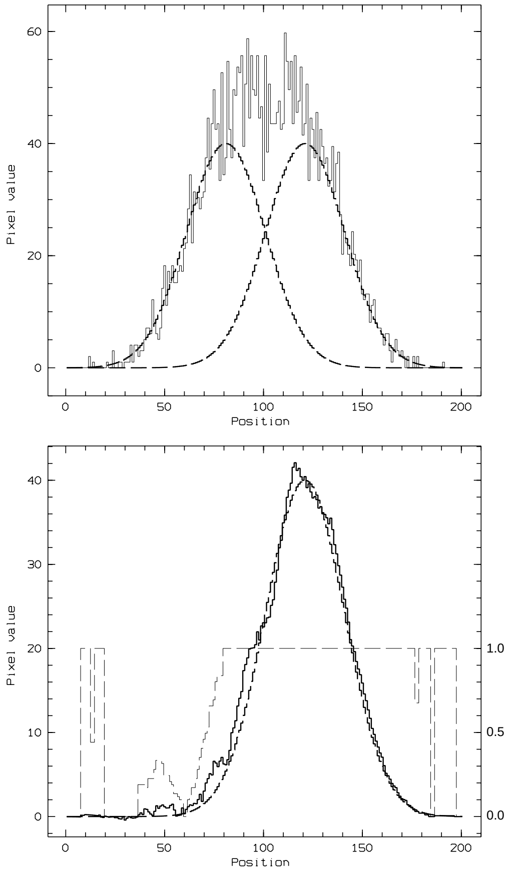

Fig. 3. Top: The data set I using two clusters of similar size together with its *building* components a_1 and a_2 (broken lines). The left component a_1 is used as "main structure". Bottom: The residual R_4 using two clusters of similar size (solid line), the *building* function a_2 (thick broken line) and the propability function (thin broken line)

The *residuum* represents extremely well the *building* component a_2. One should remember, that the algorithm does not know anything about the form of this function a_2.

4.3. The covered **E**

The second example illustrates further capabilities of the algorithm. On the one hand, different forms build the set. The superimposing main structure is built with two Gaussian distributions, whereas the covered structures are built with several *blocks*. On the other hand this example demonstrates the use of the algorithm in the two dimensional case.

A so–called *block* is a kind of a two dimensional trapecium with the following parameters:

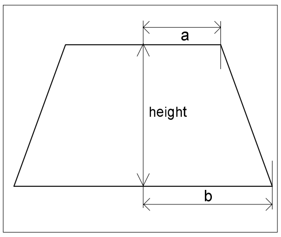

Fig. 4. The defining parameters a, b ($a \leq b$) and h (in one dimension)

The original set is shown in Fig. 5. The intensity–ratio between the main component and the subcomponent is about 90 : 1. The substructure has a constant height of 8 units (cf. Fig. 6), the maximum height of the real main component is at 409.6 units and consists of two Gaussian distributions, each of them having a height of 240.0 units. The total intensity of the main component has 3114430 units.

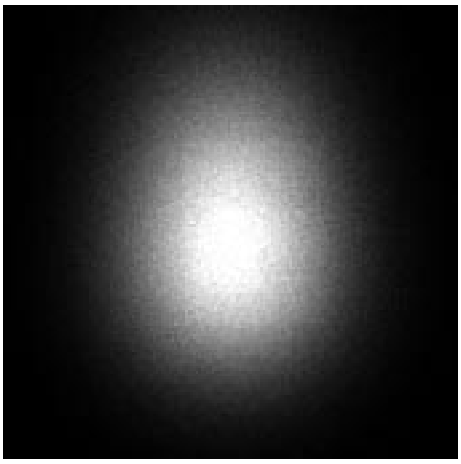

Fig. 5. The real structure I (the main component together with the substructure and the Poisson noise). A visual inspection does not show any sign of the hidden **E**

After 2 iterations, the algorithm achieves the following result: the found height of the main component has 410.4

units (+0.20%), the corresponding intensity has 3135670 units (+0.68%). The intensity of the achieved residuals is at 23464.5 units (cf. Figs. 7 and 8), while the intensity of the real substructures is at 34272 units.

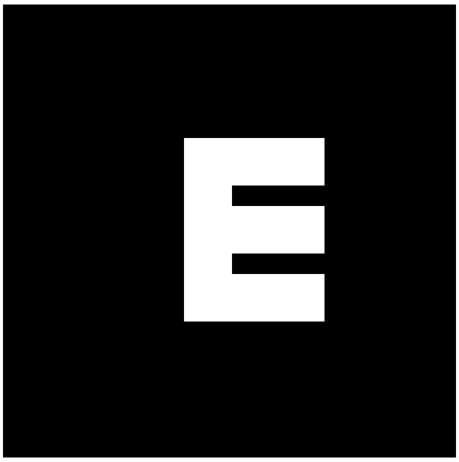

Fig. 6. The substructure built with 4 blocks

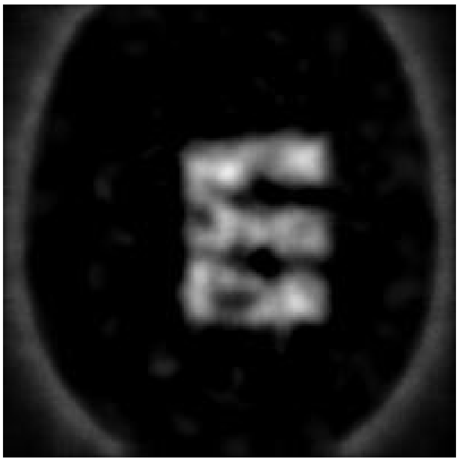

Fig. 7. The estimated residuals including the remaining noise

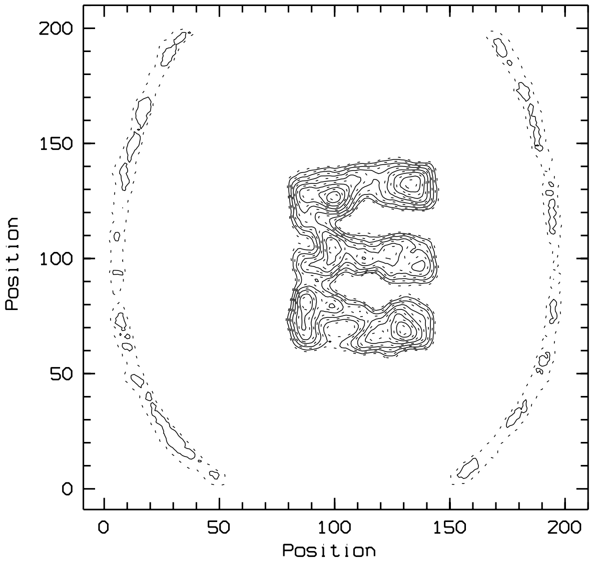

Fig. 8. The contour–plot of the estimated residuals. The substructure is very well recognizable

5. Conclusion

By examining the literature we found that many of the algorithms used in astronomy as well as in other fields of science are very complicated and need profound mathematical knowledge. In other cases we observed them to be quite simple but not really satisfying in so far as their results are concerned. This experience led to the following three main considerations an algorithm has to fulfill:

1. There does not have to be any restrictions concerning the form and dimension of the data set.
2. There must be an exact mathematical base of the algorithm.
3. The algorithm should be easy to handle in order to enable an application by scientists having a less extensive mathematical knowledge.

The resulting algorithm fully complies with these conditions. The data sets investigated with the algorithm have to fulfill two general conditions:

1. There must be a large amount of data forming a total structure.
2. There must either be the knowledge that there exists partial structures or the possibility to make an estimation about such partial structures.

Looking at these restrictions the algorithm may be applied in every field of science with large amounts of data. The methodology allows to extract well defined samples from data sets.

Acknowledgements. We want to thank R. Weinberger, J. Pfleiderer and N. Netzer for many valuable discussions. This work was supported by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung*, project number P10036–PHY (DeNIS).

Appendix

1. $G > -1\ \forall x : f(x) > 0$

Let us assume that there exists no error and that the estimated superimposing structure f_0 represents the real main structure f in order to be able to estimate the ranges of the function G.

The estimated amplitude h_e has the following form:

$$h_\mathrm{e} = h + h_\Delta = h + \frac{p}{q}h = h\left(\frac{q+p}{q}\right)$$
$$q > 0,\ p > -q.$$

The restriction $p > -q$ has the following explanation:

$$p \le -q \Leftrightarrow h_\Delta \le -h \Leftrightarrow h_\mathrm{e} \le 0,$$

but the estimated amplitude has to be larger than zero. $G(x)$ has the following form:

$$G(x) = \frac{q}{q+p}\left(1 + \frac{\sum_{i=1}^n g_i s_i(x)}{hf(x)}\right) - 1.$$

Assume that the following holds:

$$\begin{gathered} \exists x : f(x) > 0 \textit{ and } G(x) \le -1 \\ \frac{q}{q+p}\left(1 + \frac{\sum_{i=1}^n g_i s_i(x)}{hf(x)}\right) - 1 \le -1 \\ \frac{\sum_{i=1}^n g_i s_i(x)}{hf(x)} \le -1. \end{gathered}$$

This inequation cannot be fulfilled due to the restriction in Eq. (2).

2. $G < 1 + 2L\ \forall x : f(x) > 0$

The upper boundary of G depends on the ratio between the main structure $hf(x)$ and the substructures $\sum g_i s_i(x)$. If the ratio $\frac{\sum_i g_i s_i}{h_\mathrm{e} f(x)} < L$ and $h_\mathrm{e} = h + h_\Delta$, the following estimations hold:

$$\begin{aligned} h_\Delta \ge 0 \Rightarrow\quad & G(x) \le L \\ h_\Delta \in \left[-\frac{h}{2}, 0\right) \Rightarrow\ & G(x) < 1 + 2L. \end{aligned}$$

The first estimation is easy to prove:

$$h_\Delta \ge 0 \Rightarrow G(x) = \underbrace{\frac{h}{h + h_\Delta} - 1}_{\le 0} + \underbrace{\frac{\sum_i g_i s_i(x)}{(h + h_\Delta) f(x)}}_{\le \frac{\sum_i g_i s_i(x)}{hf(x)} \le L} \le L.$$

If $h_\Delta < 0$, e.g. $h_\Delta = ch, c \in [-0.5, 0), \Rightarrow$

$$\begin{aligned} G(x) &= \frac{h}{h + ch} - 1 + \frac{\sum_i g_i s_i(x)}{(h + ch) f(x)} \\ &= \frac{1}{1+c} - 1 + \frac{\sum_i g_i s_i(x)}{(1+c) hf(x)} = \\ &= \underbrace{\frac{1}{1+c}}_{1<x\le 2} \underbrace{\left(1 + \frac{\sum_i g_i s_i(x)}{hf(x)}\right)}_{<1+L} -1 \Rightarrow \\ G(x) &< 2L + 1. \end{aligned}$$

If $c \in (-1, -0.5)$, the values of G *run away*, but this happens only in case of an extremely underestimated amplitude. The case $c \le -1$ is impossible because the amplitude has always to be larger than zero.

It is only possible to estimate the value of L. If the main component superimposes for example the substructures, the ratio between the main structure and the underlying substructures has to be less than one. Therefore $G(x) < 3\ \forall\ x : h_\mathrm{e} f(x) > \sum_i g_i s_i(x)$ and $h_\mathrm{e} \ge \frac{h}{2}$.

3. $G \in (0, 1)\ \forall x : hf(x) > \sum_{i=1}^n g_i s_i(x)$

Let us assume that there exists no error and that the estimated superimposing structure f_0 represents the real main structure f in order to be able to estimate the ranges of the function G.

$$\begin{aligned} I(x) &= hf(x) + \sum_{i=1}^n g_i s_i(x) \\ G(x) &= \frac{hf(x) + \sum_{i=1}^n g_i s_i(x)}{h_\mathrm{e} f(x)} - 1 \end{aligned}$$

If the estimated amplitude h_e converges to the real amplitude h, the following estimation can be done:

$$\begin{aligned} G(x) &= \frac{hf(x) + \sum_{i=1}^n g_i s_i(x)}{h_\mathrm{e} f(x)} - 1 = \\ &= \underbrace{\frac{h}{h_\mathrm{e}} - 1}_{h_\mathrm{e} \to h \Rightarrow \to 0} + \frac{\sum_{i=1}^n g_i s(x)}{h_\mathrm{e} f(x)} = \\ &= \frac{\sum_{i=1}^n g_i s_i(x)}{h_\mathrm{e} f(x)}\,. \end{aligned} \tag{9}$$

Equation (9) has to be larger than or equal to zero because of condition 2. Due to the assumption that $h_\mathrm{e}\ f(x)\ > \sum g\ s(x)$, the equation is less than 1. The more the main structure superimposes the substructures, the lower the value of G has to be.

References

Boller Th., Meurs E.J.A., Adorf H.-M., 1992, A&A 259, 101

De Jager O.C., Swanepoel J.W.H., Raubenheimer B.C., 1986, A&A 170, 187

Epchtein N., et al., 1994, Ap&SS 217, 3

Kienel C., 1996, Thesis on Covered Data Structures

Kienel C., Kimeswenger S., 1997, A&A (submitted)

Silverman B.W., 1986, Density Estimation for Statistics and Data Analysis. Chapman and Hall

Sutherland W., Saunders W., 1992, MNRAS 259, 413

Waerden B.L. van der, 1957, Mathematische Statistik, Die Grundlagen der Mathematischen Wissenschaften, Band LXXXVII. Springer-Verlag

Walker H.J., Cohen M., 1988, AJ 95, 1801

Walker H.J., Cohen M., Volk K., Wainscoat R.J., Schwartz D.E., 1989, AJ 98, 2163