Publ. Astron. Obs. Belgrade No. 76 (2003), 19-23 *Invited lecture*

AN ALTERNATIVE APPROACH TO DATA PROCESSING

SRĐAN BUKVIĆ

Faculty of Physics, University of Belgrade, Studentski Trg 12-16, Belgrade, Serbia
E-mail: ebukvic@ff.bg.ac.yu

Abstract. We present a new form of merit function which measures an agreement between large number of data and a model function with the particular choice of parameters. We demonstrate the efficiency of introduced merit function on the common problem of finding base line of the spectrum. Also, we discuss efficiency of existing minimization algorithms in discrete topography which is concomitant to proposed merit function in general case.

1. INTRODUCTION

In the past two decades experimental technique was moved from analog to digital domain. Consequently, the amount of data acquired in experiments has increased significantly requiring reliable automated algorithms for data processing. A frequent task in data processing is to find best parameters of some model function, in the sense of the "best fit", related to the particular data set. It is common practice to use the least-squares as a merit function for various data fittings. Unfortunately, the measurement process is not free of errors which cause that some experimental points are occasionally just way off. Also, data of interest are frequently spoiled by small amount of points due to some undesirable process which can not be avoided. This can easily turn the least-squares into nonsense. To overcome these problems various robust techniques have been proposed.

In astrophysics and physics a prominent example which requires the rough approach is the task of spectrum baseline and continuum estimation. In Gabel *et al.* (2002) is in details described the complex procedure of continuum estimation in the presence of superimposed spectral lines. It is shown that a prior knowledge of spectral line positions is necessary for continuum estimation. Moreover, the described procedure assumes that parts of recorded signal, taken for continuum estimation, are not spoiled by unknown spectral lines of low intensity. In general case it is difficult to satisfy all requirements necessary for reliable continuum estimation. Here, spectrum lines act as an undesirable feature of the signal and the points belonging to the spectral lines can be considered as outliers.

The intention of this paper is to introduce a simple and efficient method for rough estimations, insensitive to outlying points. Proposed method is suitable for large data sets only, i.e. it can not replace least-squares for data sets containing just a few points.

2. MERIT FUNCTION

Consider a particular data set of x_i 's and y_i 's (i=1, 2, .. n) and a model function $y(x; a_1, a_2, ... a_m)$ depending on x and some parameters $a_1, ... a_m$. Let's define quantity d in the following way. At a distance $y(x_i; a_1, a_2, ... a_m) - y_i < d$ we will say that the point (x_i, y_i) is *close* to the model function $y(x_i; a_1, a_2, ... a_m)$ for given set of parameters $\{a\}$. If so put f_i =1, otherwise f_i=0. Now we will define quantity χ:

$$\chi = n - \sum_{i=1}^{n} f_i \qquad (1)$$

The purpose of χ is to be used as a merit function. Thus, we will search for a set of parameters $\{a\}$ for which the merit function χ has a minimum in respect to the chosen value d. In other words, we will search for set of parameters $\{a\}$ for which the maximal number of data points will be *close* to the model function, i.e. the model function will closely resemble the data in respect to d.

The explained approach will be termed Close Points Concept (CPC). The aim of the CPC is to quantify our ability to recognize trace-like set of points as a line and, at the same time, to ignore outlying points. The distance d defines what will be 'close' and which points will be disregarded. A straightforward way to apply CPC is to incorporate merit function (1) in some of the existing minimization algorithms.

3. ESTIMATION OF THE SPECTRUM BASE LINE

First we will consider most common and most simple case of a spectrum when the base line is supposed to be horizontal i.e. parallel to the y axis. This is justified when a narrow range of the spectrum is recorded. In such a case we can avoid the use of the merit function in general form (1) and explicit use of a minimization algorithm.

Suppose that the spectrum contains n points obtained by means of k bit linear A/D converter. It means that spectrum intensity in each point is expressed via one of $q=2^k$ different levels of the A/D converter. In order to obtain the base line we will make an auxiliary histogram consisting of exactly 2^k bins, each bin corresponding to the one of 2^k values of the A/D converter. We will proceed as follows: starting with the point 1 we incerese the content of bin y_1 by 1, similarly for point 2 content of bin y_2 is also incresed by 1 and so on, up to the last point n. Finally, each bin will contain the number of occurrences of appropriate level of the A/D converter within chosen part of the spectrum. It will be assumed that obtained histogram has just one maximum, i.e. bin j has a maximal count, see Fig. 1. It is obvious that the line $y_b = j$ is a base line of the spectrum in respect to the merit function (1). In this case the distance d, which is used to distinguish "close" and "far" points, was set to be one half of the A/D converter step. Also, the search for the best y was done within discrete set of values defined by the A/D converter. The proposed procedure is extremely simple and works successfully practically with any spectrum or spectrum like signals (Djeniže and Bukvić 2001, Spasojevic *et al.* 1996)

It is of interest to estimate discrimination level, parts of the signal above this level will be considered as spectral lines, everything below discrimination level will be considered as a noise. The key assumption is that the signal below the spectrum base line originates only due to noise of any kind which is present in the system, while the signal

above the base line consists of a useful signal spoiled by the noise. It is a matter of choice how to represent the noise distribution which is present in our histogram below spectrum base line, however one standard deviation (σ) appears as a most common manner. Accordingly we can take the value σ as a discrimination and $y_d=j+\sigma$ as a discrimination level, see Fig. 1.

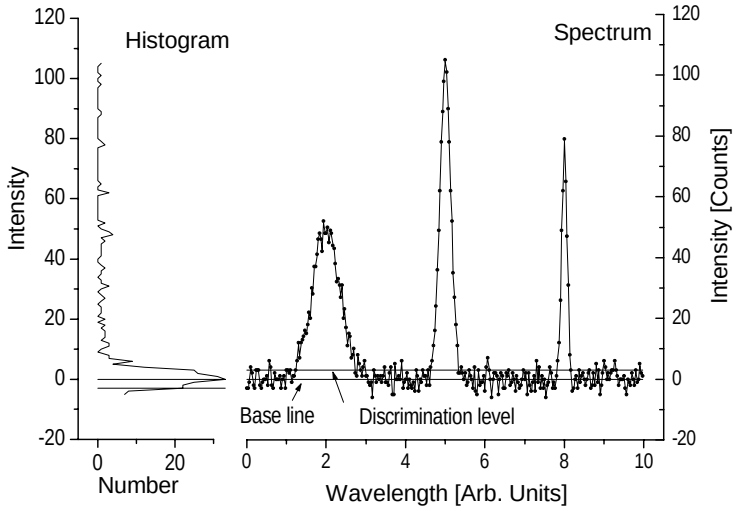

Fig. 1: Spectrum base line estimated by CPC approach. See text for details.

The purpose of our second example is to demonstrate the potentials of the CPC in numerically more complex situation with explicit use of the merit function (1) and appropriate minimization algorithm. To control the numerical process we will consider an artificial set of data generated according to the following relation:

$$y = P_1 + P_2 + P_3 + B \tag{2}$$

where peaks P_1, P_2, P_3 have the same, Gaussian, form:

$$P_i = a_i \exp\left[\frac{(x-c_i)^2}{d_i^2}\right]$$

and B is a broad line also of the Gaussian form which mimics a nonlinear base line:

$$B = a_b \exp\left[\frac{(x-c_b)^2}{d_b^2}\right] + b$$

In Fig. 2 is shown a graph generated by function (2) with the following coefficients:

a_1=5, c_1=20, d_1=10; a_2=1, c_2=30, d_2=12; a_3=1, c_3=70, d_3=10
a_b=1, c_b=50, d_b=1000, b=0

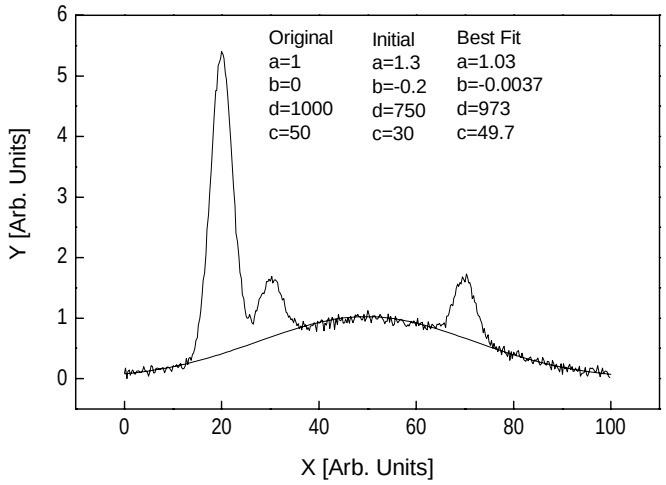

Fig. 2: Non linear base line of the spectrum estimated by CPC approach. See text for details.

Each point is randomized with normally distributed errors in order to simulate a noise concomitant to experimental data. Our intention is to obtain coefficients a_b, c_b, d_b, and b for the function B using the *whole data set* relying on the Close Points Concept. In other words we will fit our data to the function B considering sharp peaks P_1, P_2, and P_3 just as a deviation (outliers) of the main flow of the data given by term B. This task requires minimization of the merit function (1) by some of existing algorithms. We have chosen downhill simplex method (Press *et al.* 1988) applied in the routine AMOEBA, as the most suitable.

In Fig. 2 is shown the graph of the function B for parameters found by fitting for distance d set to 0.1. The best fit values are: a_b=1.03, c_b=49.7, d_b973, b=-0.0037. One can notice that these values are reasonably close to the original ones. Initial values for AMOEBA algorithm were: a_b=1.3, c_b=30, d_b=750, b=-0.2.

4. DISCUSSION

Initially we will discuss the influence of the distance d on the results. As it has been explained the role of d is to discriminate close and far points. If data are obtained by A/D converter and data set contains over a thousand points it is most simple to accept one step of A/D converter as a reasonable value for d. For data sets with several hundred points the distance d is most suitable to be slightly less than the magnitude of the noise present in the

system. A very low *d* value can cause that just a few points, or no one, are close to the model function making any algorithm unusable. Similarly, too high value for *d* can cause that all points become close to the model function for too broad range of parameters making the result useless. Generally, the distance *d* should be similar to the width of the line we are going to fit. Such a value provides sufficient number of close points in the sense of CPC.

An important feature of the CPC is that topography of a multidimensional space defined by (1) is not smooth; rather it is stepwise, discrete, topography. On the contrary, practically all minimization algorithms are supposed to be used in smooth, continual space. Some of them, however, work in stepwise topography, and we have applied downhill simplex method as a most rough one. Due to discrete nature of the topography, the choice of initial guess is more difficult then usual. Too bad choice can cause that no one experimental point is close to the model function moving initial point into completely flat region of the multidimensional space where not any minimization algorithm can work. Also, the usual problem of finding just a local minimum is more emphasized here due to discrete nature of the merit function (1). For very large data set the merit function becomes pseudo continual which facilitates the use of standard algorithms. But, for small number of points the discrete nature of (1) make it useless in the sense of finding the best fit. It is interesting to relate this feature to our ability to recognize trace which is consisting of many points as a line, while just a few points are very difficult to be thought of as a line.

A few words more about uncertainties concomitant to the parameters obtained by fitting. Unfortunately this problem is related to the non statistical approach of the CPC. Namely, we suppose that outliers *are not* distributed according to the Gaussian model, therefore, we can not apply standard procedures to estimate uncertainties of the best fit parameters. AMOEBA minimization algorithm is suitable because it is not based on the assumption of normally distributed errors and, consequently, it does not produce uncertainties for the best fit parameters.

5. CONCLUSION

We have introduced a new form of merit function (1) based on close points concept suitable for problems where rough approach is necessary. The efficiency has been shown on two typical examples of the introduced merit function. In the first example, which is related to the common problem of finding spectrum base line, we have explained a simple numerical procedure avoiding explicit use of the merit function (1) and appropriate minimization algorithm. In the second example a standard minimization algorithm has been used to fit artificially generated data to the nonlinear function of the Gauss type over which three sharp peaks were superimposed. Comparing original and values obtained by fitting we have demonstrated the potentials of CPC. Also, we have discussed the influence of introduced parameter *d* and possible problems related to the discrete character of merit function (1). Finally, we would like to emphasize the necessity for appropriate minimization algorithm developed for the use in discrete space.

References

Gabel, J.R., Crenshaw, D.M., *et al.*: 2003, *Apstrophys. J.*, **583**, 178.

Spasojević, Dj., Bukvić, S., Milošević, S., Stanley, E.: 1996, *Phys. Rev. E*, **54**, 2531.

Djeniže, S., Bukvić, S.: 2001, *Astron. Astrophys.*, **365**, 252.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: 1988, *Num. Rec.*, Cambridge University Press, Cambridge.